STAT 243Z Statistics I
Course Packet 24548

Lead Instructor: Cara Lee

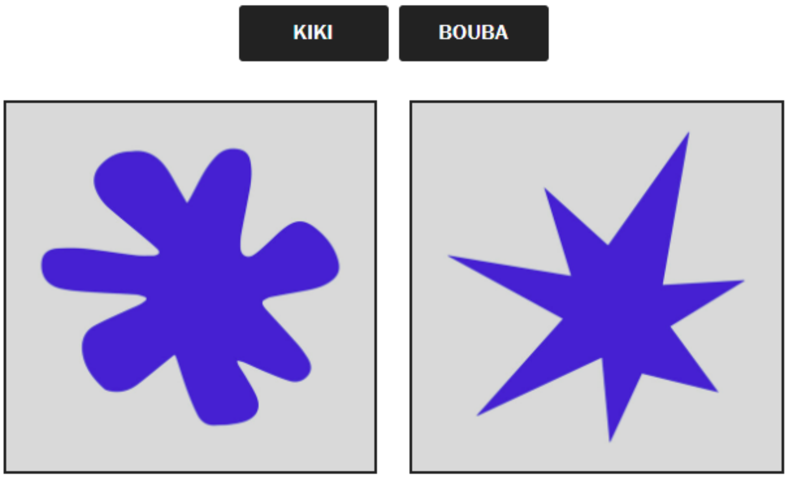
Contents

Blank page

Intro to Statistics – Statistical Significance

A Statistical Test

Which of these shapes is a KIKI and which is a BOUBA? Please take the survey right above this video if you haven't already. If you've seen this test before, try to answer like you did the first time.



Example 1. For this study,

a. What **research question** is being explored?

b. Who/what are the **subjects, cases, or observational units**?

c. What is the **variable**?

d. Is the variable **categorical or numerical**?

e. What additional factors should be considered in this study?

Let's suppose there is no association between the words and the shapes. Then making the choice would be like flipping a coin. To test our results against this assumption, we'll flip a coin 25 times. You can grab a coin and flip it 25 times or use a google coin flipper or random.org/coins.

Your coin flip results:

Number of heads:

Total flips:

Proportion of heads:

One class result:

Number who said bouba is on the left: 18

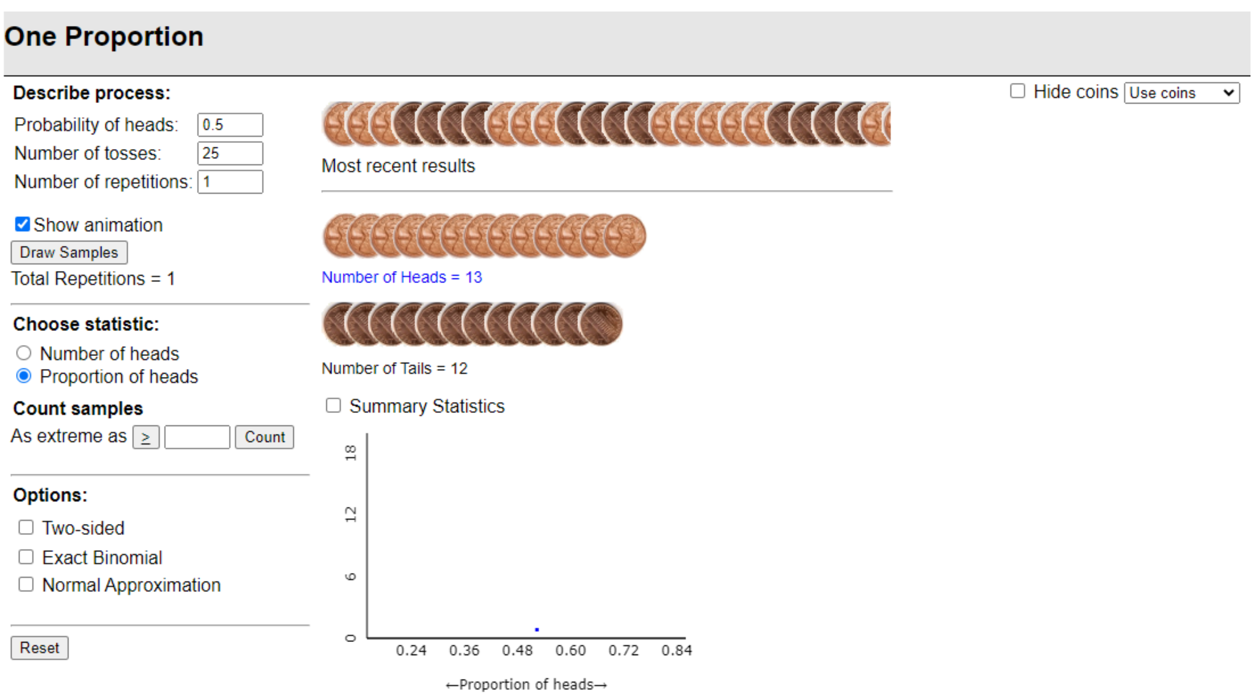
Total students: 25

Sample proportion:

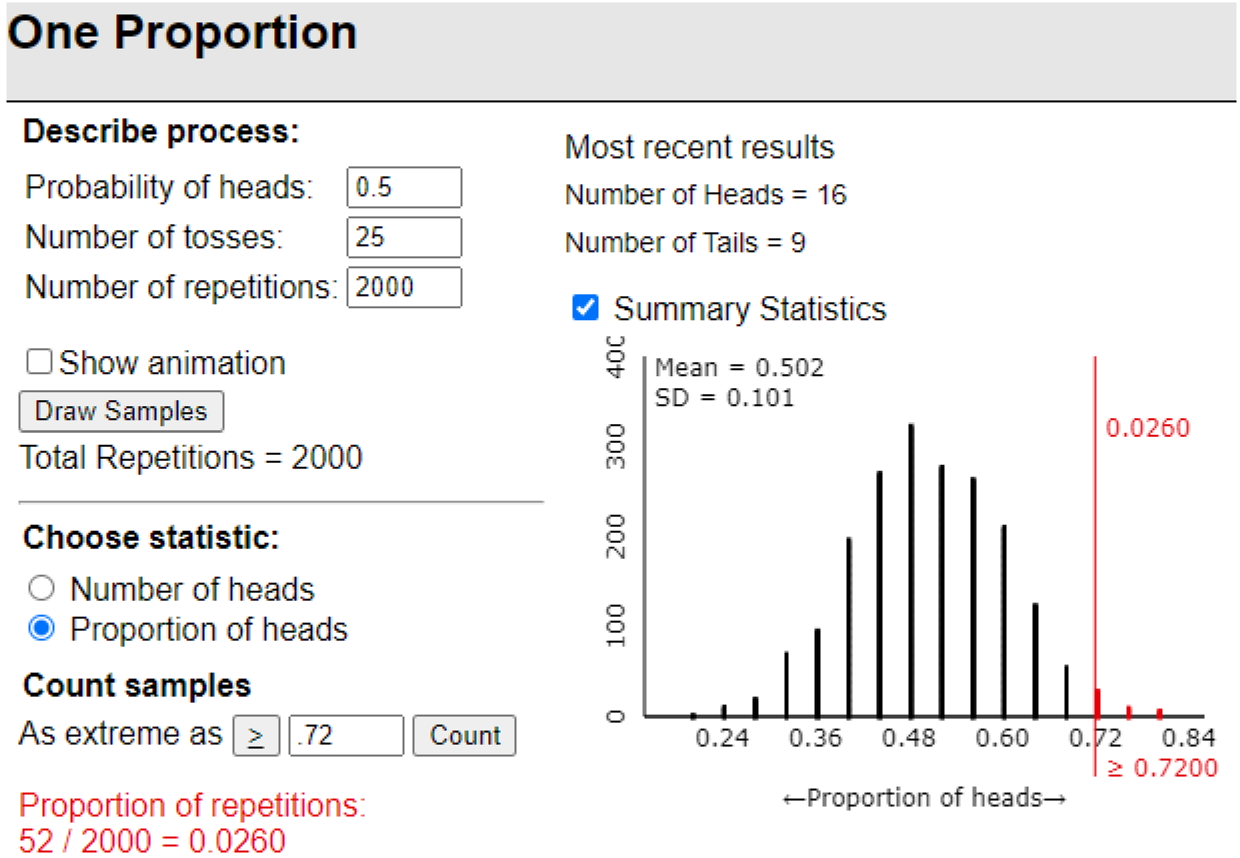
Does your coin flip result seem very different from the class results? We want to know whether the class result could have happened by chance or whether there is a significant association between the words and shapes.

Simulation

Let's flip 25 coins many times and see what the range of reasonable values are for a 50-50 choice. This would get very tedious so we're going to use technology to do a **simulation**. We will use this applet many times throughout the class. Please click on this applet: Rossman/Chance One Proportion



Proportion of heads in samples of 25-coin tosses



Proportion of heads in samples of 25-coin tosses

What can we conclude from this simulation? Is our result statistically significant?

Further information about the association between the words and shapes.

[Video: The Bouba-Kiki Effect]

[Research Article: The bouba/kiki effect is robust across cultures and writing systems]

Intro to the Statistical Process and Context

> *Data is never a raw, truthful input – and it is never neutral.*
>
> *-Dr. Catherine D'Ignazio, co-author of Data Feminism*

What is/are statistics?

The field of statistics is the science of collecting, summarizing, and drawing conclusions from data.

A statistic is a number calculated from data with units and context.

What are data?

Data are plural, datum is singular.

Any collection of numbers, characters (words), images, or other items that provide information - along with the lens through which they were gathered.

Context and Data Justice

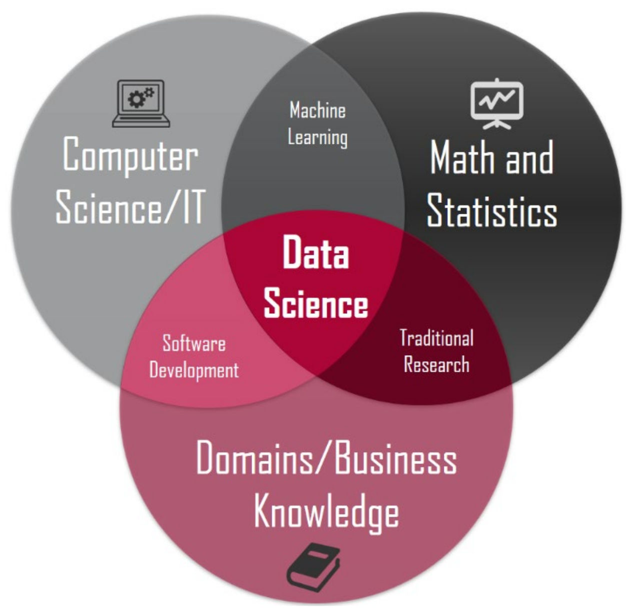
Traditionally data has been treated as neutral and objective, but they reflect and reinforce the lens or system through which they were gathered.

More resources in D2L

Statistical Process

- Examine impacts and implicit bias. Who is harmed, who is served?
- Center nondominant voices and impacted communities. "Nothing about us, without us," *-South African disability rights movement*

1. Identify a question about a target population
2. Design a study and collect data from a sample
3. Analyze data and draw inferences
4. Form a conclusion, review, and repeat



Descriptive Statistics

Summarizing a set of data.

Inferential Statistics

Generalizing from a sample to a population or determining statistical significance.

What is data science?

The field of data science is a multidisciplinary field using statistics, computer science and domain information to visualize and make meaning from data[1].

---

[1] Barber, Michael. "Data Science Venn Diagram." *Data Science Concepts You Need to Know! Part 1*, Medium.com, 14 Jan. 2018, https://towardsdatascience.com/introduction-to-statistics-e9d72d818745. Accessed 7 July 2023.

Introducing Data Justice

Watch the video Introducing Data Justice, by the Alan Turing Institute, and answer these questions.

a. What is datafication?

b. What are some issues with datafication? Who is the most impacted and/or harmed by it?

c. Why is data justice important?

d. Which voices does the video suggest centering?

e. What are data?

f. Why are data not objective or neutral?

