

# AP Statistics Notes

anastasia

Fall 2024 & Spring 2025

# Contents

<b>1</b>	<b>Exploring One-Variable Data</b>	<b>2</b>
1.1	Representing Categorical and Quantitative Variables with Graphs . . . . .	2
1.2	Representing Quantitative Variables with Graphs . . . . .	4
1.3	Describing Distributions of Quantitative Variables . . . . .	7
1.4	Comparing Distributions of Quantitative Variables . . . . .	10
1.5	Z-Scores and the Empirical Rule . . . . .	13
1.6	The Standard Normal Curve . . . . .	16
<b>2</b>	<b>Exploring Two-Variable Data</b>	<b>20</b>
2.1	Two Categorical Variables . . . . .	20
2.2	Scatterplots and Correlation . . . . .	23
2.3	Linear Regression . . . . .	28
2.4	Influential Points and Departure from Linearity . . . . .	32
<b>3</b>	<b>Collecting Data</b>	<b>37</b>
3.1	Planning a Study . . . . .	37
3.2	Selecting a Random Sample . . . . .	40
3.3	Experimental Design . . . . .	42
<b>4</b>	<b>Probability, Random Variables, and Probability Distributions</b>	<b>49</b>
4.1	Basic Probability and Simulations . . . . .	49
4.2	The Addition Rule . . . . .	53
4.3	Venn Diagrams and the Multiplication Rule . . . . .	57
4.4	Conditional Probability and Tree Diagrams . . . . .	61
4.5	Discrete and Continuous Random Variables . . . . .	65
4.6	Combining Random Variables . . . . .	69
4.7	The Binomial Distribution . . . . .	72
4.8	The Geometric Distribution . . . . .	75
<b>5</b>	<b>Sampling Distributions</b>	<b>77</b>
5.1	Sampling Distributions of Sample Proportions . . . . .	77
5.2	Sampling Distributions of Sample Means . . . . .	80
5.3	Combining Sample Proportions and Sample Means . . . . .	82
<b>6</b>	<b>Inference for Categorical Data: Proportions</b>	<b>85</b>
6.1	Constructing a One Proportion z-Interval . . . . .	85
6.2	Constructing a One Proportion z-Test . . . . .	87
6.3	Relating Confidence Intervals and Significance Tests . . . . .	91
6.4	Inference for Comparing Two Population Proportions . . . . .	93
6.5	Errors & Power . . . . .	96
<b>7</b>	<b>Inference for Quantitative Data: Means</b>	<b>99</b>
7.1	Constructing a One Sample t-Interval . . . . .	99
7.2	Constructing a One Sample t-Test . . . . .	102
7.3	Inference for Paired Data . . . . .	104
7.4	Inference for Comparing Two Sample Means . . . . .	106
<b>8</b>	<b>Inference for Categorical Data: Chi-Square</b>	<b>110</b>
8.1	Chi Square Test for Goodness of Fit . . . . .	110
8.2	Chi Square Test for Homogeneity . . . . .	113
8.3	Chi Square Test for Independence . . . . .	116
<b>9</b>	<b>Inference for Quantitative Data: Slopes</b>	<b>119</b>
9.1	Sampling Distributions and Confidence Intervals for Slope . . . . .	119
9.2	Significance Test for a Slope . . . . .	122

# 1 Exploring One-Variable Data

## 1.1 Representing Categorical and Quantitative Variables with Graphs

Statistics is the study of data.

Data contains information about a group of individuals. This information is organized using variables.

Individuals are objects described by a set of data. Individuals may be people but may be animals or inanimate objects.

Variables are characteristics of individuals. A variable may take on different values of different variables. Variables can be split into two types: categorical or quantitative.

Categorical variables place individuals into specific groups.

Quantitative variables take on numerical values for which it makes sense to do arithmetic operations like adding and averaging. Quantitative variables fall into two categories: discrete and continuous.

Careful: just because it is a number does not automatically make it quantitative.

Discrete variables are numerical values where counting makes sense; in other words, decimals would not be an appropriate way to record the data.

Continuous variables are numerical values where decimals are appropriate; it usually involves some form of measuring.

One of the easiest ways to display categorical data is with a table. Here is a list of the first 10 US presidents, their political party and their state of birth.

George Washington	Federalist	Virginia
John Adams	Federalist	Massachusetts
Thomas Jefferson	Democratic-Republican	Virginia
James Madison	Democratic-Republican	Virginia
James Monroe	Democratic-Republican	Virginia
John Quincy Adams	Democratic-Republican	Massachusetts
Andrew Jackson	Democrat	South Carolina
Martin Van Buren	Democrat	New York
William H. Harrison	Whig	Virginia
John Tyler	Whig	Virginia

A one-way table could look like this:

Category	Count	Relative Count
Federalist	2	20%
Democratic-Republican	4	40%
Democrat	2	20%
Whig	2	20%

If you wanted to display two categorical variables at a time, you can make a two-way table.

To better visualize the data, we can also make graphs from our data. Visualizing graphs helps get a better idea of the distribution.

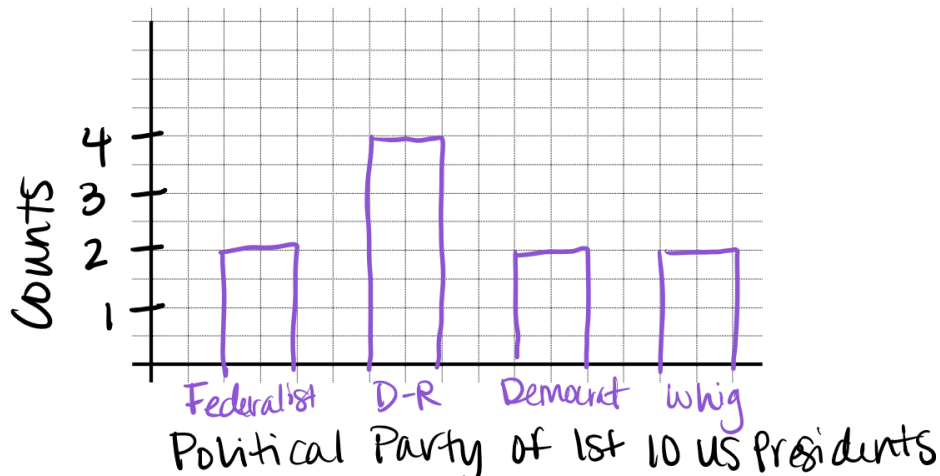
The distribution of a variable tells us what value the variable takes and how often it takes these values.

Bar graphs have the following characteristics:

- Label each axis clearly
- The  $x$ -axis will contain the categorical variable and the  $y$ -axis will display the counts
- Each category has its own bar and the bars cannot touch

- Order is not important when creating the  $x$ -axis.

Creating a bar graph of the four political parties of the first 10 US presidents gives us the following.



We can answer questions such as which political party was the most affiliated with, which was least affiliated with, and what the individuals and the variables of the data set are.

Talking about histograms now, let's use the table below which contains the combined scores for two weeks of NFL games.

16	22	23	24	25	32	33	35	35	37	37	37	38	39	39	40
41	44	44	44	45	46	47	48	49	50	56	59	59	65	72	80

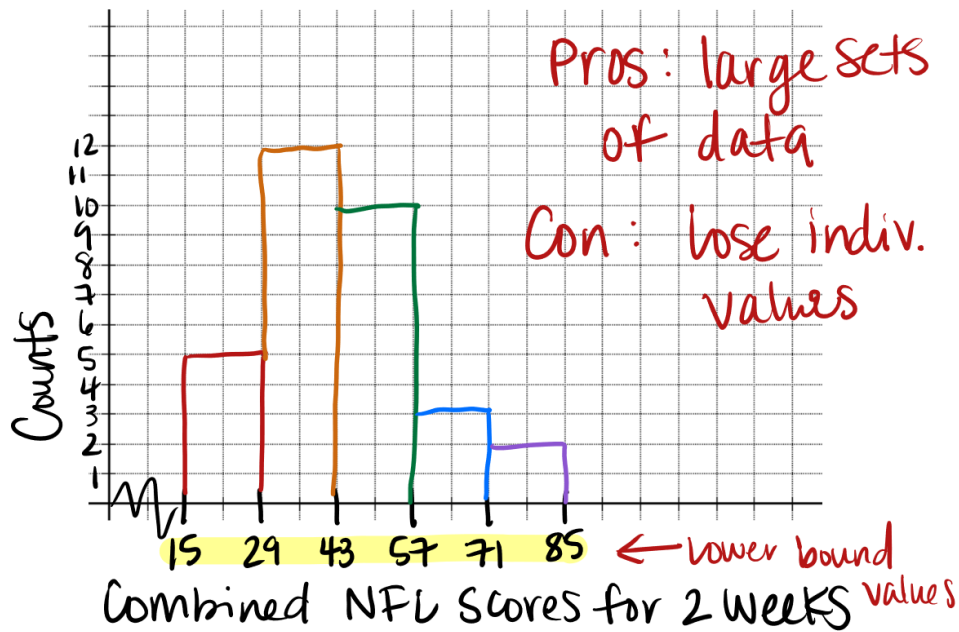
To make a histogram, we need to put the data into "bins" (even intervals that capture our data). We will do this first by hand by counting how many data scores are in each bin.

The lowest value is 16, the highest value is 80, so the bin width should be 14, if we want 5 bins.

Interval	Count
15-28	5
29-42	12
43-56	10
57-70	3
71-84	2

To make a histogram:

- Draw rectangles for each bin with height representing the count
- Bars must touch
- Label the  $x$ -axis with the lower bound values of your bins.



## 1.2 Representing Quantitative Variables with Graphs

Using the following data: we can create a stem plot. This data is the gender and average pulse rate for a set of students.

Gender	APR	Gender	APR	Gender	APR	Gender	APR	Gender	APR
M	54	F	81	M	90	F	70	M	62
F	73	F	55	F	65	F	67	M	66
F	60	F	57	F	60	F	68	M	98
F	81	M	70	F	97	M	77	M	59
F	85	M	62	F	65	M	64	M	61
F	60	M	69	F	83	M	60	F	64
F	55	M	80	F	60	M	60	F	98

Stemplots are an alternate way of illustrating data using a semi-graph. It is similar to a histogram but, unlike a histogram, the data isn't lost. If the data has two digits, the stem is the first digit and the leaf is the second. If the data has 3 digits, the stem is the first two digits and the leaf is the third.

Stem	Leaf
5	4 5 5 7 9
6	0 0 0 0 0 0 1 2 2 4 4 5 5 6 7 8 9
7	0 0 3 7
8	0 1 1 3 5
9	0 7 8 8

Key: 5|4 is a pulse of 54 bpm

Back-to-Back stemplots are created when you can separate the data into two categories. Using the same data, we can split the data up into male and females. The stem is still the tens digit and the leaf is the ones digit.

Male Leaf	Stem	Female Leaf
9 4	5	5 5 7
9 6 4 2 2 1 0 0	6	0 0 0 0 4 5 5 7 8
7 0	7	0 3
0	8	1 1 3 5
8 0	9	7 8

Key: 4|5 is a male pulse of 54 bpm

Key: 5|5 is a female pulse of 55 bpm

Split stemplots are the last type of stem and leaf plots. When you have too many data values on a single stem, it can be helpful to split the stem; the same way we would create more bins on a histogram if our bin width resulted in a skyscraper.

Let's say these are the test scores for an Advanced Algebra 2 Unit 1 Test:

80	88	78	93	69	80	92	90	88	84	82	92
88	62	84	75	74	96	88	88	90	94	92	79

The split stem and leaf plot looks like this.

Stem	Leaf
6	2
6	9
7	4
7	5 8 9
8	0 0 2 4 4
8	8 8 8 8 8
9	0 0 2 2 2 3 4
9	6

Key: 6|2 is a 62% on A2A Unit 1 Test

A dot plot is a very simple type of graph that involves plotting the data values, with dots, above the corresponding values on a number line.

To construct a dot plot:

1. Label your axis and title your graph. Draw a horizontal line and label it with the variable.
2. Scale the axis based on the values of the variable.
3. Mark a dot above the number on the horizontal axis corresponding to each data value.

Consider the Algebra 2 Advanced quiz scores.



Cumulative relative frequency graphs (ogives) display percentiles.

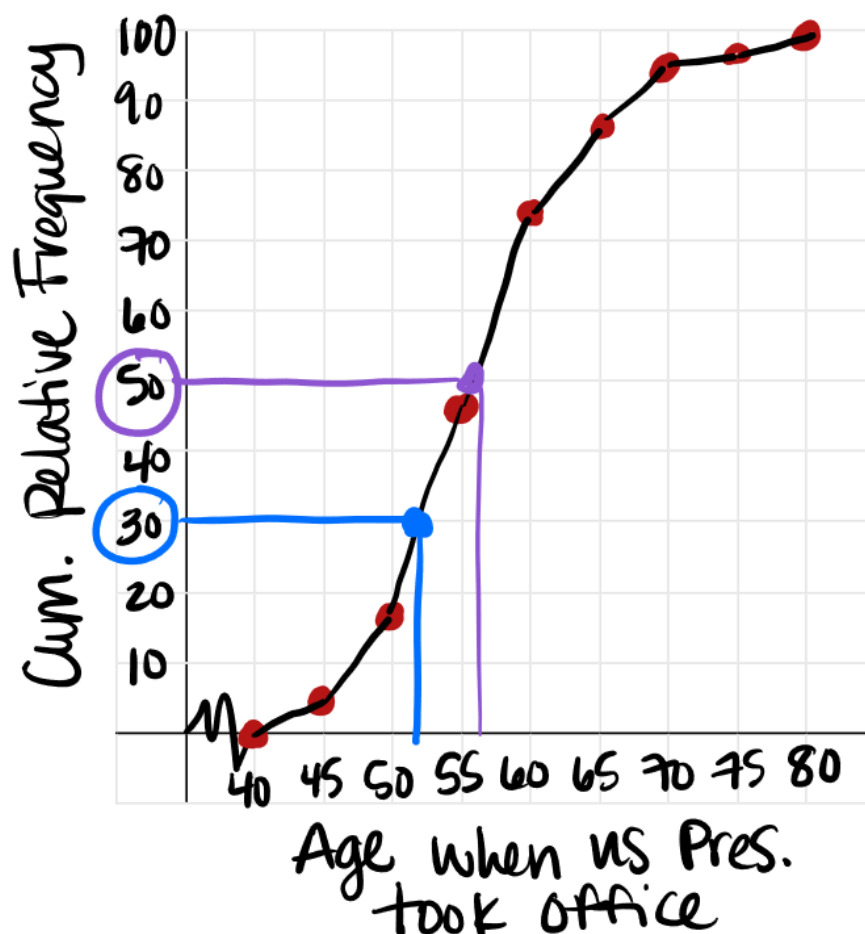
A percentile will tell you what percent of data falls below a value.

You first must make a table of the cumulative relative frequencies in order to graph it. This can be done by finding the relative frequencies and then find the cumulative frequencies.

The following is the ages of the 46 US presidents and their relative and cumulative frequencies.

Age	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
Frequency	2	7	13	12	7	3	1	1
Relative Frequency	$\frac{2}{46}$ 4%	$\frac{7}{46}$ 15%	$\frac{13}{46}$ 28%	$\frac{12}{46}$ 26%	$\frac{7}{46}$ 15%	$\frac{3}{46}$ 7%	$\frac{1}{46}$ 2%	$\frac{1}{46}$ 2%
Cumulative Relative Frequency	4%	19%	47%	73%	88%	95%	97%	99%

The following is what the graph looks like



Creating an ogive:

- The first interval is at 40-44, with a cumulative relative frequency of 4%. Since 40 is the starting interval, that starts at 0 and 4% is graphed at 45.
- The next interval is 45-49, with a cumulative relative frequency of 20%. At 45, 4% is graph, so at 50, the 20% will be graphed.
- This pattern continues until 100% is graphed at 80.

Answer the following questions with the ogive:

1. At approximately what age is the 30th percentile? What does this mean in the context of the problem?

2. Richard Nixon was the 37th president and was 56 years old when he was inaugurated. Approximately what percentile is this and what does it mean in the context of the problem?

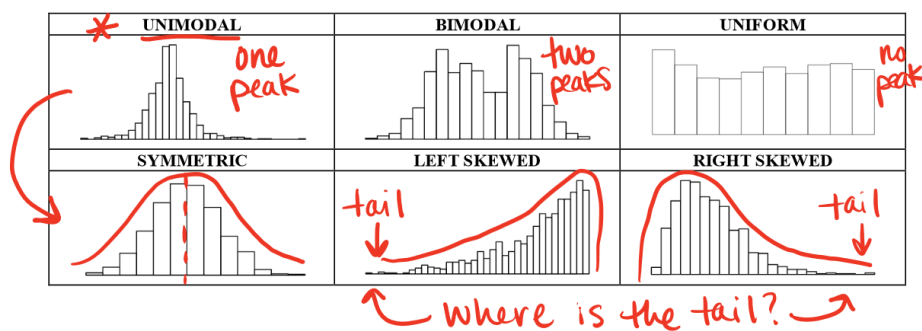
### 1.3 Describing Distributions of Quantitative Variables

In AP Statistics, to describe a distribution of a quantitative variable, we use the acronym SOCS:

- S: Shape
- O: Outliers
- C: Center
- S: Spread

After we graph, describing what we see helps identify patterns and answers questions about the data.

Once the distribution is graphed, the first thing we identify is the shape.



There are three measures of center in statistics: mean, median, and mode.

The following data is the number of siblings for a group of students.

5	2	1	3	1	1	2	2	2	2	4	0	1	1	1	2	0	5
1	6	7	7	1	3	1	1										

Mean is symmetric. The formula for mean is

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

Where  $\bar{x}$  is the mean of the sample,  $x_i$  is each individual observation,  $n$  is the number of observations, and  $\sum$  is notation for a summation.

The mean of this data would be 2.38 siblings.

The reason we do not always use the mean to describe the center is the inclusion of outliers in the data set.

For example, if we had 7 scores of 90, 92, 94, 98, 86, 88, and 0, the mean would be 78.29%. Removing the outlier, the mean increases to 91.33%. The outlier is bringing the mean down in this instance.

The mean is called non-resistant. This means that the mean is strongly influenced by extreme values.

Median - skewed. Luckily, we have another measure of center that is resistant, meaning that if there are extreme values in a data set, the measure of center will not be affected by it.

The median is found by ordering the data and then finding the middle value in that list.

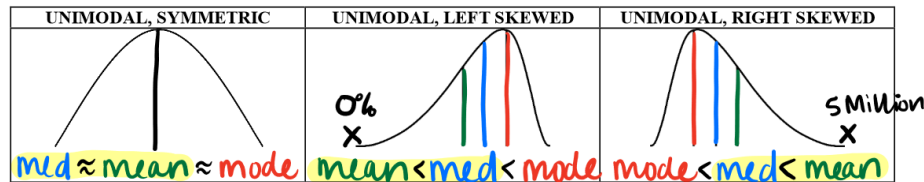
Looking at the set of 7 scores above, the median would be 90%. This shows that the median is greater than the mean.

Whenever there are extreme values or outliers, the median is the better measure of center compared to the mean.



The last measure of center is perhaps the most useless: the mode. It finds the most occurring value in a data set. The idea of having this as a measure of center comes from having a symmetric, unimodal distribution, where the most occurring value happens in the middle. However, as we have seen, there are many shapes to different distributions and the most occurring doesn't always occur in the middle.

Summary:



Spread (variability): There are three common measures of spread in statistics: range, standard deviation, and IQR.

The range is the difference between the maximum value and minimum value.

The standard deviation is the average deviation of an observation from the mean of the data set.

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

where  $s_x$  is the standard deviation of the sample,  $\bar{x}$  is the mean of the sample,  $x_i$  is each individual observation,  $s_x^2$  is the variance of the sample,  $n$  is the number of observations, and  $\sum$  is the summation notation.

Within the standard deviation, we calculate something called the variance which is the average squared deviation. We can find the variance from the standard deviation by squaring both sides to eliminate the square root.

In context the standard deviation is “The quantitative variable typically varies from the mean by standard deviation units.”

Standard deviation should only be used as your measure of spread when the mean is your chosen measure of center.

The standard deviation has the following additional properties:

- The standard deviation is always positive.
- The standard deviation is 0 when all observations are equal.
- The standard deviation has the same units of measure as the original variable measured.
- The standard deviation is non-resistant, meaning that a few outliers will make it large.
- The greater the standard deviation, the greater the distribution.

The last measure of spread is Inter-Quartile Range (IQR) and uses percentiles to describe the spread of the distribution.

Here are some other percentiles:

- 0th percentile: minimum - lowest value in a data set
- 25th percentile: Quartile 1 or Q1 - 25% of the data set is below this value
- 50th percentile: Median - middle value in a data set
- 75th percentile: Quartile 3 or Q3 - 75% of the data set is below this value
- 100th percentile: maximum - highest value in a data set

These 5 numbers make up what is called the Five Number Summary. To calculate these values by hand, place the observation in ascending order and find the median. Then find the middle value to the left and right of your median to identify your quartiles.

**Example**

Here is the data from a previous class about how many hours of sleep they got before the first day of school.

2 4.5 5 5 6 6 6.5 7 7 7 7 7 7 7.5 8 8 8 8 8 8.5

From this, the minimum is 2, the Q1 value is 6, the median is 7, the Q3 value is 8, and the max value is 8.5.

The IQR is 2 hours.

An outlier is an individual piece of data that falls outside the overall pattern of the distribution.

When an outlier occurs, we must find out why it occurs. Many times, it occurs because of a mistake. Outliers can be eliminated from the data if there is a good reason. However, if you are unsure then you cannot just remove the data point.

Most of the time, the AP exam wants you to comment on visual outliers from a graph of quantitative variable.

Using the five-number summary and the IQR, we also have a numerical way of determining if a point is an outlier.

1. Find the five-number summary.
2. Find the IQR.
3. Compute  $Q1 - (1.5 \cdot IQR)$ . Any data below that number is an outlier.
4. Compute  $Q3 + (1.5 \cdot IQR)$ . Any data above that number is an outlier.

**Example**

Literary scholars sometimes use the distribution of word lengths in a work as a test of authenticity. Here are the words lengths for the first 26 words on a randomly-selected page from Toni Morrison's *Song of Solomon*.

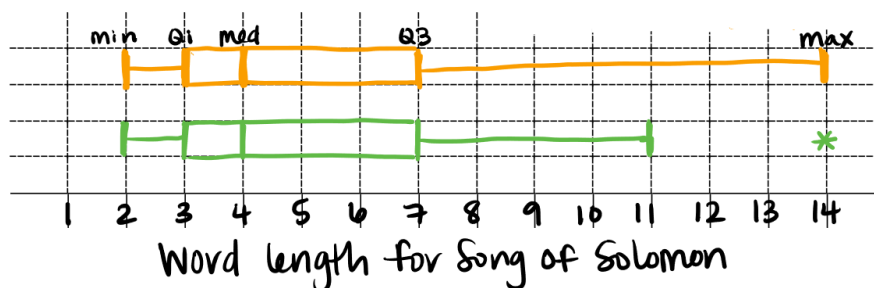
2	3	4	10	2	11	2	8	4	3	7	2	7
5	3	6	4	4	2	5	8	2	3	4	4	14

- (a) Mathematically check for outliers in your data.

The IQR is 4, so any numbers outside the bound of -3 to 13 are outliers. There is an outlier at 14 words.

The five number summary can also be used to create a boxplot or a modified boxplot (one that shows outliers).

- (b) Create a boxplot and a modified boxplot for your data.

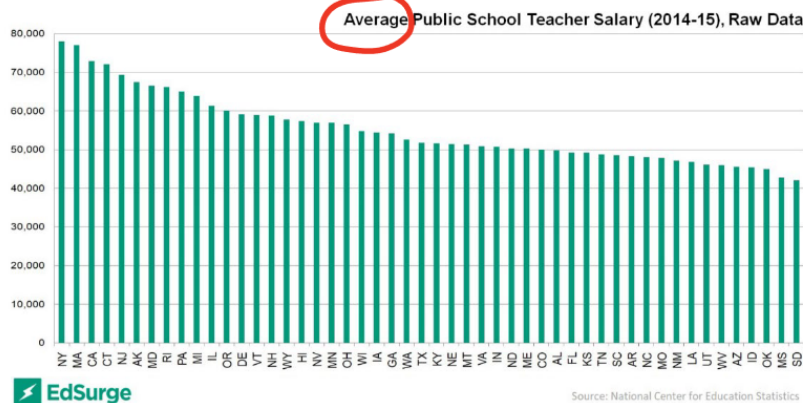


## 1.4 Comparing Distributions of Quantitative Variables

In the world of statistics, it is not enough to just report the graph or just report the summary statistics. The graph alone might not give us all the information we need to make a valid conclusion.

### Example

Do teachers in New York get paid too much? The graph below shows the average teacher salary for each state.



The graph shows that teachers in New York State make almost double than teachers in South Dakota. Is this a valid conclusion or is the graph not telling the whole story?

The cost of living is higher in New York and we are looking at the “average” which is non-resistant.

The summary statistics might not give us all the information we need to make a valid conclusion.

### Example

The mean teacher salary in the state of California is \$84,531. This lets us know that teachers are making above the median income for everyone in California, which is \$78,672. Is this a valid conclusion or are the summary statistics not telling the whole story?

The mean is non resistant which means some places have more well paying districts or more experienced teachers.

This is skewed right.

Being a statistics student means you must report both a visual display of the data as well as detailed summary statistics when asked to “describe the distribution”.

First: Create your Graph (If it is not given)

- Is your data categorical? Use a bar graph!
  - Two variables will require a segmented bar graph or mosaic plot.
- Is your data quantitative?
  - Discrete? Use a dot plot, stem plot, or boxplot.
  - Continuous? Use a histogram or boxplot.
  - Two variables will require a scatterplot.

Cumulative frequency graphs will usually take too long to create on the AP exam, but remember if you are interested in percentile positions, then you can create these graphs as well!

Second: Summarize your findings.

- Shape - Skewed, Symmetric, Bimodal, etc.
  - Be as specific as possible and combine shape terms if you can.
- Outliers - an individual observation that falls outside the overall pattern of the graph.
  - You will just have to comment on if there are any visual outliers. You do not have to do the outlier test unless instructed.
- Center - Mean and Median
  - Use the mean, unless the distribution is skewed or has outliers.
  - Unless asked for otherwise, a verbal description of the center is fine.
- Spread - Range, Standard Deviation, IQR
  - If you describe the center with the mean, comment on the standard deviation.
  - If you describe the center with the median, comment on the range or IQR.

When the problems asks to “compare the distributions”, follow the steps above, but make sure you are using comparison words to compare the distributions.

**Example**

The following data are for two popular songs on the Billboard Top 100. The length of each word in the song was recorded and below shows the number of words with the corresponding number of letters.

<i>Sweet Child O'Mine by Guns and Roses</i>										
Length of Word	1	2	3	4	5	6	7	8	9	10
Number of Words	9	109	42	51	47	2	6	2	1	1

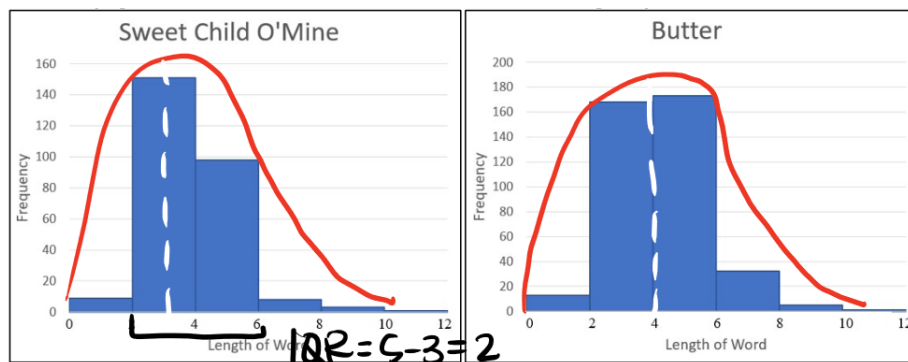
<i>Butter by BTS</i>										
Length of Word	1	2	3	4	5	6	7	8	9	10
Number of Words	13	77	91	137	36	26	6	3	2	1

(a) Determine the five number summary for each data set.

Sweet Child: min = 1, Q1 = 2, med = 3, Q3 = 4, max = 10

Butter: min = 1, Q1 = 3, med = 4, Q3 = 4, max = 10

(b) Below are the graphs of the two distributions. Write a few sentences comparing the distributions.



- Both distributions of word lengths are right skewed.
- Neither distribution has visible outliers.
- The median of Sweet Child is 3 letters which is smaller than the median of Butter which is 4 letters.
- The IQR of Sweet Child is 2 letters which is larger than the IQR of Butter which is 1 letter.

(c) There are two “rules” we used to mathematically determine outliers. Method A is the  $1.5 \times IQR$  rule and Method B is the two standard deviation rule.

(i) Using method A, determine the outliers that are present in the Sweet Child O'Mine distribution. Justify your answer.

The lower bound is -1, upper bound is 7. There are four outliers of 8, 8, 9, and 10 letters.

(ii) The mean number of letters in the Butter distribution is 3.62 and the standard deviation is 1.42. Using method B, determine the outliers that are present in the Butter distribution. Justify your answer.

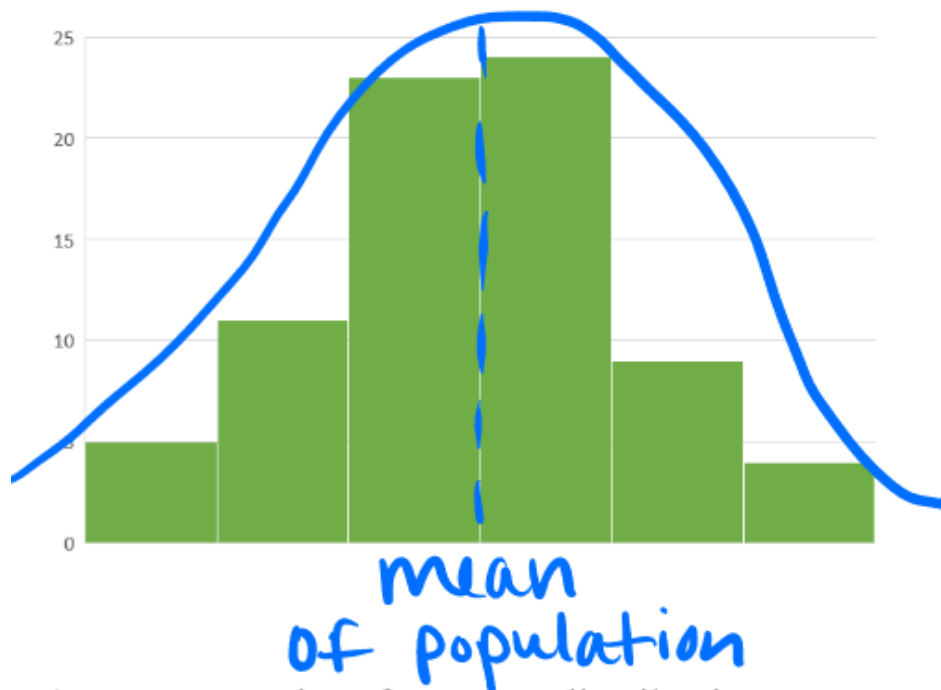
The lower bound would be 0.78 and the upper bound would be 6.46 in this. There are 12 outliers, 6 with 7 letters, 3 with 8 letters, 2 with 9 letters, and 1 with 10 letters.

(d) Explain why method A is better for determining outliers than method B in these distributions.

Method A is better because both distributions are right skewed. Method B uses statistics that are strongly affected by outliers.

## 1.5 Z-Scores and the Empirical Rule

The normal distribution is bell-shaped, and we draw a density on curve on the graph below. You can see that this curve will approximate the data, not describe it perfectly.



Density curves, like distributions, come in many shapes. A density curve is often a good description of the overall pattern of a distribution.

Normal distributions are appropriate for many distributions whose shapes are unimodal and approximately symmetric.