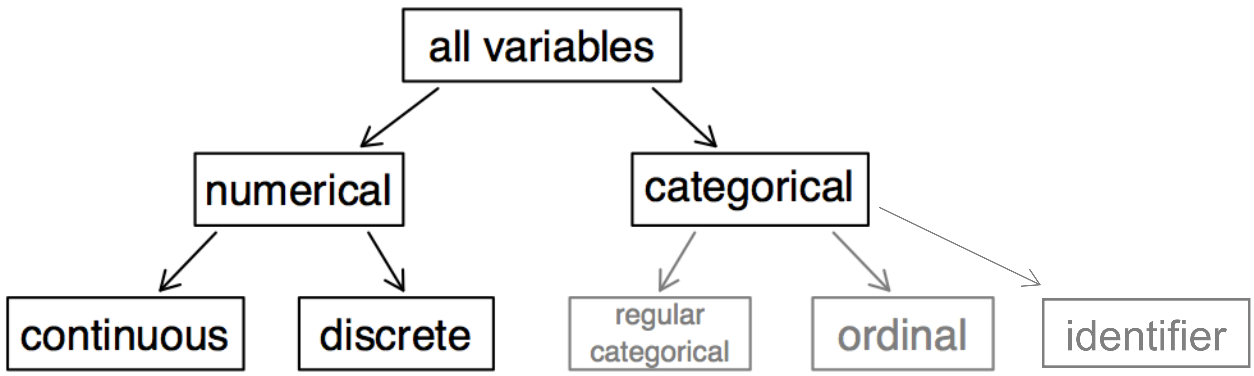
g. What is data justice? Many definitions are given, feel free to summarize in your own words.

h. Reflecting on this video, what impacted you the most?

Data Organization and Variables

Variable – A characteristics being recorded or measured

Types of Variables



Example 2. Use the data table to explore the ideas of cases and variables.

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	NO	OLYMPIC	10	1	3.13	54	4
2	F	YES	ACADEMY	4	7	2.5	66	2
3	Nonbinary	NO	NOBEL	14	5	2.55	130	1
4	M	NO	NOBEL	3	1	3.1	78	1

a. What is represented by the rows? What is represented in the columns?

b. List the numerical or quantitative variables. Also specify whether each is discrete or continuous.

c. List the categorical or qualitative variables. Are any of them ordinal or identifiers?

Population Parameters and Notation

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

Population, Sample, Parameter and Statistic

Example 3.

a. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 100 videos for you. You find that 2% of these videos are cat videos. Draw a picture to represent the population, sample, parameter and statistic.

b. Match the vocabulary word to each part of the study.

Element of the Study	Vocabulary Number	
i. Percentage of all videos on YouTube that are cat videos		1. Sample statistic
ii. 2%		2. Population
iii. A video in your sample		3. Variable
iv. Whether or not a video is a cat video		4. Population parameter
v. All YouTube videos		5. Sample
vi. The 100 videos		6. Case or subject

Sampling Methods

Surveys and observational Studies: In an observational study, researchers gather data without interacting with the subjects. *We cannot infer causation.*

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

Method	Description	Example
Census		
Simple Random Sample		
Stratified		
Cluster		
Systematic		
Multistage		

Biased Methods and Types of Bias

Representative Sample:

Bias: Any systematic failure of a sampling method to represent the population. There is no way to fix biased data, so it is better to design a good survey to begin with.

Methods that are Usually Biased

Method	Description	Example
Voluntary or Self-Selected Sampling Voluntary Response Bias		
Convenience Sampling Convenience Bias		

Additional Types of Bias

	Description	Example
Selection bias or Under coverage		
Non-response bias		
Response bias		

Experiments

In a **controlled, randomized experiment**, researchers assign treatments to groups of subjects and measure a response variable. *If the results are significant, we can infer causation.*

Example 4. Does the use of stents reduce the risk of stroke?

A stent is a wire support placed in a blood vessel. 451 at-risk patients volunteered for the study. Researchers randomly assigned 224 participants to the treatment group and 227 to the control group. They studied the effects at two time points: after 30 days and after 365 days.

Treatment group: Participants received a stent and medical management. The medical management included medications, management of risk factors and help in lifestyle modification.

Control group: Participants received the same medical management as the treatment group, but they did not receive stents.

a. Why did the researchers use a control group?

b. Why is it important to use **random assignment** of the participants to the groups? How is this different from **random sampling**?

c. Was a **placebo** used for **blinding**?

d. **Response variable**: What proportion of patients in each group did not have a stroke within a year?

Treatment group:

Control group:

	Stroke in 0-365 days	No stroke in 0-365 days	Total
Treatment Group	45	179	224
Control Group	28	199	227
Total	73	378	451

e. Does it look like the results of this experiment are statistically significant or due to random variation?

Rounding Review

Rounding

Step 1. Determine the place to which the number is to be rounded. Circle or underline it.

Step 2. If the digit to the right of the number to be rounded is less than 5, replace it and all the digits to the right of it by zeros. If the digit to the right of the underlined number is 5 or higher, increase the underlined number by 1 and replace all numbers to the right by zeros. If the zeros are decimal digits, you may eliminate them.

Place value chart

Ten thousands	Thousands	Hundreds	Tens	Ones	Decimal Point	Tenths	Hundredths	Thousandths	Ten Thousandths	Hundred Thousandths
10,000	1,000	100	10	1	.	.1	.01	.001	.0001	.00001

Example 5. Round each number to the place value given:

a. 126.745 inches to the nearest tenth

b. 5.68932 feet to two decimal places

c. 0.038594 to three decimal places

d. $ 43.893 to the nearest cent

e. 0.00125 to four decimal places

f. 0.00199 to four decimal places

Percentage Review

Percent means per 100, so depending on which way we are converting, we move the decimal 2 places to the left or to the right.

Percent to decimal

50% means 50/100. When we divide this we get 0.50 or 0.5. Notice how the decimal is now 2 places to the left.

Decimal to percent

0.25 is read as 25 hundredths, which can be written as 25/100. This is 25%. Notice how the decimal place is now 2 places to the right.

Memory aid

D **P**

Example 6. Convert each percentage to a decimal:

a. 31%

b. 130%

c. 3%

d. 0.3%

e. 1.3%

f. 1.23%

Convert each decimal to a percentage:

g. 0.97

h. 0.09

i. 0.009

j. 2.41

k. 0.3294

l. 0.0354

Divide and round each proportion to four decimal places.

m. $\frac{9}{25}$

n. $\frac{5}{36}$

o. $\frac{9}{11}$

p. $\frac{19}{570}$

Describing One Numerical Variable

A framework to describe data from a quantitative variable:

Describe the Shape, Center and Spread, and Unusual Features. Include units and the context.

Shape – How are the data distributed? We need to see a picture to determine the shape.

Four types of graphs for one quantitative variable. Always label your graph with the variable with units.

- Dot Plot
- Stem-and-leaf Plot
- Histogram
- Boxplot

Example 1. Here is a set of 15 exam scores for a fictional Stat 243 class at PCC.

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

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

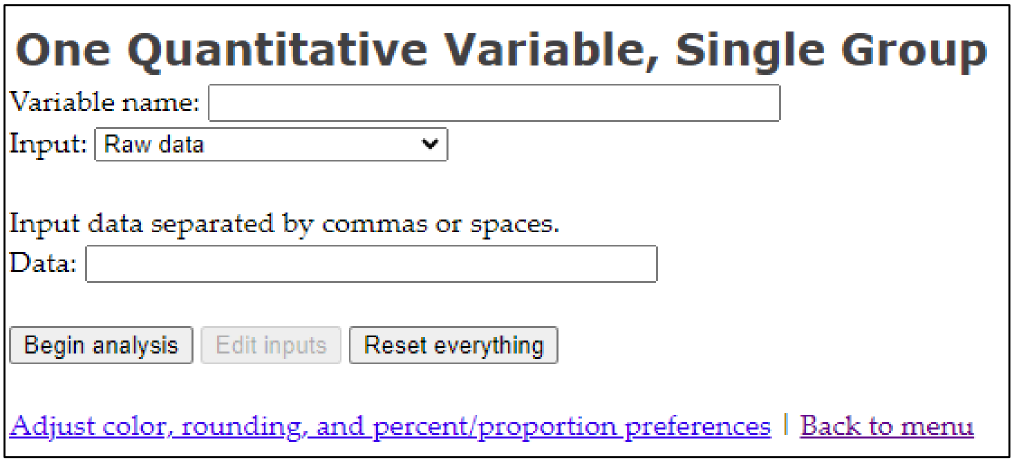
b. Draw a **stem-and-leaf plot** using the tens digits as the stem and the ones digits as the leaves.

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

Technology – Summarizing Stapplet

Once you feel comfortable with the calculations, please use technology. We will focus on using technology and interpretations in this class.

- Visit Stapplet.com
- Under Data Analysis, select 1 Quantitative Variable (Single Groups)
- Input the Variable name with units, this will be your title
- Copy and paste your data into the data box



- Click on Begin analysis.
  - Most often we'll want to select histogram and check the boxplot box.
  - You can type in an interval or bin width
- The summary statistics will come up automatically
- Use a screenshot or snipping tool to add graphs to your labs and assignments.

From now on we will graph using this applet: Stapplet.com.

Stapplet output - Dotplot

One Quantitative Variable, Single Group

Variable name: Exam Scores for a Fictional Stat 243 Class (points)

Input: Raw data

Input data separated by commas or spaces.

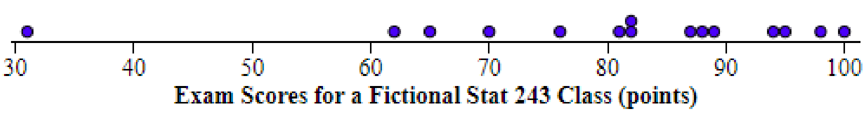
Data: 31 62 65 70 76 81 82 82 87 88 89 94 95 98 100

Begin analysis | Edit inputs | Reset everything

Graph Distribution

Graph type: Dotplot

☐ Show boxplot



Stapplet output - Stemplot

Graph type: Stemplot

Split stems: No

Shift stem 0 additional decimal places to the left, truncating as needed.

Collapse groups of empty stems? No

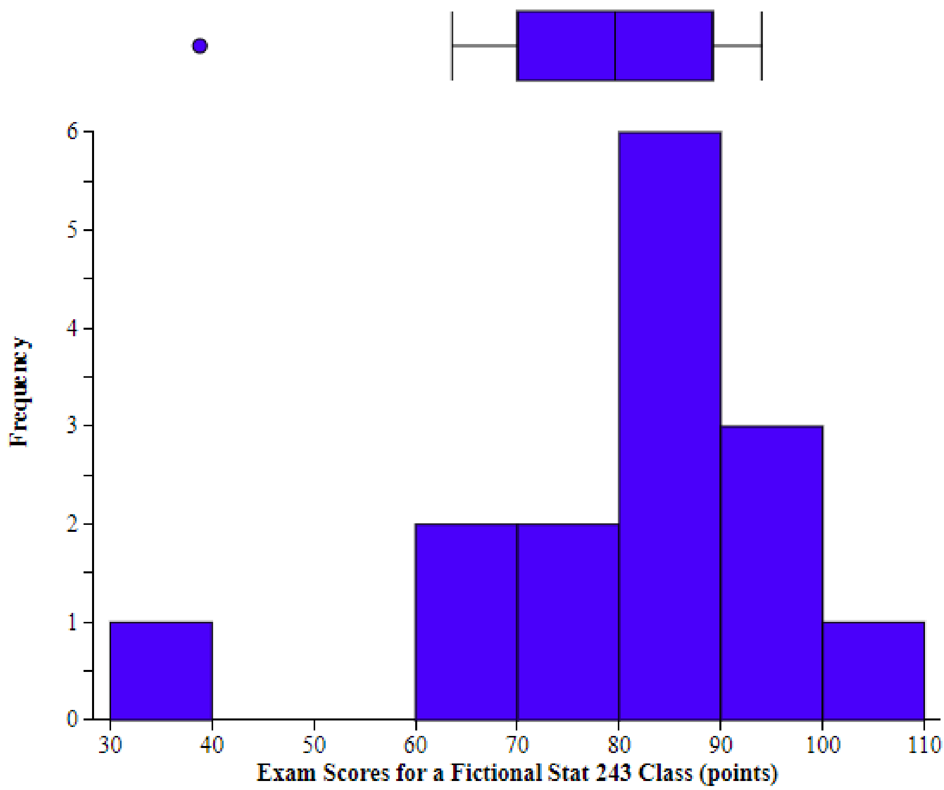
```
 3 | 1
 4 |
 5 |
 6 | 2 5
 7 | 0 6
 8 | 1 2 2 7 8 9
 9 | 4 5 8
10 | 0
```

Exam Scores for a Fictional Stat 243 Class (points)

KEY: 10|0 = 100

Stapplet output – Histogram, Boxplot and Summary Statistics

Graph type: Histogram Label histogram with: Frequency
Enter interval width: 10 Enter boundary value: Reset intervals to default
Show boxplot