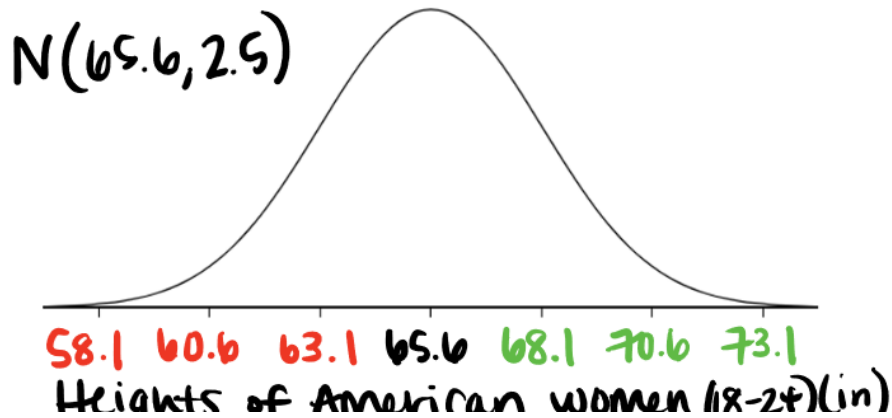
The mean is written as  $\mu$ , the standard deviation is  $\sigma$  and the normal distribution is described as  $N(\mu, \sigma)$

A normal distribution density curve has the following properties:

- Symmetric around the mean
- Mean = med = mode
- 50% of observations are less than the mean
- 50% of observations are greater than the mean
- The standard deviation tells us how measurements for a group of observations are spread out from the mean
- In a normal distribution, approximately all of the data lies within 3 standard deviations above and below the mean

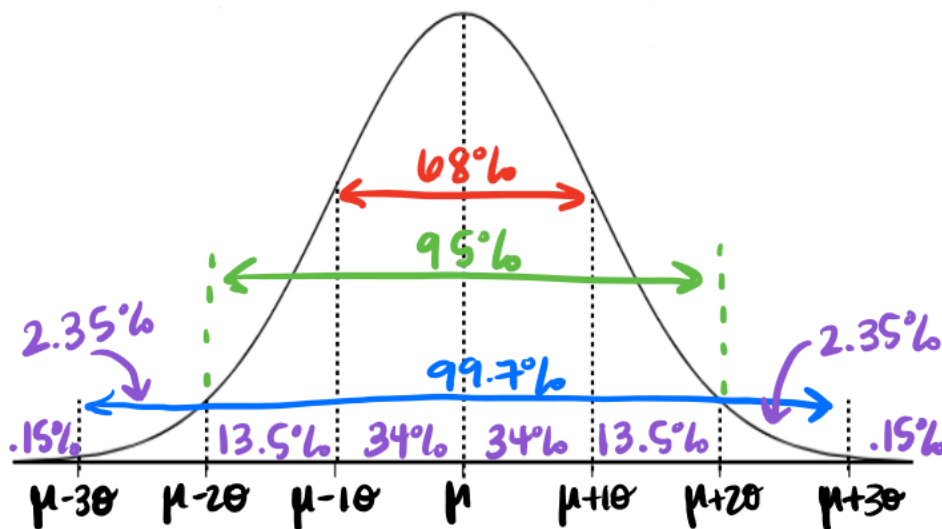
**Example**

The heights of American women aged 18 to 24 years old with a mean of 65.6 inches and a standard deviation of 2.5 inches. Assuming the population is normally distributed, draw a normal curve and label 3 standard deviations out on either side.



In a normal distribution, we can describe the variability of the data in terms of probabilities.

- 68% of data is likely to be within 1 standard deviation around the mean
- 95% of data is likely to be within 2 standard deviations around the mean
- 99.7% of data is likely to be within 3 standard deviations around the mean



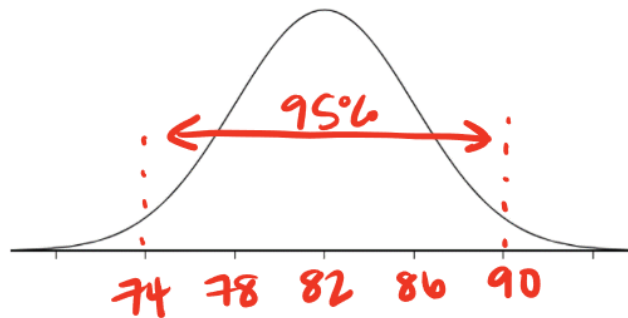
These three percentages 68-95-99.7 describe the empirical rule. These approximations can be used with the symmetry of a normal distribution to approximate proportions of data.

**Example**

For the following draw a normal distribution and use the empirical rule to answer the question.

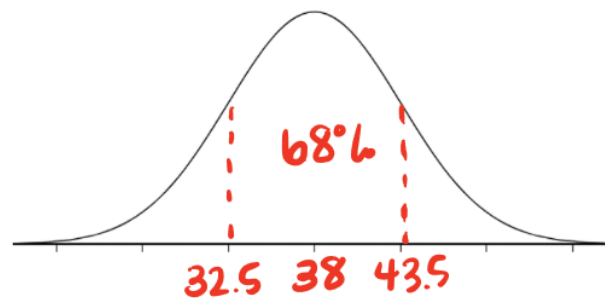
(a) The mean test score on the Unit 1 Test was an 82 with a standard deviation of 4. 95% of the test scores were between what two scores?

74 and 90.



(b) The mean age at the concert was 38 years old with a standard deviation of 5.5 years. What percent of the concert goers were between 32.5 years old and 43.6 years old?

68%

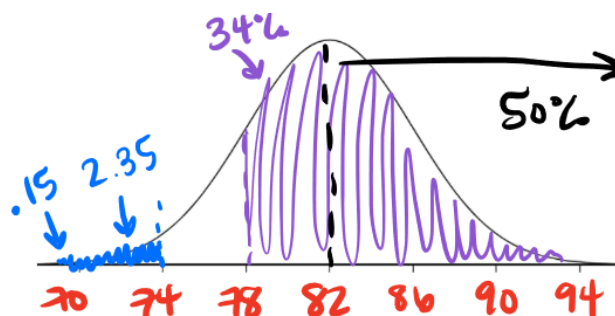


(c) The mean test score on the Unit 1 Test was a 82 with a standard deviation of 4. (i) What percent of test scores fell below a 74?

2.5%

(ii) What percent of test scores fell above a 78?

84%



The location of an observation in relationship with its mean is one that is asked often. So often, in fact, we have a method of calculating how far away an observation is from its mean. This method is called standardizing, and it is the process of counting how many standard deviations above or below the mean an observation falls.

Let's say you scored a 86% on a test, and 83% was the class mean with a standard deviation of 4.5%.



Subtracting your score from the mean gives the distance from the mean, and to determine if this distance is “typical”, divide by the standard deviation to see how many standard deviations it is away from the mean. The number you get is the number of standard deviations from the mean an observation is, or a z-score.

The z-score would be 0.67 in this case, or your score would be 0.67 standard deviations above the mean.

If  $x$  is an observation from a distribution that has known mean and standard deviation, the z-score for  $x$  is:

$$z = \frac{\text{value-mean}}{\text{SD}} = \frac{x - \mu}{\sigma}$$

- A z-score tells us how many standard deviations from the mean an observation falls and in what direction.
- Observations larger than the mean have positive z-scores.
- Observations smaller than the mean have negative z-scores.

#### Example

Female shoe sizes are approximately normally distributed with a mean of 8 and a standard deviation of 0.75. Find and interpret the z-score of a female with a size 6 shoe.

The z score can be calculated as -2.67. A female with a size 6 shoe is 2.67 standard deviations below the mean women shoe size.

#### Example

Joe is 74" tall. The average height of an adult male is 70" with a standard deviation of 2.5". Find and interpret the z-score for Joe's height.

The z score is 1.6. Joe is 1.6 standard deviations above the mean height of adult males.

#### Example

Corbin scored 1100 on the SAT in 2020. The distribution of SAT scores in 2020 was normally distributed, with a mean score of 1083 and a standard deviation of 194. Corbin also took the ACT in 2020 and earned a score of 23. The 2020 ACT scores were normally distributed with a mean of 21 and a standard deviation of 1.62. On what standardized test did Corbin scored better on?

The z-score for the SAT is 0.09 and for the ACT it is 1.23. Corbin scored better on the ACT because he was more standard deviations above the mean.

## 1.6 The Standard Normal Curve

How do we determine percentages if they do not follow the Empirical Rule? We standardized the normal curve.

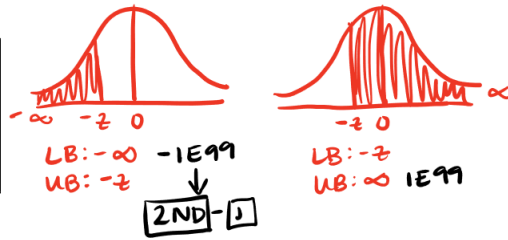
The curve we created is called the standard normal curve and any data set that is normally distributed can be standardized into the standard normal curve.

If you have a standard curve, we have a standard set of measurements we can use to determine what percentage of data falls above, below, or between specific z-scores.

We can determine what percentage of data falls above, below, or between specific z-scores using our TI-84 calculator.

2ND → VARS → 2:normalcdf(

- Lower: ~~z-score~~ or  $-\infty$
- Upper: ~~z-score~~ or  $\infty$
- $\mu$  (mean): 0
- $\sigma$  (standard deviation): 1

**Example**

- (a) What percent of data is below a z-score of 0.54?  
70.54%
- (b) What percent of data is above a z-score of 0.17?  
43.25%
- (c) What percent of data lies between  $z = -1.64$  and  $z = 2.33$ ?  
93.96%
- (d) What percent of data lies outside  $z = -1.64$  and  $z = 2.33$ ?  
6.04%

**Example**

On a typical Saturday, Kroger reports that the mean amount of money spent by customers is \$27.21 and a standard deviation of \$7.93. The distribution is approximately normal.

- (a) What percentage of people spend less than \$20?  
The z score is -0.91, so this is the upper bound. Normalcdf gives us 18.14%.
- (b) What percentage of people spend above \$40?  
The z-score is 1.61, so the percentage is 5.37%.
- (c) How many people spend between \$25 and \$30?

Note: On a MCQ question, you do not need to translate these values to z-scores. If you use the original values for the mean and standard deviation and use the numbers on this question as the bounds, you get an equivalent answer. The percentage is 24.73%.

**Example**

On a typical school night, students get an average of 8 hours of sleep with a standard deviation of 1.2 hours. How many hours of sleep would a student need to get to be in the 75th percentile?

Step 1: Find the z-score.

The 75th percentile means at or below 75%, the z-score would be expected to be positive.

2ND → VARS → 3:invNorm(

- Area (to the left):
- $\mu$  (mean): 8
- $\sigma$  (standard deviation): 1.2

**Note:** This returns the z-score that corresponds to this percentile.

The z-score is 0.67.

Step 2: Find the raw data.

Take your z-score and the mean and standard deviation of your distribution, and use the z-score formula to solve for  $x$ , the raw data value.

Solving for  $x$  in  $0.67 = \frac{x-7}{1.2}$  gives  $x = 7.804$  hours of sleep.

**Example**

If  $z = -2.3$ ,  $\mu = 23$ , and  $\sigma = 5$ , what is your  $x$ -value?

Do the same process as above, and you get  $x = 11.5$ .

**Example**

(a) If you have 35% of the data lying symmetrically in the middle of your distribution, what are your z-scores?

If you have 35% of the data in the middle, the z-scores will be the same value, but opposite signs due to symmetry. The percentages of each tail is  $1 - 0.35 = 0.325$ .

Using invNorm to get the z-score gives  $z = -0.45$  and  $z = 0.45$ .

**Example**

(a) On the most recent test, a student scored in the 40th percentile. The mean of the test scores was an 85 and the standard deviation was 1.5. What was the student's score?

Using `invNorm` gives  $z = -0.25$ , and finding for  $x$  gives  $x = 84.625$ .

(b) An automobile dealer finds that the average price of a previously owned vehicle is \$8,256. He decides to sell cars that will appeal to the middle 60% of the market in terms of price. Find the maximum and minimum prices of the cars the dealer will sell. The standard deviation is \$1,150 and the variable is normally distributed.

The tails would be  $1 - 0.60 = 0.40/2 = 0.20$ . Using `invNorm` for this gives  $z$  scores of  $-0.84$  and  $z = 0.84$ , so the  $x$  values found would be 7290 and 9222.

The minimum price is \$7290 and the maximum price is \$9222.

## 2 Exploring Two-Variable Data

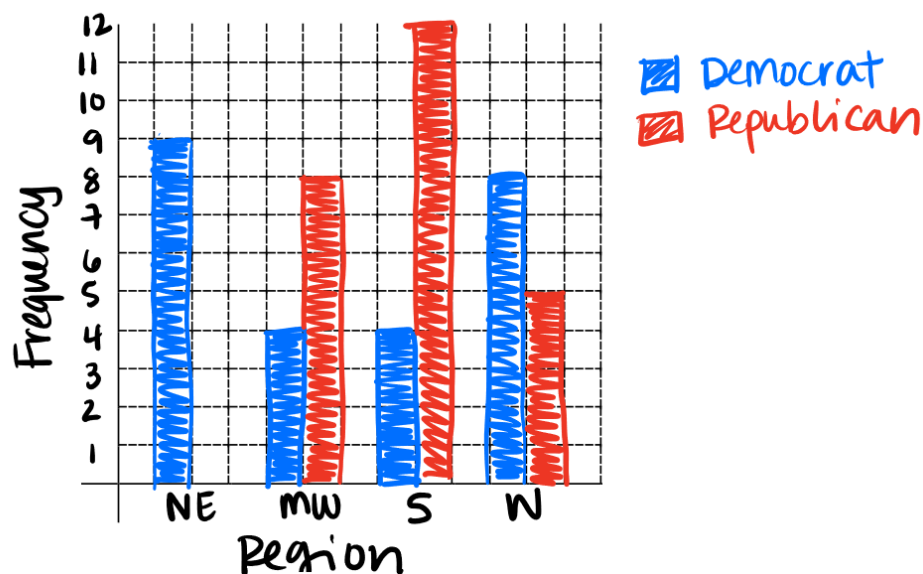
### 2.1 Two Categorical Variables

In Unit 1, we explored how bar graphs are a commonly used way to display categorical data. When we have two categorical variables to compare, we still use bar graphs, but they become a segmented bar plot or a mosaic plot.

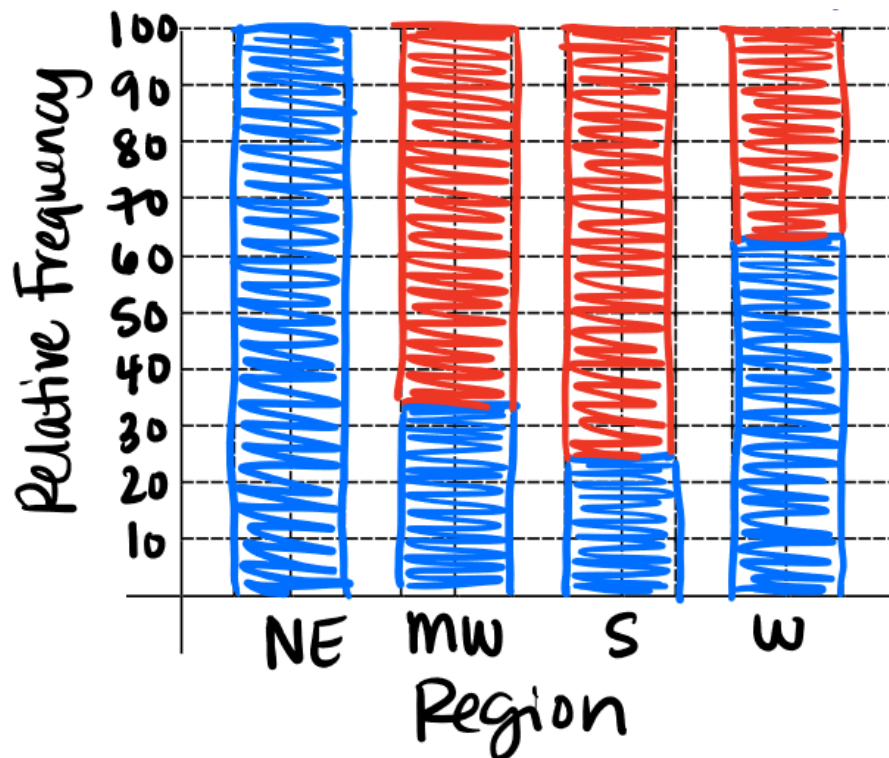
Here is data on the breakdown of state locations and how the state voted in the 2020 election.

	Region				Total
	Northeast	Midwest	South	West	
Democrat	9	4	4	8	25
Republican	0	8	12	5	25
Total	9	12	16	13	50

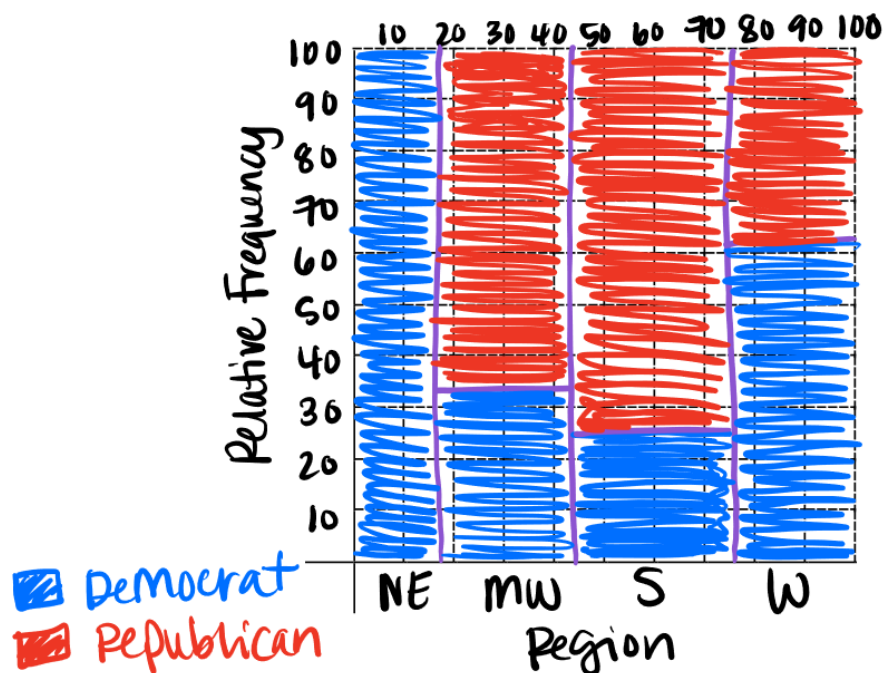
A side by side bar graph merges two bar graphs into one, in an attempt to compare the distributions of the two categorical variables.



A segmented bar graph is another way to display the data, where each group is split up by its relative frequency.



A mosaic plot is similar to a segmented bar graph, but it draws attention to the sizes of each group.



When you are working with two categorical variables, most of the time, the raw data is given in the form of a two way table - it is a table listing two categorical variables whose values have been paired. Each set of numbers in a two-way table has a specific name.

Joint relative frequencies: ratio of the frequency in a cell and the total number of data values.

Marginal relative frequencies: ratio of the sum in a row or column and the total number of data values.

Conditional relative frequencies: ratio of a joint relative frequency and related marginal relative frequency.

**Example**

	Region				Total
	Northeast	Midwest	South	West	
Democrat	9	4	4	8	25
Republican	0	8	12	5	25
Total	9	12	16	13	50

Joint Relative Frequency: What are the percent of states that are in the Midwest and voted "Democrat"?  
 Answer: 8%

Marginal Relative Frequency: Percent of states that are in the South. Answer: 32%

Conditional Relative Frequency: Of the states that voted "democrat", the percent that is from the West. Answer: 32%

**Example**

The direction of a technical school was curious about whether there is a relationship between students who complete one of the school's most popular health sciences certificate programs and whether those students go on to complete more advanced studies in the health sciences within two years of completing the certificate program. She randomly selected 100 students who completed the program. Data collected on these students is shown in the table below.

		Completed more Advanced Studies		Total
		Yes $A$	No $A^c$	
Completed Most Popular Health Science Certificate Program	Yes $B$	35	25	60
	No $B^c$	5	35	40
	Total	40	60	100

$\cap$  = and

Note we are letting  $A$ ,  $A^c$ ,  $B$  and  $B^c$  correspond to the 4 different possibilities.

The four following questions are true or false questions.

(a) Being a person who completed more advanced studies is more likely than being a person who did not complete more advanced studies.

False, 40 is not greater than 60.

(b) Being a person who completed the program is less likely than being a person who did not complete the program.

False, 60 is greater than 40.

(c) Being a person who completed the program and completed more advanced studies is less likely than being a person who did not complete the program and did not complete more advanced studies.

False, 35 is equal to 35.

(d) Being a person who did not complete the program but completed more advanced studies is less likely than being a person who completed the program and completed more advanced studies.

True

(f) Being a person who completed the program but did not complete more advanced studies is more likely than being a person who did not complete the program and did not complete more advanced studies.

False, 25 is not more than 35.

**Example**

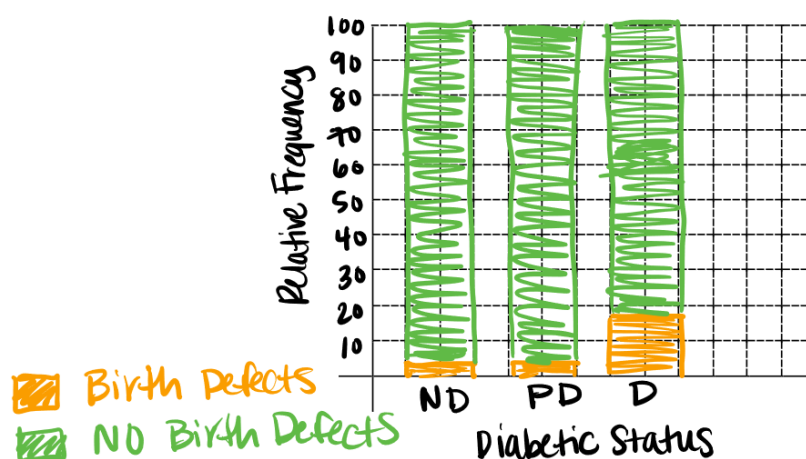
A 1968 sample study among the Pima Indians of Arizona investigated the relationship between a mother's diabetic status and the appearance of birth defects in her children. The results appear in the two-way table below.

		Diabetic Status			Total
		Nondiabetic	Prediabetic	Diabetic	
Birth Defects	No	754	362	38	1154
	Yes	31	13	9	53
Total		785	375	47	1207

(a) Compute the conditional distributions of birth defects for each diabetic status.

For nondiabetic it is  $31/785$ , for prediabetic it is  $13/375$ , for diabetic it is  $9/47$ .

(b) Display the conditional distributions in a segmented bar graph. Don't forget to label your graph completely. Comment on any clear associations you see.



There is no association.

## 2.2 Scatterplots and Correlation

We have worked with data for bar graphs, box plots, dot plots, and histograms. The type of data that is represented with these graphs are univariate data.

When we compare two variables (bivariate data), we are exploring the relationship between them.

Most statistical studies involve more than one variable. Often in the AP Statistics exam, you will be asked to compare two data sets by using side by side boxplots or histograms etc. However, there are times where we want to examine relationships among several variables for the same group of data.

When you examine the relationship between two variables you need to start with a scatterplot.

A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot fixed by the values of both variables for that individual.

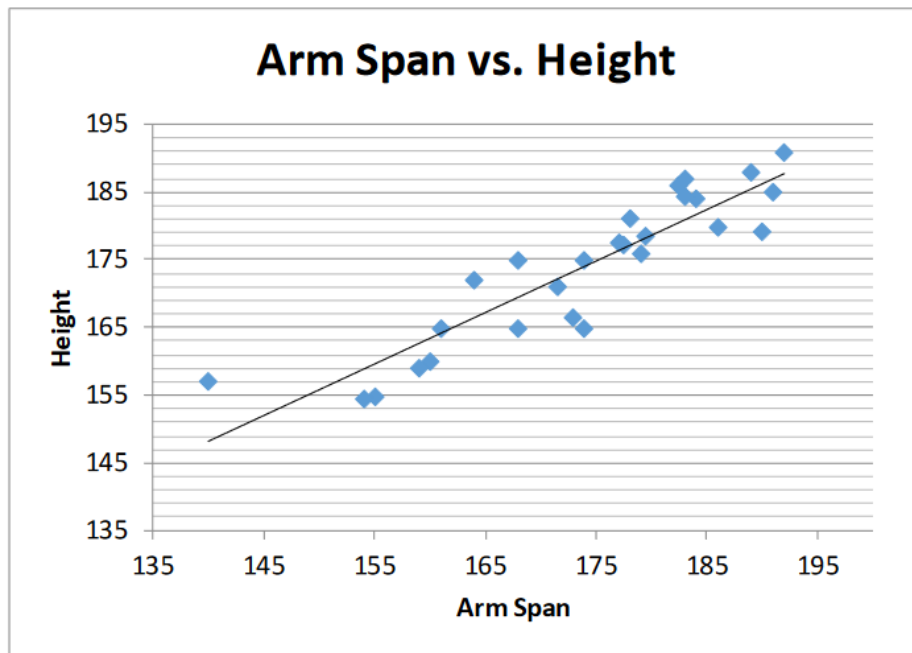
Here is a scatterplot representing Arm Span vs Height:

First, we have to identify and name the correct variables used in this study.



A response variable measures the outcome of a study or an observation (y variable, dependent)

An explanatory variable helps explain or influences change in a response variable. (x variable, independent)



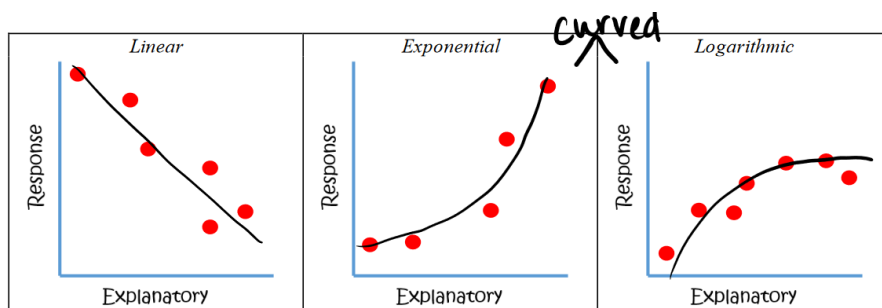
You will often find explanatory variables called independent variables, and response variables called dependent variables. The idea behind this language is that the response variable depends on the explanatory variable. Because the words independent and dependent have other, unrelated meanings in statistics, we won't use them here.

- In any graph of data, look for overall pattern and for striking deviations from that pattern.
- You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship.
- An important kind of deviation is an outlier, and individual value that falls outside the overall pattern of the relationship.

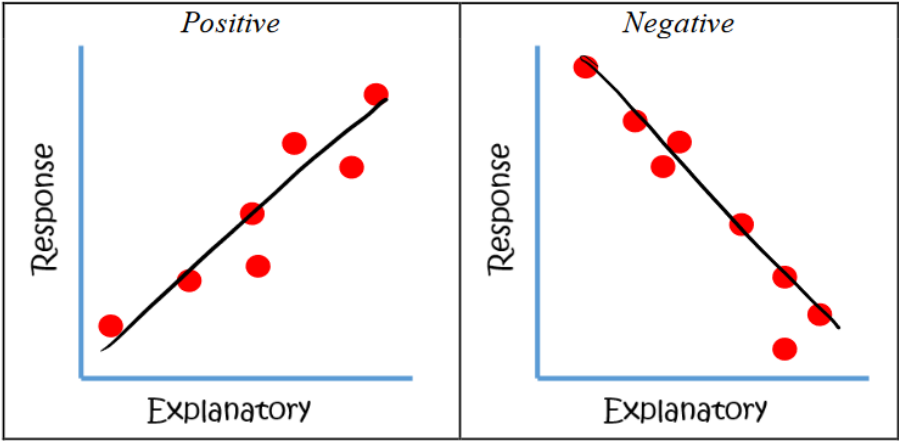
Things to look for in a scatterplot:

- Form: Overall pattern or deviations from the pattern
- Direction: Positive or negative slope
- Strength: How close do the points lie to a simple form

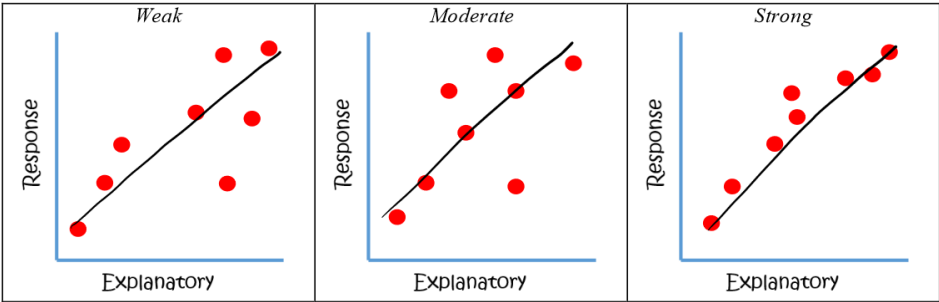
Examples of form:



Examples of direction:



Examples of strength:

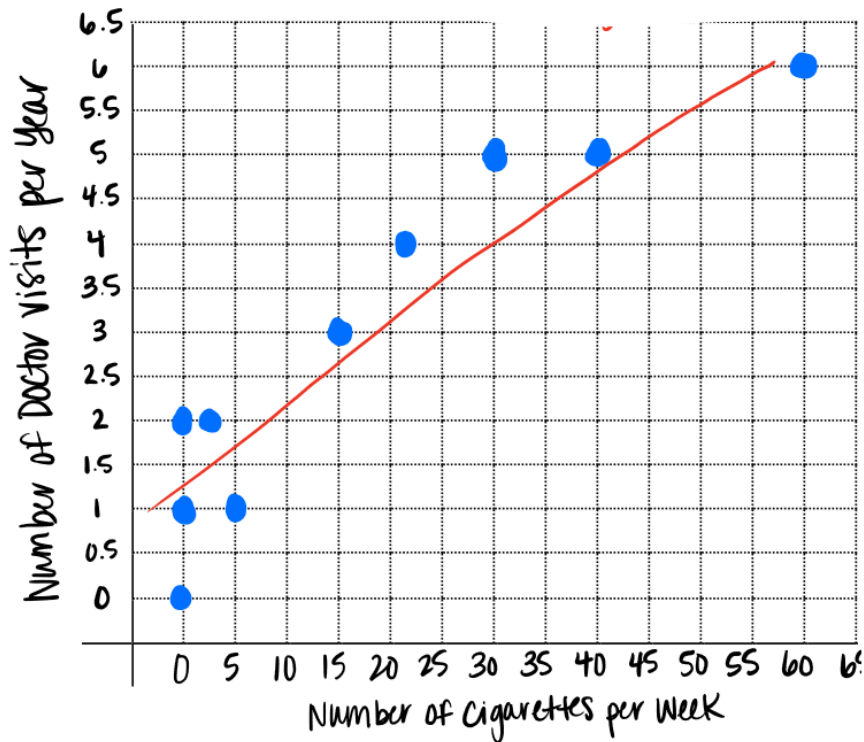


**Example**

Suppose we hypothesize that the number of doctor visits a person has can be explained by the amount of cigarettes they smoke. So we want to see if there is a relationship between the number of cigarettes one smokes a week and the number of times per year one visits a doctor. We ask 10 random people and get the following information:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0

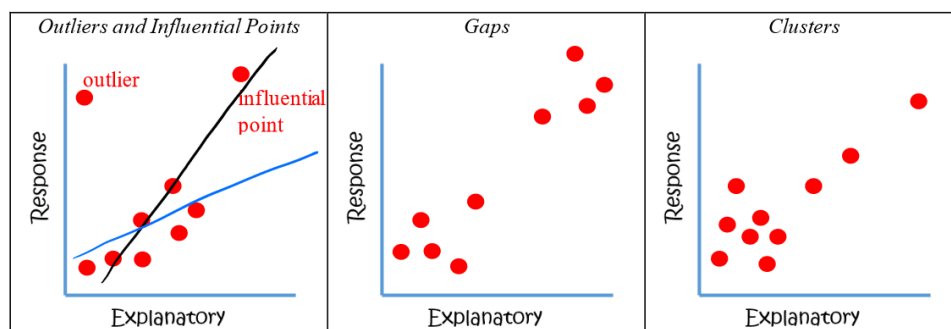
Creating a scatterplot gives us the following.



There is a strong, positive linear relationship between the number of cigarettes smoked per week and number of doctor visits per year.

If you wanted to create the previous example in a calculator, follow these steps.

1. Load the  $x$ -values into list 1 and the  $y$ -values into list 2.
2. Using StatPlot - highlight the mini-scatterplot: XList:L1 and YList:L2
3. Press "graph" but you will have to "Zoom 9" to fit the scatterplot on the screen

**Unusual Features**

In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient ( $r$ ), to describe the linear relationship between two variables. This number helps remove any subjectivity in reading a linear scatter plot.

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

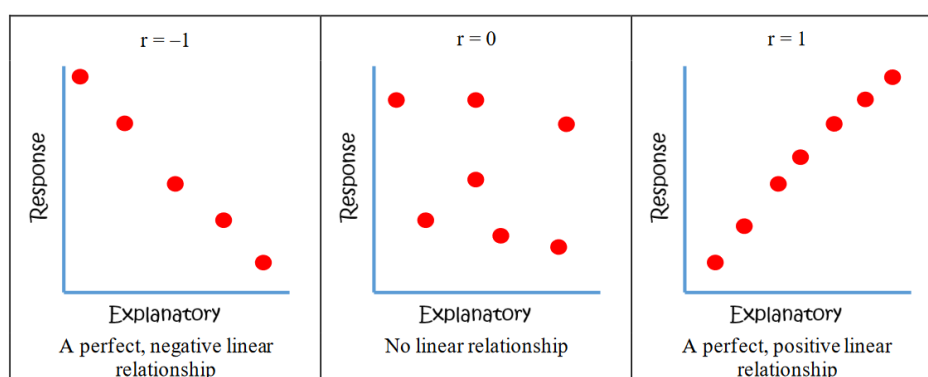
While we will never have to find correlation by hand, the formula is provided to us on the AP Statistics formula sheet. There are a few facts about the correlation that the formula can help us remember.