Summary Statistics

n	mean	SD	min	Q_1	med	Q_3	max
15	80	17.7563	31	70	82	94	100

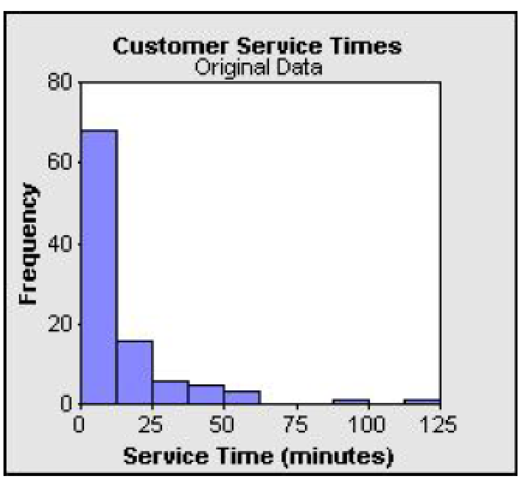
Describing Graphs of Numerical Data

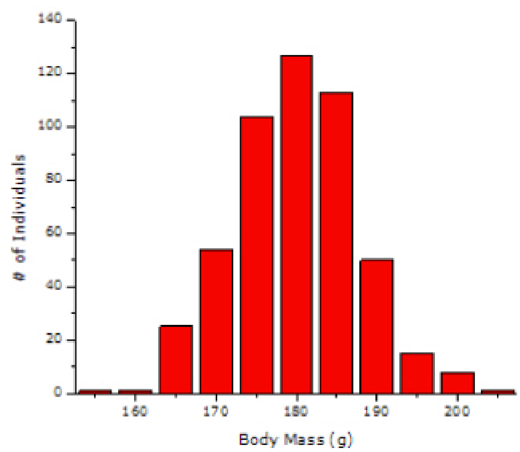
Shape

Symmetric skewed left skewed right

no mode unimodal bimodal multimodal

Example 2. Describe the shape of each distribution. What are the modes (if any)?





Describing Center and Spread

Center and Spread – These should always be reported together, with units.

Center is the position or location of the data (The average or typical value).
Spread is how much variation is in the data or how spread out the data is.

There are two different sets of measures for center and spread:

For symmetric distributions:

- Center: Mean, $\bar{x}$
- Spread: Standard Deviation, s

For skewed distributions:

- Center: Median
- Spread: IQR (Interquartile Range)

Mean and Standard Deviation

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

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

$$\bar{x} = \frac{\sum x}{n}$$ or "the sum of the values divided by n"

Example 3. Calculate the mean of the three data sets:

a. Data Set A
3 8 12 15 18

b. Data Set B
3 8 12 15 18 20

c. Data Set C
3 8 12 15 18 205

Standard Deviation: A measure of spread used for symmetric data. The “average deviation from the mean.”

Population standard deviation = σ Sample standard deviation = s

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

Example 4. The table below shows a sample of 6 Stat 243 student’s heart rates, measured in beats per minute (bpm). The mean of this set is 70. Calculate the deviation and squared deviation by hand. Then calculate the variance and the standard deviation.

Heart Rate (in bpm)	Deviation from the mean	Squared Deviation
52		
68		
70		
72		
73		
85		
	Sum of the Squared Deviations	

The **variance, s^2,** is the sum of the squared deviations divided by (n-1) called the degrees of freedom.

s^2 =

The **standard deviation, s**, is the square root of the variance.

s =

Let’s check our standard deviation using Stapplet.com.

Describe in words the center and spread of this data set.

Median & Interquartile Range

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

Two Cases:

- Odd # of values (Data Set A)

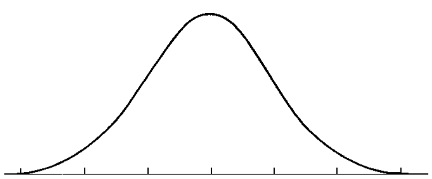
3 8 12 15 18

- Even # of values (Data Set C)

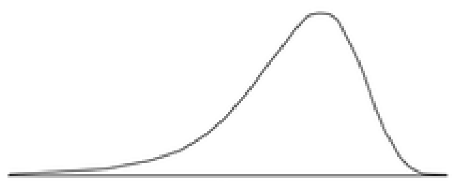
3 8 12 15 18 205

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

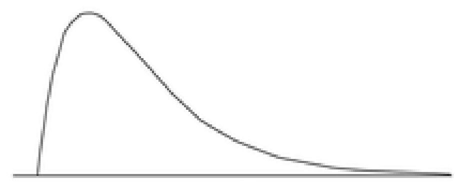
Symmetric Distributions



Skewed Distributions



Skewed Left



Skewed Right

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

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

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

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

5-Number Summary – Described by the minimum, Q1, median, Q3, and the maximum values for a data set.

Range: Describes the distance between the minimum and maximum value

Range = Max – Min

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

IQR = Q3 – Q1

Boxplots

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

Example 1. continued: The data set below represents 15 exam scores for a fictional Stat 243 Statistics class at PCC

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

1. **Collect statistics:** Collect the 5 number summary and calculate the IQR

2. **Draw the Box:** Determine the scale and draw vertical lines at the Median, Q1, and Q3. Connect these to form the box. Label your horizontal axis and include the scale.

3. **Determine Outliers:** We use 1.5 times the interquartile range on each side of the box to determine the fences. Any data outside the fences are considered outliers. The whiskers are drawn to the nearest ***data*** values inside each fence.

   Upper Fence = Q3 + 1.5 * IQR Lower Fence = Q1 – 1.5 * IQR

   The fences are invisible, so don't draw them. Fences are not data, just bounds to determine outliers.

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

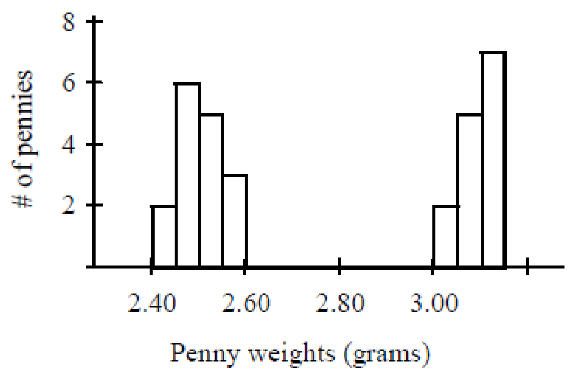
Compare this boxplot to the one we made earlier using Stapplet.com.

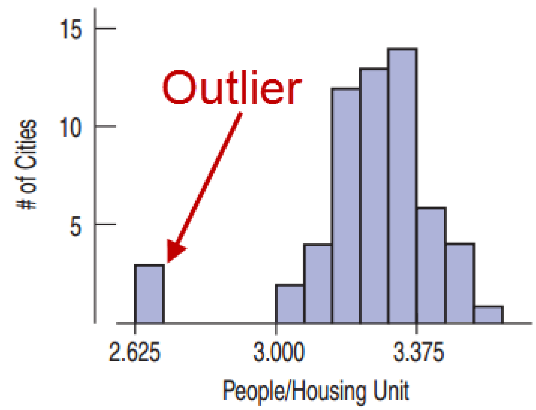
Putting it all Together in a Summary

Unusual Features

Mention anything unusual about the data or state that there aren't any unusual features

- Multiple modes (look for subpopulations)
- Gaps and Outliers





Writing a paragraph to go with the histogram and boxplot

Example 1. continued: Continuing with our test score data set, write a paragraph describing the distribution. Be sure to talk about the shape, center and spread, and any unusual features (or say that there are none). Include the context and units.

If the data were symmetric, we would use the mean and standard deviation for center and spread.

Shape Test: If you're not sure whether the data are symmetric or skewed, compare the mean and median**.**

- Mean and median about the same:
- Mean smaller than the median:
- Mean larger than the median:

Determining Relative Standing Using Z-Scores

It can be useful to know where a particular data value falls compared to the rest of the data values. The standard deviation is the average deviation from the mean, so we use this as a measure.

- 1 standard deviation from the mean is considered average
- 2-3 standard deviations from the mean is considered unusual
- 3 or more standard deviations from the mean is considered rare

A Z-score calculates how many standard deviations a data value is away from the mean.

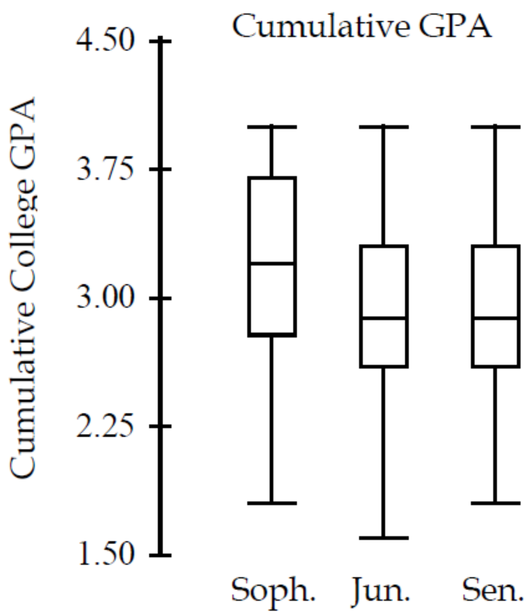
$$\boldsymbol{Z} = \frac{\boldsymbol{x} - \text{mean}}{\text{standard deviation}}$$

Example 4. Continued: From our pulse rate data set with 6 students, we found the mean of the group was 70 bpm and the standard deviation was 10.64 bpm.

a. Find the Z-score for the pulse rate of 52 bpm.

b. Find the Z-score for the pulse rate of 73 bpm.

Comparing Multiple Groups with Side-By-Side Boxplots



Example 5. The side-by-side boxplots show the cumulative college GPAs for sophomores, juniors and seniors taking an intro stats course.

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

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

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

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

Displaying a Single Categorical Variable

For a single categorical variable, we make a frequency table to tabulate the results. A frequency table uses category names for each row and records the total count of each value. A relative frequency table gives the percentage in each category.

Two types of graphs for one categorical variable.

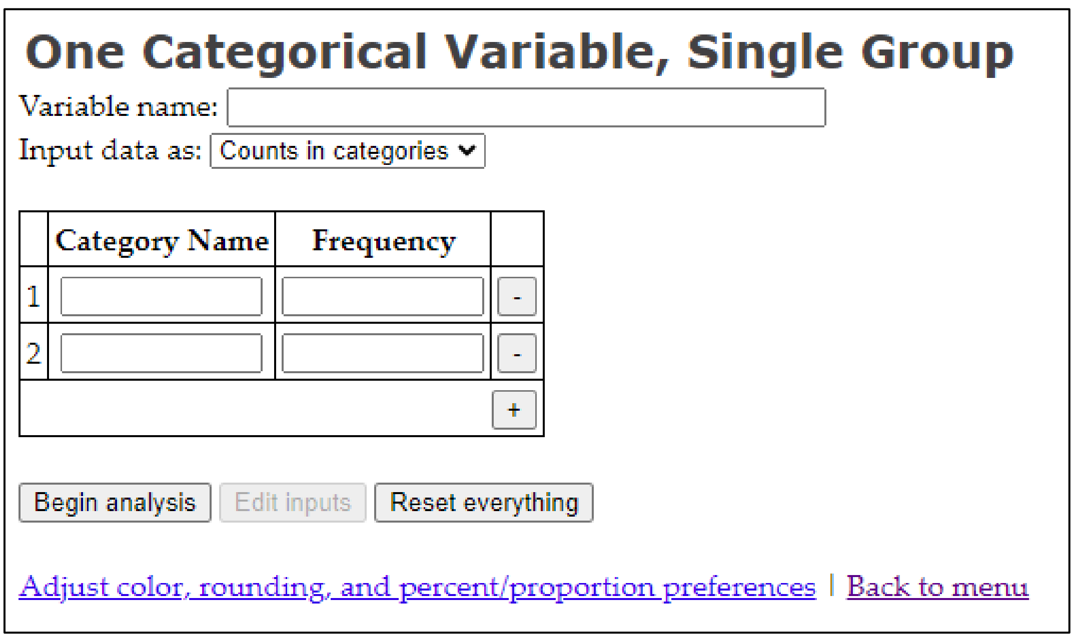
- Bar charts
- Pie charts

Example 1: Here are some results from a student survey on eye color.

a. Using the data given, find the relative frequency for each category.