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

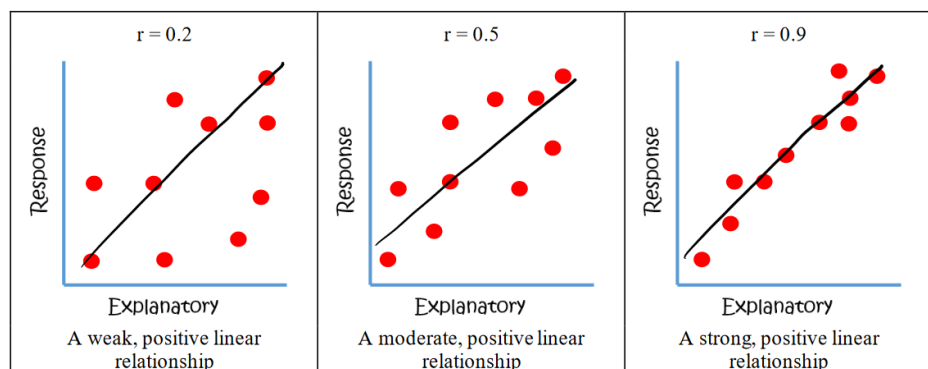
where  $r$  is the correlation,  $x_i$  is each x-value,  $y_i$  is each y-value,  $\bar{x}$  is the mean of the x values,  $\bar{y}$  is the mean of the y-values,  $s_x$  is the standard deviation of x values,  $s_y$  is the standard deviation of y values.

Essentially, the correlation coefficient,  $r$ , finds the average of the product of the standardized scores.

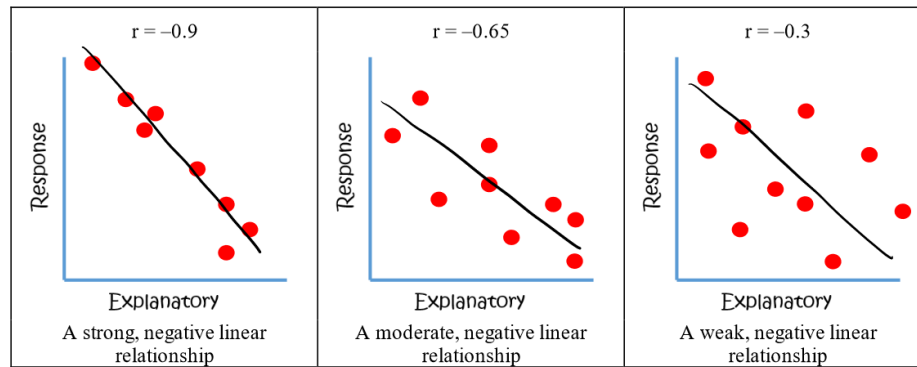
Correlation is a number that is between -1 and 1.



Positive correlations between 0 and 1 have varying strengths, with the strongest positive correlations being closer to 1.



Negative correlations between -1 and 0 have varying strengths, with the strongest negative correlations being closer to -1.



Correlation describes only linear relationships between two variables. For example, a quadratic relationship would have a correlation of 0 because it is not a linear relationship.

Correlation does not have units and changing units on either axis will not affect correlation.

If you look at the formula from above for correlation, since we are standardizing all the  $x$  and  $y$  values, it does not matter what the units are. We take the product of their standardized scores.

Switching the explanatory and response variables on the axes will not change the correlation.

From the formula, this is because the order of the multiplication does not matter. Correlation makes no distinction between explanatory and response variables. It makes no difference which variable you call  $x$  and which you call  $y$  when calculating the correlation.

Correlation is very strongly affected by outliers.

Use correlation with caution when outliers appear in your scatter plot. Don't rely on correlation alone to determine the linear strength between two variables - graph a scatter plot first.

## 2.3 Linear Regression

Least Squares Regression or linear regressions allows you to fit a line to a scatterplot in order to be able to better interpret the relationship between two variables, as well as make predictions about our response variable.

The fitted line is called the line of best fit, linear regression line, or least squares regression line, (LSRL) and has an equation in a form that should look very familiar:

$$\hat{y} = a + bx$$

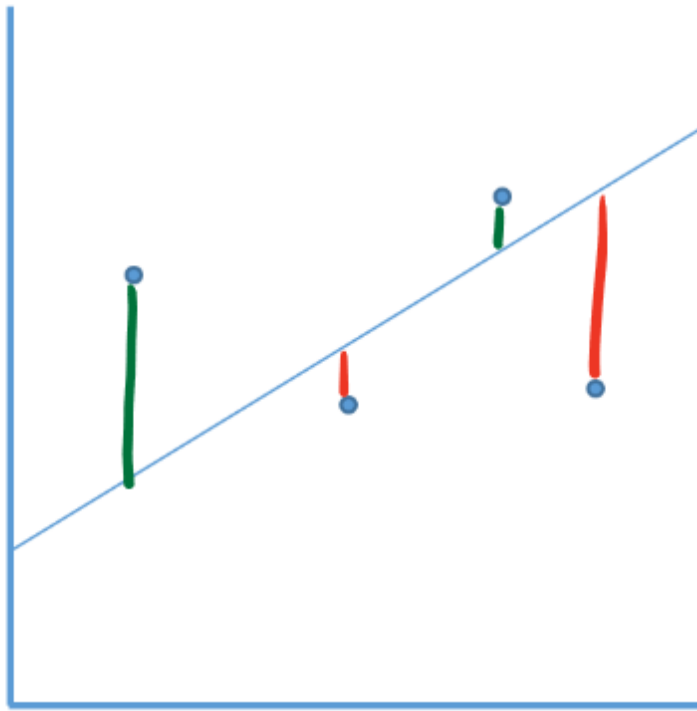
where  $\hat{y}$  is the predicted  $y$ -value,  $x$  is the explanatory variable,  $b$  is the slope,  $a$  is the  $y$ -intercept.

Another way the LSRL is written is  $\hat{y} = ax + b$ .

Slope is always the coefficient of  $x$  and it may be called "a" or "b".

The way the line is fitted to the data is through a process called the method of least squares. The main idea behind this method is that the square of the vertical distance between each data point and the line is minimized.

- If you add all of the vertical distances from the point to the line you get 0.
- We square each distance to make it positive, and then add it all up.



**Slope:** The slope of the regression line is important in the sense that it gives us the change of  $y$  with respect to  $x$ . In other words, it gives us the amount of change in  $y$  when  $x$  increases by 1.

**Intercept:** The intercept is statistically meaningful only when  $x$  can actually take values close to zero. When it does make sense to have a  $x$ -value of zero, the  $y$ -intercept is the  $y$ -value we would expect.

When we have a data set  $(x, y)$ , we can calculate the LSRL by hand or with technology.

**Example**

Many schools require teachers to have evaluations done by students. A study investigated the extent to which student evaluations related to grades. Teacher evaluations and grades are both given on a scale of 100. The evaluation score ( $y$ ) of a teacher from 10 of her students are given below with the average for each student ( $x$ ).

<b>x</b>	40	60	70	73	75	68	65	85	98	90
<b>y</b>	10	50	60	65	75	73	78	80	90	95

(a) Create the LSRL by hand.

Step 1: Enter the data into your calculator.  $x$  values go into L1 and  $y$  values into L2.

Step 2: Find the slope of your LSRL. The slope =  $r \left( \frac{s_y}{s_x} \right)$ . In this case, it will be 1.33.

Step 3: Find the y-intercept of your LSRL. The y-intercept =  $\bar{y} - \text{slope} * \bar{x}$ . In this case, the y-intercept is -28.69.

Step 4: Create your equation:

$$\text{teacher eval score} = -28.69 + 1.33(\text{student score})$$

(b) Use your equation to predict what evaluation Mrs. H will get from a student who scored a 81.

When you plug in 81 into the equation you made for student score, the estimated score will be 79.04.

(c) Interpret the slope in the context of the problem.

As student score increases by 1 point, we predict teacher eval. score to increase by 1.33 points.

(d) Do you think student grades and evaluations students give their teachers are related? Explain.

The correlation of 0.90 indicates a strong, positive, linear relationship between student score and TES.

LSRL on Calculator: There are two ways to get the line and it depends on how you like to write the line.

1. Stat  $\rightarrow$  Calc  $\rightarrow$  4: LinReg(ax+b)

2. Stat  $\rightarrow$  Calc  $\rightarrow$  8: LinReg(a+bx)

Be careful when making predictions beyond what the data shows.

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable  $x$  used to obtain the line. Such predictions are not accurate.

Correlation does not imply causation. For example, there was a study that showed a strong positive linear relationship between ice cream sales and homicides in New York City. Does this mean that if we stop selling ice cream, we will have no more homicides?

Coefficient of Determination

- The strength of a prediction which uses the LSRL depends on how close the data points are to the regression line.
- The mathematical approach to describing this strength is via the coefficient of determination ( $r^2$ )
- The coefficient of determination gives us the proportion of variation in the values of  $y$  that is explained by least-squares regression of  $y$  on  $x$ .
- The coefficient of determination turns out to be the correlation coefficient squared.

In the last example, the  $r$  value was 0.90. The coefficient of determination would be this number squared, so 81%.

Whenever you use the regression line for prediction, also include a measure of how successful the regression is in explaining the response.

This means that 81% of the variation in teacher evaluations can be explained by the linear relationship it has with the student class average.

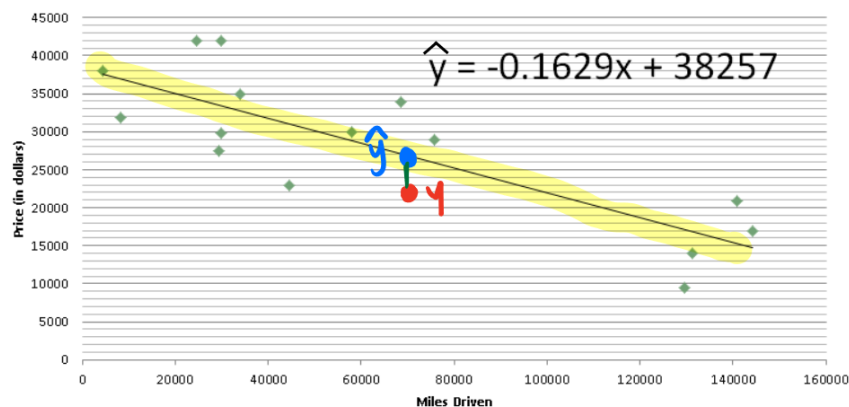
### Residuals

- In most cases, no line will pass exactly through all the points. This means that even if we use the LSRL to make predictions about our dependent variable, there will still be some error from the actual  $y$ -value.
- Because we use the line to predict  $y$  from  $x$ , the prediction errors we make are errors in  $y$ , the vertical direction in the scatterplot.
- A good regression makes the vertical deviations of the points from the line as small as possible.
- A residual is the difference between an observed value of the response variable and the value predicted by the regression line.
- Residual = observed - predicted OR  $\text{residual} = y - \hat{y}$ .
- If the residual is positive, the observed point lies above the least squares regression line.
- If the residual is negative, the observed point lies below the least squares regression line.

If you add up all the residuals from your data, you will get "0". That is why the LSRL involves squaring the residuals then adding them up and minimizing that value.

### Example

Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4x4 if we know how many miles it has on the odometer? A random sample of 16 used trucks was selected from autotrader.com. Here is a graph of the data.



Find and interpret the residual for the Ford F-150 that has 70,583 miles driven and a price of \$21,994.

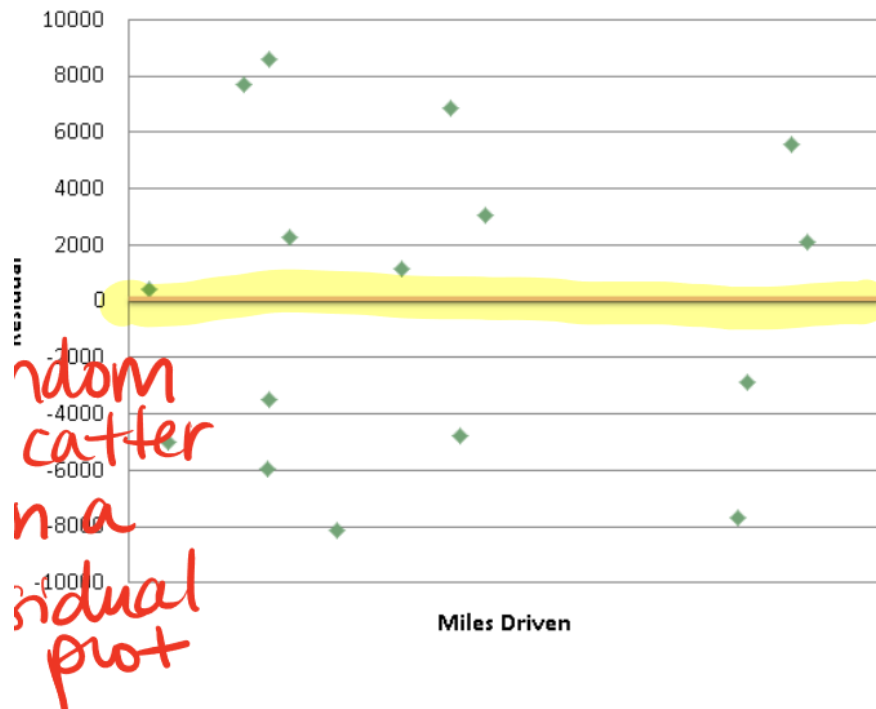
Use the LSRL equation given, and we can get  $\hat{y} = 26759.03$ . The residual is  $21994 - 26759.03 = -4765.03$ .

For a truck with 70,583 miles, our linreg overestimates the cost by \$4,765.03.

### Residual Plots

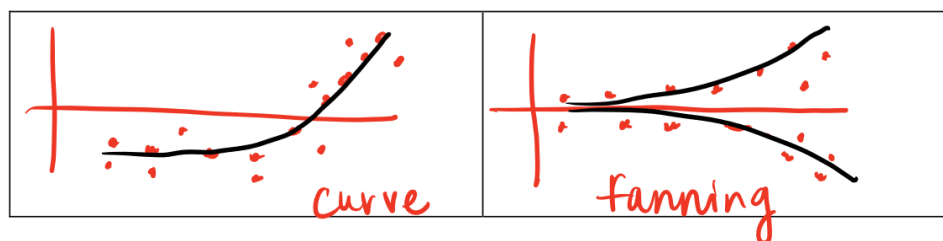
- A residual plot makes it easy to study the residuals by plotting them against the explanatory variable.
- Residual plots help us assess whether a linear model is appropriate.
- In the truck example, if we calculate all the residuals for each plot, we can then plot the miles driven vs. residuals.





- Essentially, a residual plot turns the regression line horizontal.
- Residual plots magnify the deviations of points from a line.
- This makes it easier to see an unusual pattern, so it helps us determine if a linear model is appropriate.

When an obvious pattern exists in a residual plot, the model we are using is not appropriate. For example



The TI-83/84 will generate a complete list of residuals when you perform a LinReg. They are stored in a list called RESID which can be found in the LIST menu. RESID stores only the current set of residuals. That is, a new set of residuals is stored in RESID each time you perform a new regression.

In order to draw a residual plot on the TI-83/84, first enter your data and perform a LinReg. Next, create a STAT PLOT where XList is L1 and YList is RESID (get this by pressing 2nd → STAT → 9:)

## 2.4 Influential Points and Departure from Linearity

Sometimes called “MiniTab”, computer output on the AP exam is a quick way to give statistical information on linear regression. The key is to remember what all the information stands for.

The image shows a regression output table with several handwritten annotations in green and red. Green arrows point from text labels to specific parts of the table: 'y intercept' points to the Constant's coefficient, 'slope (x variable identified)' points to the Age's coefficient, 'Standard deviation of residuals' points to the 's' value, and 'Coefficient of determination' points to the R-sq value. A red arrow points from 'SE(coef)' to the St Dev column. Two text boxes provide additional context: one states that St Dev, t ratio, and P values will be used in Unit 9, while the other states that R-sq(adj) should always be ignored in AP Statistics.

Predictor	Coef	St Dev	t ratio	P
Constant	3.371	1.337	2.52	.065
Age	2.1143	0.2321	9.11	.001

Below the table, the following statistics are listed:

- $s = 0.9710$
- R-sq = 95.4%
- ~~R-sq(adj) = 94.3%~~

Using the above information, we can also get the following:

LSRL:  $\hat{y} = 3.371 + 2.1143(\text{age})$  and  $r = \sqrt{0.954} = 0.9767$ .

The standard deviation of the residuals (" $s$ ") will be used more in Unit 9, but we can interpret it here. This value gives the approximate size of a "typical" prediction error (residual). Larger values of " $s$ " means our line is expected to give larger residuals. The units for the standard deviation of residuals is the same of the residuals (and of the  $y$ -values). So it depends on the data for what is considered a "large" residual error.

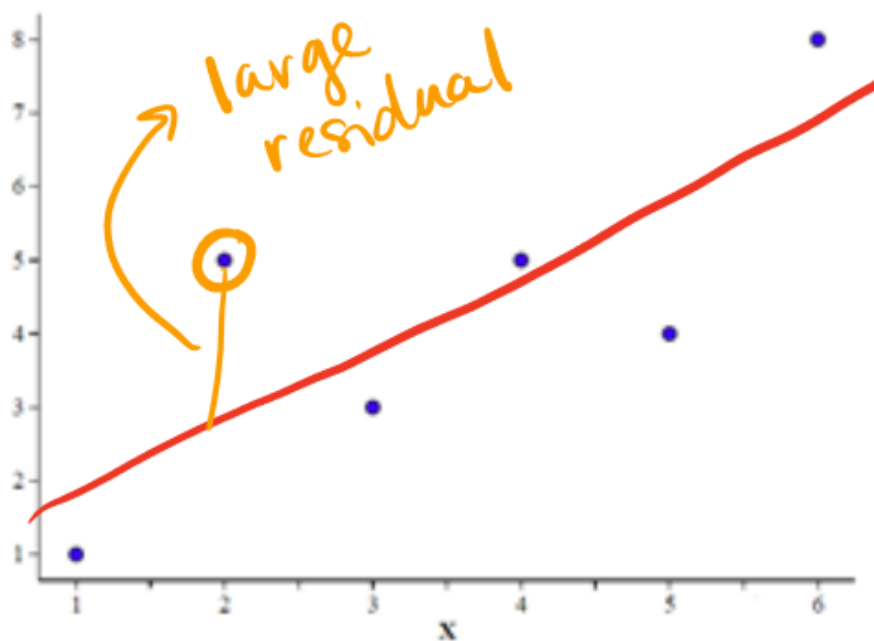
For example, if " $s$ " was 100, that would be a large value if our  $y$ -value was trying to predict age in years but it would be a small residual if we were trying to predict age in seconds. Context is important when interpreting this value.

An influential point in a data set, is a point that has leverage on the correlation and regression line. In other words, when removed, this point changes the regression line substantially. An influential point might be considered an outlier if it does not fit with the overall pattern of the data, but an influential point might also fit the data, but change the regression line and/or correlation significantly when removed.

**Example**

Here is a set of hypothetical data and its scatterplot.

X	1	2	3	4	5	6
Y	1	5	3	5	4	8



The LSRL:  $\hat{y} = .9333 + .9714x$  and the  $r$  is .777.

Suppose we removed (2,5). The point is not considered an outlier because it fits the rest of the data, but we will examine the impact on the LSRL and correlation to see that it is influential.

The LSRL becomes  $\hat{y} = -0.473 + 1.2297x$  and  $r = 0.914$ .

The slope increased since that point has leverage and was pulling our LSRL towards it.

The y-intercept decreased since the slope increased, it moved the  $y$ -intercept lower.

The correlation ( $r$ ) also increased. The point (2,5) is not considered an outlier but since it had a large residual originally, removing it made the LSRL "fit" better.

The AP Statistics course deals only with two-variable data that can be modeled by a line OR nonlinear two-variable data that can be transformed in such a way that the transformed data can be modeled by a line.

To know what transformation to use, graph the data first. If the scatterplot does not show a linear pattern, or the residual plot pattern is not random, consider a transformation.

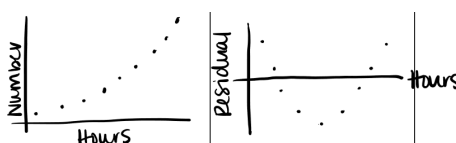
You can transform the independent variable, dependent variable, or both. There are a wide range of transformations you can do, but the AP exam will keep it simple. Here is an example of one of the most common transformations you will encounter.

**Example**

The number of a certain type of bacteria present (in thousands) after a certain number of hours is given in the following table.

Hours	Numbers
1.0	1.8
1.5	2.4
2.0	3.1
2.5	4.3
3.0	5.8
3.5	8.0
4.0	10.6
4.5	14.0
5.0	18.0

The following are the scatterplot and the residual plot.



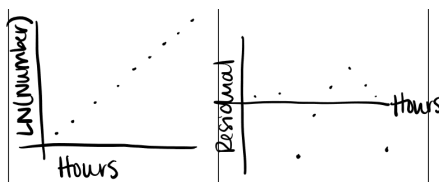
A linear model is not appropriate here because the scatterplot shows an exponential form and the residual plot is curved.

- What parent function does this scatterplot look like? Exponential!
- How do we undo exponents? Logs
- For this problem the natural logarithm will be the best way to transform the data to achieve linearity.
- On the AP exam, to save time, if you needed to do this by hand, they would tell you the transformation.

The data is now

Hours	Numbers	LN(Number)
1.0	1.8	0.59
1.5	2.4	0.88
2.0	3.1	1.13
2.5	4.3	1.46
3.0	5.8	1.76
3.5	8.0	2.08
4.0	10.6	2.36
4.5	14.0	2.64
5.0	18.0	2.89

The scatterplot and residual plot:



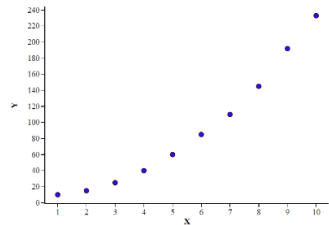
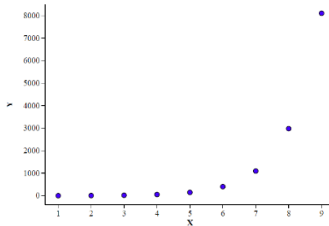
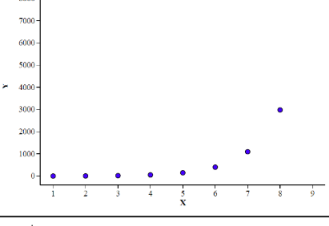
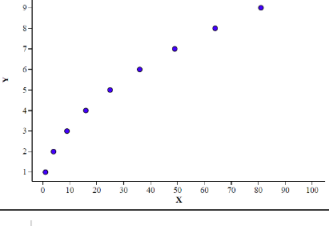
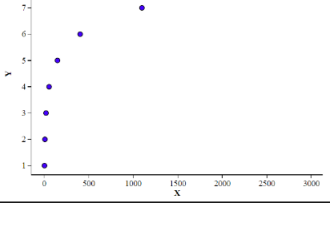
This is a better model for linear regression because the scatterplot shows a straight-line pattern and the residual plot shows random scatter.

The regression equation for the transformed data is  $\ln(\text{number}) = -0.0047 + .586(\text{hours})$ .

If we wanted to find the predicted quantity of bacteria after 3.75 hours, we can plug that number in and get 8,906.3 bacteria as the answer.

There are many types of transformations available and you can even get really crafty and add in some constants.

For AP Statistics, we keep it simple. Here are some common patterns.

Contains (0, 0) and appears to have a slight curve.	$(x, y) \rightarrow (\ln x, \ln y)$	
Contains a non-zero y-intercept and appears exponential.	$(x, y) \rightarrow (x, \ln y)$	
Contains a non-zero y-intercept and appears exponential.	$(x, y) \rightarrow (x, \ln y)$	
Contains (0, 0) and appears logarithmic.	$(x, y) \rightarrow (\sqrt{x}, y)$	
Contains a non-zero y-intercept and appears logarithmic.	$(x, y) \rightarrow (\ln x, y)$	

To test and see which model would be appropriate, you would compare the scatterplots,  $r$  and  $r^2$ , and the residual plots of both the original and transformed data.

# 3 Collecting Data

## 3.1 Planning a Study

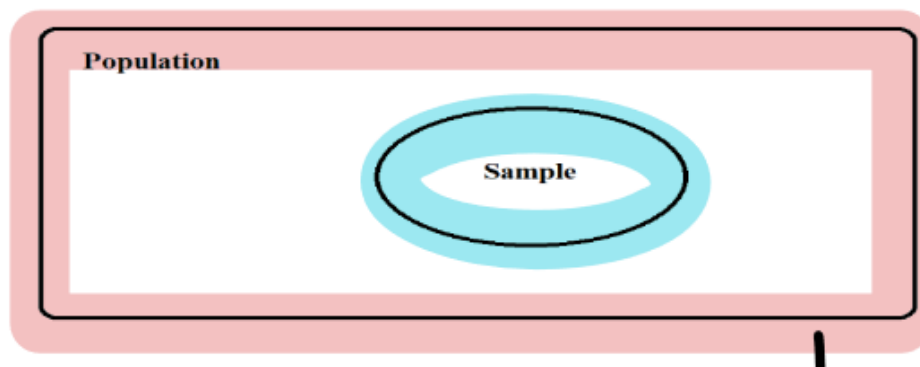
In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and make inferences based on the summary results from the sample.

Population - the entire group we want information about.

A population can be huge like “all women” or small like “top 200 grossing movies in 2023.”

Sample - a part of the population we actually examine.

The size of the sample can vary and depends on several factors we will examine through the course.



Census - collects data from every individual in the population.

The way we collect data influences what we can and cannot say about a population.

Observation Study - observes individuals and measures variables of interest but does not attempt to influence the responses.

- In an observational study, treatments are not imposed
- Investigators examine data for a sample of individuals (retrospective) or follow a sample of individuals into the future collecting data (prospective) in order to investigate a topic of interest about the population.
- A sample survey is a type of observational study that collects data from a sample in an attempt to learn about the population which the sample was taken

Experiment - deliberately imposes some treatment on individuals to measure their responses.

- We will discuss experiments in a later topic in this unit.

For now, let's discuss the various ways we plan an observational study.

It is only appropriate to make generalizations about a population based on samples that are randomly selected or otherwise representative of that population.

- A sample is only generalizable to the population from which the sample was selected.
  - For example, if you poll a sample of women asking about their shopping habits, you cannot take that result and apply it to men, since they were not represented in the sample that was taken.
- It is not possible to determine casual relationships between variables using data collected in an observational study.

- While we observe variables and gather data in an observational study, we cannot make inferences between the variables. As long as it is a well-designed observational study, we can only apply the findings from the sample to the population.

Because we make inferences about a population from the sample, it is very important that the sample is collected appropriately and that it is representative of the population being studied.

Convenience Sample:

- Definition: Uses subjects that are readily available.
- Advantage: Easy and less costly to collect
- Disadvantage: Not representative of the population
- Example: In order to get an idea of how students think of the new school policy, the principal stands outside the library and asks a few students their opinions.

Voluntary Response Sample:

- Definition: A sample obtained by allowing subjects to decide whether or not to respond.
- Advantage: Easy to collect
- Disadvantage: Overrepresents people with strong opinions
- Example: After the State of the Union speech, ABC tells its audience to call a 1-900-555-1234 if they thought the speech was good and 1-900-555-7890 if they thought the speech was bad (there is a \$0.50 charge for the call).

Simple Random Sample (SRS)

- Definition: Consists of  $n$  individuals from the population chosen in such a way that every set of  $n$  individuals has an equal chance in the sample selected.
- Advantage: Easy to accomplish using a table of random digits; likely to produce samples that are good representatives of the population
- Disadvantage: Cost or time could be an issue
- Example: In order to determine how happy students are with their education at a high school, the principal assigns each student a number from 1 to 1230 and then uses a random number generator to choose 50 numbers between 1 and 1230. He then surveys all the students with the chosen numbers.

Stratified Random Sampling

- Definition: Divide the population into groups of similar individual (strata) then select an SRS within each strata. Combine the SRSs from each strata to form your full sample.
- Advantage: Can produce more exact information (especially in large populations) by taking advantage of the fact that individuals in the same strata are similar to one another
- Disadvantage: Not appropriate unless strata are easily defined
- Example: In order to get a better idea of what a high school's athletes thought about homecoming last year, the director divides all the athletes into the teams they play for, and then selects a random sample from each sports team. His full sample consists of aggregating the random samples from each team.

Cluster Sampling:

- Definition: Divide the population into sectors (clusters) then randomly choose a few of those clusters. Each member of the cluster becomes your sample.
- Advantage: Don't need a list of entire population
- Disadvantage: More variability between samples depending on how clusters are determined.
- Example: A psychologist at the University of Texas collects a sample by first dividing up the students into their respective schools (engineering, nursing, arts and sciences) then by the departments that their major is in, and then she selects a few departments that their major is in, and then she selects a few departments at random and surveys every student within those chosen departments.

### Systematic Random Sampling

- Definition: Randomly select an arbitrary starting point and then select every  $k$ th member of the population.
- Advantage: Every member has an equal probability of being selected.
- Disadvantage: Not every sample of size  $n$  has an equal chance of being selected.
- Example: HP selects every 200th computer off the assembly line and inspects it for quality control.

When an item from a population can be selected only once, this is called without replacement. When an items from the population can be selected more than once, this is called with replacement.

Samples are biased if they are systematically not representative of the desired population.

- Bias occurs when certain responses are systematically favored over others.

Voluntary Response: When a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population (voluntary response bias).

- Example: A radio talk show host asks listeners to call in with their opinions of making wearing masks in public space mandatory.

Undercoverage: Occurs when some groups in the population are left out of the process of choosing a sample.

- Because they are generally fearful of government intrusion, many immigrants from Latin America did not return their census questionnaire during the 1990 census.

Non-response: Occurs when an individual chosen for a sample can't be contacted or refuses to respond. Non-response is a big problem in mail surveys.

- Example: Our administration sends out 100 survey questions to a sample of parents in order to gauge their attitudes towards returning to school in 2020. Only 23 respond.

Response Bias - bias caused by the behavior of the respondent or interviewer.

Untruthful Answers: people give untruthful answers for several reasons..

1. Sensitive Questions: How often do you drink alcohol?
2. Socially Acceptable Answers: Do you use corporal punishment with your children?
3. Interview Bias: One year after the Detroit race riots of 1967, interviewers asked a sample of black residents in Detroit if they felt they could trust most white people, some white people, or none at all. When the interviewer was white, 35% answered "most", when the interviewer was black, 7% answered "most".

Ignorance: people will give silly answers just so they appear to know something about the subject.

- Example: In a study, educators were asked how they would rank Princeton's undergraduate business program. In every case, it was rated among the top 10 departments in the country, even though Princeton doesn't offer an undergraduate business major.

Lack of Memory: giving a wrong answer simply because respondent doesn't remember the correct answer.

- Example: Students were asked to report their grade point averages. Researchers then determined the actual GPA's. Over 17% of the students reported a GPA that was .4 or more above their actual average, and about 2% reported a GPA more than .4 below their actual GPA. (most inflated their GPA's!)

Timing: When a survey is taken can have an impact on the answers

- Example: In January, the National Football League reported a poll that revealed football as the nation's favorite sport (this is at the time of the Super Bowl;)

Phrasing: Subtle differences in phrasing make large differences in the results.

Example:

- Should the president have the line-item veto to eliminate waste? 97% said "yes"
- Should the president have the line-item veto? 57% said "yes"



When drawing a sample, two types of errors may occur:

**Sampling Error:** The difference between a sample result and the true population result. This error results from chance variation.

**Example:** Place 50 red and 50 green balls in a bag. Mix the balls thoroughly and randomly sample 30 balls. In your sample you find that 12 balls are red and 18 are green. Your sample result is different than the true population ratio of 1 to 1. The difference is due to sampling error. Virtually any experiment involving a sample will have sampling error. We can minimize sampling error through various statistical techniques, the most obvious is to increase sample size.

**Non-sampling error:** Occurs when the sample data is incorrectly collected, recorded, or analyzed. Usually occurs when the sample is selected in a non-random fashion.

**Example:** In order to gauge student opinion on a new grading policy, an administrator stands outside the library during common time and asks a sample of 50 students if they agree with the new policy. The administrator finds that 25 out of the 50 students sampled agree with the new policy. When the entire student body is asked for their opinion, however, the results were 30% in favor 70% against. The difference between the sample percentage and the true population percentage is due to non-sampling error, because the sample was collected in such a way that a lot of bias was involved (convenience sampling).