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

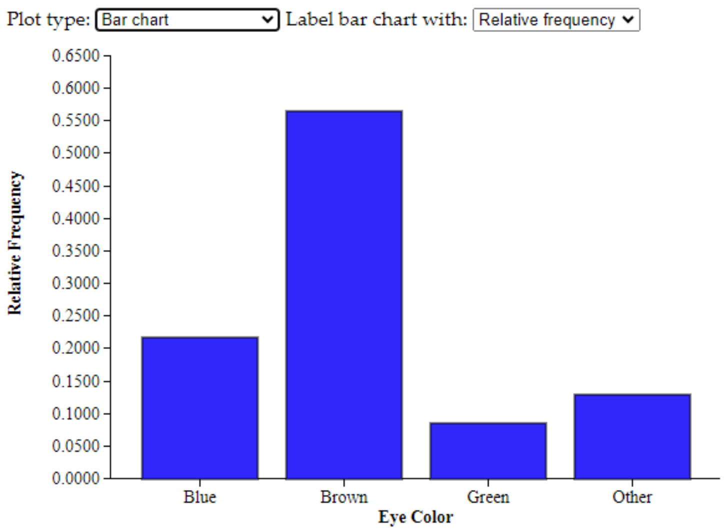
b. Using [stapplet.com]() or a spreadsheet, make a bar chart and a pie chart for this data set.



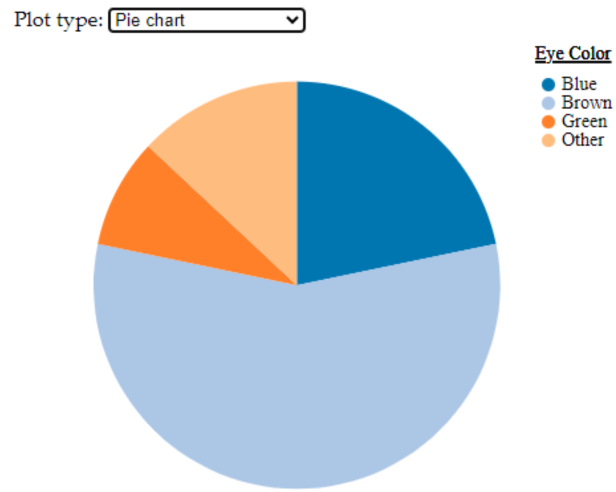
At stapplet.com, choose the applet for 1 categorical Variable, Single Group

Enter the variable name, each category and the frequency of each category, then click on Begin Analysis

Bar Chart from Stapplet.com



Pie Chart from Stapplet.com



Summary Statistics

Category Name	Frequency	Relative Frequency
Blue	5	0.2174
Brown	13	0.5652
Green	2	0.087
Other	3	0.1304
Total	23	1

Writing Sample Proportions

Example 1 continued: Answer the following questions using the data.

c. What proportion of the group has blue eyes?

d. What proportion of the group has brown eyes?

e. What proportion of the group doesn't have brown eyes?

f. What proportion of the group has green or blue eyes?

Two Categorical Variables and Empirical Probability

Example 2. A study on treatments for addiction to cocaine. Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
Total	48	24	72

Marginal Probabilities (Margins or Totals)

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

b. What is the percentage of participants who were given Lithium in the study?

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

Joint Probabilities (And)

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

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

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

"Or" Probabilities

"Or" probabilities depend on whether the events are disjoint or not

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

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

If A and B are disjoint events, **P(A or B) = P(A) + P(B)**

If two events A and B are non-disjoint, **P(A or B) = P(A) +P(B) – P(A and B)**

	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
Total	48	24	72

e. If we select a participant at random, what is the probability that they were given desipramine or lithium? Are these disjoint events or not?

f. What is the probability that a participant had lithium or relapsed? Are these disjoint or not?

Empirical vs. Theoretical Probabilities

An **empirical probability** is calculated from data, an experiment or simulation.

Examples: the probabilities we just calculated, flipping coins, or running a computer simulation

A **theoretical probability** is calculated using a mathematical model or formula.

Conditional Probability

Example 2 continued:

	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
Total	48	24	72

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

P(Relapse | Desipramine)=

h. Given that a person had a relapse, what is the probability that they were in the placebo group? Write the probability statement and the answer.

i. What is the probability that someone had a relapse if they were in the placebo group? Write the probability statement and the answer.

These are called conditional probabilities because we are given one of the variable values. We only use a single row or column to find the conditional probability.

Conditional Probability Formula: For events A and B,

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

Note: this is exactly how we calculated the conditional probabilities using the table, so you don't need to use this formula unless you want to.

Independence Test – Conditional Test

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

For **theoretical probabilities**, if the two sides of the equation are equal, the two events are independent. If the two sides are not equal, they are dependent.

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

Note: this is a basic test for now to understand the concept of independence. There is a significance test using the whole table in Math 244.

Example 2 continued: Is having a relapse independent of the type of treatment?

We want to choose one row or column of the **response variable** (outcome) and compare the conditional probability given each category in the **explanatory variable** (the one that might affect the response variable).

	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
Total	48	24	72

Probability Practice

Example 3. How are the smoking habits of students related to their parents' smoking? Here is a contingency table of data from a survey of students in 8 Oregon high schools.

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

a. P(student smokes)

b. P(no parent smokes)

c. P(at least 1 parent smokes)

d. P(student smokes and 1 parent smokes)

e. P(student smokes or no parent smoke)

f. What is the probability that a student who smokes has no parents that smoke?

g. What is the probability that if two parents smoke, their child will smoke?

h. Does student smoking seem to be independent of their parents smoking? Show your test and explain your conclusion.

Overview of Statistical Inference

Inferential statistics

Estimation – Confidence Intervals

Statistical Significance – Hypothesis Testing

Population, Parameter, Sample Statistic

Variables and Point Estimates

Type of Variable	Quantity of Interest	Population Parameter Fixed but unknown	Sample Statistic or Point Estimate Known but varies by sample
Categorical (yes/no)	Proportion	p	$\hat{p}$
Numerical	Mean	μ	$\bar{x}$
2 Categorical	Difference of 2 Proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
2 Numerical	Difference of 2 Means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

Example 4. Identify the parameter and statistic of interest in each situation with units for numerical variables.

a. In a study of 100 YouTube videos, 2% of the videos were cat videos.

b. A phone company wants to know whether teenagers send more texts than adults. They collect the number of texts sent per day for a month from 1000 randomly selected adults and 1000 randomly selected teens.

c. A large company is interested in the average number of hours its employees spend on email each week. They collect data from a random sample of 200 employees.

Sampling Distribution of a Mean

Sampling Distribution – A distribution for a statistic from many random samples of the same size from the same population. The statistic can be a count, proportion, mean, median, difference of two means or proportions, etc.

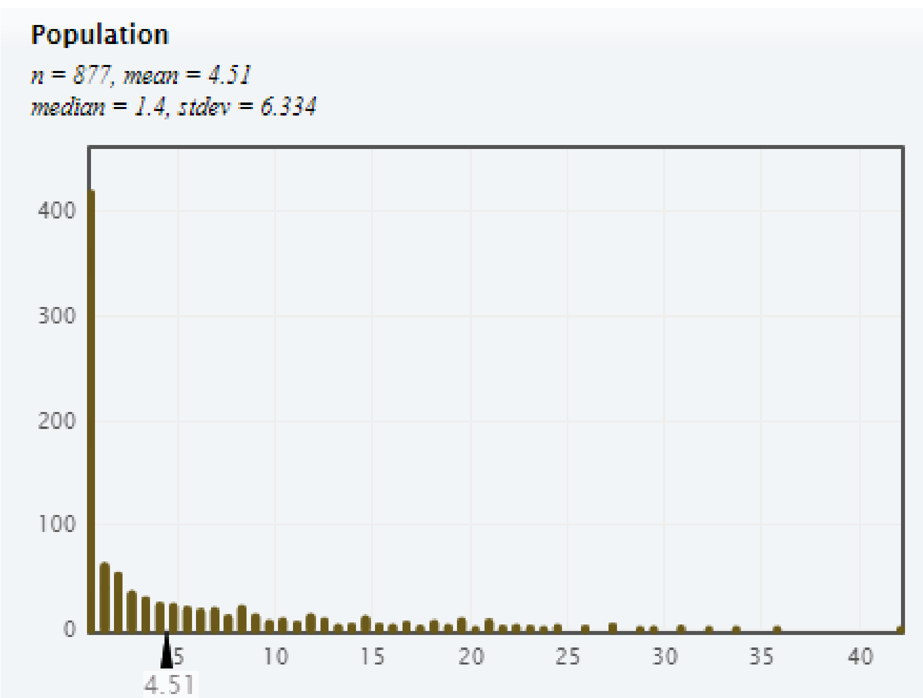
Note: We are pretending to know the population here so we can understand what the distribution of random samples looks like. Usually, we don't have the population information – that's why we're doing statistics in the first place!

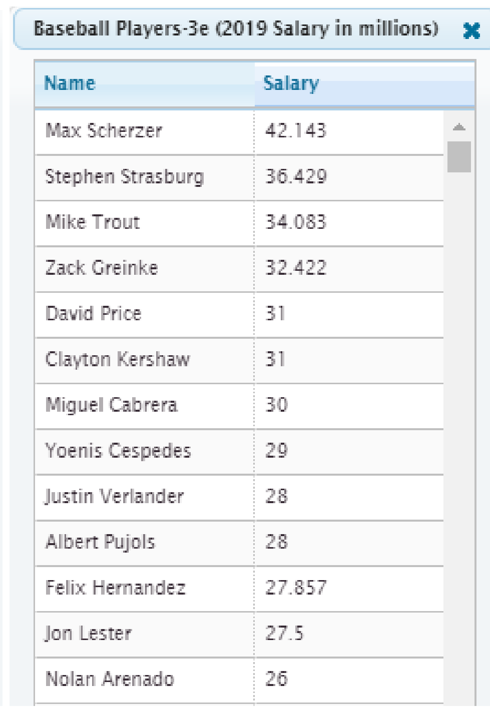
Go to the Rossman/Chance Sampling from a Finite Population applet.

StatKey Sampling Distribution of a Mean applet and click on Mean in the Sampling Distribution row.

Example 5. In the upper left corner, select Baseball Players-3e (2019 Salary in millions).

You'll see this population graph on the right and if you click on Show Data Table it will list the data set.





Baseball Players-3e (2019 Salary in millions)

Name	Salary
Max Scherzer	42.143
Stephen Strasburg	36.429
Mike Trout	34.083
Zack Greinke	32.422
David Price	31
Clayton Kershaw	31
Miguel Cabrera	30
Yoenis Cespedes	29
Justin Verlander	28
Albert Pujols	28
Felix Hernandez	27.857
Jon Lester	27.5
Nolan Arenado	26

First, let's examine this population.

a. Who or what are the cases or subjects? What is the size of the population?

b. What is the shape of the population?

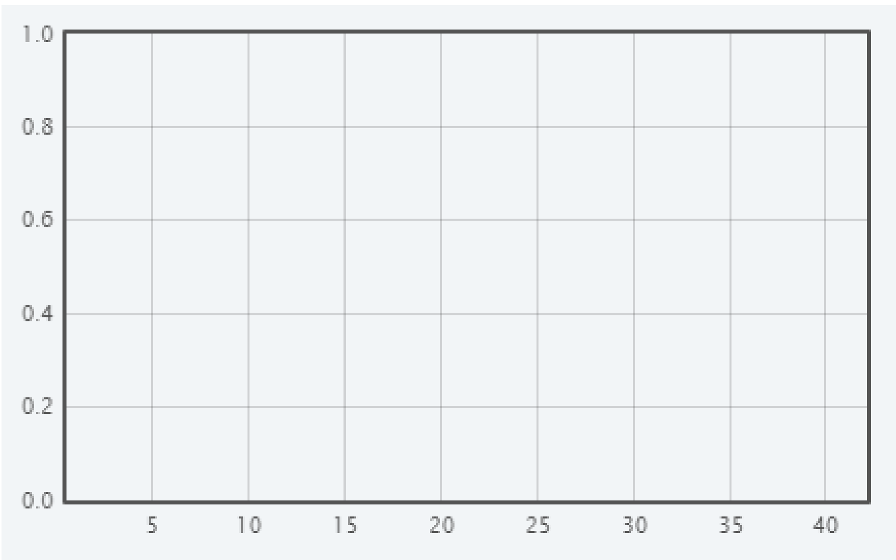
c. What is the mean of the population with notation and units?

d. What is the standard deviation of the population with units?

Now, we're going to use the applet to select one sample of size n=10. You can see the sample in the lower right window, and the average is placed on the large dot plot. Do a few of these until you understand what's happening. Then generate 1000 twice to get roughly 2000 random samples of size 10.