## 3.2 Selecting a Random Sample

The Hat Method

- This is a classic description of an SRS that can be used on the AP exam to describe selecting a random sample but it is rarely done in practice due to it being so time consuming.
- Script: "Write down all the names or numbers on their own slip of paper. Then put all the pieces of paper into a hat, mix well in-between selections, and pull out the desired amount of slips (mention with or without replacement)"

Calculator - Random Number Generator

- MATH - PROB - 5:randInt(lower, upper, n)
- Can be used to describe on the AP Exam but not really used for "doing" random samples on the AP Exam because there is no way to "check" for it.
- In practice, computer generated numbers are the least time consuming.
- If using this method, make sure you "seed" your calculator, otherwise everyone will get the same "random numbers".
  - MMDDYYYY - STO→ - MATH - PROB - 1:rand-ENTER

Random Digit Table

- Given as a long string of digits 0-9 within the question. The digits are grouped in 5s to make it easier to read but it has no significant meaning.

**TABLE B** Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335

Choosing SRS with a Random Digit Table

1. Label: Assign a number label to every individual in the population.
2. Random Digits: Start at the very first number and identify how many digits you will take at a time.

3. Stop: Indicate when you should stop sampling (toss out repeated numbers or numbers out of your range).
4. Identify Sample: Use the random numbers to identify subjects to be selected from your population. This is your sample!

### Example

The school newspaper is planning an article on family-friendly places to stay over spring break at a nearby beach town. The editors intend to call 4 randomly chosen hotels to ask about their amenities for families with children. They have an alphabetized list of all 28 hotels in the town. Starting at Line 140 (given below), find an SRS of 4 hotels. Describe how you would select your SRS and then collect your sample.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 <u>Banyan Tree</u>	11 Diplomat	18 <u>Sandpiper</u>	25 Tradewinds
05 <u>Beach Castle</u>	12 <u>Holiday Inn</u>	19 <u>Sea Castle</u>	26 Tropical Breeze
06 Best Western	13 <u>Lime Tree</u>	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda

12975	13258	13048	45144	72321	81940	00360
✓ X	X X X	✓ ✓ X	X X	X X ✓		

1. Each hotel is given a number 01 to 28.
2. Choose 2 digits at a time. If the number is 01-28, it will represent a hotel in the sample.
3. Ignore numbers 29-99 and repeats. Repeat process until 4 hotels are chosen.
4. The checkmarks in the image are those to include in the sample, and the x's are skips.

The four hotels are 04, 12, 13, and 18.

Observational Study: Observes individuals and measures variables of interest but does not attempt to influence the response.

Experiment: Deliberately imposes some treatment on individuals to measure their response.

Experimental Unit: the things on which the experiment is done

Subjects: experimental units that are human beings

Treatment: a specific experimental condition applied to the units

**Example**

Are the following scenarios an experiment or an observational study? Explain your answer.

(a) A medical team examines the records of 5 large hospitals and compares the survival times of those cancer patients who had surgery versus those who had chemotherapy.

Observational Study - medical team did not impose any treatments

(b) In a gym class, the effect of exercise on blood pressure is studied by requiring that half of the students walk a mile each day while the other students run a mile each day.

Experiment - treatments are running or walking a mile.

(c) The relationship between weights of bears and their lengths is studied by measuring bears that have been anesthetized.

Observational Study - weights and lengths are recorded, no treatments

(d) People who smoke are asked to halve the number of cigarettes consumed each day so that any effect on pulse rate can be measured.

Experiment - treatment is halving cigarettes

**Example**

Determine if it is an observational study or an experiment, and then identify the explanatory and response variables in each situation.

(a) One effect of alcohol is a drop in body temperature. To study this effect, researchers give several amounts of alcohol to mice, and then measured the change in each mouse's body temperature.

Experiment, explanatory: amount of alcohol, response: body temperature

(b) A study is done to try and find the correlation between verbal and math SAT scores. The scientist wants to use the verbal score to predict the math score.

Observational Study, explanatory: verbal SAT score, response: math SAT score

(c) Some breast cancer patients were given each a new treatment. The patients were closely followed to see how long they lived following surgery.

Experiment, explanatory: "new treatment", response: length of life

(d) To find out how well a child's height predicts their age a study was done where they measured the heights of a group of children at age 6, wait until they are 16 and then measure their heights again.

Observational study, explanatory: height, response: age

### 3.3 Experimental Design

Two advantages of an experiment over an observational study:

1. We can study the specific factors we are interested in while controlling the effects of lurking variables.
2. Experiments also allow us to study the combined effects of several factors.

How to design an experiment

- Factor: The explanatory variables in an experiment
- Level: the various groups the factors take

For example, if an experiment compared the drug doses 50 mg, 100 mg, and 150 mg, then the factor "drug dosage" would have three levels: 50 mg, 100 mg, and 150 mg

Principles of Experimental Design

### 1. Comparison

- We want to make sure we are using a design that compares two or more treatments
- We need to make sure the groups we are comparing don't differ greatly before our experiment begins or bias can result.

### 2. Randomization

- The most important element of any experiment. It must be incorporated either in the selection process of experimental units and/or the distribution of experimental units into treatment and control groups.
- You can use your calculator (random number generator), random digit table, the hat method, or flipping a coin to randomize an experiment.
- Randomization produces groups of experimental units we expect to be similar in all respects before treatment is applied. Therefore, measured differences must be due either to treatment or by change of random assignment.

### 3. Control

- Control group is treated identically in all respects to the group receiving the treatment except that the members of the control group do not receive the treatment.
- The control group is our baseline and our experimental group has only one thing changed - the explanatory variable.
- This reduces variability in the response variable. If one group is controlled, we would expect their responses to be controlled as well.

### 4. Replication

- Use enough experimental units in each group so that any difference in the effects of the treatments can be distinguished from chance differences between groups.
- Even with control, natural variability occurs among experimental units.
- We would like to see units within a treatment group responding similarly to one another, but differently from units in other treatment groups (then we can be sure that the treatment is responsible for the differences).
- If we assign many individuals to each treatment group, the effects of chance (and individual differences) will average out.

### Experimental Terms

- Placebo: treatment designed to have no therapeutic value
- Placebo Effect: subjects receiving the placebo have a response that is similar to what we would expect from the treatment
- Single Blind: subject does not know what treatment they are receiving
- Double Blind: subject and administrator do not know who receives the treatment

The design of an experiment is crucial. Experiments are designed with the purpose of isolating the effect of the treatment on the response variable and removing any confounding effects.

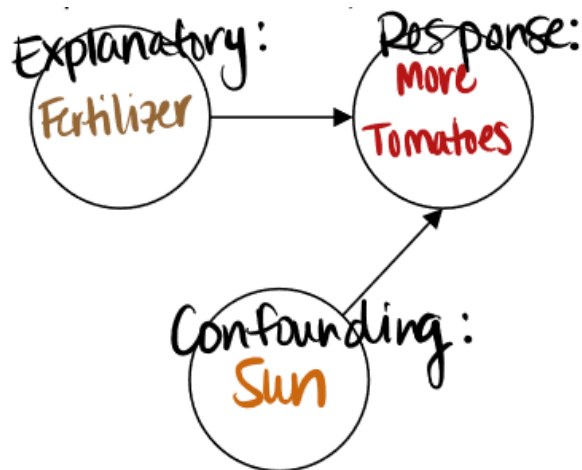
- In a poorly designed experiment, it might be difficult to tell if the explanatory variable causes a change or if it was another variable that wasn't measured
- Confounding variables are variables that might affect the outcome, but we did not control or account for them in our experiment.

One way to remove the effect of any confounding variable is to randomly assign subjects to the treatment or control group. This allows for any possible bias in the population to be evenly spread among the treatment and control groups. Sometimes instead of relying on randomization to make groups as even as possible we actually force the groups to be similar.

An extraneous variable is one that is not an explanatory variable in the study but is thought to affect the response variable. There are two types of extraneous variables present in studies:

Confounding variable refers to another variable that may affect the response and is in some way tied together with the factor under investigation. It leaves us unable to tell which of the two variables (or perhaps some interaction) caused the observed response.

For example, we plant tomatoes in a garden that's half-shaded. We test a fertilizer by putting it on the plants in the sun and apply none to the shaded plants. Months later the fertilized plants grow more and better tomatoes. Why? Well, maybe it's the fertilizer, maybe it's the sun, maybe we need both. We're unable to conclude that the fertilizer works because any effect of fertilizer is confounded with any effect of the extra sunshine.



Lurking variable refers to a variable that drives each of the two variables under investigation, making it appear that there's some association between them.

For example, there is a strong association between the number of firefighters who respond to a fire and the amount of damage done. One shouldn't conclude that the firefighters may be responsible for the damage; the lurking variable is the size of the fire.

Lurking variables are the risk we face in sampling and observational studies. In an experiment, though, the factor under consideration isn't being driven by some lurking variable, because we are the ones in control there.



**Example**

State whether the relationship between the two variables involves a lurking or confounding variable.

(a) Does watching TV make you live longer? Measure the number of television sets per person,  $x$ , and the average expectancy,  $y$ , for the world's nation. There is a high positive correlation: nations with many TV sets have high life expectancies. Could we lengthen the lives of people in Rwanda by shipping them TV sets? Justify your answer.

Lurking variable is money.

Money pays for food and healthcare (increasing life expectancy). More TVs generally mean more money.

(b) Do artificial sweeteners cause weight gain? People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

The confounding variable is diet.

People who use artificial sweeteners could be trying to lose weight so they may be heavier to begin with.

Inference is drawing conclusions beyond the data at hand.

Let's take a look at two different studies:

Study 1: The U.S. Census Bureau carries out a monthly Current Population Survey of about 60,000 households. Their goal is to use data from these randomly selected households to estimate the percent of unemployed individuals in the population.

Study 2: Scientists performed an experiment that randomly assigned 21 volunteer subjects to one of two treatments: sleep deprivation for one night or unrestricted sleep. The experimenters hoped to show that sleep deprivation causes a decrease in performance two days later.

- Random selection of individuals allows inference for the population
- Random assignment in an experiment allows inference for cause and effect

For the U.S. Census Bureau, individuals were randomly chosen to participate in the survey. The Bureau would be safe in making an inference about the population.

In the sleep deprivation experiment, subjects were randomly assigned to their treatments. If there is a large difference in the results, then we can assume it is not due to chance variation between the groups alone and must be due to sleep deprivation.

Well-designed experiments randomly assign individuals to treatment groups, but most don't select experimental units at random from the larger population, so their findings are limited to just cause and effect.

		Were individuals randomly assigned to groups?	
		YES	NO
Were individuals randomly selected?	YES	Inference about the Population: <b>YES</b> Inference about Cause and Effect: <b>YES</b>	Inference about the Population: <b>YES</b> Inference about Cause and Effect: <b>NO</b>
	NO	Inference about the Population: <b>NO</b> Inference about Cause and Effect: <b>YES</b>	Inference about the Population: <b>NO</b> Inference about Cause and Effect: <b>NO</b>

Both random sampling and random assignment introduce chance variation into a statistical study. When performing inference, statisticians use the laws of probability to describe this chance variation.

In some cases, it is not practical or ethical to do an experiment to establish a cause and effect relationship. Consider the following examples:

- Does texting while driving increase the risk of having an accident?
- Does going to church regularly help people live longer?
- Does smoking cause lung cancer?

There are laws now that must be followed when dealing with human subjects:

- Reviewed by an institutional review board - the board's purpose is "to protect the right and welfare of human subjects recruited to participate in research activities."
- Informed consent - subjects must be aware of the harm and danger a study could inflict and must be given written permission to participate.
- Confidentiality - protect individuals' privacy by keeping their identity separate from their results.

#### Writing up an experiment

1. Determine what type of design is best for your experiment. This will depend on the context of the question.
2. Diagram your experiment (time permitting).
3. Tell exactly how you will randomly assign your treatments.
  - Random Digit Table
  - The Hat Method
  - Dice, Coin, or Playing Cards
4. Explain exactly what you are comparing once you gather the data

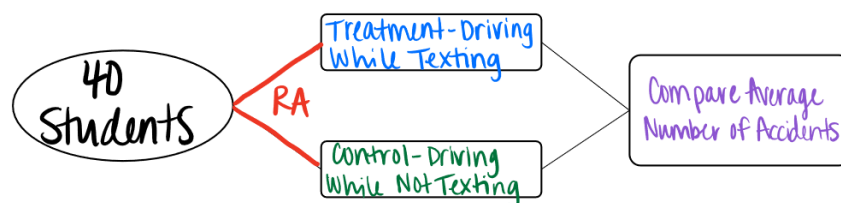
#### Completely Randomized Design

- The experimental units are assigned to treatments completely by chance.
- Treatment groups and control groups will be about equal in size in a completely randomized design.
- There are mathematical reasons for having groups of equal sizes, which we will discuss later.

#### Example

Is texting while driving causing accidents? There are 40 students that have volunteered for a study to determine if texting while driving causes more accidents. The county sheriff's department has given a driving simulator to use in the experiment. Design a completely randomized experiment.

- Write each student's name on identical slips of paper and place in a hat.
- Select 20 names from the hat without replacement and mixing well in between
- The 20 names selected will receive the treatment - driving while texting
- The remaining 20 students will receive the control - driving while not texting
- Compare the average number of accidents between the two groups



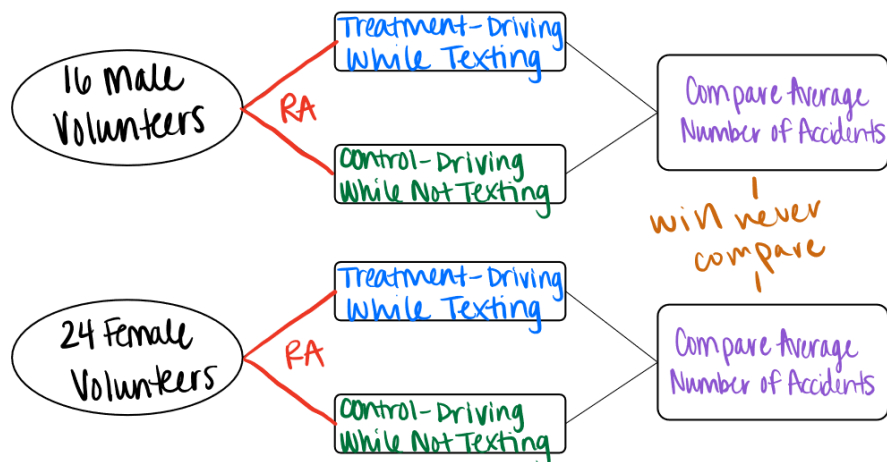
#### Randomized Block Design

- When groups of experimental units are similar, it's often a good idea to gather them together in blocks.
- Blocking isolates the variability due to the differences between the blocks so that we can see the differences due to the treatments more clearly.
- When randomization occurs only within the blocks, we call the design a randomized block design.
- Control what you can, block on what you can't control, and randomize to create compatible groups.

**Example**

It is brought to a teacher's attention that gender could be another contributing factor to the number of accidents people get into. Design an experiment to address this new information.

- Separate volunteers into two blocks based on gender.
- Number males 01-16. Use a random digit table to select 8 unique two digit numbers, ignoring 00 and 17-99.
- The 8 names selected will receive the treatment - driving while texting.
- The remaining 8 names will receive the control - driving while not texting.
- Repeat for females.
- Compare the average number of accidents between the two groups within each block.

**Matched Pairs Design**

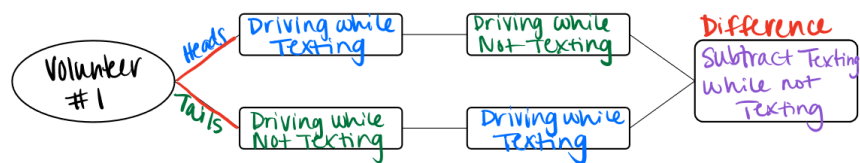
- These are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control.
- Often the “pair” in a matched pairs design is just one experimental unit which serves as its own control.
- In the case where an individual receives both the treatment and the control, the order in which this happens should be random.



**Example**

A student now brings up the fact that each student has different driving styles with many other variables that can influence the number of accidents. Design an experiment that would help address the other variables present for individual drivers.

- Each volunteer will do both treatments.
- Randomly assign order of treatments by flipping a coin.
- Heads - driving while texting first
- Tails - driving while not texting first
- Perform remaining treatment the next day
- Repeat for other 39 volunteers
- Compare the difference in number of accidents for each of the volunteers



# 4 Probability, Random Variables, and Probability Distributions

## 4.1 Basic Probability and Simulations

What is Probability?

- Outcomes are governed by chance, but in many repetitions a pattern emerges.
- Sample Space: all of the possible outcomes
- Event: a specific, desired outcome or set of outcomes
- Notation: probability of Event A  $\rightarrow P(A)$
- Range of Probabilities: the probability of an event is between 0 and 1
- Sum of Probabilities: the probability of the whole sample is 1

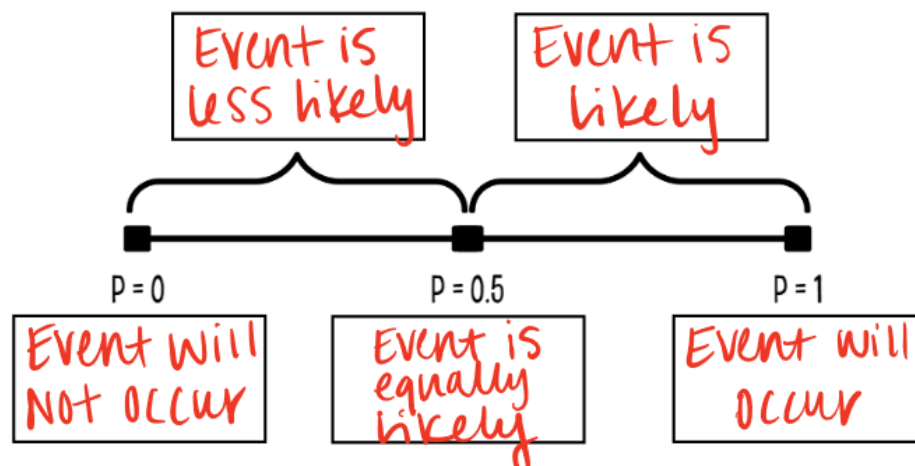
Theoretical Probability - What should happen given the sample space.

This is the desired outcomes divided by the total outcomes.

Empirical Probability - When you are performing a simulation or experiment, it is what does happen given the trials.

This is the number of successes divided by trials.

The law of large numbers states that simulated (empirical) probabilities tend to get closer to the true (theoretical probability) as the number of trials increases.



Probability can be written in the forms of fractions, decimals, or percentages.

For our class, we will write probabilities as decimals, rounded to four decimal places.

**Example**

A bag contains 10 marbles - 2 black, 3 blue, 1 red, and 4 white.

(a) What is the probability you will pull out a white marble?

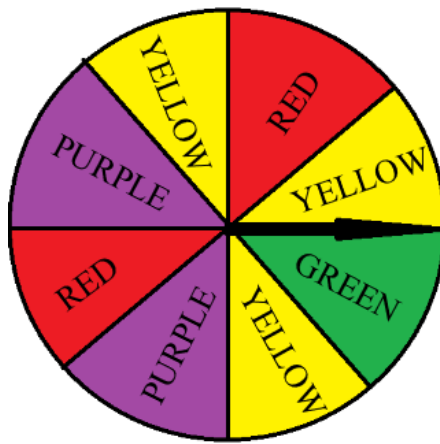
$$P(\text{White}) = 4/10 = 0.40$$

(b) What is the probability you will pull out a blue marble?

$$P(\text{Blue}) = 3/10 = 0.30$$

**Example**

In the spinner below, each wedge is equal in area.



(a) What is the probability that the spinner will land on a red wedge?

$$P(\text{red}) = 2/8 = 0.25$$

(b) What is the probability that the spinner will land on a yellow wedge?

$$P(\text{yellow}) = 3/8 = 0.375$$

The notation for the probability that an event does not occur is  $P(A^c)$ . We also refer to this as “Not A” in words.

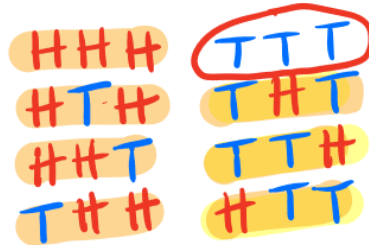
The probability that an event does not occur is  $P(A^c) = 1 - P(A)$ .

Sometimes, it is easier to compute the complement of an event, than the event itself.

**Example**

A coin is flipped three times and the results of each flip is noted.

(a) Draw a sample space for this event.



(b) What is the probability a series of three flips will produce exactly one head?

$$P(1 \text{ head}) = 3/8 = 0.375$$

(c) What is the probability that a series of three flips will produce at least one head?

$$P(\text{at least 1 Head}) = 1 - P(\text{no Heads}) = 7/8 = 0.875$$

Myth of “Law of Averages”

- The idea of probability is that randomness is predictable in the long run.
- Probability does not allow us to make short run predictions.
- Probability tells us random behavior evens out in the long run.
- Future outcomes are not affected by past behavior.

**Example**

A group of 50 high school students were interviewed and asked what their favorite leisure activity is. Use the results in the table below to find and interpret the following probabilities.

	Dance	Sports	Read	Total
Male	2	10	8	20
Female	16	6	8	30
Total	18	16	16	50

(a) What is the probability of picking a male whose favorite leisure activity is dance?

$$P(\text{male and dance}) = 2/50 = 0.04$$

(b) What is the probability of picking a person whose favorite leisure activity is reading?

$$P(\text{reading}) = 0.32$$

(c) What is the probability of picking someone who does not consider dance as their favorite leisure activity?

$$P(\text{not dance}) = 1 - P(\text{dance}) = 0.64$$

(d) Given that you’ve selected a female, what is the probability she likes reading?

$$P(\text{read} \mid \text{female}) = 8/30 = 0.2667$$

The imitation of chance behavior, based on a model that accurately reflects the situation, is called a simulation. Simulations are usually done with a table of random digits, random number generator, dice, deck of cards,

spinner, etc.

Four Principles of Simulation:

State - Identify the probability calculation.

Must include:

- Identify variable.
- Statement of probability in symbols or words.

Plan - Describe how to use your chance process.

Must include:

- What tool?
- What values are you assigning?
- How many values are you picking each time?
- How many times do you run the simulation?
- What about repeat digits or ignored digits?
- What are you recording?

Do - Perform the simulation (trials typically indicated, perform at least 10 if not).

Must include:

- Simulation data, if number of trials is 10 or less.
- Summary of data for larger trials.

Conclude - Use the results of your simulation to answer the question.

Must include:

- Statement of probability.
- Answer to question.
- Usually about being surprised/reasonable/expected, etc.

### Example

What is the probability that a student gets 6 out of 6 questions correct on a true/false quiz written in Greek? (Assume the exam taker does not know any Greek.) Should the instructor be concerned about cheating? Perform a simulation to answer this question.

- What is the probability of guessing 6/6 correct on a T/F quiz? 50% correct or 50% incorrect
- Label digits 0-4 as correct and 5-9 as incorrect. Using a random digit table, select 1 number at a time (repeats allowed) and record if correct or incorrect. Perform this simulation 6 times to represent one quiz. Count how many questions are "correct". Repeat for 20 trials.

# Correct	# Trials
0	0
1	2
2	3
3	8
4	5
5	2
6	0

After conducting the simulation, we see that we had no trials where 6/6 is correct by guessing. We should be suspicious of the result the student got.

## 4.2 The Addition Rule

### Mutually Exclusive Events

- When two events have no outcomes in common, we refer to them as mutually exclusive events.
- $P(A \text{ and } B) = 0$
- $P(A \text{ or } B) = P(A) + P(B)$

#### Example

Let's revisit our "rolling the dice" game, where the italicized numbers represent the sum of the dice rolled.

Dice	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Let A be rolling a sum of 5 and B be rolling two even numbers.

These events are mutually exclusive because they have no outcomes in common.

$P(A)$  is  $4/36$ , and  $P(B)$  is  $9/36$ , so  $P(A \text{ or } B)$  is  $13/36$ .

#### Example

Identify if the following events are mutually exclusive. If they are not, given an example of where they overlap.

(a) Rolling a sum of 4 and rolling doubles.

Not mutually exclusive. Ex. 2, 2

(b) Rolling an odd sum and rolling a 3.

Not mutually exclusive. Ex. 3, 4

(c) Rolling a 4 and rolling a sum of 12

Mutually exclusive

(d) Rolling an odd number and rolling a sum of 10

Not mutually exclusive. Ex. 5, 5

**Example**

Find the probability of rolling doubles or a sum of 8.

Dice	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Let A be rolling doubles, and B be rolling a sum of 8.

They are not mutually exclusive events.

$P(A) = 6/36$  and  $P(B) = 5/36$ . We have to subtract  $P(A \text{ and } B) = 1/36$  from this, and we get  $P(A \text{ or } B) = 10/36$ .

If A and B are any two events resulting from some chance process, then

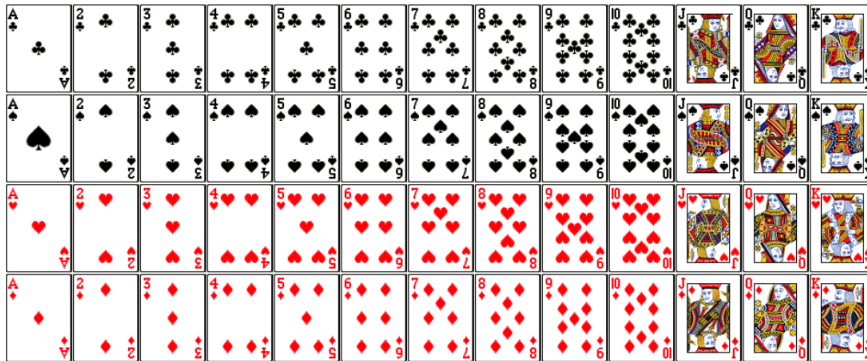
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Notice how if A and B are mutually exclusive,  $P(A \text{ and } B) = 0$  and  $P(A \text{ or } B) = P(A) + P(B) - 0$

How do we know the value of  $P(A \text{ and } B)$ ? For now, we will use common sense and drawing out the sample space, but the general multiplication rule will allow us to find this mathematically later.

**Example**

A card is selected at random from a deck of 52 cards. Determine if the following events are mutually exclusive and then find the probability of the events.



(a)  $P(\text{red or Queen})$

Not mutually exclusive.  $26/52 + 4/52 - 2/52 = 28/52$ .

(b)  $P(\text{Ace or King})$

Mutually exclusive.  $4/52 + 4/52 = 8/52$

(c)  $P(\text{Queen or even})$

Mutually exclusive.  $4/52 + 20/52 = 24/52$

(d)  $P(\text{black or odd})$

Not mutually exclusive.  $26/52 + 16/52 - 8/52 = 34/52$



**Example**

You were interested in how many students at your school regularly eat breakfast. You conducted a survey, asking, “Do you eat breakfast on a regular basis?” Your random sample consisted of 600 students at the school and the results are shown in the table below.

	<b>Male</b>	<b>Female</b>	<b>Total</b>
<b>Eats Breakfast Regularly</b>	190	110	300
<b>Doesn’t Eat Breakfast Regularly</b>	132	168	300
<b>Total</b>	322	278	600

If we select a student from this sample at random, what is the probability that the student is

(a) A female?

$$P(\text{female}) = 278/600$$

(b) Someone who eats breakfast regularly?

$$P(\text{Breakfast}) = 300/600$$

(c) A female and eats breakfast regularly?

$$P(\text{female and breakfast}) = 110/600$$

(d) A female or east breakfast reguarly?

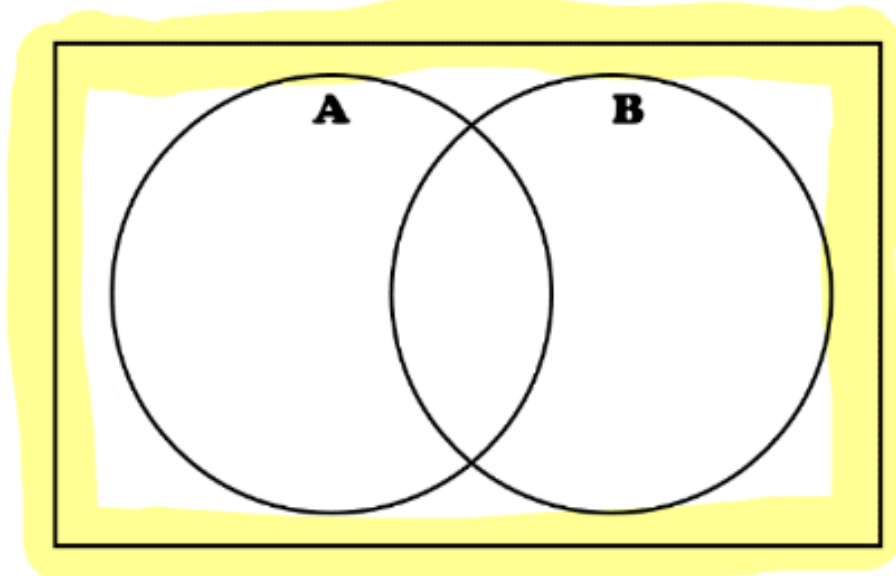
$$P(\text{female or breakfast})$$

$$278/600 + 300/600 - 110/600 = 468/600$$

### 4.3 Venn Diagrams and the Multiplication Rule

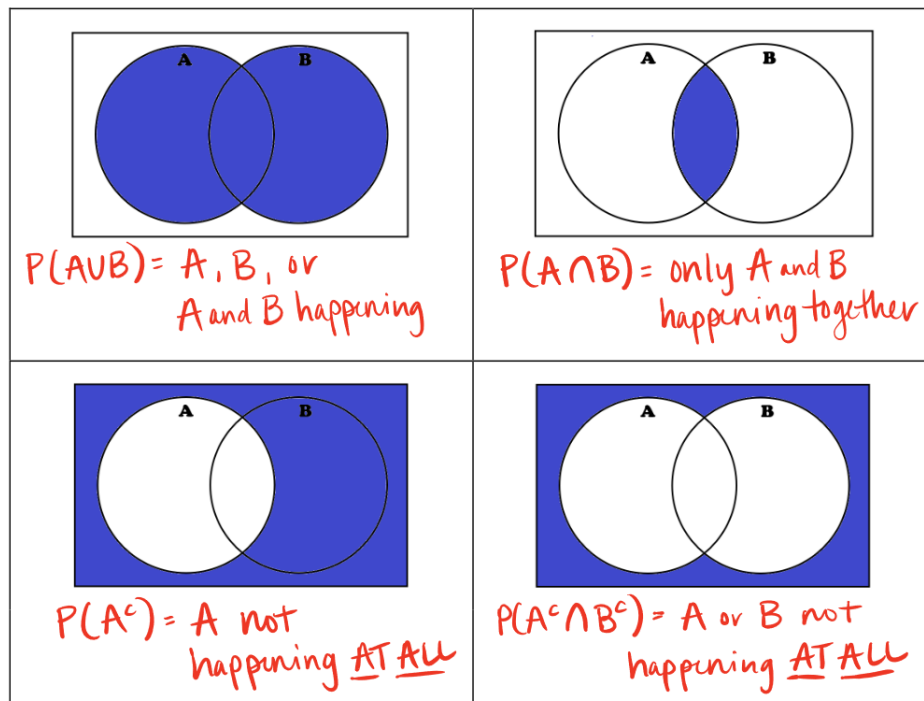
#### Union and Intersection

- We can represent events with a Venn Diagram - a display of potential probabilities.
- The box around the Venn diagram represents the total sample space where  $P(S) = 1$ .
- The circles themselves represent the probability of each event.



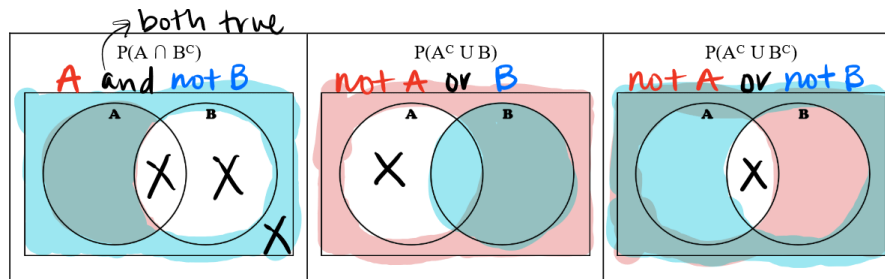
Union: OR = probability that either occurs. The symbol is  $\cup$

Intersection: AND = probability both occur. The symbol is  $\cap$



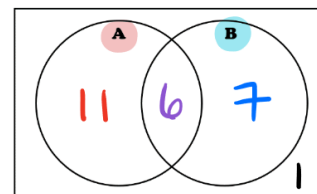
**Example**

Shade the probabilities for the following notations.

**Example**

In your statistics class, you are interested in exploring the taste preference of your classmates. You have two tastes, sweet and salty. 25 students were asked if they prefer both, only sweet, only salty, or neither. The following two-way table shows the data.

	Yes Sweet	No Sweet	Total
Yes Salty	6	7	13
No Salty	11	1	12
Total	17	8	25



Let  $A$  = sweet and  $B$  = salty. Find and interpret the following probabilities.

(a)  $P(A \cap B)$

$\frac{6}{25} = 0.24$ . The probability a student likes both salty and sweet foods is 24%.

(b)  $P(A \cup B)$

$\frac{11+6+7}{25} = 0.96$ . The probability a student likes either salty or sweet foods is 96%.

(c)  $P(A \cap B^c)$

$\frac{11}{25} = 0.44$ . The probability a student likes sweet but not salty foods is 44%.