Noting the original scale at the bottom, draw a rough sketch outline of the distribution of 2,000-ish random sample means of size n=10 and list the mean and standard error. Repeat for n=50. How do these compare with the population?

e. 2000 samples of size n=10

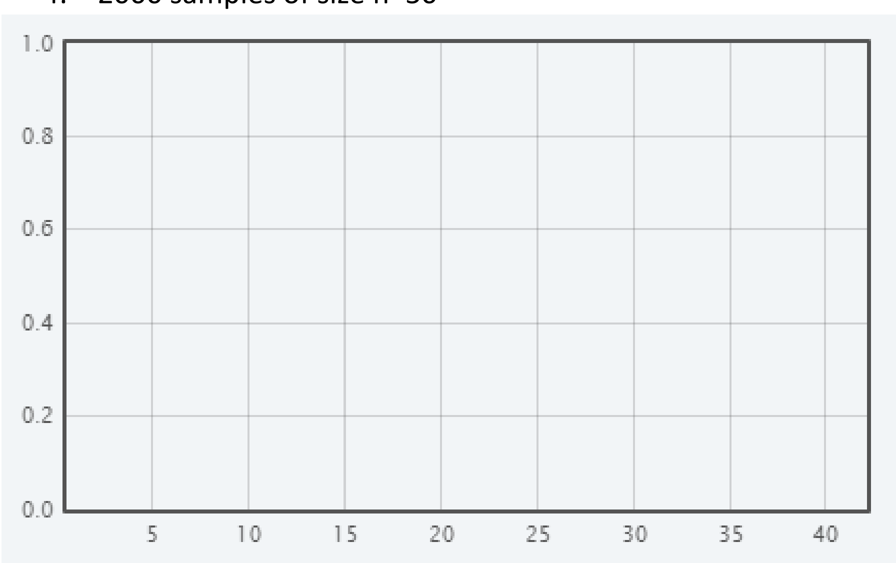


Shape:

Mean =

Standard Error =

f. 2000 samples of size n=50



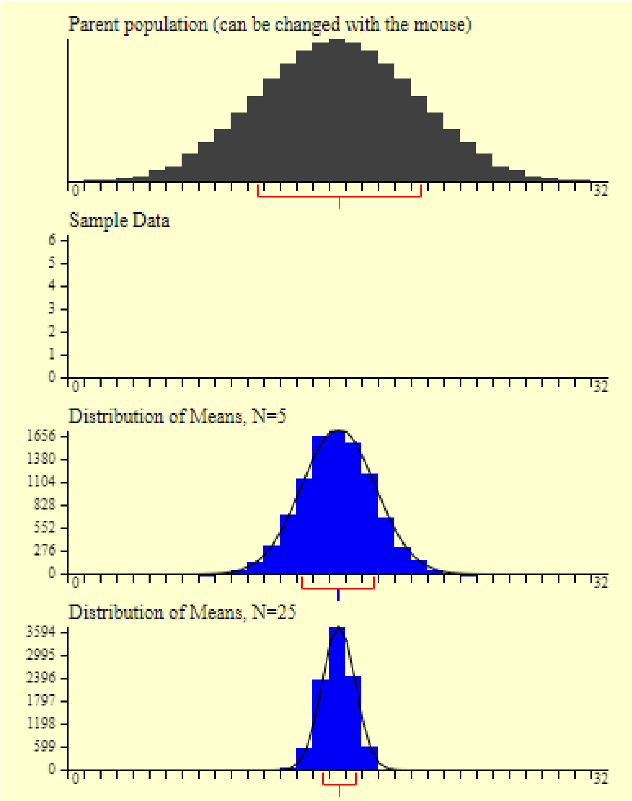
Shape:

Mean =

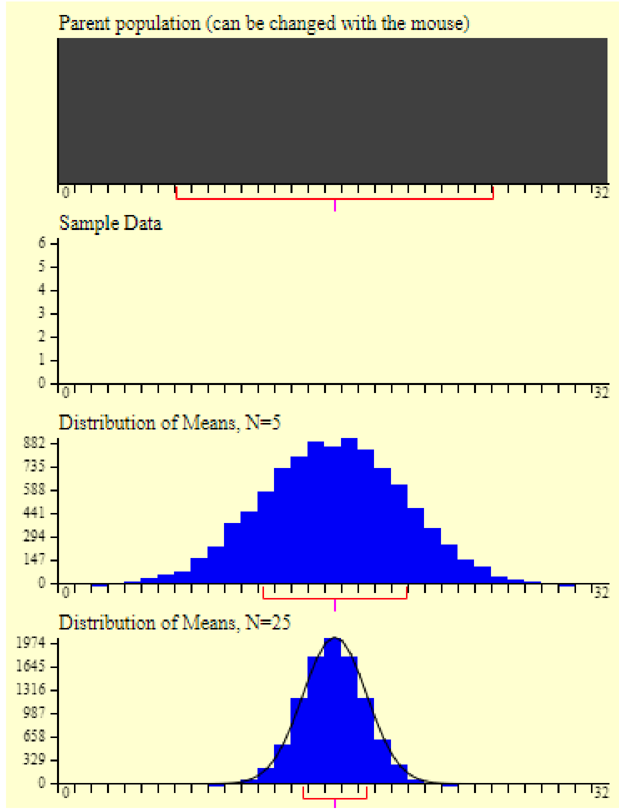
Standard Error =

What about other population shapes? We will explore that with the [OnlineStatbook Sampling Distribution of a Mean applet](). Choose the parent population shape, select mean in the lowest two graphs, with n=5 and n=25. Then click on 10,000. What do you notice?

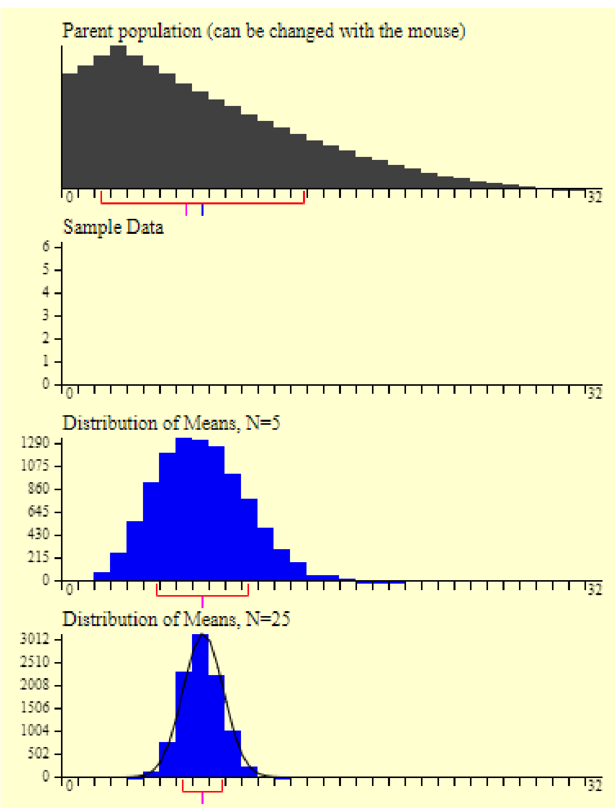
Normal Population



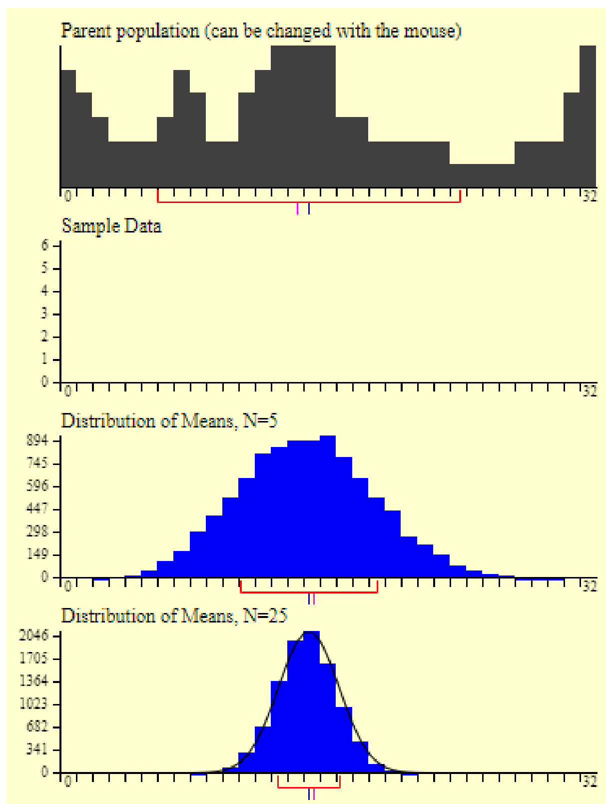
Uniform Population



Skewed Population



Multimodal Population



Standard Error and Central Limit Theorem

Standard Error

The **Standard error, SE**, is the standard deviation of a sample statistic. It is the sample-to-sample variation. We use a slightly different term to distinguish it from individual-to-individual variation.

Central Limit Theorem

Regardless of the population shape, for large enough random samples from a large population, the distribution of the mean of random samples follows a normal distribution, centered at the population mean. The larger the sample size, the smaller the _______________.

Example 6. State whether the quantity described is a parameter or statistic and give the correct notation.

a. Average number of cigarettes smoked per person for all smokers in the United States.

b. Average household income for a random sample of 5000 households in the American Community Survey.

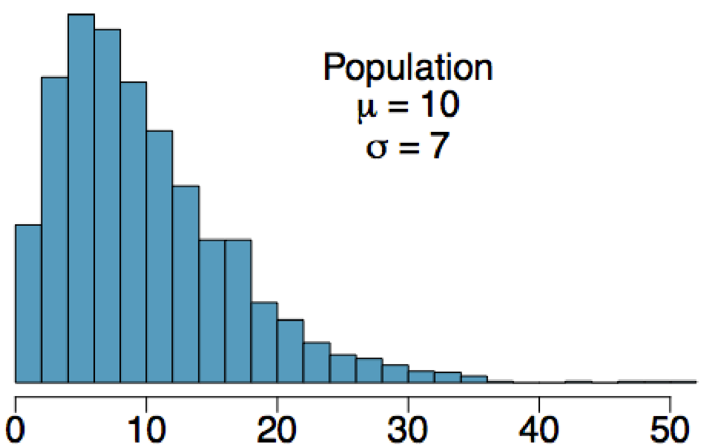
Example 7. Matching sample sizes to graphs. A population histogram is shown with a mean, μ = 10, and standard deviation, σ = 7.

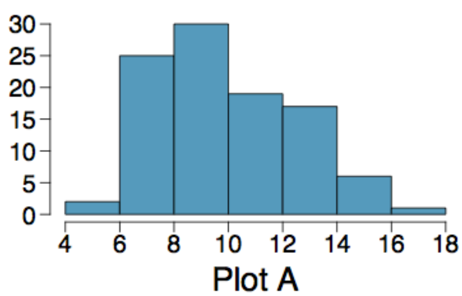
Determine which plot (A, B, or C) goes with each of the following. Note: look at the horizontal scales as well as the shape.

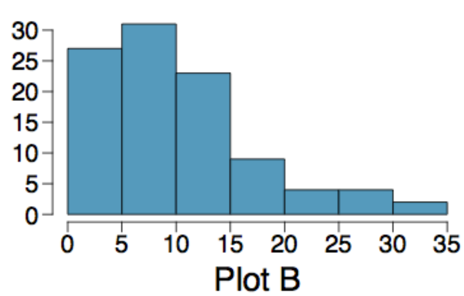
a. a single random sample of 100 observations from this population,

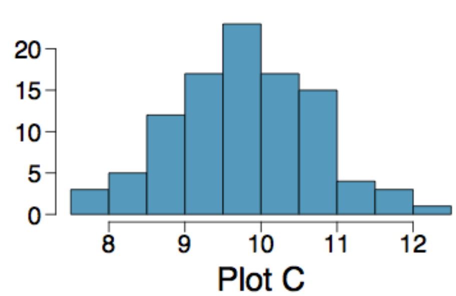
b. a distribution of 100 sample means from random samples with size n=7,

c. a distribution of 100 sample means from random samples with size n=49.









Finding Probabilities for a Sampling Distribution of a Mean

Example 8. Use the Rossman/Chance One Variable with Sampling Applet, select the Sleep 1 data set and click on Use Data. Check the box for Show Sampling Options.

a. Simulate 2000 random samples from the population of size n=100. Choose mean for the mean number of hours of sleep per night. What is the mean and SE of the sampling distribution?

b. What proportion of samples have a mean of 8.4 hours of sleep or greater?

c. Find the probability $P(\overline{X} \geq 8.4)$.

d. What proportion of samples have a mean of 7.75 hours of sleep or less?

e. Find the probability $P(\overline{X} \leq 7.75)$.

Sampling Distribution of a Proportion

Sampling Distribution – A distribution for a statistic from many random samples of the same size from the same population. The statistic can be a count, proportion, mean, median, difference of two means or proportions, etc.

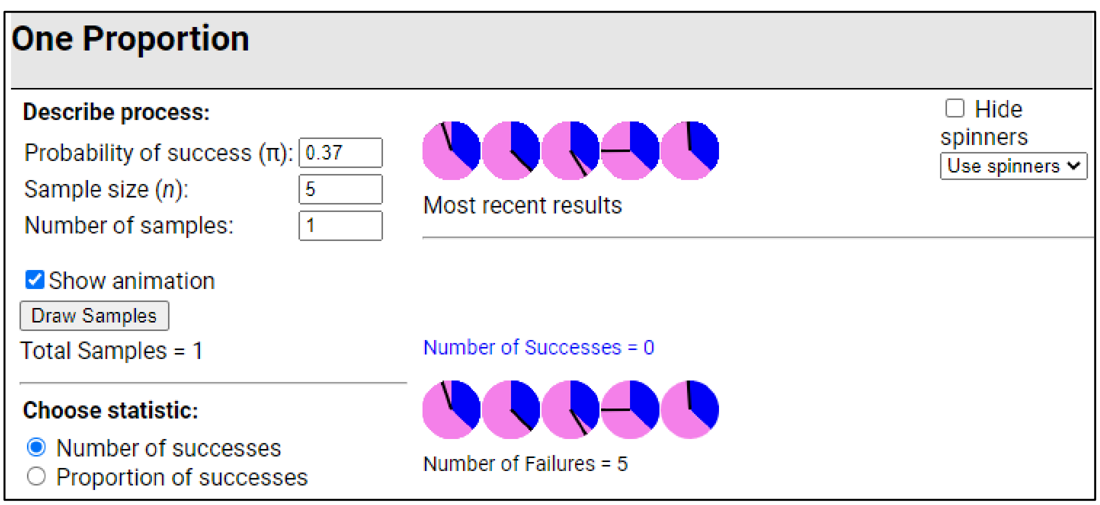
Note: We are pretending to know the population here so we can understand what the distribution of random samples looks like. Usually, we don't have the population information – that's why we're doing statistics in the first place!

Now we'll look at the sampling distribution for a categorical variable, blood type. The type O+ is the most common blood type. In the United States, approximately 37% of people have type O+ blood according to the Red Cross website.

Let's use a simulator applet for the sampling distribution of a proportion. The link is also in D2L above this video. Rossman/Chance One Proportion Inference

Enter 0.37 in the "Probability of Heads" box and you'll see it change to the "Probability of Success." We've seen with a probability of 0.5 that coins are used, and with a different proportion it changes to spinners.

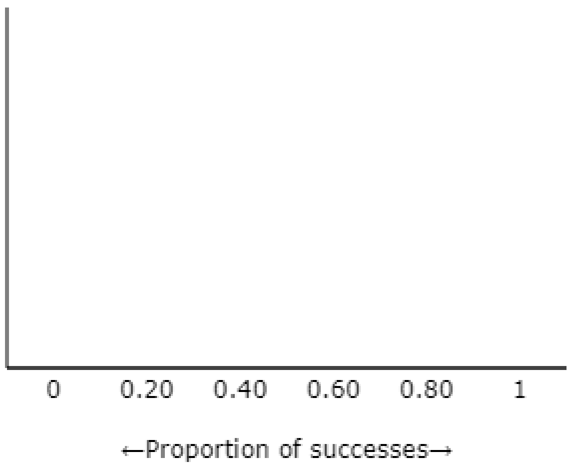
Enter 5 for the sample size, 1 sample and check show animation. Then click on draw samples. You'll see that the blue area of each spinner represents 0.37 and the pink area is 0.63. The black line shows which region the spinner landed in.



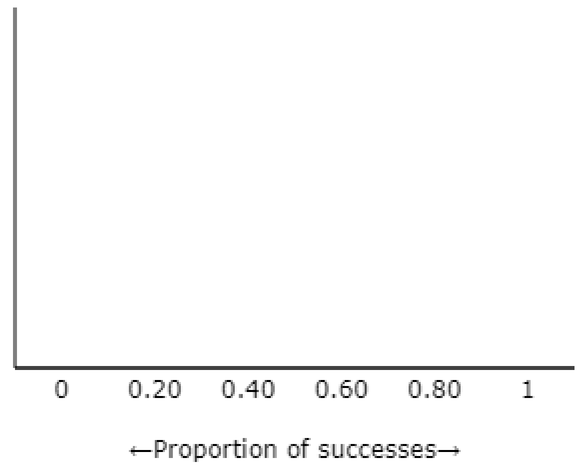
Click on draw samples a few more times and notice the dot for each sample placed on the dot plot.

Then generate about 2000 samples, change the statistic to the proportion and sketch the outline of the distribution on the graph below. Check the summary statistics box and write the mean and standard error for the graph. Repeat for each sample size below.

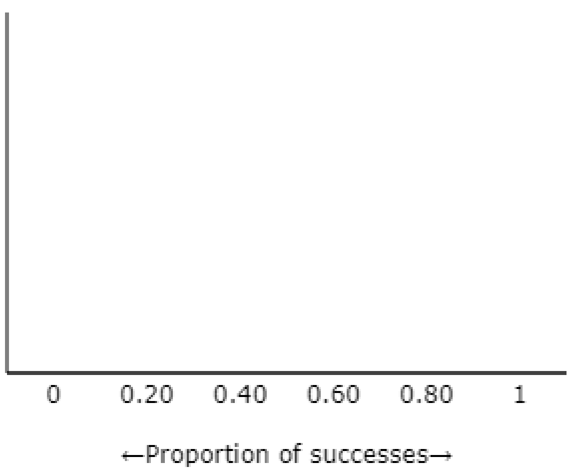
$\boldsymbol{p = 0.37, n = 5}$



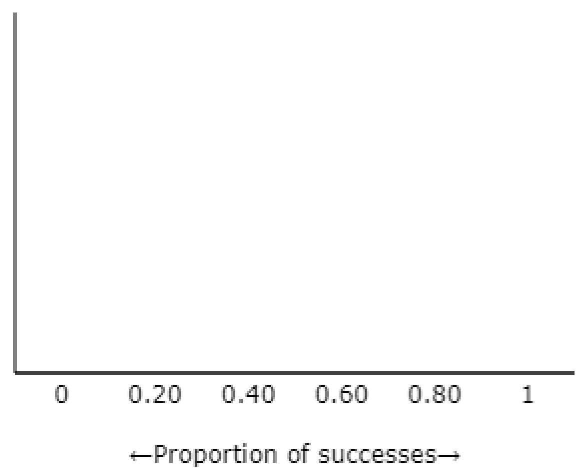
$\boldsymbol{p = 0.37, n = 10}$



$\boldsymbol{p = 0.37, n = 50}$



$\boldsymbol{p = 0.37, n = 100}$



What do you notice about the shape as the sample size gets larger?

What do you notice about the standard error as the sample size gets larger?

Standard Error and the Central Limit Theorem for Proportions

Standard Error

The **Standard error, SE**, is the standard deviation of a sample statistic. It is the sample-to-sample variation. We use a slightly different term to distinguish it from individual-to-individual variation.

Central Limit Theorem

Regardless of the population proportion, for large enough random samples from a large population, the distribution of a proportion follows a normal distribution, centered at the population proportion. The larger the sample size, the smaller the ____________.

The value of p and n

Does the value of the proportion affect the shape of the sampling distribution? Check the show sliders box and move the success probability slider from left to right. What do you notice about the shape?

Increase the sample size slider to 1000 all the way to the right. Then move the probability slider again. What changed?

Finding proportions and probabilities

Example 1. Simulate a sampling distribution with p=0.37 for the proportion of people in the US with type O+ blood. Simulate 2000 random samples of 100 people.

a. In what proportion of random samples do you get a proportion of 0.42 or greater who have a blood type of O+?

b. Using your simulation, estimate $P(\hat{p} \geq 0.42)$.

c. In what proportion of samples do you get a proportion with type O+ blood less than 0.25?

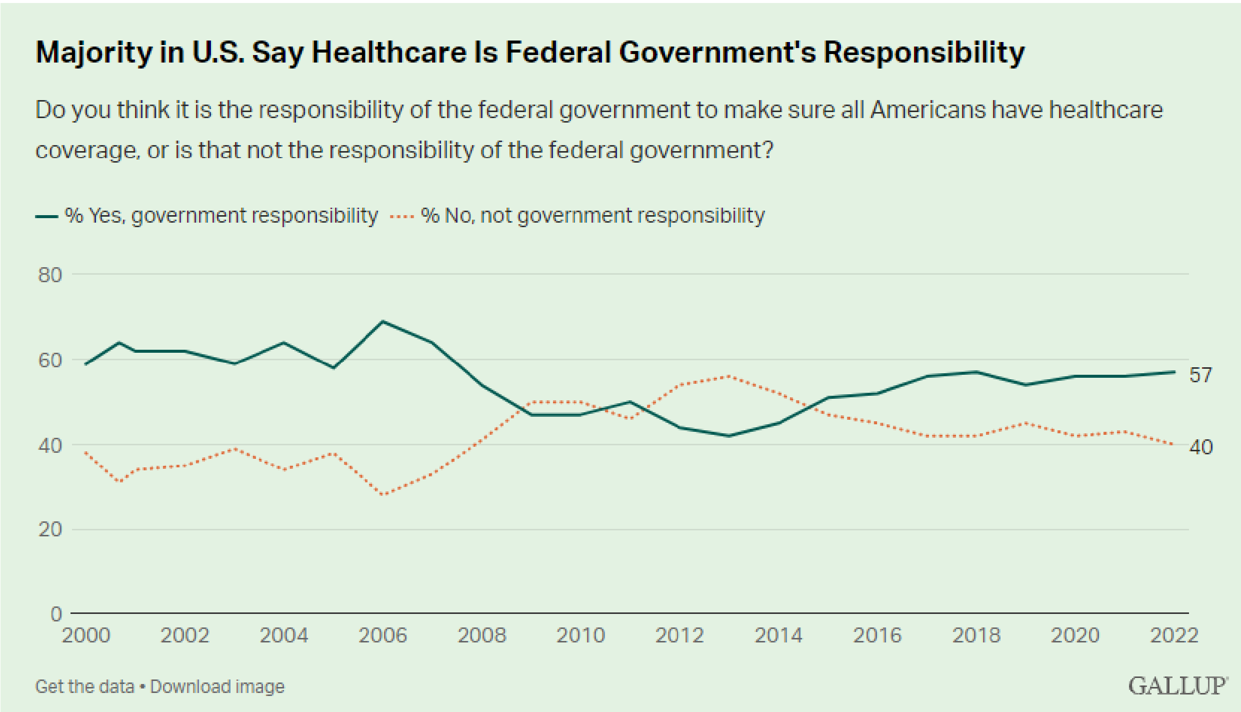
d. Estimate $P(\hat{p} \leq 0.25)$.

Introduction to Confidence Intervals

Statistical Polls

Go to this Gallup article where they describe polling results on U.S. healthcare. Find the latest sample proportion of people who think ensuring healthcare is the government's responsibility. Click on the methodology link at the bottom of the article to find the sample size, and the margin of error. Then use the margin of error to write the confidence interval and give the interpretation.

https://news.gallup.com/poll/468401/majority-say-gov-ensure-healthcare.aspx



Sample Proportion or point estimate, $\hat{p} =$

Margin of Error, ME =

Sample Size, n =

Confidence Level:

A **confidence interval** gives a plausible range for the values for the population parameter we are trying to estimate. The range is used to account for sample-to-sample variation.

Calculate the confidence interval:

Interpret the confidence interval: We are ________% confident that the true proportion of ________________

____ is between __________% and ____________%.

Confidence Interval vs. Point Estimate

Now that we have studied the sampling distribution of a proportion, $\hat{p}$, we can begin to look at the estimation part of **inferential statistics**. That is, we want to take a single random sample and make an estimate of the population parameter, which we do not know.