(d)  $P(A^c \cup B)$

$\frac{7+1+6}{25} = 0.56$ . The probability a student likes no sweet foods or likes salty foods is 56%.

(e)  $P(A \cup B^c)$

$\frac{11+6+1}{25} = 0.72$ . The probability a student likes sweet foods or does not like salty foods is 72%.

(f)  $P(A^c \cap B^c)$

$\frac{1}{25} = 0.04$ . The probability a student does not like sweet foods and does not like salty foods is 4%.

### Multiplication rule for independent events

- Two events are independent if they do not influence one another.
- The occurrence of one has no effect on the occurrence of the other.
  - Toss a coin twice.
  - The gender of each child born to the same mother.
  - Trials with replacement.

- If A and B are independent events, then the probability that both A and B occur is found using the multiplication rule.

$$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B)$$

### Example

Find the probabilities.

- (a) I flip a coin and then roll a dice. What is the probability I flip a head and then roll a 5?

$$\left(\frac{1}{2}\right) \left(\frac{1}{6}\right) = 0.0833$$

- (b) What is the probability that a mother will have 4 girls in a row?

$$\left(\frac{1}{2}\right)^4 = 0.0625.$$

- (c) I have a standard deck of 52 cards. What is the probability I pull a red card, replace it and shuffle, and then pull a black card?

$$\left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = 0.25$$

- (d) You spin the following spinner 5 times. What is the probability you land on yellow all 5 times?



$$\left(\frac{3}{8}\right)^5 = 0.0074.$$

### Multiplication Rule for Dependent Events

- Two events are dependent if they influence one another.
- The occurrence of one affects the occurrence of the other.
  - Trials without replacement.
- If A and B are dependent events, then the probability that both A and B occur is found using:

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

- Where  $P(B|A)$  is the probability of B "given" that A has occurred.
- The conditional probability of an event is the probability that one event will happen, if it is known that another event has happened.

**Example**

Find the probabilities.

(a) I have a standard deck of 52 cards. What is the probability of pulling a spade and then a 5?

$$\left(\frac{13}{52}\right)\left(\frac{4}{51}\right) = 0.0196$$

(b) I have a standard deck of 52 cards. What is the probability of pulling a jack and then another jack?

$$\left(\frac{4}{52}\right)\left(\frac{3}{51}\right) = 0.0045$$

(c) In your statistics class, you are interested in exploring the taste preference of your classmates. You have two tastes, sweet and salty. 25 students were asked if they prefer both, only sweet, only salty, or neither. The following two-way table shows the data.

	Yes Sweet	No Sweet	Total
Yes Salty	6	7	13
No Salty	11	1	12
Total	17	8	25

What is the probability you select someone who likes salty food then select another person who likes salty food?

$$\left(\frac{13}{25}\right)\left(\frac{12}{24}\right) = 0.26$$

What is the probability you select someone who does not like sweet food then select a person who likes sweet food?

$$\left(\frac{8}{25}\right)\left(\frac{17}{24}\right) = 0.2267$$

## 4.4 Conditional Probability and Tree Diagrams

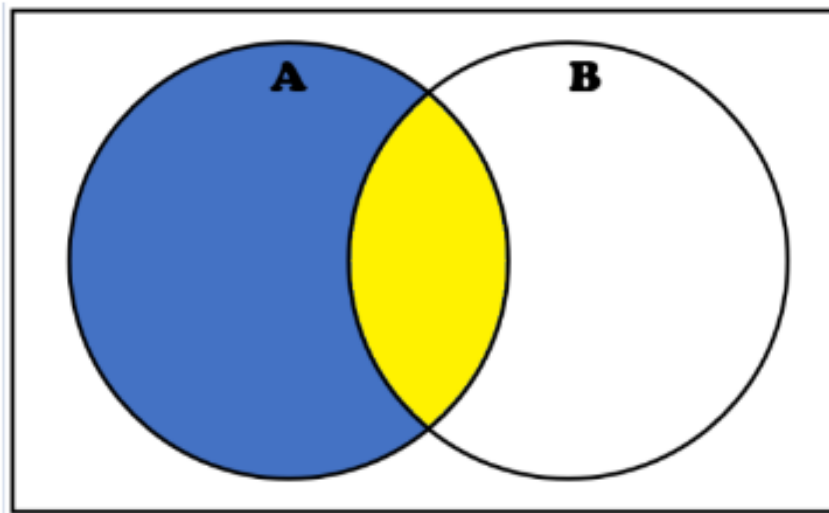
Rule for conditional probability:

- If A and B are dependent events, then the probability that both A and B occur is found using  $P(A \cap B) = P(A) \cdot P(B|A)$  where  $P(B|A)$  is the probability of B “given” that A has already occurred
- The conditional probability of an event is the probability that one event will happen, if it is known that another event has happened.
- We can rearrange the multiplication rule a little and get a formula for conditional probability:

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

Why does this formula make sense?

- When you know something has happened, it changes the sample space - specifically, it shrinks to the given event.
- If we know A has already occurred, that becomes your “sample space”.
- Then, for B to occur, it has to occur with A, which is the intersection of A and B.



### Example

Use the two-way table below to calculate the indicated probabilities.

		Eye Color				
Hair Color		Brown	Blue	Hazel	Green	Total
	Black	36	9	5	2	52
	Brown	66	34	29	14	143
	Red	16	7	7	7	37
	Blonde	4	64	5	8	81
	Total	122	114	46	31	313

(a) If we select a woman and find she has brown hair, what is the probability that she will also have hazel eyes?

$$P(\text{Hazel Eyes} | \text{Brown Hair}) = \frac{P(H \cap B)}{P(B)} = \frac{29/313}{143/313} = 0.2028$$

(b) If we select a woman and find she has brown eyes, what is the probability that she will also have black hair?

$$P(\text{Black Hair} | \text{Brown Eyes}) = \frac{P(B \cap H)}{P(H)} = \frac{36/122}{114/122} = 0.2951$$

Showing Independence:

- If we draw two cards with replacement, we know that those events are independent.
- If we draw two cards without replacement, we know that those events are dependent.

However, in the real world, it can be difficult to determine if two events are independent or dependent. We do have a way to prove independence, however, using either of the following formulas.

$$P(A \cap B) = P(A) \cdot P(B) \text{ or } P(A) = P(A|B)$$

### Example

Below is a two-way table that shows produce and how many packages are purchased based on whether it is fresh or frozen.

	Fresh	Frozen	Total
Blueberries	20	36	56
Pineapple	21	39	60
Avocado	9	15	24
Total	50	90	140

(a) Are “blueberries” and “fresh” independent?

$P(B) = P(B|Fresh)$ ,  $0.4=0.4$ , so independent.

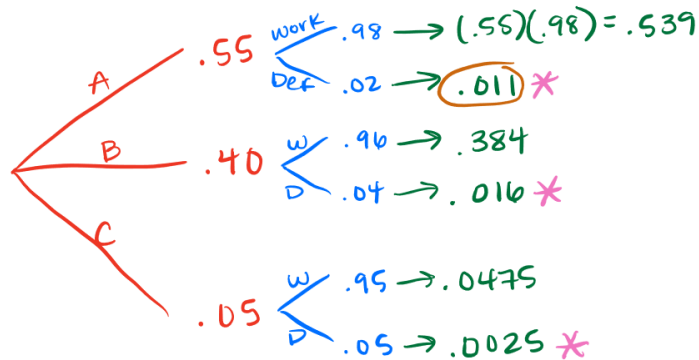
(b) Are “avocados” and “frozen” independent?

$P(Frozen \cap Avocados) = P(Frozen) \cdot P(A)$ , and  $\frac{15}{140} \neq \frac{90}{140} \cdot \frac{24}{140}$ . They are not independent.

Tree Diagrams

**Example**

A company manufacturing electronic components for home entertainment systems buys electrical connectors from three suppliers. The company prefers to use supplier A because only 2% of those connectors prove to be defective, but supplier A can deliver only 55% of the connectors needed. The company must also purchase connectors from two other suppliers, 40% from supplier B and the rest from supplier C. The rates of defective connectors from B and C are 4% and 5%, respectively. You buy one of these components.



(a) What is the probability that you use supplier B to and the connector is defective?

$$P(B \cap \text{Defective}) = (0.40)(0.04) = 0.016$$

(b) What's the probability that you find that the connector is defective?

$$P(\text{Defective}) = 0.011 + 0.016 + 0.0025 = 0.0295$$

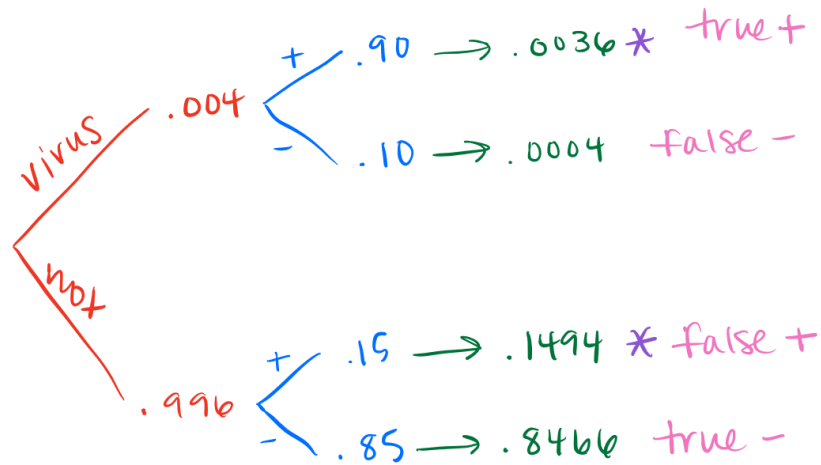
(c) Given that the component is defective, what is the probability it came from supplier A?

$$P(A|\text{Defective}) = \frac{P(A \cap \text{defective})}{P(\text{defective})} = \frac{0.011}{0.0295} = 0.3729$$



**Example**

The probability that a person has a particular virus is 1 out of 250. If they have the virus, the probability they test positive for the virus is 0.90. If they do not have the virus, the probability they test positive for the virus is 0.15.



(a) Given that a person tests positive for the virus, what is the probability they have the virus?

$$P(\text{virus}|+) = \frac{P(\text{virus} \cap +)}{P(+)} = \frac{0.0036}{0.0036 + 0.1494} = 0.0235$$

(b) What is the probability of a person testing and getting a false positive?

$$P(\text{false}+) = P(\text{no virus} \cap +) = 0.1494$$

## 4.5 Discrete and Continuous Random Variables

A random variable takes numerical values that describe the outcomes of some chance process. Random variables involve the same probability rules we have already learned, but we will extend those rules to be able to model more events.

### Example

Suppose we toss a coin 3 times.

This is the sample space.

$HHH$   $THH$   $TTH$   
 $HTH$   $TTT$   $HTT$   
 $HHT$   $THT$

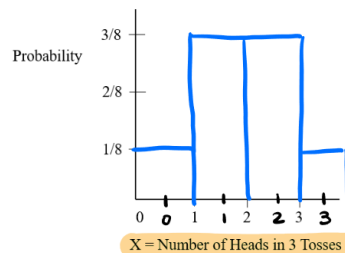
The total number of outcomes is 8, the probability of any one outcome is  $1/8$ .

Let  $X$  be the number of heads out of 3 coin tosses.

Symbols	Meaning	Probability
$X = 0$	no heads out of 3 tosses	$1/8$
$X = 1$	1 heads out of 3 tosses	$3/8$
$X = 2$	2 heads out of 3 tosses	$3/8$
$X = 3$	3 heads out of 3 tosses	$1/8$

With these 4  $X$  values, and their corresponding probabilities, we can create a frequency table and a histogram to describe this event.

$X$	0	1	2	3
$P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



- The frequency table and histogram above each represent a probability distribution.
- A probability distributions tells us the value that our random variable can take and the probability associated with each value.

Every probability distribution must satisfy each of the following requirements.

1. There is a numerical (not categorical) random variable  $X$ , and each of the values of  $X$  has an associated probability with it.
2. The sum of all the probabilities in the distribution  $\sum P(X)$  must equal 1.
3.  $0 \leq P(X) \leq 1$  for all probabilities in the distribution.

Notation:

Notation	What it means	Example					
$P(X = 2)$	What is the probability that the random variable equals 2?	$P(X = 2)$ $\boxed{\frac{3}{8}}$	$X$	0	1	2	3
$P(X \geq 2)$	What is the probability that the random variable is greater than or equal to 2?	$P(X \geq 2)$ $P(X=2) + P(X=3)$ $\frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \boxed{\frac{1}{2}}$					
$P(X < 1)$	What is the probability that the random variable is less than 1?	$P(X < 1)$ $P(X=0) = \boxed{\frac{1}{8}}$					
			$P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

A discrete random variable takes on a fixed set of possible values with whole number outcomes.

- Random variables are usually capital letters (X, Y, Z, etc.)
- Random variables can be discrete or continuous (continuous is next lesson)
- Random variables must be numeric in value.
  - In our example, even though a result of Head or Tails is categorical, we described it numerically.

The mean of a discrete random variable X, is the mean outcome for infinitely many trials.

We think of this as the “expected value” because it is the average value we would expect to get if the trials could continue indefinitely. (Not what we expect to get in a single trial)

To find the mean,  $\mu_x$ , or expected value,  $\sum(x)$ , of a discrete random variable X, multiply each possible value by its probability, then add all the products.

<b>X</b>	$x_1$	$x_2$	$\dots$	$x_n$
<b>P(X)</b>	$p_1$	$p_2$	$\dots$	$p_n$

$$\mu_{u_x} = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots x_n \cdot p_n = E(x) = \sum x_i \cdot p_i$$

**Example**

Northwestern University posts the grade distributions for its courses online. Students in Statistics 101 in a recent semester received 26% A's, 42% B's, 20% C's, 10% D's, and 2% F's.

(a) Create a probability distribution for the random variable  $X$ , which represents the student's grade on a four point scale (with A = 4)

$X$	4	3	2	1	0
$P(X)$	.26	.42	.20	.10	.02

(b) Find and interpret  $P(X \geq 3)$  if  $X$  represents a randomly selected Statistics 101 student.

$P(X \geq 3) = P(X=3) + P(X=4) = 0.68$ . The probability that a randomly selected Stats 101 student gets a grade of A or B is 68%.

(c) Write the event "the student got a grade worse than C" in terms of values of the random variable  $X$ . What is the probability of this event?

$$P(x < 2) = P(x=1) + P(x=2) = 0.10 + 0.02 = 0.12$$

(d) Compute the mean of the random variable  $X$ . Interpret this value in context.

$\mu_x = (4)(0.26) + (3)(0.42) + (2)(0.20) + (1)(0.10) + (0)(0.02) = 2.8$ . If many students took this course, we would expect an average grade of 2.8, which is between a B and C.

The standard deviation of a random variable  $X$  is a measure of how much the values of the variable typically vary from the expected value.

$X$	$x_1$	$x_2$	$\dots$	$x_n$
$P(X)$	$p_1$	$p_2$	$\dots$	$p_n$

$$\sigma_x^2 = (x_1 - \mu_x)^2 \cdot p_1 + (x_2 - \mu_x)^2 \cdot p_2 + \dots + (x_n - \mu_x)^2 \cdot p_n$$

$$\sigma_x^2 = \text{Var}(X) = \sum (x_i - \mu_x)^2 \cdot p_i$$

$$\sigma_x = \sqrt{\sum (x_i - \mu_x)^2 \cdot p_i}$$

**Example**

For the Northwestern example, find and interpret the standard deviation of the grade probability distribution.

$$\sigma_x = \sqrt{(4 - 2.8)^2(.26) + (3 - 2.8)^2(.42) + (2 - 2.8)^2(.2) + (1 - 2.8)^2(.1) + (0 - 2.8)^2(.02)} = 1$$

If many students took this course, we would expect the typical deviation from the mean to be about 1 letter grade.

**Continuous Random Variable**

- Continuous Random Variables take on all possible values in an interval of numbers. The probability distribution of  $X$  is described by a density curve.
- The probability of any event is the area under the density curve and above, below, or between the values  $x$  that define the event.

- Area under a density curve is always equal to 1.
- The probability is on the  $y$ -axis in a distribution, so finding the area equates to finding the probability of getting an interval of numbers.
- The probability at an event is 0 for a continuous random variable because the area under a point is 0.
- The mean and standard deviation of a continuous random variable are found using integration, so we will not discuss their formulas here.

Continuous random variables are usually represented as functions in practice. The continuous random variable we will explore is the normal distribution, and we will revisit it at the start of Unit 5.

Why do we have random variables?

This is an important question to keep in mind as we continue to look at different random variables. We need random variables because they represent more things than a single measurement can. One-variable statistics used data, where random variables can use theory and apply it to real situations.

Calculator: Mean and Standard Deviation from a Probability Distribution Table

- Enter outcomes into L1 and probabilities into L2
- Run 2nd - VARS - 1VarStats with List:L1 and FreqList:L2
- Mean will be represented by  $\bar{x}$
- Standard deviation will be represented by  $\sigma_x$

## 4.6 Combining Random Variables

### Example

Let  $X$  and  $Y$  be two independent variables with the following probability distribution.

$X$	1	2	3
$P(X)$	0.3	0.2	0.5

$Y$	5	10
$P(Y)$	0.4	0.6

If we let  $S = X + Y$ , create the probability distribution below.

$S$	6	7	8	11	12	13
$P(S)$	.12	.08	.2	.18	.12	.3

Now we can find the following:

- $E(S) = 10.2$
- $E(X) = 2.2$
- $E(Y) = 8$
- $\sigma_S = 2.6$
- $\sigma_X = 0.87$
- $\sigma_Y = 2.45$

There is a shortcut for finding means and standard deviations of combined random variables without having to create a new distribution!

If  $X$  and  $Y$  are two independent random variables:

$$E(X + Y) = E(X) + E(Y)$$

$$E(X - Y) = E(X) - E(Y)$$

$$\sigma_{X+Y} = \sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

With the problem above, if you insert the values, you can see that they are the same.

**Example**

The head folks at Google decided to try a random salary assignment for their workers this year! They have the following probability distribution, where  $X$  = salary.

$X$	50,000	60,000	70,000	80,000
$P(X)$	0.2	0.3	0.4	0.1

(a) What is the expected salary for a Google employee? What is the standard deviation?

$$\mu_x = \$64,000, \sigma_x = \$9165.15$$

(b) Google decides to give everyone a \$500 bonus. Create a new probability distribution and find the expected value and standard deviation. Notice anything?

$X+500$	50500	60500	70500	80500
$P(X+500)$	.2	.3	.4	.1

The expected value increased by \$500, but the standard deviation remained the same.

(c) Google decides that it will pay cash at the end of the year for this salary. They will give out \$100 bills. Let  $X$  = number of \$100 bills they will give out. Create a new probability distribution (use the bonus distribution) and find the expected value and standard deviation. Notice anything?

$\frac{X+500}{100}$	505	605	705	805
$P\left(\frac{X+500}{100}\right)$	.2	.3	.4	.1

In this case, both the mean and standard deviation were divided by 100.

If  $X$  is a random variable and  $a$  and  $b$  are both constants:

$$E(a + bX) = a + bE(X)$$

$$\sigma_{a+bX} = |b|\sigma(X)$$

**Example**

For an upcoming concert, each customer may purchase up to 3 child tickets and 3 adult tickets. Let  $C$  be the number of child tickets purchased by a single customer. The probability distribution of the number of child tickets purchased by a single customer is given in the table below.

$C$	0	1	2	3
$P(C)$	0.4	0.3	0.2	0.1

(a) Compute the mean and the standard deviation of  $C$ .

$$\mu_C = 1 \text{ child ticket}, \sigma_C = 1 \text{ child ticket.}$$

(b) Suppose the mean and the standard deviation for the number of adult tickets purchased by a single customer are 2 and 1.2, respectively. Assume that the number of child tickets and adult tickets purchased are independent random variables. Compute the mean and the standard deviation of the total number of adult and child tickets purchased by a single customer.

$$\mu_A = 2, \sigma_A = 1.2.$$

$$T = A + C, \mu_T = \mu_A + \mu_C = 3 \text{ tickets.}$$

$$\sigma_T = \sqrt{\sigma_A^2 + \sigma_C^2} = \sqrt{(1.2)^2 + (1)^2} = 1.562 \text{ tickets.}$$

(c) Suppose each child ticket costs \$15 and each adult ticket costs \$25. Compute the mean and the standard deviation of the total amount per purchase.

$$S = 25A + 15C,$$

$$\mu_S = 25\mu_A + 15\mu_C = \$65$$

$$\sigma_S = \sqrt{(25 \cdot \sigma_A)^2 + (15 \cdot \sigma_C)^2} = \$33.54$$

**Example**

Each full carton of Grade A eggs consists of 1 randomly selected empty cardboard container and 12 randomly selected eggs. The weights of such full cartons are approximately normally distributed with a mean of 840 grams and a standard deviation of 7.9 grams.

The weights of the empty cardboard containers have a mean of 20 grams and a standard deviation of 1.7 grams. It is reasonable to assume independence between the weights of the empty cardboard containers and the weights of the eggs. It is also reasonable to assume independence among the weights of the 12 eggs that are randomly selected for a full carton.

Let the random variable  $X$  be the weight of a single randomly selected Grade A egg.

(a) What is the mean of  $X$ ?

$$1 \text{ container} + 12 \text{ eggs, full carton has } \mu = 840, \sigma = 7.9, \text{ empty container has } \mu = 20, \sigma = 1.7$$

$$840 = 20 + x_1 + \cdots + x_{12} \text{ so, } 840 = 20 + 12X, \text{ or } X = 68.33 \text{ grams.}$$

(b) What is the standard deviation of  $X$ ?

$$7.9 = \sqrt{(1.7)^2 + \sigma_{x_1}^2 + \cdots + \sigma_{x_{12}}^2} \implies 62.41 = (1.7)^2 + 12\sigma_x^2.$$

Solving for  $\sigma_x$  gives 2.227 grams.



## 4.7 The Binomial Distribution

Bernoulli Trials:

Requirements for a Bernoulli Trial:

1. Two Possible Outcomes
2. Probability of Success is the Same for Each Trial
3. Trials are Independent

If we take a Bernoulli Trial and then we are interested in the number of successes in a specific number of trials, we create a Binomial Probability Distribution.

There are four requirements for a setting to follow the binomial probability distribution. As long as these four things are true, we can use the binomial probability model for calculating the probabilities.

- Binary:
  - Each trial falls into one of two categories - we call them “success” or “failure”.
  - Success does not necessarily mean a positive outcome, but instead, the outcome we are looking for.
- Independent Trials
  - Each trial is independent of the next.
  - Reasonable Assumption for coins, cards with replacement, spinners, rolling a die, etc.
  - What about sampling without replacement, when it can't be avoided? This is where the “10% condition” comes in.
    - \* The 10% condition says that if you are sampling from a large enough population, you can proceed with sampling without replacement as though the trials were independent.
    - \* Ex. Sampling from a deck of cards without replacement would violate this but sampling people from a large town would be okay.
- Number of Trials is Fixed
  - We are counting the number of successes from a set number of trials (called  $n$ ).
- Same Probability
  - The probability of success (called  $p$ ) is the same for each trial.

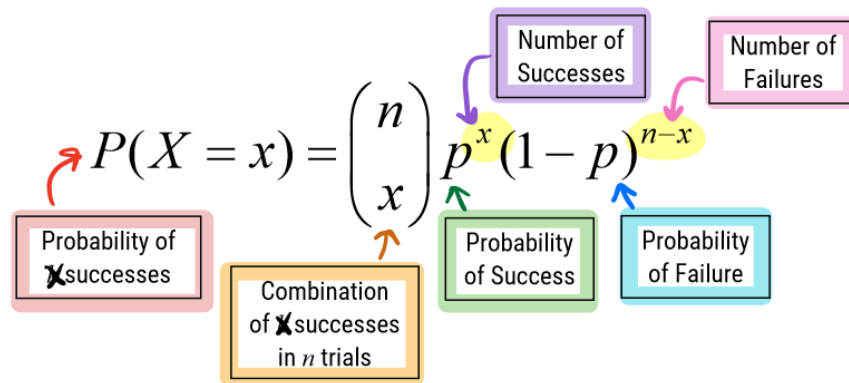
If all four points are satisfied, you can compute the probability you get a certain number of successes in a specified number of trials:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where  $\binom{n}{x} = nC_x = \frac{n!}{x!(n-x)!}$ . (MATH - PROB - 3:nCr) and  $n$  is the number of independent trials,  $p$  is the probability of success, and  $x$  is the number of successes.

The mean is  $\mu_x = np$ , the standard deviation is  $\sigma_x = \sqrt{np(1-p)}$

Why this formula?

**Example**

Explain how each of the following situations follows the binomial probability distribution model.

(a) You toss a coin 10 times and count the number of tails you get.

$X$  is the number of tails.

- Binary: Tails/Heads
- Independence: Coin flips are independent
- Number of Trials: 10
- Probability: 0.50

Find and interpret  $P(X=4)$ .

$P(X = 4) = \binom{10}{4} (0.50)^4 (0.50)^{10-4} = 0.2051$ . The probability you get 4 tails in 10 coin flips is 20.51%.

(b) Type A blood is found in 43% of the population. You have a group of 40 students.

$X$  is the number of students with Type A blood.

- Binary: Type A/Not Type A
- Independence: Assume many students in population
- Number of Trials: 40
- Probability: 0.43

Find and interpret  $P(X=15)$

$P(X = 15) = \binom{40}{15} (0.43)^{15} (0.57)^{25} = 0.1008$ . The probability you get 15 students out of 40 who have Type A blood is 10.08%.

Distributions in your calculator often have “pdf” and “cdf” after them and there is an important difference.

- pdf = probability distribution function and is used when you are trying to calculate  $P(X = k)$
- cdf = cumulative distribution function and is used when you are trying to calculate  $P(X \leq k)$

Specifically for binomial distributions, we want

- 2nd - VARS - A:binompdf(trials, probability of success) for  $P(X = k)$
- 2nd - VARS - B:binomcdf(trials, probability of success) for  $P(X \leq k)$

A binomial random variable is discrete so how you set up your inequality matters.

**Example**

Let  $X$  = number of students that pass the AP Statistics Exam out of a class of 20. The probability that a random student will pass is 0.62. Find the probability that

(a) All students pass the exam.

$$P(x=20) = \text{binompdf}(\text{trials: } 20, p: 0.62, x: 20) = 0.0007$$

(b) Six or fewer students pass the exam

$$P(x \leq 6) = \text{binomcdf}(\text{trials: } 20, p: 0.62, x: 6) = 0.0037$$

(c) At least 12 students pass the exam

$$P(x \geq 12) = 1 - P(x \leq 11) = 1 - \text{binomcdf}(\text{trials: } 20, p: 0.62, x: 11) = 0.6659$$

**Example**

Suppose the probability that any random freshman girl will agree to try out for a high school dance team is 10% and one of the team captains asks 25 random freshman girls to try out.

(a) What is the probability that exactly 1 will say yes?

$$P(x = 1) = \text{binompdf}(\text{trials: } 25, p: 0.10, x: 1) = 0.1994$$

(b) What is the probability that at least 1 will say yes?

$$P(x \geq 1) = 1 - P(x = 0) = 1 - \text{binompdf}(\text{trials: } 25, p: 0.10, x: 0) = 0.9282$$

(c) On average, how many freshman girls will agree to try out for the team? Interpret in context.

mean =  $np = 2.5$  girls. On average, if we select many groups of 25 freshman girls, we would expect 2.5 girls to try out.

(d) What is the standard deviation of this binomial distribution? Interpret in context.

$\sigma_x = \sqrt{np(1-p)} = 1.5$  girls. The number of girls who try out typically differs from the mean by 1.5 girls.

## 4.8 The Geometric Distribution

When we perform many independent trials of the same chance process and are interested in the occurrence of the first success, a geometric setting arises.

Before attempting to calculate a geometric setting, check to make sure all of the following conditions have been met.

- Binary - the possible outcomes can be classified as a success or failure
- Independent - trials must be independent (the result of one trial tells us nothing about the result of another trial)
- First - you are interested in the number of trials until the first success.
- Success - the probability of success remains consistent trial to trial

### Example

You roll a six-sided fair die. Let  $X$  = number of trials until you roll a six. Complete the probability distribution table below.

$X$	1	2	3	4
$P(X)$	$\frac{1}{6}$	$(\frac{5}{6})(\frac{1}{6})$	$(\frac{5}{6})(\frac{5}{6})(\frac{1}{6})$	$1 - (.1667 + .1389 + .1157)$
	.1667	.1389	.1157	.5787

If  $X$  has a geometric distribution with a probability of success  $p$  on each trial and  $x$  represents the trial that you get your first success. The probability that  $X$  equals  $x$  is given by

$$P(X = x) = (1 - p)^{x-1}(p)$$

Where  $P(X = x)$  is the probability of first success on  $x$  trial,  $(1 - p)$  is the probability of failure,  $p$  is the probability of success, and  $x - 1$  is the number of failures.

### Example

You roll a six-sided fair die. Let  $X$  = number of trials until you roll a six.

(a) What is the probability that you get your first six on the 5th roll?

$$P(X = 5) = \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right) = 0.0804$$

(b) What is the probability that you get your first six within 3 rolls?

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.1667 + 0.1389 + 0.1157 = 0.4213.$$

(c) What is the probability that it takes at least 4 rolls to get your first six?

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.4213 = 0.5787.$$

Specifically for geometric distributions, we want

- 2nd - VARS - E:geometpdf(probability of success,  $k$ ) for  $P(X = x)$
- 2nd - VARS - F:geometcdf(probability of success,  $k$ ) for  $P(X \leq x)$

Note: Your calculator can only do less than or equal to so all inequalities must be written in this manner.

**Example**

Let  $X$  = number of basket attempts needed for a college basketball player to make his first free throw. The player has an 82% chance of making a random free throw. Find the probability that

(a) He makes his first basket on his fifth attempt.

$$P(X = 5) = \text{geompdf}(p = 0.82, x = 5) = 0.0009$$

(b) It takes less than 4 attempts to make his first basket.

$$P(X < 4) = \text{geomcdf}(p = 0.82, x: 3) = 0.9942$$

(c) It takes at least 3 attempts to make his first basket.

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \text{geomcdf}(p: 0.82, x: 2) = 0.0324$$

(d) He will make 6 baskets in a row before he misses.

$$P(X = 7) = (0.82)^6(0.18) = \text{geompdf}(p: 0.18, x: 7) = 0.0547$$

Like all probability distributions, the geometric distribution has a mean and a standard deviation. If  $X \sim G(p)$  then:

$$\mu_x = \frac{1}{p}$$

$$\sigma_x = \frac{\sqrt{1-p}}{p}$$

**Example**

Drew decides to place a \$10 bet on Number 3 in consecutive spins of a roulette wheel until he wins. On any spin, Drew has a  $1/38$  chance that the ball will land in the Number 3 slot. Let  $X$  = number of spins until first win.

(a) How many spins do you expect it to take until Drew wins? What does this typically vary by?

$$\mu_x = \frac{1}{1/38} = 38 \text{ spins.}$$

$$\sigma_x = \frac{\sqrt{1 - \frac{1}{38}}}{\frac{1}{38}} = 37.4967 \text{ spins}$$

Recall: if an event has a probability less than 5%, we consider it to be rare/unusual/surprising if it were to actually happen.

(b) Would you be surprised if it took 10 or fewer spins for Drew to win? Calculate a probability to support your answer.

$$P(X \leq 10) = \text{geomcdf}(p: \frac{1}{38}, x: 10) = 0.2341, \text{ non unusual.}$$

The probability of this occurring is approximately 23.41%, so it would not be surprising.

(c) Would you be surprised if it took more than 13 spins of roulette wheel until he won? Calculate a probability to support your answer.

$$P(X > 13) = 1 - P(X \leq 13) = 1 - \text{geomcdf}(p: \frac{1}{38}, x: 13) = 0.7070.$$

The probability of this occurring is approximately 70.70%, so it would not be surprising.

# 5 Sampling Distributions

## 5.1 Sampling Distributions of Sample Proportions

I have a bag of beads - we are interested in the true proportion of purple beads in your population (bag).

- $n$  = population size = 50
- $p$  = population proportion = true prop. of purple beads = ?

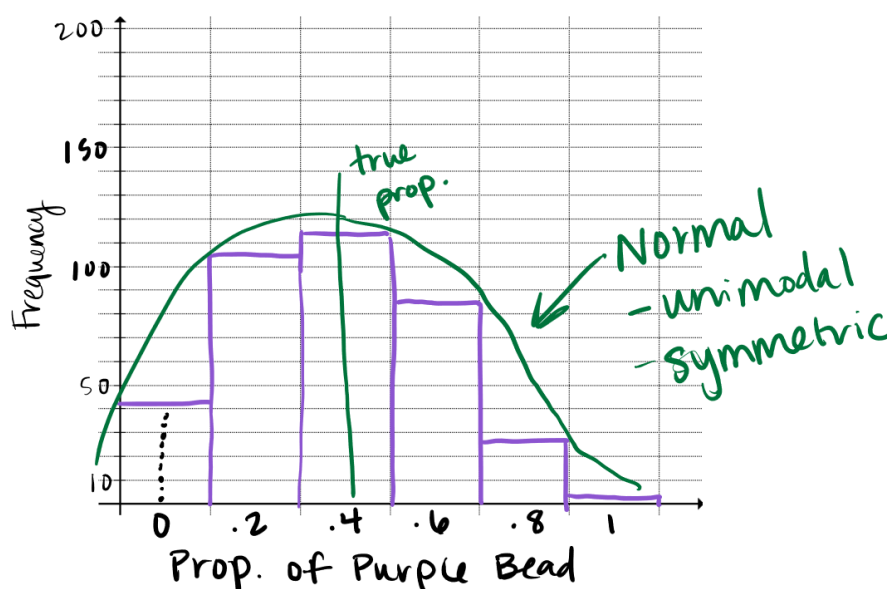
You received a bag of 50 beads. You will perform 20 trials of this experiment.