Using only a sample statistic to estimate a parameter is like fishing in a lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

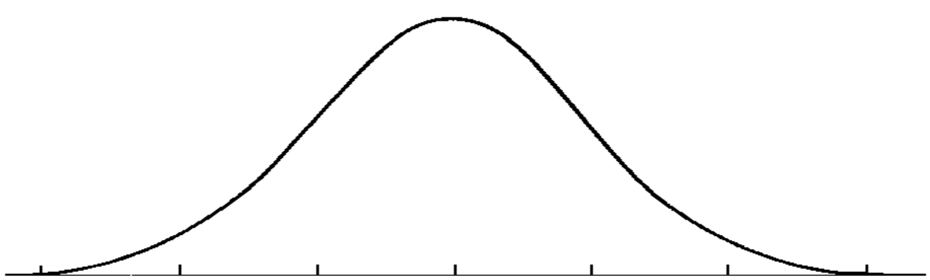
Spear fishing vs. Net fishing[1]

Confidence Level

The width of our net is determined by the confidence level and the standard error (affected by sample size).

Empirical Rule or 68-95-99.7% Rule for Normal Distributions.

In a normal distribution, about 68% of the values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean. Label the bell curve to show these key features. To be 95% confident of capturing the true parameter, we go out **2 standard errors from the point estimate**. This is called the **margin of error (ME)**.



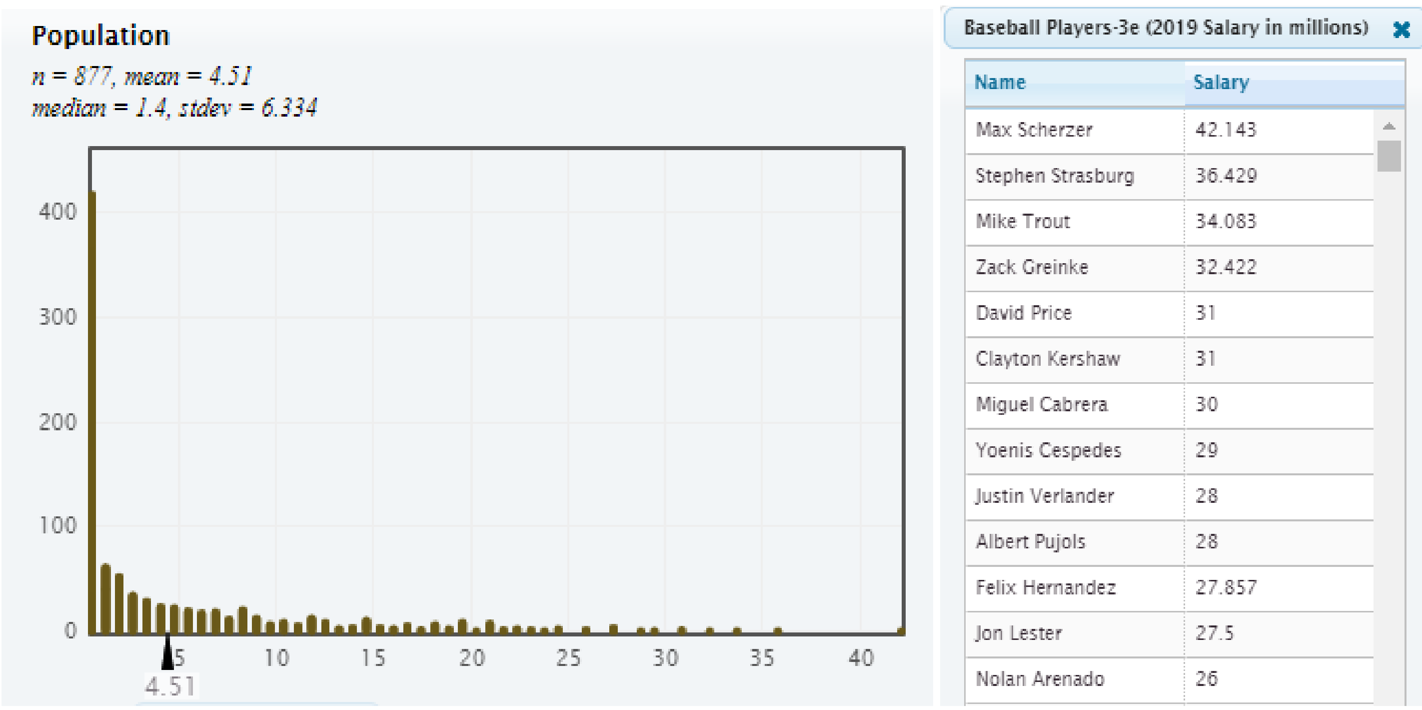
[1] Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence Interval Formula

$$\boxed{\begin{gathered}\text{point estimate} \pm \text{margin of error}\\ \hat{p} \pm 2 \cdot \text{SE}\\ \bar{x} \pm 2 \cdot \text{SE}\end{gathered}}$$

Example 2. Let's continue the baseball salary example. Please go to the StatKey Sampling Distribution of a Mean applet. In the upper left corner, select Baseball Players-3e (2019 Salary in millions).

We're still pretending to know the population characteristics so we can learn about confidence intervals, but we're only taking one sample. Then we will take this information away.



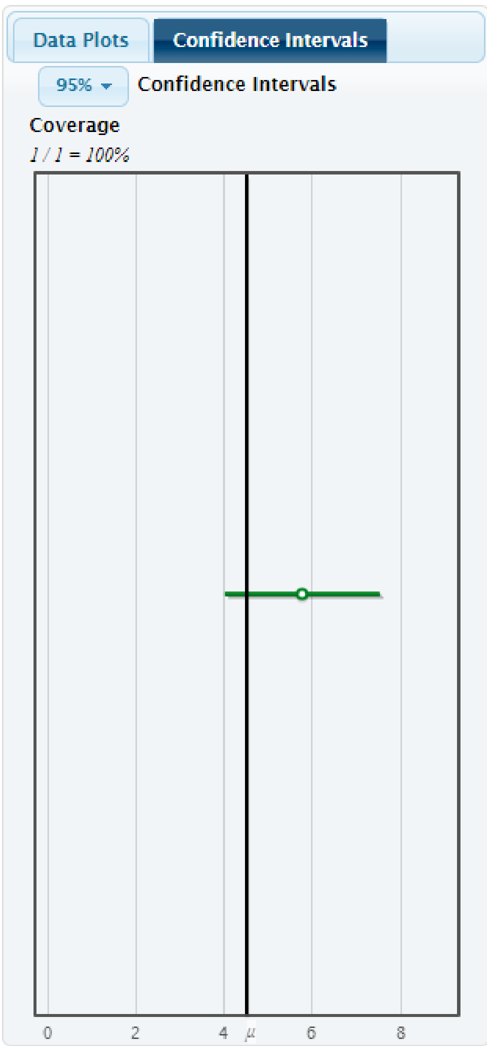
Change the sample size to n=50. Then generate 1 sample and write down $\bar{x}$. When we simulated a sampling distribution for samples of 50 players before, we got a standard error of $0.906 million.

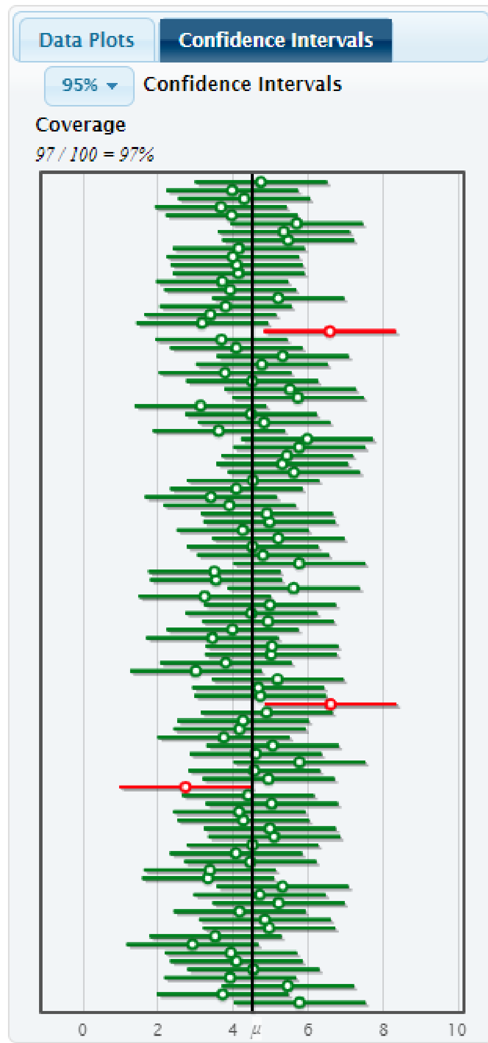
Using the formula above, calculate the 95% confidence interval. Does it contain the true mean?

The Meaning of the Confidence Level

Remember the parameter is fixed but unknown, and the point estimate is known but varies. We use the word confidence to convey that the uncertainty is in the confidence interval, not in the population parameter. The interval varies from sample-to-sample, not the parameter.

Next, we will generate 100 confidence intervals to understand the meaning of confidence intervals and the confidence level.





If we were to take random samples over and over, with the same sample size:

- Each time we would get a different sample statistic (point estimate) and a different confidence interval.
- About 95% of these confidence intervals would capture the true mean.
- About 5% would miss the true mean.

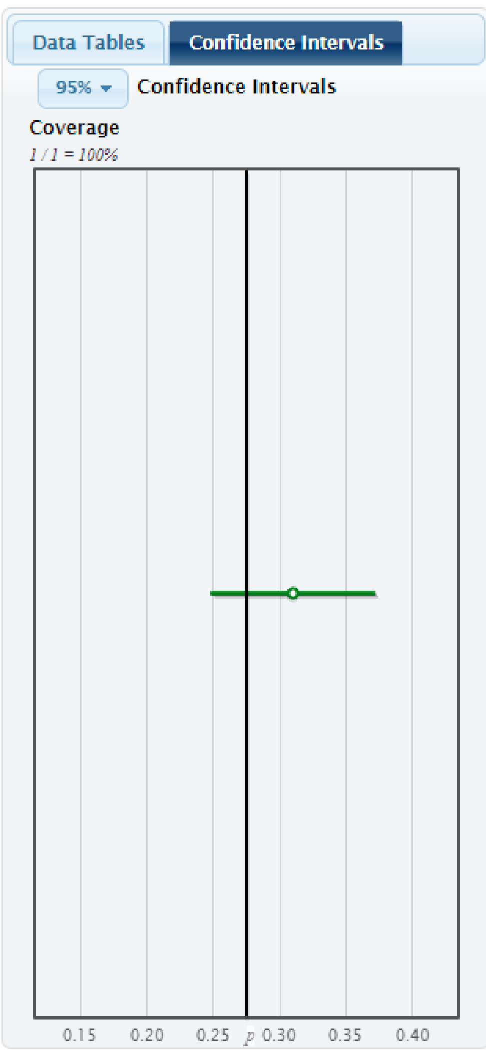
Example 3. Now let's continue the blood type example where 37% of the U.S. population has type O+ blood. Please go to the StatKey Sampling Distribution for a Proportion applet. Click on Edit proportion, change it to 0.37, and change the sample size to n=100. Generate 1 random sample of 100 people.

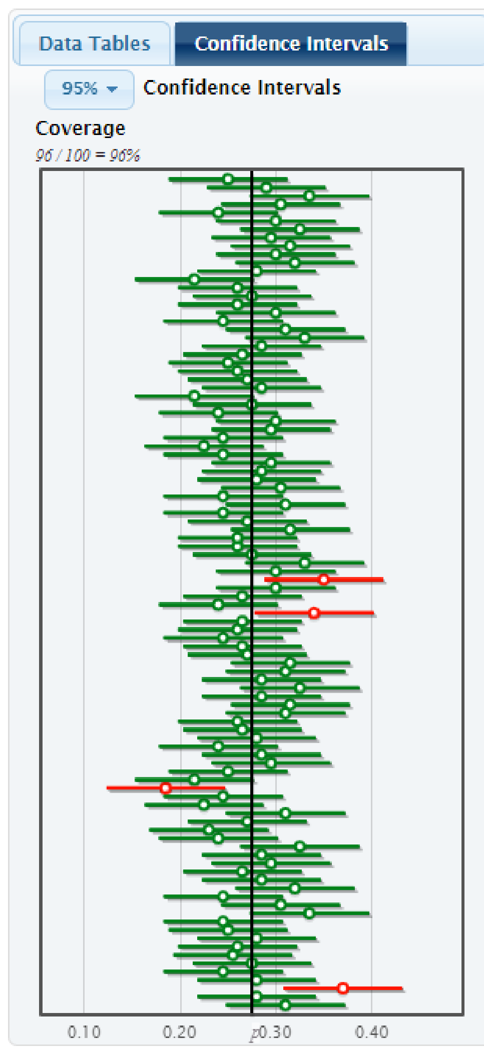
We're still pretending to know the population characteristics so we can learn about confidence intervals, but we're taking only one sample. Then we will take this information away.