1. Mix the bag well before collecting each sample.
2. Without looking, randomly select 5 beads from your bag (without replacement).
3. Count the number of purple beads in your sample and record a tally in the appropriate column below.

Here is the data

Proportion of Purple Beads	0	.20	.40	.60	.80	1
Count	41	104	112	83	29	6

Sampling Distribution of the data



As we begin to use sample data to draw conclusions about a larger population, we must be clear about whether a number describes a sample or a population.

- Parameter: is a number that describes some characteristic of a population
- Statistics: is a number that describes some characteristic of a sample
- Unbiased Estimator: a sample proportion or sample mean that is equal to the population proportion or population mean

SAMPLE STATISTIC			POPULATION PARAMETER	
$\hat{p}$	(sample proportion)	estimates	$p$	(population proportion)
$\bar{x}$	(sample mean)	estimates	$\mu$	(population mean)
$s_x$	(sample standard deviation)	estimates	$\sigma$	(population standard deviation)

Rather than showing real repeated samples, imagine what would happen if we were to actually draw many samples. Now imagine what would happen if we looked at the sample proportions for these samples. The histogram we'd get if we could see all the proportions from all possible samples is called the sampling distribution of the sample proportions.

What would the histogram of all the sample proportions look like?

- We would expect the histogram of the sample population to center at the true population,  $p$ , in the population.
- The spread is calculated as standard deviation based on the true proportion,  $p$ , and the sample size,  $n$ . As the sample size gets larger the standard deviation will get smaller.
- The shape of the histogram would be unimodal and symmetric.
- More specifically, a normal model is just the right one for the histogram of sample proportions.

Assumptions and Conditions

- Randomness: The sample should be a simple random sample of the population. This allows us to calculate mean.
- Independence (10% Condition): The sample size,  $n$ , must be no larger than 10% of the population.

Basically, population  $\geq 10n$ . This allows us to calculate standard deviation.

- Normality (Large Counts Condition): The sample size has to be big enough so that both number of successes and number of failures is at least 10. We also refer to this as the Success/Fail Condition.

Success:  $np \geq 10$ , Fail:  $n(1 - p) \geq 10$ . This allows us to call our sampling distribution approximately normal.

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of  $p$  is modeled by a normal model with:

Sample Proportions:  $\hat{p} = \frac{\text{\#successes}}{\text{sample size}}$

Mean of Sample Proportions:  $\mu_{\hat{p}} = p$

Standard Deviation of Sample Proportions:  $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Where  $p$  is population proportion, and  $n$  is sample size.

**Example**

A recent study looked at the percentage of young adult women who get less than 7 hours of sleep a night. In this study, they say that 45% of women get less than 7 hours of sleep a night. What is the probability that a random sample of 50 women will result in a sample proportion of 50% or higher?

(a) Communicate the parameter in the context of the problem.

$p$  = true proportion of women who get less than 7 hours of sleep per night.

(b) Create the sampling distribution by finding the center, spread, and shape. Make sure to check each condition first.

Random: Random sample of 50 women,  $\mu_{\hat{p}} = 0.45$

Independent: Assume  $n = 50 \leq 0.10$  (all young women),  $\sigma_{\hat{p}} = \sqrt{\frac{0.45(1-0.45)}{50}} = 0.0704$

Normal:  $50(0.45) = 22.5 \geq 10$ ,  $50(0.55) = 27.5 \geq 10$ , sampling distribution is approximately normal.

(c) Calculate the probability by converting your data to a z-score and using normalcdf to find the probability.

$P(\hat{p} \geq 0.50) = P(z \geq 0.7102)$ . normalcdf(lower: 0.7102, upper:  $\infty$ ,  $\mu$ : 0,  $\sigma$ : 1) = 0.2388.

(d) Conclude the problem by answering the question in a complete sentence with the context of the problem.

When the true proportion of young women who sleep less than 7 hours is 45%, the probability a random sample of 50 women results in a sample proportion of 50% or higher is approximately 23.89%.

**Example**

Percy fails to study for his chemistry final. The final has 100 multiple choice questions, each with five choices. Assume all questions are independent and that Percy has given himself over to the gods of chance and randomly guesses on each of the 100 questions.

(a) What is the parameter in the context of the problem.

$p$  = true proportion of questions Percy gets correct

(b) Identify  $n$  and  $p$  for this problem.

$n = 100$ ,  $p = 0.20$

(c) Is the sampling distribution approximately normal? Why or why not?

$100(0.20) = 20 \geq 10$ ,  $100(0.80) = 80 \geq 10$ . Sampling distribution is approximately normal.

(d) Is the sampling distribution unbiased? Why or why not?

Because each question is randomly guesses, it is unbiased.

(e) Do we have to check the 10% condition here? Why or why not?

It is stated that each guess is independent, so there is no need to check.

(f) Calculate the mean and standard deviation for this problem.

$\mu_{\hat{p}} = 0.20$ ,  $\sigma_{\hat{p}} = \sqrt{\frac{0.20(1-0.20)}{100}} = 0.04$

(g) What is the probability that Percy will get between a 25% and a 50% on the test? Show all work.

z-score for 0.25 = 1.25, z-score for 0.50 = 7.5.

$P(1.25 \leq z \leq 7.5) = \text{normalcdf}(\text{lower: } 1.25, \text{upper: } 7.5, \mu : 0, \sigma_1) = 0.1056$ .

When the true proportion of correct questions is 0.20, the probability Percy will get between a 25% and 50% on the test is 0.1056.



## 5.2 Sampling Distributions of Sample Means

The notation for means is the following:

- Parameters:  $\mu$  and  $\sigma$
- Statistics:  $\bar{x}$  and  $s$
- Sampling Distribution Mean:  $\mu_{\bar{x}}$
- Sampling Distribution Standard Deviation:  $\sigma_{\bar{x}}$

Conditions for Sample Means

- Random - As long as the sampling method is random, our mean is an unbiased estimator.
- Independent - When sampling, we have to make sure the 10% condition is satisfied.
- Normal - No longer checking Large Counts (Success/Fail)

Central Limit Theorem:

- The central limit theorem (CLT) states that when the sample size is sufficiently large, a sampling distribution of the mean of random variable will be approximately normally distributed.
- The central limit theorem requires that the sample values are independent of each other and that  $n$  is sufficiently large.

Therefore:

- If the population distribution is normal, then so is the sampling distribution of  $\bar{x}$ . This is true no matter what the sample size  $n$  is.
- If the population distribution is not normal (or has an unknown shape), the central limit theorem tells us that the sampling distribution of  $\bar{x}$  will be approximately normal in most cases of  $n \geq 30$ .

Sampling Distributions for Sample Means:

Shape: Approximately Normal, as long as

- If  $n < 30$ , the population distribution is normal.
- If  $n \geq 30$ , the CLT tells us the sampling distribution will be approximately normal.

Center:  $\mu_{\bar{x}} = \mu$ , as long as

- You randomly selected from the population of interest (sampling) or randomly assigned treatments (experiments).

Spread:  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ , as long as

- Your sample size is less than 10% of the population of interest.
- You only use 10% condition when sampling otherwise we will have to assume independence.

**Example**

Suppose the heights of young women are normally distributed with  $\mu = 64.5$  inches and  $\sigma = 2.5$  inches. What is the probability that the mean height of an SRS of 10 young women is greater than 65 inches?

(a) Communicate the parameter in the context of the problem.

$\mu$  = true average height of a young woman

(b) Create a sampling distribution by finding the center, spread, and shape. Make sure to check conditions!

Random: SRS of 10 young women,  $\mu_{\bar{x}} = 64.5$

Independence: Assume  $n = 10 \leq 0.10$ (all young women).  $\sigma_{\bar{x}} = \frac{2.5}{\sqrt{10}} = 0.7906$

Normal: Population is normally distributed therefore the sampling distribution is normal.

(c) Calculate the probability by converting your data to a z-score and finding the probability.

$$P(z > \frac{65-64.5}{.7906}) = P(z > 0.6324).$$

$$\text{normalcdf}(\text{lower: } 0.6324, \text{upper: } \infty, \mu : 0, \sigma : 1) = 0.2636.$$

(d) Conclude by answering the question in a complete sentence with the context of the problem.

When the mean height of young women is 64.5 inches, the probability of selecting an SRS of 10 young women with a mean height greater than 65 inches is 26.36%.

**Example**

A restaurant is suspected of undercooking its burgers. The restaurant claims that its burgers are cooked to a medium doneness, with an average internal temperature of 160 degrees Fahrenheit and a standard deviation of 5 degrees Fahrenheit. For the questions that follow, assume that the restaurant's claim is accurate and that the distribution of internal temperatures follows a normal distribution.

(a) What is the probability that a single randomly selected burger is cooked to an internal temperature of 158 degrees or less.

$$P(x \leq 158) = P(z \leq \frac{158-160}{5}) = P(z \leq -0.4).$$

$$\text{normalcdf}(\text{lower: } -\infty, \text{upper: } -0.4, \mu : 0, \sigma : 1) = 0.3446.$$

(b) A secret shopper comes in at random times throughout the data to test the internal temperature of the burgers. They select an SRS of 25 burgers and calculate the sample mean. What are the mean and standard deviation of the resulting sampling distribution?

Random: SRS Of 25 burgers.  $\mu_{\bar{x}} = 160$

Independence: Assume  $n = 25 \leq 0.10$ (all burgers).  $\sigma_{\bar{x}} = \frac{5}{\sqrt{25}} = 1$

(c) The secret shopper in part (b) obtains a sample mean of  $\bar{x} = 158$  degrees Fahrenheit. What is the probability that a random sample of 25 burgers produces a sample mean amount of 158 degrees or less? Assume all conditions are met.

$$P(\bar{x} \leq 158) = P(z \leq -2) = \text{normalcdf}(\text{lower: } -\infty, \text{upper: } -2, \mu : 0, \sigma : 1) = 0.0228.$$

(d) Why is there a difference between the results of (a) and (c)? Explain why one is more likely to happen than the other.

In Part (a) we are looking at the probability a single burger is 158 degrees or less, and since the distribution has more variability in it, we have a higher probability of it happening randomly. In part (c) we are looking at the probability that 25 burgers have an average temperature of 158 degrees or less, and since averages have less variability than individual values, we have a lower probability of it happening randomly.

### 5.3 Combining Sample Proportions and Sample Means

Creating sampling distributions of a difference in sample proportions will set us up for success when we study inferences in Unit 6.

These sampling distributions can be used to answer the common question of “which is better?” For example:

- Which of two popular drugs - Lipitor or Pravachol - helps lower “bad cholesterol” more?
- Researchers used 4000 people with heart disease as subjects in a completely randomized experiment. They were randomly assigned to one of two treatment groups: Lipitor or Pravachol.
- At the end of the study, researchers compared the proportion of subjects in each group who had died, had a heart attack, or suffered other serious consequences within two years.
- For the subjects assigned to Pravachol, 0.263 suffered a serious consequence.
- For the subjects assigned to Lipitor, 0.224 suffered a serious consequence.

Does this mean that Lipitor is better at lowering bad cholesterol? Is this difference a result of Lipitor being better or is it merely due to a chance involved in the random assignment of treatments? Answers to these questions require a sampling distribution of a difference in sample proportions.

$\hat{p}_1$  will denote the sample proportion from the first group.

Center:  $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ , as long as:

- Both samples must have been randomly selected from the population (or involve random assignment for an experiment).

Spread:  $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}$  as long as

- Both samples satisfy the 10% condition:

$$n_1 < 0.10N_1$$

$$n_2 < 0.10N_2$$

- If this is an experiment, check if it is safe to assume independence.

Shape: Approximately normal, as long as

- Both samples must satisfy the success/fail condition.

$$\begin{aligned} n_1 p_1 &\geq 10 \text{ and } n_1(1 - p_1) \geq 10 \\ n_2 p_2 &\geq 10 \text{ and } n_2(1 - p_2) \geq 10 \end{aligned}$$

**Example**

A recent study examined the percentage of young adult women and men who get less than 7 hours of sleep a night. In this study, it was found that 45% of women and 38% of men get less than 7 hours of sleep a night. What is the probability that a random sample of 50 women and 75 men will result in a difference of 15% or more between the groups?

(a) State the parameters in the context of the problem.

$p_w$ : proportion of women who get less than 7 hours of sleep

$p_m$ : proportion of men who get less than 7 hours of sleep.

(b) Create the sampling distribution by finding the center, spread, and shape. Make sure to check each condition first.

Random: Random sample of 50 women and 75 men.  $\mu_{\hat{p}_w - \hat{p}_m} = 0.45 - 0.38 = 0.07$

Independent: Assume  $n_w = 50 \leq 0.10(\text{all women})$  and  $n_m = 75 \leq 0.10(\text{all men})$ .

$$\sigma_{\hat{p}_w - \hat{p}_m} = \sqrt{\frac{.45(.55)}{50} + \frac{.38(.62)}{75}} = 0.0900$$

Normal:  $50(.45) = 22.5 \geq 10$ ,  $75(.38) = 28.5 \geq 10$ ,  $50(.55) = 27.5 \geq 10$ ,  $75(.62) = 46.5 \geq 10$ . Sampling distribution is approximately normal.

(c) Calculate the probability by converting your data to a z-score and using normalcdf to find the probability.

$$P(p_w - p_m \geq 0.15) = P(z \geq 0.8889) = 0.1870$$

$$P(p_w - p_m \leq -0.15) = P(z \leq -2.4444) = 0.0073$$

(d) Conclude the problem by answering the question in the context of the problem.

The probability that a random sample of 50 women and 75 men will result in a difference of 15% or more getting less than 7 hours of sleep between the two groups is approximately 19.43%.

Sampling Distribution of  $\bar{x}_1 - \bar{x}_2$ :

Center:  $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$  as long as

- Both samples must have been randomly selected from the population (or involve random assignment for an experiment)

Spread:  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$  as long as

- Both samples satisfy the 10% condition:

$$n_1 < 0.10N_1$$

$$n_2 < 0.10N_2$$

- If this is an experiment, check if it is safe to assume independence.

Shape: Approximately normal, as long as

- $n_1 \geq 30$  and  $n_2 \geq 30$  to use the Central Limit Theorem (CLT), and if one or both are less than 30, each population should be normal.

**Example**

A study is conducted to compare the effectiveness of two study techniques, A and B, in improving test scores. For Technique A, a random sample of 80 students yields an average score of 85 with a standard deviation of 10. For Technique B, a random sample of 100 students yields an average score of 82 with a standard deviation of 8. Calculate the probability that the difference in the average scores between Technique A minus Technique B is greater than 4.

(a) State the parameters in the context of the problem.

$\mu_A$ : mean score of students using Technique A

$\mu_B$ : mean score of students using Technique B

(b) Create the sampling distribution by finding the center, spread, and shape. Make sure to check each condition first.

Random: Random sample of 80 Technique A and 100 Technique B Students.  $\mu_{\bar{x}_A - \bar{x}_B} = 85 - 82 = 3$

Independent: Assume  $n_A = 80 \leq 0.10$ (all technique A students) and  $n_B = 100 \leq 0.10$ (all technique B students).  $\sigma_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{10^2}{80} + \frac{8^2}{100}} = 1.3748$ .

Normal:  $n_A = 80 \geq 30$ ,  $n_B = 100 \geq 30$ . CLT says sampling distribution will be approximately normal.

(c) Calculate the probability by converting your data to a z-score and using normalcdf to find the probability.

$$P(\mu_{\bar{x}_A} - \mu_{\bar{x}_B} > 4) = P(z > 0.7274) = 0.2335.$$

(d) Conclude the problem by answering the question in context of the problem.

The probability that the difference in the average scores between Technique A and B is greater than 4 is approximately 23.35%.

# 6 Inference for Categorical Data: Proportions

## 6.1 Constructing a One Proportion z-Interval

Definition: A confidence interval for a population parameter is an interval of plausible values for that unknown parameter.

It is constructed in such a way so that, with a chosen degree of confidence, the value of the parameter will be captured inside the interval.

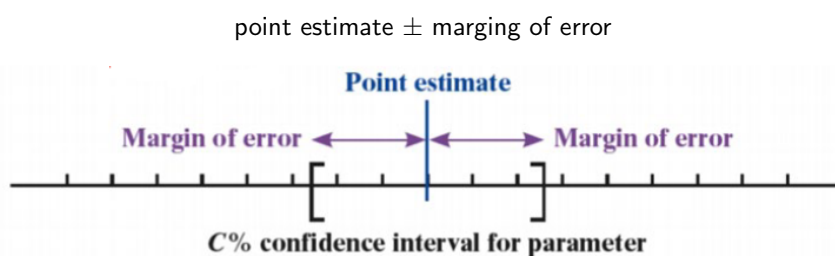
The chosen degree of confidence is called the confidence level. The confidence level gives information about how much “confidence” we have in the method used to construct the interval.

Interpreting a Confidence Level:

- If we were to select many random samples  $n$  in context and construct a \_\_\_\_\_% confidence interval using each sample, about \_\_\_\_\_% of the intervals would capture the parameter in context.

To create an interval of plausible values for a parameter, we need two components:

- A point estimate is a single value used to estimate the population parameter such as a sample proportion.
- A margin of error represents the maximum expected difference between the two population parameter and the sample estimate.



Constructing a Confidence Interval

P	Define the Parameter
A	Check the Assumptions and Conditions
N	Name the Inference Method
I	Calculate the Interval
C	Write your Conclusion in Context

- Define the parameter

$p$  = true proportion of parameter in context

- Check the Assumptions and Conditions

- Random Condition: The sample should be a random sample of the population.
- 10% Condition: The sample size,  $n$ , must be no larger than 10% of the population.
- Success/Fail Condition: The sample has to be large enough so that there are at least 10 success and 10 failures.

$$n\hat{p} \geq 10 \quad n(1 - \hat{p}) \geq 10$$

- The value of the true population proportion ( $p$ ) is unknown so we use the sample proportion ( $\hat{p}$ )

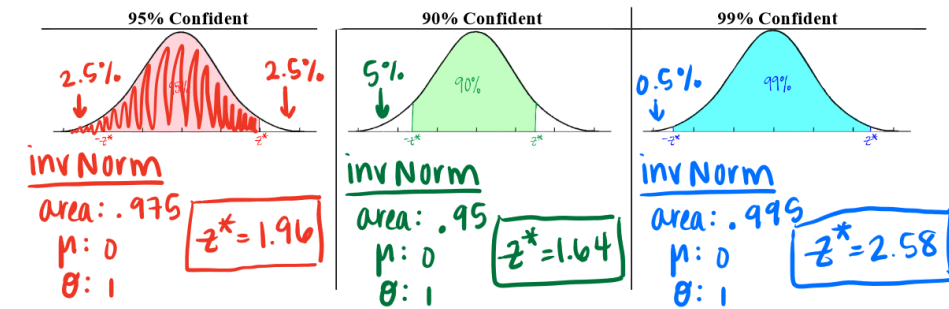
- Name the Inference Method:
  - Method: One Proportion z-Interval
- Calculate the Interval

point estimate  $\pm$  margin of error

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

Recall: The value of the true population proportion ( $p$ ) is unknown so we use the sample proportion ( $\hat{p}$ ).

How do we calculate the Critical Value ( $z^*$ )



- Write your Conclusion in Context
  - We are \_\_\_\_\_% confident that the interval from \_\_\_\_\_ to \_\_\_\_\_ captures the true proportion of parameter in context.

### Example

A New York Times poll asked a random sample of 400 US adults the question, "Do you favor an amendment to the Constitution that would permit organized prayer in public schools?" Based on this poll, the 95% confidence interval for the population proportion who favor such an amendment is (0.63, 0.69).

(a) Interpret the confidence interval in context.

We are 95% confident that the interval from 0.63 to 0.69 captures the true proportion of US adults who favor the amendment.

(b) Interpret the confidence level in context.

If we were to select many random samples of 400 US adults and construct a 95% confidence interval using each sample, about 95% of them would capture the true proportion of US adults who favor the amendment.

(c) What is the point estimate that was used to capture the interval? What is the margin of error?

The point estimate is  $\hat{p} \pm$  margin of error.

The margin of error is  $\frac{0.69 - 0.63}{2} = 0.03$ .

The point estimate point estimate is therefore  $\hat{p} = 0.63 + 0.03 = 0.66$ .

(d) Based on this poll, a reporter claims that more than two-thirds of US adults favor such an amendment. Use the confidence interval to evaluate this claim.

Their claim is incorrect, our confidence interval states that the true proportion could also be below 0.67.

**Example**

In a random sample of 250 Texas high school students, it was found that 210 of them later graduated. Calculate and interpret a 95% confidence interval to estimate the proportion of all Texas high school students who graduate.

$p$  = true proportion of Texas HS students who graduate

Random sample of 250 TX HS students:  $\hat{p} = \frac{210}{250} = 0.84$

Independence:  $n = 250 \leq 0.10(\text{all TX HS Students})$ .  $\sigma_{\hat{p}} = \sqrt{\frac{0.84(1-0.84)}{250}} = 0.0232$

Normal:  $210 \geq 10$ ,  $250 - 210 = 40 \geq 10$ , sampling dist. is approx. normal.

One proportion z-interval.

$\hat{p} \pm z^*(\text{std error}) = .84 \pm 1.96(0.0232)$ .

The interval is (0.7945, 0.8855).

We are 95% confident the interval from 0.7945 to 0.8855 captures the true proportion of Texas high school students who graduate.

Using your calculator you can do the following.

STAT-Tests-A: 1-PropZInt

- $x$ : number of successes
- $n$ : sample size
- C-Level: Confidence Level

Important: Your number of successes must be whole number otherwise your calculator will give you an error.

## 6.2 Constructing a One Proportion z-Test

A significance test is another inference method that assesses evidence provided by data about a claim. Significance tests tell us if sample data gives us convincing evidence against a null hypothesis.

- A null hypothesis ( $H_0$ ) is the claim being assessed in a significance test. Usually, the null hypothesis is a statement of “no change from the expected value”.
- An alternative hypothesis ( $H_A$ ) proposes what we should conclude if we find the null hypothesis to be unlikely.

Hypotheses always refer to the population, not the sample. We use  $p$  and not  $\hat{p}$ .

A p-value is the probability of getting results as extreme or more extreme in the direction of the null hypothesis by random chance alone assuming the claim of the null hypothesis is true.

- Small p-values give convincing evidence against the null hypothesis since the result we got is unlikely to occur.
- Large p-values fail to give convincing evidence against the null hypothesis since the result we got is likely to occur.

Interpreting a P-value.

- Assuming that the true proportion of parameter in context is null hypothesis, there is a p-value probability of getting a sample proportion of alternative hypothesis just by chance in a random sample of  $n$  w/ units.



**Example**

Jason claims he makes 85% of his free throws. Thatcher believes that he makes less than 85% and just wants to test Jason's claim. Jason shoots 100 free throws and makes 82 of them. A significance test is performed and a p-value of 0.2004 is obtained. Interpret the p-value in context.

Assuming that the true proportion of free throws Jason makes is 85%, there is a 20.04% probability of getting a sample proportion of 82 free throws or less just by chance in a random sample of 100 free throws.

The significance level ( $\alpha$ ) is a fixed value that we will regard as the decisive value that determines if the p-value is small or large.

- Typically we choose  $\alpha = 0.05$  which says we need data so strong that it would happen by chance less than 5% of the time.

Constructing a Significance Test:

- P: Define the Parameter
- H: State the Hypotheses
- A: Check the Assumptions and Conditions
- N: Name the Inference Method
- T: Calculate the Test Statistics
- O: Obtain the P-Value
- M: Make a Decision
- S: State the Conclusion in Context

Define the parameter:

$p$  = true proportion of parameter in context

State the Hypotheses:

- Null Hypothesis:  $H_0 : p = p_0$
- Alternative Hypothesis:
  - $H_A : p < p_0$
  - $H_A : p > p_0$
  - $H_A : p \neq p_0$

If you are not given a claimed proportion, we use a conservative estimate which is 0.50.

Check the Assumptions and Conditions

- Random Condition: The sample should be a random sample of the population.
- 10% Condition: The sample size  $n$ , must be no larger than 10% of the population.
- Success/Fail Condition: The sample size has to be large enough so that there are at least 10 success and 10 failures.

$$np_0 \geq 10 \quad n(1 - p_0) \geq 10$$

Name the Inference Method:

Method: One Proportion z-Test

Calculate the Test Statistic

On the formula chart:

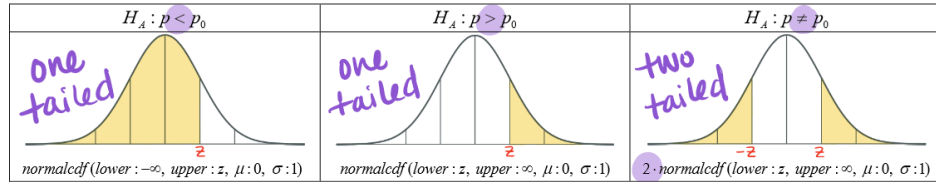
$$\text{test statistic} = \frac{\text{statistics-parameter}}{\text{standard deviation of statistic}}$$

## One Proportion z-Test

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where  $\hat{p}$  is sample proportion,  $p_0$  is the null proportion, and  $n$  is sample size.

Obtain the P-Value



## Make a Decision

- If the p-value is less than  $\alpha$ , we reject the null hypothesis.
- If the p-value is greater than  $\alpha$ , we fail to reject the null hypothesis.

## Write your Conclusion in Context

Since our p-value of \_\_\_\_\_ is (less/greater) than  $\alpha = \text{_____}$ , we (reject/fail to reject) the null hypothesis. There (is/is not) convincing evidence that alternative hypothesis.

**Example**

According to the U.S. Census Bureau, the proportion of students in high school who have a part-time job is 0.25. An administrator at a local high school suspects that the proportion of students at her school who have a part-time job is less than the national figure. The administrator selects a random sample of 200 students from the school and finds that 39 of them have a part-time job. Is there convincing evidence that the proportion of students at the administrator's school who have a part-time job is less than the national figure?

$p$  = true proportion of students at admin's school with a part time job.

$$H_0 : p = 0.25, H_A : p < 0.25$$

- Random: Random sample of 200 students at admin's school
- Independence:  $n = 200 \leq 0.10(\text{all students at admin's school})$
- Normal:  $200(0.25) = 50 \geq 10$ ,  $200(0.75) = 150 \geq 10$ , sampling dist. is approx. normal.

## One Proportion z-Test

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = -1.7963.$$

$$P(z < -1.7963) = 0.0362.$$

Since the p-value of 0.0362 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that the proportion of HS students with part time jobs is less than the national claim at admin's school.

**Example**

According to the Centers for Disease Control and Prevention, 68% of high school students have never smoked a cigarette. Serene wonders if this national figure holds true at her high school. Serene takes an SRS of 150 students from her school. She gets responses for all 150 students and 62% say that they have never smoked a cigarette. Is there convincing evidence that the CDC's claim does not hold true at Serene's school?

$p$  = true prop. of students at Serene's high school who never smoked a cigarette

$H_0 : p = 0.68, H_A : p \neq 0.68$

- Random: SRS of 150 students at Serene's school
- Independence:  $n = 150 \leq 0.10(\text{all high school students at Serene's school})$
- Normal:  $150(0.68) = 102 \geq 10, 150(1 - 0.68) = 48 \geq 10$ . Sampling dist. is approx. normal.

One proportion z-Test

$z = -1.5753, P(z < -1.5753) = 0.0576$ .

Multiply this by two, since the alternative is two tailed, so  $p = 0.1152$ .

Since the p-value of 0.1152 is greater than  $\alpha = 0.05$ , we fail to reject the null. There is not convincing evidence that the proportion of students at Serene's high school who never smoked a cigarette is different than the CDC's claim.

Calculator Steps:

STAT-Tests-5:1-PropZTest:

- $p_0$ : null proportion
- $x$ : number of successes
- $n$ : sample size
- prop: alternative hypothesis

Important: Your number of successes must be a whole number otherwise your calculator will give you an error.

### 6.3 Relating Confidence Intervals and Significance Tests

#### Example

A recent study suggested that 77% of teenagers have texted while driving. A random sample of 50 teenage drivers at the school was taken and 33 admitted to texting while driving. Assume all conditions have been met. Use your calculator answer the following questions.

(a) Construct a 99% confidence interval to estimate the true parameter of teens who text while driving.

$p$  = true prop. of teens who have texted while driving at this school.

one proportion z-interval:  $(0.4874, 0.8326)$ ,  $\hat{p} = 0.66$ ,  $n = 50$ .

We are 99% confident the interval from 0.4874 to 0.8326 captures the true prop. of teens who have texted while driving at this school.

(b) Conduct a significance test to determine if the proportion of teens who text while driving is different from 77%. Use a 1% significance level.

$H_0 : p = 0.77$ ,  $H_A : p \neq 0.77$

one proportion z-test.

$z = -1.8483$ ,  $\hat{p} = 0.66$ ,  $p = 0.0646$ ,  $n = 50$ .

Since the p-value of 0.0646 is greater than  $\alpha = 0.01$ , we fail to reject the null. There is not convincing evidence that the true prop. of teens who text and drive is different than 77%.

(c) What type of results did you get for the confidence interval and significance test?

- Confidence interval: Interval of plausible values for  $p$
- Significance Test: said 77% is still a plausible value for  $p$ .

Confidence interval contains 0.77 which agrees with failing to reject  $H_0 : p = 0.77$ .

100(1 - $\alpha$ )% confidence interval	$\leftrightarrow$	significance level $\alpha$ with $H_0: p = p_0$ $H_a: p \neq p_0$
90% Confidence Interval	$\leftrightarrow$	<del>10%</del> Significance Level
<del>95%</del> Confidence Interval	$\leftrightarrow$	5% Significance Level
99% Confidence Interval	$\leftrightarrow$	<del>1%</del> Significance Level

- If the confidence interval contains the null hypothesis from a two-sided test, we would fail to reject the null.
- If the confidence interval does not contain the null hypothesis from a two-sided test, we would reject the null.

**Example**

Scott's Turf Lawn Builder makes the claim that, when a user plants their grass seeds, 88% of seeds will germinate. The germination rate of seeds is defined as the proportion of seeds that, when properly planted in the fertilizer and watered, sprout and grow. The company regularly tests this claim, to make sure its product is efficient. They plant 500 grass seeds in their fertilizer and find that 412 of the seeds germinate. Assume that all conditions are met.

(a) What is the 95% confidence interval for this observation?

(0.7906, 0.8574)

(b) Suppose the company conducted a test of  $H_0 : p = 0.88$  against the alternative  $H_A : p \neq 0.88$ , using  $\alpha = 0.05$ . Use the confidence interval to determine whether this test would reject or fail to reject the null hypothesis. Explain your reasoning.

Because our 95% confidence interval does not contain 0.88, we reject the null. There is convincing evidence that the germination rate of seeds is not 0.88.

(c) Find the p-value for the test. Explain what the p-value measures in the context of the problem.

$z = -3.8534$ ,  $p = 0.001$ . Assuming the germination rate is 0.88, there is a 0.0001 prob. of getting  $\hat{p} = 0.824$  or more extreme just by random chance alone.

point estimate  $\pm$  margin of error

$$\text{Statistic} \pm (\text{critical value})(\text{std. error})$$

$$\hat{p} \pm (z^*) \left( \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \rightarrow \text{margin of error}$$

When increasing confidence level

- Critical value increases
- Margin of error increases
- Wider interval

When decreasing confidence level

- Critical value decreases
- Margin of error decreases
- Narrower interval

When increasing sample size

- Standard error decreases
- Margin of error decreases
- Narrower interval

When decreasing sample size

- Standard error increases
- Margin of error increases
- Wider interval

Keep in mind

- The margin of error in a confidence covers only accounts for sampling variability
- The margin of error does not account for bias in the sampling methods

Often time, researchers need to both limit the margin of error to a fixed amount and still maintain a certain confidence level. In such cases, the only thing that can vary is the sample size, since the value for  $\hat{p}$  is something that is out of our control.

If we know our margin of error and our confidence level, we can solve for  $n$  to determine the sample size needed to obtain a certain margin of error size.

However, we don't know  $\hat{p}$  until we actually conduct the study.

We can then do one of two things:

1. Use an estimated  $\hat{p}$  based on previous studies (this would be the wording used in the problem)
2. Use a "conservative"  $\hat{p} = 0.5$ . Using a  $\hat{p}$  of 0.5 gives the largest possible margin of error for any given  $z$  or  $n$ . We will use this option when a value of  $\hat{p}$  is not given in the problem.

### Example

In 2009 a survey of Internet usage found that 79 percent of adults age 18 years and older in the United States use the Internet. A broadband company believes that the percent is greater now than it was in 2009 and will conduct a survey. The company plans to construct a 98 percent interval to estimate the current percent and wants the margin of error to be no more than 2.5 percentage points. Assuming that at least 79 percent of adults use the Internet, how many people must they random sample to achieve this?

The  $z^*$  for a 98% confidence interval is 2.33.

We have  $0.025 \geq 2.33 \sqrt{\frac{0.79(1-0.79)}{n}}$ , so  $n \geq 1441.0472$ .

We need at least 1442 people.

## 6.4 Inference for Comparing Two Population Proportions

Constructing a two proportion z-interval

- Define the parameter
  - $p_1$  = true proportion of parameter in context for Population/Treatment 1
  - $p_2$  = true proportion of parameter in context for Population/Treatment 2
- Check the Assumptions and Conditions
  - Random Condition: Each sample should be a random sample of both populations.
    - \* If dealing with treatment groups, you must have "random assignment of treatments."
  - Independence (10% Condition): The sample size,  $n$ , must be no larger than 10% for both populations.
  - Normality (Success/Fail Condition): The sample has to be large enough to have at least 10 successes/fails

$$\begin{aligned} n_1(\hat{p}_1) &\geq 10 & n_1(1 - \hat{p}_1) &\geq 10 \\ n_2(\hat{p}_2) &\geq 10 & n_2(1 - \hat{p}_2) &\geq 10 \end{aligned}$$

- Name the inference method: Two Proportion z-Interval for  $p_1 - p_2$
- Calculate the interval

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

Calculating the  $z^*$  still remains the same as previously shown.

- Write your conclusion in context

- We are \_\_\_\_\_% confident that the interval from \_\_\_\_\_to \_\_\_\_\_captures the true difference between the proportion of parameter in context.

**Example**

A survey was conducted to compare the proportion of adults and teens who use social networking sites daily. The first survey took an SRS of 950 U.S. teenagers (aged 13-19). The second survey took an SRS of 2975 U.S. adults. In the two studies, 83% of teens and 72% of adults used social media daily. Construct and interpret a 95% confidence interval for the difference between the proportion of teens and adults who use social media daily.

- $P_A$  = true proportion of adults who use social media daily
- $P_T$  = true proportion of teens who use social media daily
- Random: SRS of 950 US teens and 2975 US Adults
- Independence:  $n_T = 950 \leq 0.10(\text{all US teens})$ ,  $n_A = 2975 \leq 0.10(\text{all US adults})$
- Normal:  $2975(0.72) = 2142 \geq 10$ ,  $2975(1 - 0.72) = 833 \geq 10$ ,  $950(0.83) = 788.5 \geq 10$ ,  $950(1 - 0.83) = 161.5 \geq 10$

Sampling Dist. is approx. Normal.

Two Proportion z-Interval for  $p_A - p_T$

You can either use the formula above and you will get  $(-0.1388, -0.0812)$ , or the following calculator steps.

STAT-Tests-B:2-PropZInt

- x1: the number of successes in the sample population of Population 1
- n1: sample size from Population 1
- x2: number of successes in sample of Population 2
- n2: sample size from Population 2
- C-Level: Confidence Level

and using this you get  $(-0.1393, -0.0817)$ .

We are 95% confident that the interval from  $-0.1393$  to  $-0.0817$  captures the true difference between the proportion of US teens and adults who use social media daily.

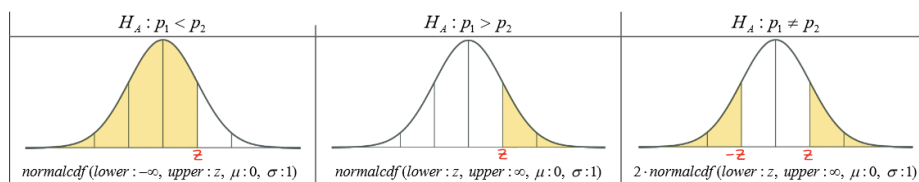
**Constructing a Two Proportion z-Test**

- Define the parameter - same as the two proportion z-interval
- State the hypothesis
  - Null hypothesis:  $H_0 : p_1 = p_2$
  - Alternative hypothesis:  $H_A : p_1 < p_2$ ,  $H_A : p_1 > p_2$ ,  $H_A : p_1 \neq p_2$
- Check the Assumptions and Conditions - Same as Confidence Interval except Success/Fail is checked using  $\hat{p}_c$
- Name the Inference Method: Two Proportion z-Test
- Calculate the Test Statistic: The null hypothesis states that there is no difference between the two population proportions. If this is true, the observations really come from a single population. So instead of using  $\hat{p}_1$  or  $\hat{p}_2$  separately, we use  $\hat{p}_c$  = p-combined.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c(1 - \hat{p}_c) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where  $\hat{p}$  is the sample proportion,  $\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$ ,  $x$  is the number of successes, and  $n$  is the sample size.

- Obtain the p-value



- Make a Decision
  - If the p-value is less than  $\alpha$ , we reject the null hypothesis.
  - If the p-value is greater than  $\alpha$ , we fail to reject the null hypothesis.
- State the conclusion:
  - Since the p-value of \_\_\_\_\_ is (less/greater) than  $\alpha = \text{_____}$ , we (reject/fail to reject) the null hypothesis. There (is/is not) convincing evidence that alternative hypothesis.

### Example

Jacob, an AP Statistics student, is curious to know if more male students than female students own an iPhone at his high school. He takes an SRS of 90 females and 120 male students. The survey found that 61 females and 97 males owned an iPhone. Is there convincing evidence at the  $\alpha = 0.05$  level that more male than female students own an iPhone at Jacob's school?

- $P_M$  = true proportion of males at Jacob's HS with iPhones
- $P_F$  = true proportion of females at Jacob's HS with iPhones
- $H_0 : p_M = p_F$
- $H_A : p_M > p_F$
- Random: SRS of 90 females and 120 males at Jacob's high school
- Independence:  $n_M = 120 \leq 0.10$  (all males at Jacob's HS),  $n_F = 90 \leq 0.10$  (all females at Jacob's HS)
- Normal:  $120(0.7524) = 90.29 \geq 10$ ,  $120(1 - 0.7524) = 29.71 \geq 10$ ,  $90(0.7524) = 67.62 \geq 10$ ,  $90(1 - 0.7524) = 22.28 \geq 10$ . Sampling dist. is approx. normal

To find the p-value:  $\hat{p}_m = \frac{97}{120} = 0.8083$  and  $\hat{p}_F = \frac{61}{90} = 0.6778$ . Plug this into the formula given above, and we get  $z = 2.1683$ . Note that  $\hat{p}_c = 0.7524$  to do this calculation.

Using normalcdf gives us a p-value of 0.0151.

We can also use a calculator:

STAT-Tests-6:2-PropZTest:

- x1: number of successes in sample of Population 1
- n1: sample size from Population 1
- x2: number of successes in sample of Population 2
- n2: sample size from Population 2
- p1: alternative hypothesis

This will give the z-score and the p-value.

Since the p-value of 0.0151 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that more male than female students at Jacob's high school own an iPhone.



## 6.5 Errors & Power

The courtroom analogy

		The Truth	
Jury's Decision		Not Guilty	Guilty
	Guilty	Innocent Person Goes to Jail	✓
	Not Guilty	✓	Guilty Person Goes Free

What would be considered the bigger mistake?

- If the crime was murder, would a guilty person going free be the bigger mistake?
- If the crime was shoplifting, would an innocent person going to jail be the bigger mistake?
- It really comes down to the situation - statisticians have to decide what the bigger error would be before they conduct their tests.
- A jury is going to decide whether the defendant is guilty or not guilty → fail to reject or reject the null
- Always start with the assumption that the defendant is innocent → assume the claim  $H_0$  is true
- The court presents evidence and a decision is made → Find p-value based on a sample.





		The Truth About the Population	
Your Decision		Fail to Reject Null (Stick with $H_0$ ) <del><math>H_0</math> True</del>	Reject Null (Evidence for $H_a$ ) <del><math>H_a</math> True</del>
	Reject Null (Evidence for $H_a$ )	Type I Error	✓ Power
	Fail to Reject Null (Stick with $H_0$ )	✓	Type II Error FIR

**Example**

$H_0$  : You are not pregnant.

$H_A$  : You are pregnant.

When you go into a doctor's office, it is always assumed you are not pregnant until we get tests that tell us otherwise.

		The Truth About the Population	
		$H_0$ TRUE	$H_A$ TRUE
Your Decision	$H_A$	<b>Type I Error</b> Truth: <i>you are not Pregnant (<math>H_0</math>)</i> Decision: <i>you are Pregnant (<math>H_A</math>)</i> 	<b>Correct Decision (Power)</b> Truth: <i>you are Pregnant (<math>H_A</math>)</i> Decision: <i>you are Pregnant (<math>H_A</math>)</i> 
	$H_0$	<b>Correct Decision</b> Truth: <i>you are not Pregnant (<math>H_0</math>)</i> Decision: <i>you are not Pregnant (<math>H_0</math>)</i> 	<b>Type II Error</b> Truth: <i>you are Pregnant (<math>H_A</math>)</i> Decision: <i>you are not Pregnant (<math>H_0</math>)</i> 

**Type I Error**

- Reject  $H_0$  incorrectly.
- The probability of a Type I error is equal to the significance level.
- $P(\text{Type I Error}) = \alpha$
- Our significance level tells us what p-value is "low enough".

**Type II Error**

- Fail to reject  $H_0$  incorrectly
- $P(\text{Type II Error}) = \beta$

**Power**

- Reject  $H_0$  correctly
- The probability we reject the null correctly is 1 minus the probability we reject the null incorrectly
- $\text{Power} = 1 - \beta$

**Errors and their relationships**

- Type I and Type II errors have an indirect relationship: as the probability of one increases, the probability of the other decreases.
- Our Type I error is set with our significance level.
- The higher our significance level is, the lower our probability of failing to reject becomes, which is why  $\alpha$  and  $\beta$  have an indirect relationship.
- The higher our significance level, the more likely we will reject the null, which increases the likelihood we do that incorrectly (as well as correctly)

- Therefore, Type I error and Power have a direct relationship: as the probability of Type I Error increases, the higher the Power of the test
- What type of error would you rather have? Well that depends on the problem!
- For some tests, we want a low Type I but for others we want a low Type II.
- This is why being able to interpret errors is a key skill on the AP Exam.

**Example**

A drug manufacturer claims that less than 10% of patients who take its new drug for treating gestational diabetes will experience nausea. To test this claim, researchers conduct an experiment at the 5% significance level.

(a) What the null and alternative hypothesis?

- $H_0 : p = 0.10$
- $H_A : p < 0.10$

(b) Describe a Type I error and a Type II error in this setting, and explain the consequences of each.

Type I: We say less than 10% of patients experience nausea, when really it is equal to 10%. Manufacturer won't do anything to fix the medicine because they believe it is okay.

Type II: We say 10% of patients experience nausea, when really it is less than 10%. Manufacturer will spend time to fix the drug when they don't need to.

(c) The test has a power of 0.54 to detect that the alternative is true. Explain what the power means in this setting. What is one way we can increase the power?

The probability we find less than 10% of patients experiencing nausea correctly is 0.54. We can increase the power by increasing the significance level.

(d) You know that the power of this test at the 5% significance level against the alternative is 0.54. If you decide to use  $\alpha = 0.01$  instead of the 5% significance level, with no other changes to the test, will the power increase or decrease? Explain.

Power will decrease because the probability we reject  $H_0$  completely will decrease.

# 7 Inference for Quantitative Data: Means

## 7.1 Constructing a One Sample t-Interval

Review:

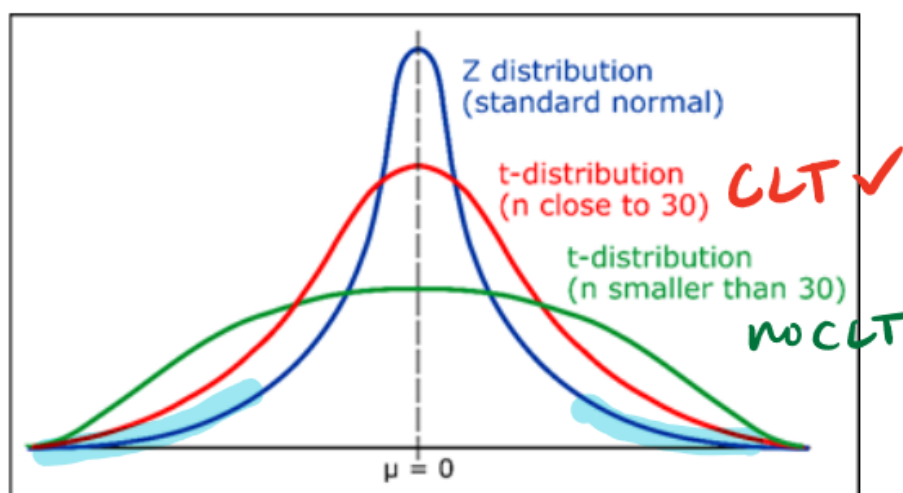
- Shape
  - Normal population indicates a normal sampling distribution
  - $n < 30$  then data needs to be stated as approximately normal.
  - $n \geq 30$  satisfies Central Limit Theorem which guarantees approximately normal
- Center
  - As long as you are taking a random sample or using random assignment,  $\mu_{\bar{x}} = \mu$
- When sampling, as long as 10% condition is met,  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

As long as the above conditions are met, the sampling distribution of  $\bar{x} : \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

When we do not know the population standard deviation (which we usually don't), we must estimate it from our sample using the sample standard deviation,  $s$ . However, when we do, the test statistic ( $z$ ) that we previously used changes. The new test statistic is called the t-statistic and has a new distribution associated with it.

The new t-distribution is not exactly like the standard normal curve, but it is very close:

- It is still centered at 0.
- It is bell shaped.
- Its spread is slightly greater than the normal distribution.
- It has more area in the tails.



The special thing about t-distributions is the fact that it is a family of distributions. There is a unique density curve for each sample size (and the dependence on sample size is taken care of by degrees of freedom:  $n - 1$ ).

The tails on a t-distribution are “fatter” than a standard normal distribution. This is true because the smaller our sample size, the more variation we have, hence the fatter tails. The larger our sample size gets, the closer the t-distribution will move towards the standard normal distribution.

Constructing a Confidence Interval:

- Define the Parameter
    - $\mu$  = true mean of population parameter in context
  - Check the Assumptions and Conditions
    - Random: The sample should be a random sample of the population or random assignment in an experiment
    - Independence (10% Condition): The sample size,  $n$ , must be no larger than 10% of the population.
    - Normality: There are multiple ways to verify this condition.
      - \* Stated in Problem: It may be stated in the problem that the sampling distribution is approximately Normal.
      - \* Central Limit Theorem: When  $n$  is large, the sampling distribution of the sample means is approximately normal.
      - \* Visual Representation: You are given a graphical representation (histogram or boxplot) that depicts a shape that is approximately normal. You may also be given data that you have to graph yourself (include a sketch). You are looking for no strong skew or outliers.
- Note: If you get a sample size less than 30, only use T-procedures if there are no outliers and no strong skew.
- Name the Inference Method: One Sample t-Interval
  - Calculate the Interval: point estimate  $\pm$  margin of error

$$\bar{x} \pm t_{n-1}^* \left( \frac{s}{\sqrt{n}} \right)$$

To calculate  $t_{n-1}^*$ :

2nd-VARS-4:invT()

- \* Enter in the desired area.
  - 90% Confidence Interval: 0.95
  - 95% Confidence Interval: 0.975
  - 99% Confidence Interval: 0.995
- \* Enter the degrees of freedom
  - Calculation:  $n - 1$  where  $n$  is the sample size

Write your conclusion in Context:

- We are \_\_\_\_\_% confident that the interval from \_\_\_\_\_to \_\_\_\_\_captures the true mean of population parameter in context.

**Example**

The Tribal Urban District Assessment (TUDA) is a government sponsored study of student achievement in large urban school districts. TUDA gives a reading test scored from 0 to 500. A score of 243 is “basic” reading level and a score of 281 is “proficient.” Scores for a random sample of 1,470 eighth graders in a district has a mean score of 240 with a standard deviation of 42.17.

(a) Construct and interpret a 99% confidence interval for the mean reading test score for all of this district's eighth graders.

$\mu$  = true mean reading test score of all the district eighth graders.

- Random: Random sample of 1470 district eighth graders.
- Independence:  $n = 1470 \leq 0.10$ (all of this district's eighth graders)
- Normal:  $n = 1470 \geq 30$ . CLT applies, so sampling dist. is approx. normal.

One Sample t-Interval

$\bar{x} \pm t^* \left( \frac{s}{\sqrt{n}} \right)$ , with  $t^* = 2.5792$  gives interval (237.1632, 242.8368).

We are 99% confident the interval from 237.1632 to 242.8368 points captures the true mean reading test score of all the eighth graders in this district.

Using a calculator we can do the following:

Calculator Steps	
[STAT] – Tests – 8: TInterval...	
Inpt: Data (Actual Data)	Stats (Summary Statistics)
List: L <sub>1</sub> (Enter Data)	$\bar{x}$ : sample mean
Freq: 1 (Don't Change)	$s_x$ : sample SD
C-Level: Confidence Level	$n$ : sample size
	C-Level: Confidence Level

(b) Based on your interval, is there convincing evidence that the mean reading test score for all the eighth graders in this district is less than the basic reading level? Justify your answer using your confidence interval.

Since our CI is entirely below 243 points there is convincing evidence that the mean reading test score is below basic reading level for all the eighth graders in this district.

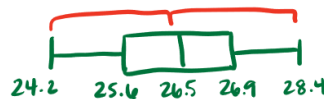
**Example**

A teacher's car records the fuel efficiency (mpg) and resets every time she fills up her gas tank. She randomly selected 20 samples of fuel efficiency from her car's computer.

26.8	24.7	26.6	27.2	28.4	27.2	27.0	26.4	24.6	26.8
26.2	26.0	24.2	25.8	25.9	26.8	26.6	27.3	25.4	24.9

Given  $\mu$  = true mean fuel efficiency for this teacher's car, verify the conditions for inference are met.

- Random: Randomly selected 20 fuel efficiency samples.
- Independence:  $n = 20 \leq 0.10$ (all car fill ups)
- Normal:



No strong skew or outliers so sampling distribution is approximately normal.

To graph a boxplot with a calculator:

- STAT-Edit-1:Edit. Enter Data into L1
- 2nd - Y= - Turn Plot1 On
- Change Type to Boxplot w/ Outliers
- ZOOM-9:ZoomStat
- TRACE - Identify the five number summary
- Sketch on your paper - looking for no strong skew or outliers

Determining Sample Size:

$$ME = t_{n-1}^* \left( \frac{s}{\sqrt{n}} \right)$$

**Example**

The health teachers at a high school want to estimate the body mass index (BMI) of students at their school. The BMI of US high school students follows a normal distribution with a standard deviation of 8.1%. How large of a sample is needed to estimate the mean BMI of this high school's students within 3% with 98% confidence?

Use invT to get  $t^* = 2.4258$ .

Plugging this into the above formula and solving for  $n$  gives  $n = 43$  people.

## 7.2 Constructing a One Sample t-Test

- Define the parameter:  $\mu$  = true mean of population parameter in context
- State the Hypotheses
  - Null Hypothesis:  $H_0 : \mu = \mu_0$
  - Alternative Hypothesis:  $H_A : \mu < \mu_0$ ,  $H_A : \mu > \mu_0$ ,  $H_A : \mu \neq \mu_0$
- Check the Assumptions and Conditions
  - Randomness: The sample should be a random sample of the population or random assignment in an experiment.

- 10% Condition: The sample size,  $n$ , must be no larger than 10% of the population.
- Approx. Normal Condition: There are multiple ways to verify this condition. (These ways are the same as the one sample t-interval)

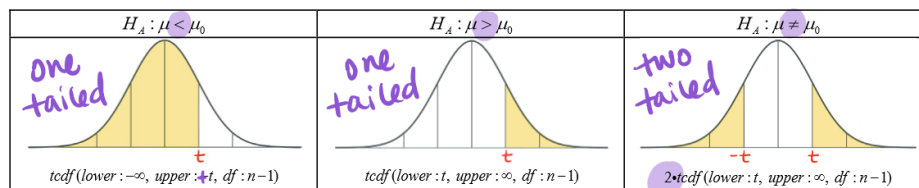
- Name the Inference Method: One Sample t-Test
- Calculate the Test Statistic

The test statistic is test statistic =  $\frac{\text{statistic-parameter}}{\text{standard deviation of statistic}}$

One Sample t-Test:  $t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$ .

Where  $\bar{x}$  is sample mean,  $\mu_0$  is null mean,  $s$  is sample standard deviation, and  $n$  is sample size.

- Obtain the p-Value



- Make a Decision
  - If the p-value is less than  $\alpha$ , we reject the null hypothesis.
  - If the p-value is greater than  $\alpha$ , we fail to reject the null hypothesis.

State the conclusion in context:

- Since our p-value of \_\_\_\_\_ is (less/greater) than \_\_\_\_\_, we (reject/fail to reject) the null hypothesis. There (is/is not) convincing evidence that alternative hypothesis.

### Example

At the Hawaii Pineapple Company, managers are interested in the sizes of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 32 ounces. A new irrigation system was installed in this field after the growing season. Managers wonder whether this change will affect the mean weight of future pineapples grown in the field. To find out, they select and weigh a random sample of 50 pineapples from this year's crop. Their sample has a mean of 31.935 ounces and a standard deviation of 2.394 ounces. Does this data give convincing evidence that the mean weight of pineapples produced in the field has changed this year?

$\mu$  = true mean weight (ounces) of pineapples produced in this field

$$H_0 : \mu = 32, H_A : \mu \neq 32.$$

- Random: Random sample of 50 pineapples from this field
- Independence:  $n = 50 \leq 0.10$ (all pineapples in this field)
- $n = 50 \geq 30$ , CLT applies to sampling dist. is approx. Normal

One Sample t-Test

$t = -0.1920$  (from the formulas above).  $P(t < -1.92) = 0.4243$ . Since this is two tailed,  $p = 0.8486$ .

Since the p-value of 0.8486 is greater than  $\alpha = 0.05$ , we fail to reject the null. There is not convincing evidence that the mean weight of pineapples produced in the field has changed this year.



**Example**

A study was conducted on whether time perception, an indication of a person's ability to concentrate, is impaired during caffeine withdrawal. Twenty randomly selected coffee drinkers abstained from caffeine for 24 hours. They were asked to estimate how much time had passed during a 45-second time period. The data is displayed below.

70	66	73	74	60	56	40	53	68	58
57	51	71	48	57	56	71	65	68	54

Is there convincing evidence at the  $\alpha = 0.05$  significance level that caffeine abstinence had a negative impact on time perception (causing elapsed time to be overestimated?)

$\mu$  = true average estimation of time elapsed (seconds)

$H_0 : \mu = 45$ ,  $H_A : \mu > 45$

- Random: 20 randomly selected coffee drinkers
- Independence:  $n = 20 \leq 0.10$ (all coffee drinkers)
- Normal:



No strong skew or outliers, sampling dist. is approx. normal.

One Sample t-Test

Calculator Steps	
[STAT] – Tests – 2: T-Test...	
Inpt: Data (Actual Data)	Stats (Summary Statistics)
$\mu_0$ : null mean	$\mu_0$ : null mean
List: L <sub>1</sub> (enter data)	$\bar{x}$ : sample mean
Freq: 1 (don't change)	$s_x$ : sample SD
$\mu$ : alternative hypothesis	$n$ : sample size
	$\mu$ : alternative hypothesis

From this we get  $t = 7.5888$  and  $p = 0.0000002$ .

Since the p-value of 0.0000002 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that caffeine abstinence had a negative impact on time perception.

### 7.3 Inference for Paired Data

Comparative studies are more convincing than single-sample investigations. For that reason, one-sample inference is less common than comparative inference. Study designs that involve making two observations on the same individual or one observation on each of two similar individuals, result in paired data.

When paired data result from measuring the same quantitative variable twice, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample t-procedures to perform inference about the mean difference:  $\mu_D$ . These methods are called matched pairs procedures.

**Example**

A researcher studied a random sample of identical twins who had been separated and adopted at birth. In each case, one twin (Twin A) was adopted by a high-income family and the other (Twin B) by a low-income family. Both twins were given an IQ test as adults. Here are their scores.

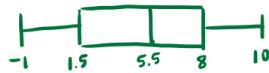
Pair	1	2	3	4	5	6	7	8	9	10	11	12
Twin A's IQ (High Income)	128	104	106	100	115	103	100	100	103	124	114	102
Twin B's IQ (Low Income)	120	99	99	94	111	97	99	94	104	113	113	100
Difference (High - Low)	8	5	9	6	5	8	1	6	-1	10	1	2

Construct and interpret a 95% confidence interval for the true mean difference in IQ scores among twins raised in high-income and low-income households.

$\mu_D$  = true mean difference in IQ scores between twins raised in high-income and low-income households.

- Random sample of 12 sets of identical twins
- $n = 12 \leq 0.10$  (all sets of identical twins)

Normal:



One Sample t-Interval for Matched Pairs.

You can calculate using  $\bar{x}_D \pm t_{n-1}^* \left( \frac{S_D}{\sqrt{n}} \right)$  or STAT-Tests-8:TInterval:

The interval is (2.7495, 7.2505).

We are 95% confident that the interval from 2.7495 to 7.2505 captures the true difference (High-Low) between the IQ scores of identical twins raised on high income and low income households.

**Example**

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from coffee, sodas, and other substances with caffeine during the duration of the experiment. During one two-day period, subjects took capsules containing their normal caffeine intake. During another two-days period, they took placebo capsules. The order in which the subjects took caffeine and the placebo is randomized. At the end of each two-day period, a test for depression was given to all 11 subjects. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression. The table displays data on the subjects' depression test scores. Higher scores show more symptoms of depression.

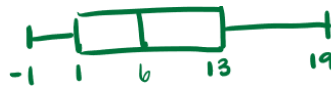
Subject	1	2	3	4	5	6	7	8	9	10	11
Depression Score (Placebo)	1	23	5	7	14	24	6	8	15	12	0
Depression Score (Caffeine)	5	5	4	3	8	5	0	0	2	14	1
Difference (Placebo - Caffeine)	11	18	1	4	6	19	6	8	13	1	-1

Does the data provide convincing evidence at the  $\alpha = 0.01$  significance level that caffeine withdrawal increases depression score, on average, for subjects like the ones in this experiment?

$\mu_D$ : true mean difference in depression test scores between normal caffeine intake and placebo capsules.

$H_0 : \mu_D = 0$ ,  $H_A : \mu_D > 0$

- Order of treatments is randomized
- Assume volunteers are independent of one another



No strong skew or outliers so sampling dist. is approx. normal.

One Sample t-Test for matched pairs.

You can calculate using  $t = \frac{\bar{x}_D - 0}{\left(\frac{s_D}{\sqrt{n}}\right)}$  and 2nd-VARS-6:tcdf or STAT-Tests-2:T-Test

The  $t$  value is 3.5304 and  $p = 0.0027$

Since the p-value of 0.0027 is less than  $\alpha = 0.01$ , we reject the null. There is convincing evidence that caffeine withdrawal increases depression score, on average, for subjects like those in this experiment.

## 7.4 Inference for Comparing Two Sample Means

Constructing a Two Sample T-Interval

- Define the Parameter:
  - $\mu_1$  = true mean of population parameter in context for Sample 1
  - $\mu_2$  = true mean of population parameter in context for Sample 2
- Check the Assumptions and Conditions: This is the same as for a one sample t-interval, just they have to apply for both populations.
- Name the Inference Method: Two Sample t-Interval for  $\mu_1 - \mu_2$
- Calculate the Interval

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n-1}^* \left( \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \right)$$

Calculating the  $t_{n-1}^*$  value is similar to earlier, the degrees of freedom is  $n - 1$  for the smaller sample size (also known as the conservative df)

- Write your conclusion in context.

We are \_\_\_\_\_% confident the interval from \_\_\_\_\_to \_\_\_\_\_units captures the true mean difference Pop 1-Pop 2 between Context of Question OR

We are \_\_\_\_\_% confident that the true mean of Pop 1 in Context is between \_\_\_\_\_and \_\_\_\_\_units (higher/lower) than Pop 2

### Example

College financial aid offices expect students to use summer earnings to help pay for college. But how large are these earnings? The University of Texas studied this question by asking random samples of 675 male and 621 female students with summer jobs how much they earned. Their data is summarized below.

Group	$n$	$\bar{x}$	$s_{\bar{x}}$
Males	675	\$1884.52	\$1368.37
Females	621	\$1360.39	\$1037.46

Construct and interpret a 90% confidence interval for the true mean difference between summer earnings of male or female students at the University of Texas.

$\mu_M$  = mean summer earnings of UT males,  $\mu_F$  = mean summer earnings of UT females.

- Random sample of 675 male and 621 female UT students.
- $n_m = 675 \leq 0.10$ (all male UT students),  $n_F = 621 \leq 0.10$ (all female UT students)
- $n_m = 675 > 30$ ,  $n_F = 621 \geq 30$ , CLT applies, sampling dist. is approx. normal

Two Sample t-Interval for  $\mu_M - \mu_F$

Using the formula above, we find that  $t^* = 1.6473$ , and the interval to be (413.54, 634.72).

We are 90% confident the interval from 413.62 to 634.64 dollars captures the true mean difference in summer earnings of male and female students at UT.

To use a calculator:

Calculator Steps	
[STAT] – Tests – 0: 2-SampTInt...	
Inpt: Data (Actual Data)	Stats (Summary Statistics)
List1: L1 (Sample 1 Data)	$\bar{x}1$ : Sample 1 Mean
List2: L2 (Sample 2 Data)	$s_{x1}$ : Sample 1 SD
Freq1: 1 (Don't Change)	$n1$ : Sample 1 Size
Freq2: 1 (Don't Change)	$\bar{x}2$ : Sample 2 Mean
C-Level: ConfidenceLevel	$s_{x2}$ : Sample 2 SD
Pooled: No (Don't Change)	$n2$ : Sample 2 Size
	C-Level: ConfidenceLevel
	Pooled: No (Don't Change)

Constructing a two sample t-Test

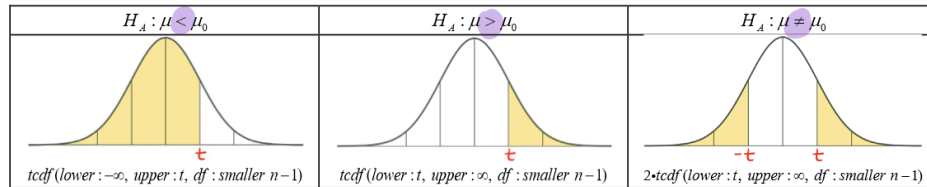
- Define the Parameter - Same as the two sample t-Interval
- State the Hypotheses:
  - Null Hypothesis:  $H_0 : \mu_1 = \mu_2$
  - Alternative Hypothesis:  $H_A : \mu_1 < \mu_2$ ,  $H_A : \mu_1 > \mu_2$ ,  $H_A : \mu_1 \neq \mu_2$
- Check the Assumptions and Conditions: Same as two sample t-Interval

- Calculate the Test Statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where  $\bar{x}$  is sample mean,  $s$  is sample standard deviation, and  $n$  is sample size.

- Obtain the P-Value



- Make a Decision: You know how to do this by now
- State your conclusion in Context

Since our p-value of \_\_\_\_\_ is (less/greater) than \_\_\_\_\_, we (reject/fail to reject) the null hypothesis. There (is/is not) convincing evidence that alternative hypothesis.

**Example**

Many people take ginkgo supplements advertised to improve memory. Are these over-the-counter supplements effective? In a study, elderly adults were random assigned to the treatment group or control group. The 104 participants who were assigned to the treatment group took 40 mg of ginkgo 3 times a day for 6 weeks. The 115 participants assigned to the control group took a placebo pill 3 times a day for 6 weeks. At the end of the 6 weeks, a memory test was administered. Higher scores indicate better memory function. Summary values are given in the following table.

Treatment	$n$	$\bar{x}$	$s_{\bar{x}}$
Ginkgo	104	5.7	0.6
Placebo	115	5.5	0.5

Based on these results, is there significant evidence that taking 40 mg of ginkgo 3 times a day is effective in increasing performance on a memory test?

$\mu_G$  = mean memory score for Ginkgo treatment group.

$\mu_P$  = mean memory score for placebo treatment group.

$H_0 : \mu_G = \mu_P$ ,  $H_A : \mu_G > \mu_P$

- Randomly assigned elders to treatment or control group
- Assume independence among elderly memory test scores
- $n_G = 104 \geq 30$ ,  $n_P = 115 \geq 30$ . CLT applies, sampling dist. is approx. normal.

Two Sample t-Test

Using the formula above to calculate  $t$ , we get  $t = 2.6642$ , and  $p = 0.0045$ .

Since the p-value of 0.0045 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that taking Ginkgo is effective in increasing memory score.

This is how to do it in your calculator:

Calculator Steps	
[STAT] – Tests – 4: 2-SampTTest...	
Inpt: Data (Actual Data)	Stats (Summary Statistics)
List1: L1 (Sample 1 Data)	$\bar{x}1$ : Sample 1 Mean
List2: L2 (Sample 2 Data)	$Sx1$ : Sample 1 SD
Freq1: 1 (Don't Change)	$n1$ : Sample 1 Size
Freq2: 1 (Don't Change)	$\bar{x}2$ : Sample 2 Mean
$\mu1$ : Alternative Hypothesis	$Sx2$ : Sample 2 SD
Pooled: No (Don't Change)	$n2$ : Sample 2 Size
	$\mu1$ : Alternative Hypothesis
	Pooled: No (Don't Change)

# 8 Inference for Categorical Data: Chi-Square

## 8.1 Chi Square Test for Goodness of Fit

### Expected Counts

- A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits to some claimed distribution.
- An example of an equal expected count might be the following:
- A fair, six-sided die is rolled 60 times. What would be the expected count of each outcome?

Outcome	1	2	3	4	5	6
Expected Count	10	10	10	10	10	10

We found the expected count by taking the total number of trials and dividing it equally among each of the outcomes.

- An example of an unequal expected count might be the following:
- An unfair, six-sided die is rolled 60 times. The die is loaded so that the number 1 turns up 50% of the time and the other five outcomes occur 10% of the time. What would be the expected count of each outcome?

Outcome	1	2	3	4	5	6
Probability	.50	.10	.10	.10	.10	.10
Expected Count	30	$60(.10) = 6$	6	6	6	6

We found the expected count by taking the total number of trials and multiplying it by the probability of the outcome.

### Observed Counts:

While we know what is expected, when we run a simulation of rolling a fair die 60 times, we do not expect each outcome to be observed exactly 10 times each due to sampling variability.

Let's say a die was given to you claimed as fair, and you observed the following counts when rolling the die 60 times:

Outcome	1	2	3	4	5	6
Expected Count	10	10	10	10	10	10
Observed Count	16	7	15	10	4	8

- Do you suspect the die given to you was a fair die?
- This is going to be the question that our new test helps us answer: Are the differences between the actual observed counts and the expected counts significant?

Note: The observed counts must be all whole numbers because they represent actual counts, but the expected counts do not need to be whole numbers.

To measure the difference between the observed and expected counts, and to determine if the difference is

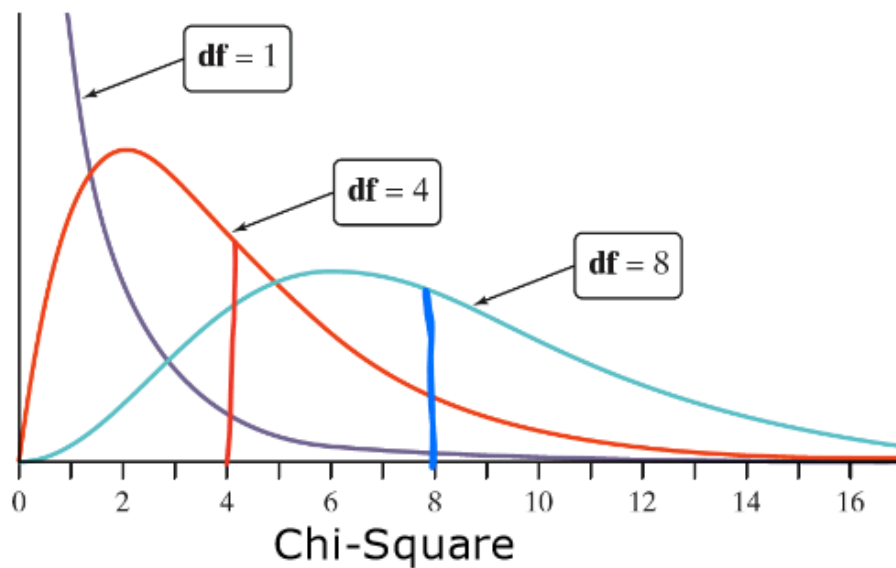
significant, we will introduce a new test statistic, called the chi-square statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

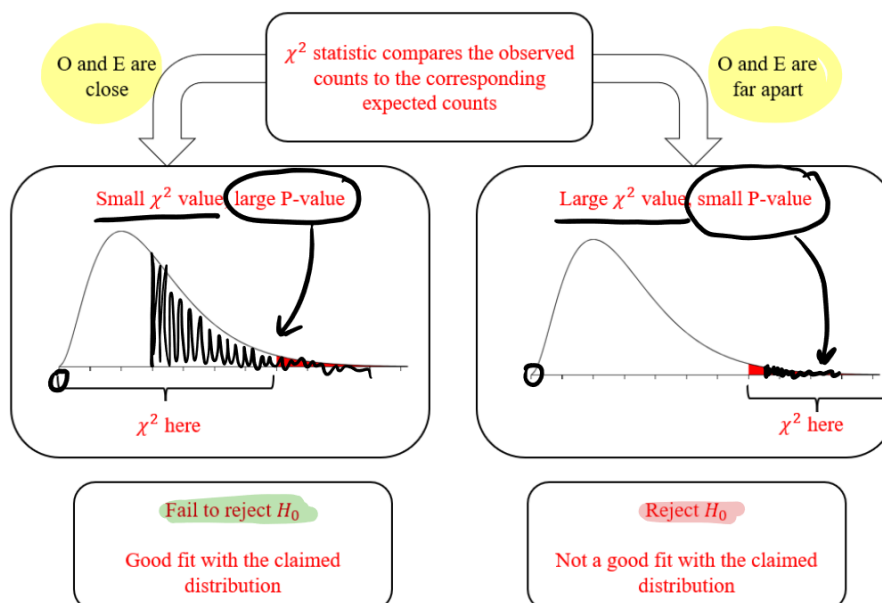
Where O represents each observed count in the distribution and E represents each corresponding expected count.

- The sampling distribution of the chi-square statistic is not a normal distribution
- It is a right-skewed distribution that allows only for positive values because the statistic cannot be negative.

When the expected counts are all at least 5, the sampling distribution of the  $\chi^2$  statistic is close to a chi-square distribution with degrees of freedom (df) equal to the number of categories minus 1.



- The chi-square distributions are a family of distributions that take only positive values and are skewed to the right.
- A particular chi-square distribution is specified by giving its degrees of freedom.
- The chi-square goodness-of-fit test uses the chi-square distribution with  $df = \# \text{categories} - 1$





## Hypotheses

- The null hypothesis in a chi-square goodness-of-fit test should state a claim about the distribution of a single categorical variable in the population of interest.
- We can write this in words or symbols; both are acceptable and used on the AP exam.
- Using our fair die as an example, we would say:

Using symbols,  $H_0 : p_1, p_2, p_3, p_4, p_5, p_6 = \frac{1}{6}$ , where  $p$  is the proportion of outcomes of each face of die.

Using words,  $H_0$ : The proportion of dice outcomes is equally distributed.

The alternative hypothesis in a chi-square goodness-of-fit test is the categorical variable does not have the specified distribution, and is easily given in words:

$H_A$ : At least one of the claimed proportions is incorrect.

## Conditions:

- Random: The data came from a well-designed random sample or randomized experiment
- Independent: When sampling without replacement, the 10% condition is met.
- Large Counts
  - All expected counts are at least 5
  - This allows us to say that the sampling distribution will follow a Chi-Square distribution

When the conditions are met, the chi-square goodness of fit test can be performed with the hypotheses:

$H_0$ : The claimed distribution is correct.  $H_A$ : At least one proportion in the claimed distribution is incorrect.

We find the expected counts assuming the claimed distribution is true, and then we calculate the chi-square statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

The p-value is the area to the right of  $\chi^2$  under the density curve of the chi-square distribution with  $k-1$  degrees of freedom, where  $k$  represents the number of categories.

$$P\text{-value} = P(\chi^2 > \text{value}) = \chi^2\text{cdf}(\text{value}, 1e99, \text{df}) \text{ on the TI-84}$$

## Beware

1. The chi-square test statistic compares observed and expected counts. Don't try to perform calculations with the observed and expected proportions in each category.
2. When checking the Large Counts condition, be sure to examine the expected counts, not the observed.

**Example**

A geneticist is studying the gene pattern of eye color in a group of white mice. He observed a random sample of mice from the lab and found that 110 had red eyes, 57 had brown eyes, 32 had pink eyes, and 13 had blue eyes. His model suggest that this distribution of eye color should occur in a 9 : 3 : 3 : 1 ratio. Is there evidence that the geneticist's model is not accurate? Use a 5% level of significance.

$H_0$ : The proportion of white mice eye color occurs in a 9 : 3 : 3 : 1 ratio.

$H_A$ : At least one of the proportion is incorrect.

- Random: Random sample of 212 white mice
- Independent:  $n = 212 \leq 0.10(\text{all white mice})$
- Large Counts:

	Red	Brown	Pink	Blue
OBS	110	57	32	13
EXP	119.25	39.75	39.75	13.25

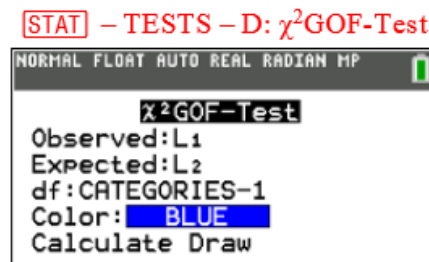
All expected counts  $\geq 5$ .

Chi Square Test for Goodness of Fit

$$\chi^2 = \frac{(110-119.25)^2}{119.25} + \frac{(57-39.75)^2}{39.75} + \frac{(32-39.75)^2}{39.75} + \frac{(13-13.25)^2}{13.25} = 9.7191.$$

Doing  $\chi^2$ cdf, with (lower: 9.7191, upper:  $\infty$ , df: 3) gives 0.0211 for the p value.

We can also use a calculator:



Since the p-value of 0.0211 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that the color of white mice eyes does not match a 9 : 3 : 3 : 1 ratio.

## 8.2 Chi Square Test for Homogeneity

- Recall: The Chi Square Test for Goodness of Fit is testing if one population “fits” a given claim.
- The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different

Constructing a Chi-Square Test for Homogeneity:

- State the Hypotheses:
  - $H_0$ : There is no difference in the distribution of categorical variable among two or more groups.
  - $H_A$ : There is a difference in the distribution of categorical variable among two or more groups.
- Check the Assumptions and Conditions
  - Randomness: The individuals whose counts are available for analysis should be a random sample of the population.
  - 10% Condition: The sample size,  $n$ , must be no larger than 10% of the population.

- Large Counts: We should expect to see at least 5 counts in each category of the categorical variable. List the counts and state “all exp counts  $\geq 5$ ”

Note: When performing an experiment, Randomness is satisfied by randomly assigning treatments to subjects and the 10% condition is satisfied if we can assume independence among the individuals in the study.

- Name the Inference Method: Chi-Square Test for Homogeneity
- Calculate the Test Statistic

$$\chi^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}$$

Observed Counts - Actual frequencies of the variable from your sample

Expected Counts - Projected frequencies of the variable if the null hypothesis is true

$$\text{EXP} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

- Obtain the P-value 2nd-Vars-8: $\chi^2\text{cdf}()$ 
  - Lower:  $\chi^2$
  - Upper:  $\infty$
  - df:  $(\# \text{Rows}-1)(\# \text{Columns}-1)$

Calculator Steps:

1. Enter Observed Values into Matrix [A]
  - 2nd- $x^{-1}$ (Matrix)-EDIT
  - #Rows x # Columns
2. Enter Expected Values into Matrix [B]
  - Will populate when you run the  $\chi^2$  Test
3. STAT-Tests-C: $\chi^2\text{Test}$ 
  - Observed: [A]
  - Expected: [B]

- Make a decision: This is the same as always
- State your conclusion in context: This is also the same

**Example**

Andrea is addicted to TikTok and she believes that most students in her high school are too. She wants to investigate if there is a difference in TikTok usage among Juniors and Seniors. She randomly samples 65 Juniors and 85 Seniors. The data she collected is given in the table below. Is there convincing evidence of a difference in the distribution of TikTok among the Juniors and Seniors in her school?

*OBS =*

↓ TikTok Use ↓	Juniors	Seniors	Total
Once a Month or Less	4	10	14
Once a Week	15	21	36
At Least Once a Day	46	54	100
Total	65	85	150

$H_0$ : TikTok usage among juniors and seniors is the same

$H_A$ : TikTok usage among juniors and seniors is different.

- Random: Randomly sampled 65 juniors and 85 seniors.
- Independent:  $65 \leq 0.10(\text{all juniors in her HS})$ ,  $85 \leq 0.10(\text{all seniors in her HS})$
- Large Counts:  $\begin{bmatrix} 6.07 & 7.93 \\ 15.6 & 20.4 \\ 43.33 & 56.67 \end{bmatrix}$  All  $\text{exp} \geq 5$
- $\chi^2$  Test for Homogeneity
- $\chi^2 = 1.5727$ ,  $p = 0.4555$ ,  $\text{df} = 2$ :  $(\# \text{col} - 1)(\# \text{rows} - 1) = 2$ .
- Since the p-value of 0.4555 is greater than  $\alpha = 0.05$ , we fail to reject the null. There is not convincing evidence that TikTok usage among juniors and seniors is different at Andrea's high school.

**Example**

Aspirin prevents blood from clotting which helps prevent strokes. A medical study (we will assume this is a well-designed experiment) asked whether adding another anti-clotting drug named Dipyridamole would be more effective for patients who already had a stroke. Here are the data on strokes during the two years of the study.

Group	Treatment	Number of Patients	Number of Patients w/Stroke
1	Placebo	1649	250
2	Aspirin	1649	206
3	Dipyridamole	1654	211
4	Both	1650	157

(a) Summarize the data into a two-way table.

	Placebo	Aspirin	Dipyridamole	Both	Total
Stroke	250	206	211	157	824
No Stroke	1399	1443	1443	1493	5778
Total	1649	1649	1654	1650	6602

(b) Is there convincing evidence of a difference in the effectiveness of the four treatments at the  $\alpha = 0.05$  significance level?

- $H_0$ : No difference in effectiveness among treatments
- $H_A$ : Difference in effectiveness among treatments.
- Random: Treatments randomly assigned - medical study
- Independent: Assume patient results are independent
- Large Counts:  $EXP = \begin{bmatrix} 205.81 & 205.81 & 206.44 & 205.94 \\ 1433.19 & 1433.19 & 1447.56 & 1444.06 \end{bmatrix}$  All  $EXP \geq 5$
- $\chi^2$  Test for Homogeneity
- $\chi^2 = 24.2428$ ,  $p = 0.00002$ ,  $df = 3$
- Since the p-value of 0.00002 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence of a difference in effectiveness among the four treatments for preventing strokes.

### 8.3 Chi Square Test for Independence

- Recall: The Chi Square Test for Goodness of Fit is testing if one population "fits" a given claim.
- Recall: The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different
- The Chi Square Test for Independence compares the distribution of two categorical variables across one population to see if they are independent (not associated)

Constructing a Chi-Square test for Independence:

- State the Hypotheses:
  - $H_0$ : Categorical Variable 1 and Categorical Variable 2 are independent (not associated) for population
  - $H_A$ : Categorical Variable 1 and Categorical Variable 2 are not independent (associated) for population
- Check Assumptions and Conditions: Same as the test for Homogeneity

- Name the Inference Method: Chi-Square Test for Independence
- Calculate the Test Statistic: Same as Homogeneity
- Obtain the P-Value: This is also the same as Homogeneity
- Make Decision: Same as always
- State your conclusion in context: This remains the same

Don't Forget: Association does not imply causation

- A small p-value is not proof of causation.
- The Chi Square Test for Independence treats the two variables symmetrically, we cannot differentiate the direction of any possible causation even if it existed.
- There is no way to eliminate the possibility that a lurking variable is responsible for the lack of independence.
- Don't say that one variable "depends" on the other just because they are not independent.

### Example

Andrew thinks there might be a relationship between angry students and GPA. He asks, "Do students who are prone to sudden bursts of anger have lower GPAs?" He took an SRS of 300 students at his high school at the beginning of the year. He had each student take the Spielberger Trait Anger Scale Test which measures how prone a person is to sudden anger. At the end of the school year, Andrew collected the data on student GPAs. Here are the results:

Anger Scale Test Results				
		Low Anger	Moderate Anger	High Anger
G	Low GPA (0.1 – 1.9)	4	14	37
P	Mid GPA (2.0 – 2.9)	122	33	3
A	High GPA (3.0 – 4.0)	79	7	1

= OBS

Does the data provide convincing evidence of an association between anger level and GPAs at Andrew's HS?

$H_0$ : Anger level and GPA are independent for students at Andrew's HS.

$H_A$ : Anger level and GPA are not independent for students at Andrew's HS.

- Random: SRS Of 300 students at Andrew's HS
- Independent:  $n = 300 \leq 0.10$ (all students at Andrew's HS)
- Large Counts:  $\begin{bmatrix} 37.58 & 9.90 & 7.52 \\ 107.97 & 28.44 & 21.59 \\ 59.45 & 15.66 & 11.89 \end{bmatrix}$  All exp counts  $\geq 5$

Chi Square Test for Independence

$$\chi^2 = 187.1097, p \approx 0 \text{ df} = 4$$

Since the p-value is approx. 0 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence that anger level and GPA are not independent for students at Andrew's HS.

**Example**

Is your index finger longer than your ring finger? Or is it the other way around? It isn't the same for everyone! To investigate if there is a relationship between gender and relative finger length, we selected a random sample of 460 U.S. high school students who completed a survey. The results are shown in the table below:

		Gender	
		Male	Female
Relative Finger Length	Index Longer	85	73
	Same Length	42	44
	Ring Longer	100	116

] = OBS

$H_0$ : Gender and finger length are not associated for US HS Students

$H_A$ : Gender and finger length are associated for US HS Students

- Random Sample of 460 US HS Students
- Independent:  $n = 460 \leq 0.10(\text{all US HS Students})$
- Large Counts:  $\text{EXP} = \begin{bmatrix} 77.97 & 80.03 \\ 42.44 & 43.56 \\ 106.59 & 109.41 \end{bmatrix}$  All exp counts  $\geq 5$
- Chi Square Test for Independence
- $\chi^2 = 2.0652$ ,  $p = 0.3561$ ,  $df = 2$
- Since the p-value of 0.3561 is greater than  $\alpha = 0.05$ , we fail to reject the null. There is not convincing evidence that gender and finger length are associated for US HS students.

# 9 Inference for Quantitative Data: Slopes

## 9.1 Sampling Distributions and Confidence Intervals for Slope

### Population Regression Line

- An “ideal” linear relationship can be described with a population regression line:  $\mu_y = \alpha + \beta x$ 
  - Where  $\mu_y$  represents the mean value of the response variable  $y$  for any given value of the explanatory variable  $x$
  - $\alpha$  represents the population  $y$ -intercept and  $\beta$  represents the population slope
- An observed linear relationship can be described with a sample regression line:  $\hat{y} = a + bx$
- If we took many LSRLs of the same size from the sample population, we can create a sampling distribution for our slope.

### Sampling Distribution for the slope of a LSRL

- For a bivariate population with a given slope,  $\beta$ , a standard deviation of residuals  $\sigma$ , and a standard deviation of  $x$ -values  $\sigma_x$ .
- If you take all samples of size  $n$  and compute the slope of each of those samples, you get the sampling distribution:
  - Shape: The distribution of sample slopes is approximately normal.
  - Mean:  $\mu_b = \beta$
  - Standard Deviation:  $\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$ , where  $\sigma$  is the standard deviation of residuals,  $\sigma_x$  is the standard deviation for the explanatory variable, and  $n$  is sample size.

Once we develop a sampling distribution for our slope, we can begin to ask and answer our inference questions:

- Is there a linear relationship between  $x$  and  $y$  in the population, or could the pattern we see happen just by chance?
- In the population, how much will the predicted value of  $y$  change for each increase of 1 unit in  $x$ ?

### Conditions for Regression Inference

- Linear:  $x$  and  $y$  have a linear relationship. Check: make sure scatter plot can be described by a line
- Independent: If sampling without replacement, check the 10% condition.
- Normal Residuals: When  $x$  is fixed,  $y$  follows a normal distribution. Check: Make a histogram of the residuals and make sure it looks approximately normal. If the graph has outliers or strong skewness,  $n$  should be larger than 30.
- Equal SD: Standard deviation of residuals doesn't vary with  $x$ . Check: Make a residual plot and check for a random pattern
- Random: Random sampling (SRS) or random assignment (experiment)

Together, this makes up the acronym LINER, which can help you remember what conditions to check when creating a confidence interval or running a significance test.

A C% Confidence interval is created to estimate the slope  $\beta$  of the population (true) regression line.

$$b \pm t^* \left( \frac{s}{s_x \sqrt{n-1}} \right) \text{ with df} = n - 2$$

- $t^*$  has C% of the area between  $-t^*$  and  $t^*$



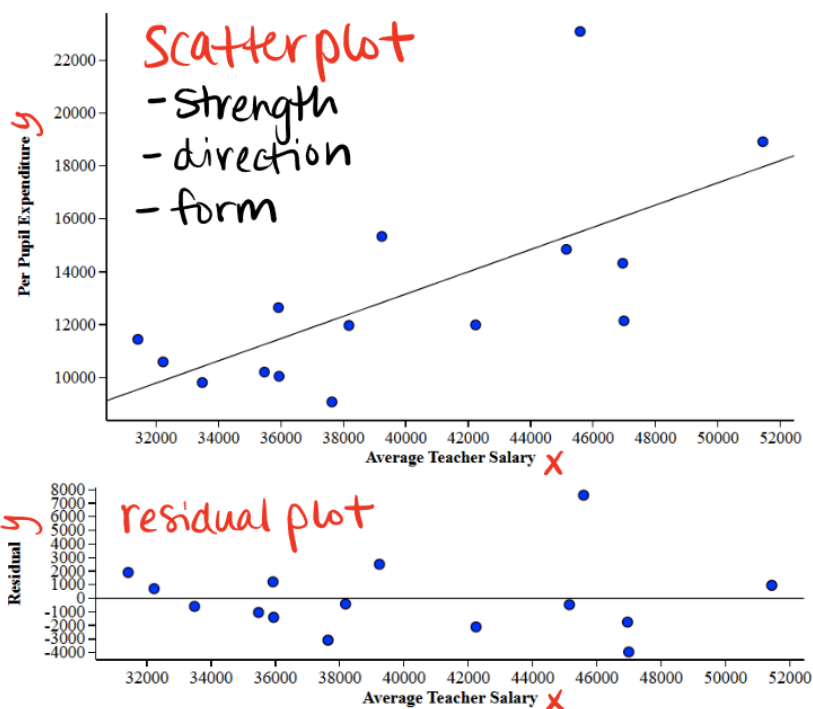
- $b$  is the point estimate (slope from our sample data)
- $SE_b = \frac{s}{s_x \sqrt{n-1}}$  is the standard error of the slope
- $s_x$  is the standard deviation of  $x$ -values
- $s$  is the standard deviation of the residuals

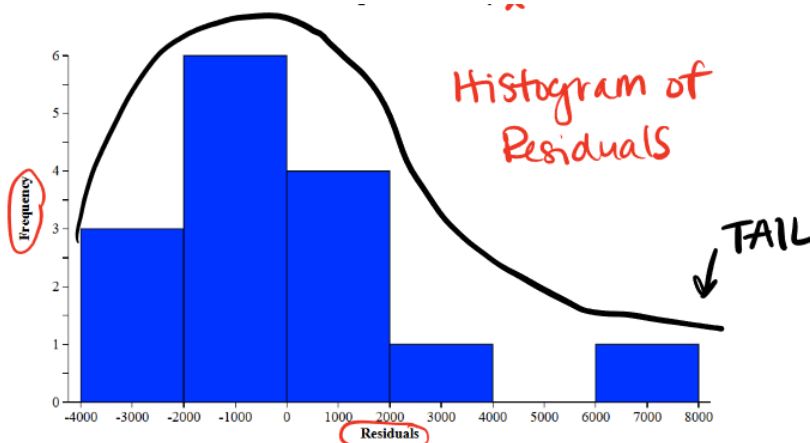
Interpretation: We are C% confident that the interval from \_\_\_\_\_ & \_\_\_\_\_ captures the true slope of the regression line between  $x$ -variable and  $y$ -variable.

### Example

The data below was obtained from 15 randomly selected large school districts from around the nation. It shows what their average teacher salary is and how much they spend per student (per pupil expenditure).

Average Teacher Salary	Per Pupil Expenditure
\$31,418	\$11,443
\$32,226	\$10,589
\$33,483	\$9,809
\$35,474	\$10,205
\$35,923	\$12,645
\$35,943	\$10,045
\$37,636	\$9,075
\$38,181	\$11,968
\$39,236	\$15,337
\$42,240	\$11,989
\$45,147	\$14,848
\$45,589	\$23,091
\$46,954	\$14,322
\$46,992	\$12,143
\$51,443	\$18,920





Data:

- $n = 15$
- $r = 0.684$
- $r^2 = 0.468$
- $s = 2865.8$
- $s_x = 6154.2$
- $s_y = 3786.1$
- $\hat{y} = -3679.1 + 0.4208x$

Construct and interpret a 95% confidence interval for the slope of the population regression line.

$\beta$  = true population slope between average teacher salary and per pupil expenditure.

- Scatterplot shows a moderately positive linear relationship.
- $n = 15 \leq 0.10$  (all school districts in the nation)
- A histogram of the residuals appears skewed right. (This is the only condition not correctly met.)
- The residual plot shows random scatter around LSRL.
- Randomly selected 15 large school districts.

Linear Regression t-Interval for Slope

Using the formula given, we can use invT to get  $t^* = 2.1604$  and the formula to get the confidence interval (0.1519, 0.6897).

Calculator Steps:

**STAT** – TESTS – G: LinRegTInt

```

NORMAL FLOAT AUTO REAL RADIAN MP
LinRegTInt
Xlist:L1
Ylist:L2
Freq:1
C-Level:0.95
RegEQ:
Calculate
  
```

We are 95% confident that the interval from 0.1519 to 0.6897 captures the true slope of the regression line between average teacher salary and per pupil expenditure. However, because our “Normal Residuals” condition was not met, we should be careful with this interpretation because it might not be correct.

Most AP Problems will not require you to do what we did in the previous example. Most inference questions come with a computer output, like what is pictured below.

Predictor	Coef	SE Coef	T	P
Constant	2.544	0.134	18.955	0.000
Caffeine (mg)	0.164	0.057	2.862	0.005

$S = 1.532$     $R\text{-Sq} = 60.032\%$     $R\text{-Sq(adj)} = 58.022\%$

*Handwritten notes:*  
 y-int slope (points to Constant)  
 NEVER use these numbers!  
 $H_0: \beta = 0$  &  $H_A: \beta \neq 0$   
 $t = \frac{b - \beta_0}{SE_b} = \frac{b}{SE_b}$   
 p-value (points to 0.005)  
 NEVER use!  
 $r^2$  (points to  $R\text{-Sq}$ )  
 Coef. of determination  
 $r = \text{Correlation} = \pm 0.6032$   
 standard deviation of residuals (points to  $S$ )

### Example

A study attempted to establish a linear relationship between IQ score and musical aptitude. The following table is a partial printout of the regression analysis based on a sample of 20 individuals.

The regression equation is  
MusApp = -22.3 + 0.493 IQ

Predictor	Coef	SE Coef	T	P
Constant	-22.26	12.94	-1.72	0.102
IQ	0.4925	0.1215	4.05	0.000

$S = 6.143$     $R\text{-Sq} = 47.7\%$     $R\text{-Sq(adj)} = 44.8\%$

*Handwritten notes:*  
 $b$  (points to 0.4925)  
 $SE_b$  (points to 0.1215)

Construct and interpret a 99% confidence interval for the slope of the regression line. Does it suggest a linear relationship? Assume all assumptions and conditions for inference have been met.

$\beta$  = true population slope between IQ score and musical aptitude

All assumptions and conditions met

Linear Regression t-Interval for Slope

$$t^* = \text{invT}(\text{area} = .995, \text{df} = 18) = 2.8784$$

$$0.4925 \pm 2.8784(0.1215) = (0.1428, 0.8422)$$

We are 99% confident that the interval from 0.1428 to 0.8422 captures the true slope of the regression line between IQ score and musical aptitude.

## 9.2 Significance Test for a Slope

- The significance test for a slope is called a Linear Regression t-Test for Slope.
- It can help us answer three different questions with the hypotheses

Is the relationship between the explanatory and response variable negative?

- $H_0: \beta = 0$
- $H_A: \beta < 0$

Is there a relationship between the explanatory and response variable?

- $H_0: \beta = 0$
- $H_A: \beta \neq 0$

Is the relationship between the explanatory and response variable positive?

- $H_0 : \beta = 0$
- $H_A : \beta > 0$

Conditions for Regression Inference: Same for the confidence interval

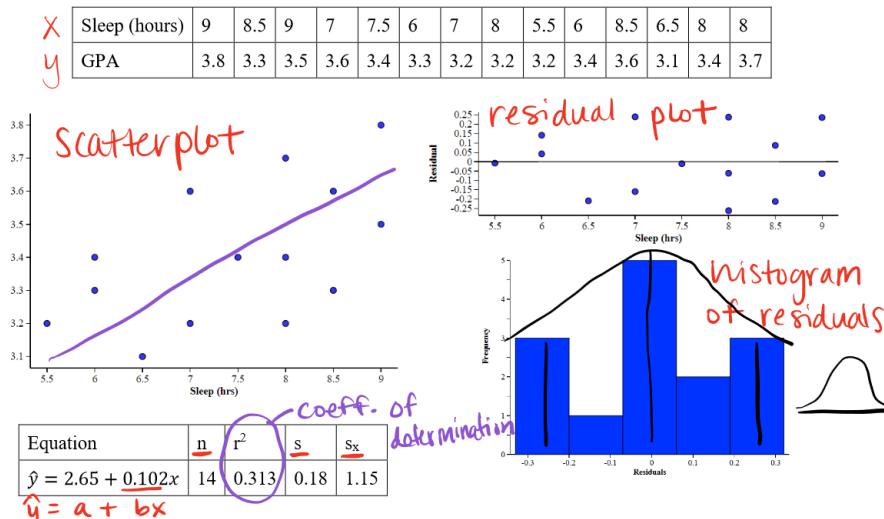
The test statistics is  $t = \frac{\text{statistic-parameter}}{\text{standard error}}$ , or  $\frac{b}{SE_b}$ , where  $SE_b = \frac{s}{s_x \sqrt{n-1}}$ .

df = n-2, p- value is calculated using your calculator, in the direction of the alternative hypothesis. p-value = tcdf(lower, upper, df)

- In your conclusion, you would state the results of your significance test (reject or fail to reject) and then interpret the findings in context
- Note: Having a low p-value and finding evidence of the alternative hypothesis of some linear association does not mean that the association is strong

**Example**

A school counselor is concerned that the number of hours of sleep his students get each night is affecting their GPA in a negative way. He selects a random sample of 14 seniors in his district and asks them how many hours of sleep they get on a typical school night. He then uses school records to determine the most recent grade-point average (GPA) for each student. His data are given below.



Do these data provide convincing evidence, at the 0.05 significance level, of a positive linear relationship between the hours of sleep students typically get and their academic performance?

$\beta$  = true pop. slope between hours of sleep students typically get and their GPA.

$H_0: \beta = 0$ ,  $H_A: \beta > 0$

- Scatterplot shows a weak positive linear relationship.
- $n = 14 \leq 10$  (all HS Seniors)
- Histogram of residuals doesn't appear normal but no strong skew or outliers
- Residual plot shows random scatter
- Random sample of 14 HS seniors

Linear Regression t-Test for Slope

$$t = \frac{0.102}{\frac{0.18}{1.15\sqrt{13}}} = 2.3496, \text{ tcdf with this gives } p = 0.0184.$$

You can also run this on the calculator

**[STAT] – TESTS – F: LinRegTTest**



Since the p-value of 0.0184 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence of a positive linear relationship between hours of sleep per night and GPA for HS seniors.

**Example**

The computer output given shows a regression analysis of an honors social science course (score in points) versus a reading comprehension score (in points) for 25 sophomores at your school.

Social Science Score = 76.56 + 0.731(Reading)					
Predictor	Coef	SE Coef	T	P	
Constant	76.56	10.168	7.53	<.0001	
Reading	0.731	0.0351	20.84	<.0001	
s = 25.83      R-Sq = 0.610      R-Sq(Adj) = 0.610					

Carry out a hypothesis test for these data to determine if there is a linear relationship and interpret your results in the context of the problem. Assume all assumptions and conditions for inference have been met.

$\beta$  = true pop. slope between social science score and reading score.

$H_0 : \beta = 0$ ,  $H_A : \beta \neq 0$

Linear Regression t-Test for Slope

$t = \frac{0.731}{0.0351} = 20.8262$ , tcdf with this gives  $p \approx 0 \times 2 \approx 0$ .

Since the p-value of approx. 0 is less than  $\alpha = 0.05$ , we reject the null. There is convincing evidence of a linear relationship between social science score and reading score for your school's sophomores.