Generate 1 sample of n=100 people and write down $\hat{p}$. When we simulated a sampling distribution for samples of 100 people before, we got a standard error of 0.048.

Using the formula above, calculate the 95% confidence interval. Does it contain the true proportion?

Generate 100 confidence intervals to help understand what the confidence level means.





Writing the Interpretation of Confidence Intervals

Let's say we calculated a 95% confidence interval for the true proportion of hiring managers who use social media to research job applicants to be (58%, 62%). Note: $\hat{p}$ must be 60% since it is right in the center of the interval. Which interpretations are correct?

Correct

- We are 95% confident that the interval from 58% to 62% captures the true proportion of hiring managers who use social media to research job applicants.

More Casual, But Fine

- We are 95% confident that 58% to 62% of hiring managers use social media to research job applicants.

Incorrect

- We are 95% confident that 58% to 62% of hiring managers in the sample use social media to research job applicants.
- There is a 95% probability that the true proportion is between 58% and 62%.
    - This makes it sound like the interval is fixed and the proportion is variable, but it's the other way around.

Example 4. A random sample of 50 college students were asked how many exclusive relationships they have been in so far. The approximate 95% confidence interval is given by: (2.7, 3.7). Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that…

a. the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

b. college students on average have been in between 2.7 and 3.7 exclusive relationships.

c. a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

d. 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Write 95% Confidence Intervals using the Margin of Error or Standard Error

$$\text{point estimate} \pm \text{margin of error}$$

$$\hat{p} \pm 2 \cdot \text{SE}$$

$$\bar{x} \pm 2 \cdot \text{SE}$$

Example 5. For each random sample, identify the point estimate, give the 95% confidence interval, and write the interpretation of the confidence interval.

a. A random sample of 80 cars had an average fuel efficiency of 28 miles per gallon. Provide a 95% confidence interval for the population mean fuel efficiency, with a margin of error of 3 miles per gallon.

b. A drug trial involved 300 patients, with 90 of them experiencing a positive outcome. Determine a 95% confidence interval for the proportion of patients who responded positively to the drug, with a margin of error of 0.09.

c. In an election poll, 320 out of 800 randomly selected respondents indicated their intention to vote for Candidate A. Provide a 95% confidence interval for the proportion of voters who intend to vote for Candidate A, given a standard error of 0.041.

d. The weights of 60 laptops chosen at random from a manufacturing facility were measured, with an average weight of 2.5 kg. Provide a 95% confidence interval for the population mean laptop weight, with a standard error of 0.12 kg. If the laptops are advertised as weighing 2.3 kg, is that a plausible value for the true mean?

Confidence Intervals with Bootstrapping

Theoretical vs. Simulation Methods for Inference

Theoretical calculations are based on a model such as the normal model or t-distribution. These methods were developed in the 1930's, before computers were available! We will study some of these models, but now that computing power has increased, simulation methods are becoming more commonly used.

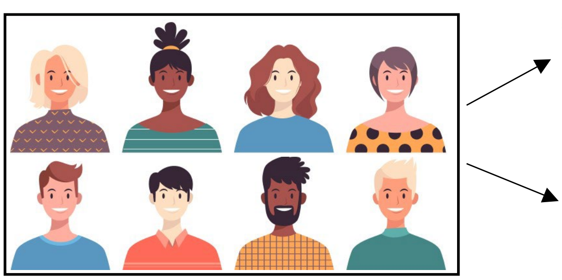
Simulation methods use a computer to run many trials quickly, as we have seen already. The beauty of simulation is you don't have to have a theoretical model. As long as you can simulate a real-world process, you can get results.

Bootstrapping was developed in 1979 by Stanford statistician Bradley Efron. Now that computing power has increased, simulation methods such as bootstrapping are being used more frequently. Bootstrapping comes from the saying, "picking yourself up by your bootstraps." In statistics it means approximating a sampling distribution and standard error with just one random sample.

Bootstrap Samples

In real life, we take one sample from our population. If we have more resources to do another sample, it would be better to just take a larger sample. We don't have the information for the full population, but, if our sample is representative, we can think of the population as many copies of our sample. If we sample many times from that we can estimate our sample-to-sample variation.

Bootstrapping is the idea that a population is many copies of that sample.

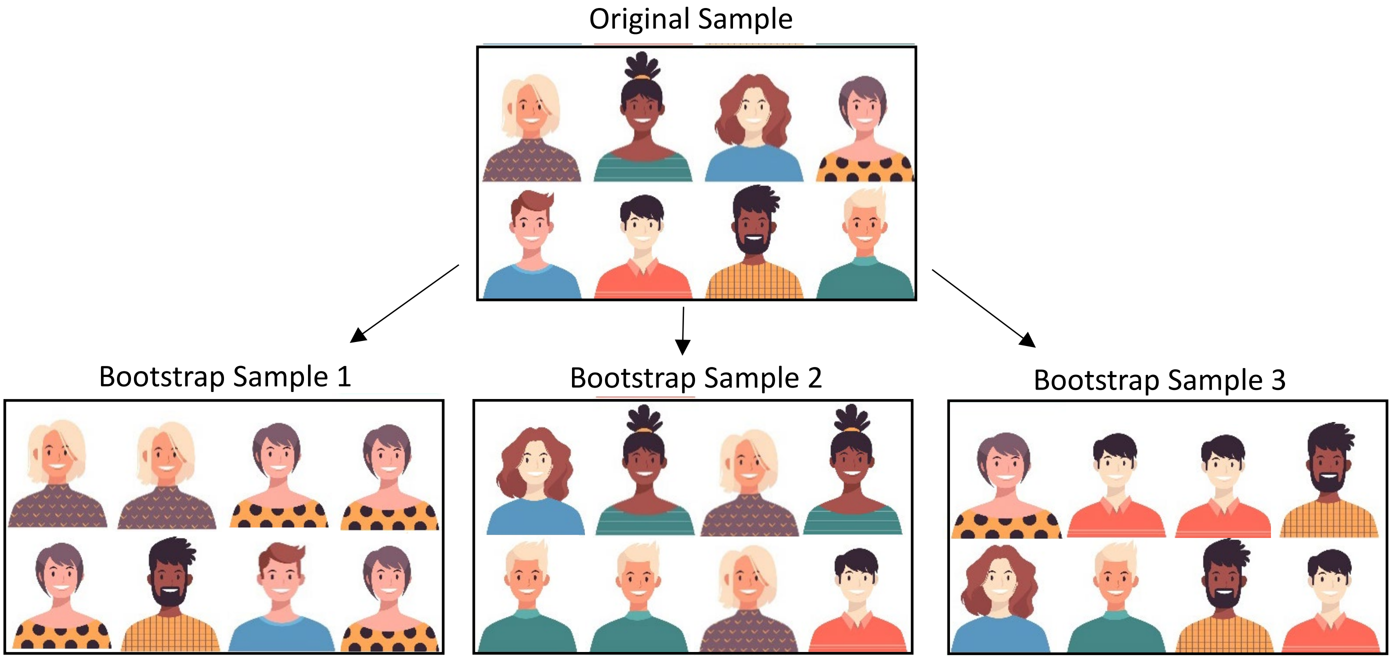


Original Sample[1]

[1] Image by pikisuperstar on Freepik

Then we take many samples from that "population." Another way to think about it is called sampling with replacement. That means each time we choose a subject, they go back in the hat and can be chosen again.

Just like with a sampling distribution, we use the distribution of random samples of the same size to estimate the sample-to-sample variation.



Generating a Bootstrap Distribution

Similarly to how we simulated sampling distributions, we will take repeated samples from one sample using specific applets to form a distribution. It has been shown that the sample-to-sample variation of the bootstrap distribution is very close to that of the sampling distribution.

How Can This Work?

If you are skeptical, I was, too. First, I'll show you how to use bootstrapping. Then we'll do an experiment to compare the methods and there is an optional video you can watch that does many more simulations. It shows that bootstrap distributions are very similar to their corresponding sampling distributions.

Bootstrap 95% Confidence Intervals Using Standard Error

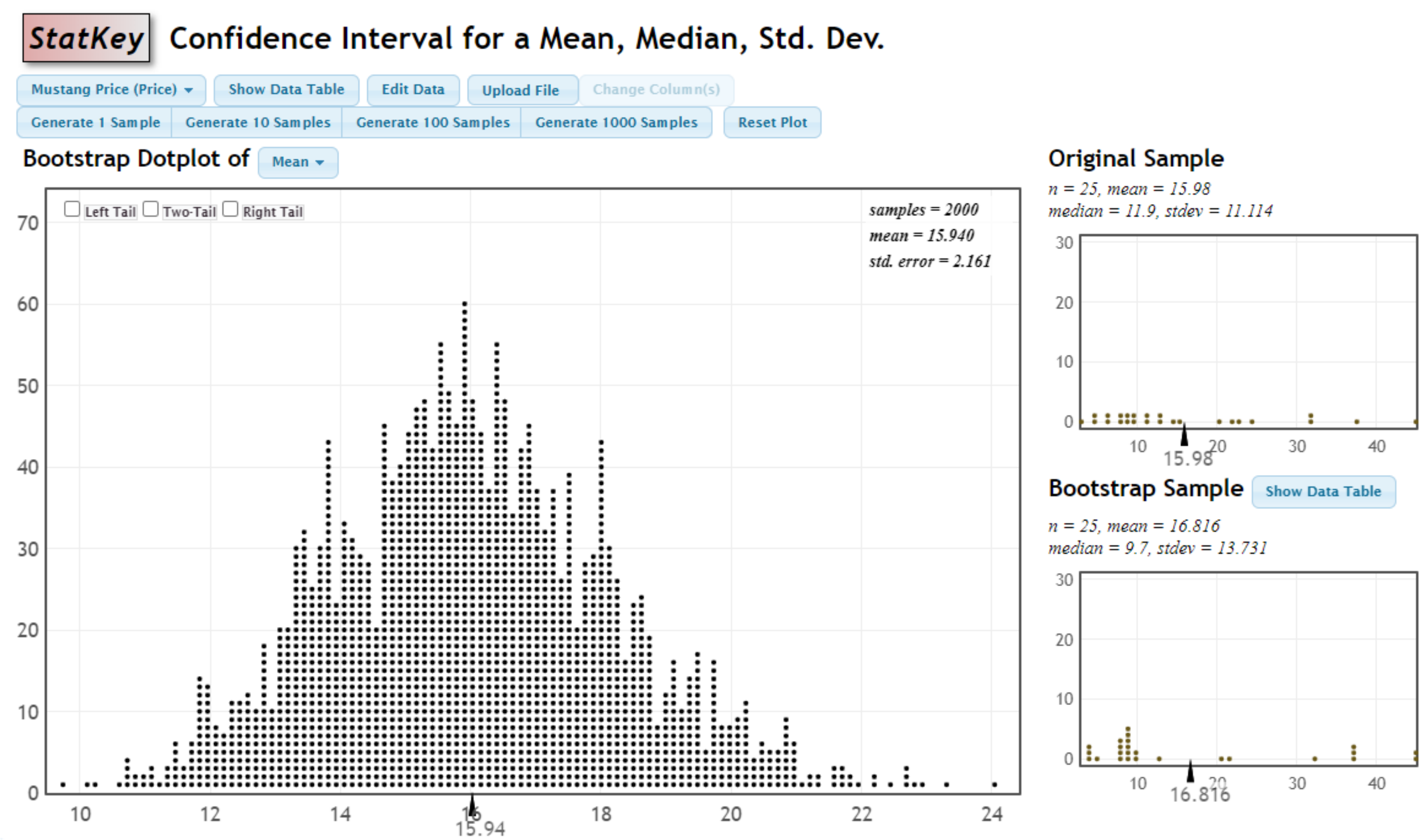
$$\text{point estimate} \pm \text{margin of error}$$
$$\hat{p} \pm 2 \cdot \text{SE}$$
$$\bar{x} \pm 2 \cdot \text{SE}$$

We use the point estimate from our original sample, and the estimate of the Standard Error from our bootstrap distribution. The sample-to-sample variation in a bootstrap distribution is close to the sample-to-sample variation in the population.

Example 1. Find a 95% bootstrap confidence interval for the mean price of a Ford Mustang.

Using StatKey, choose the Bootstrap Confidence Interval for a Single Mean applet. Select the Mustang Price data set. View the original sample and create 2000 resamples from the original sample.

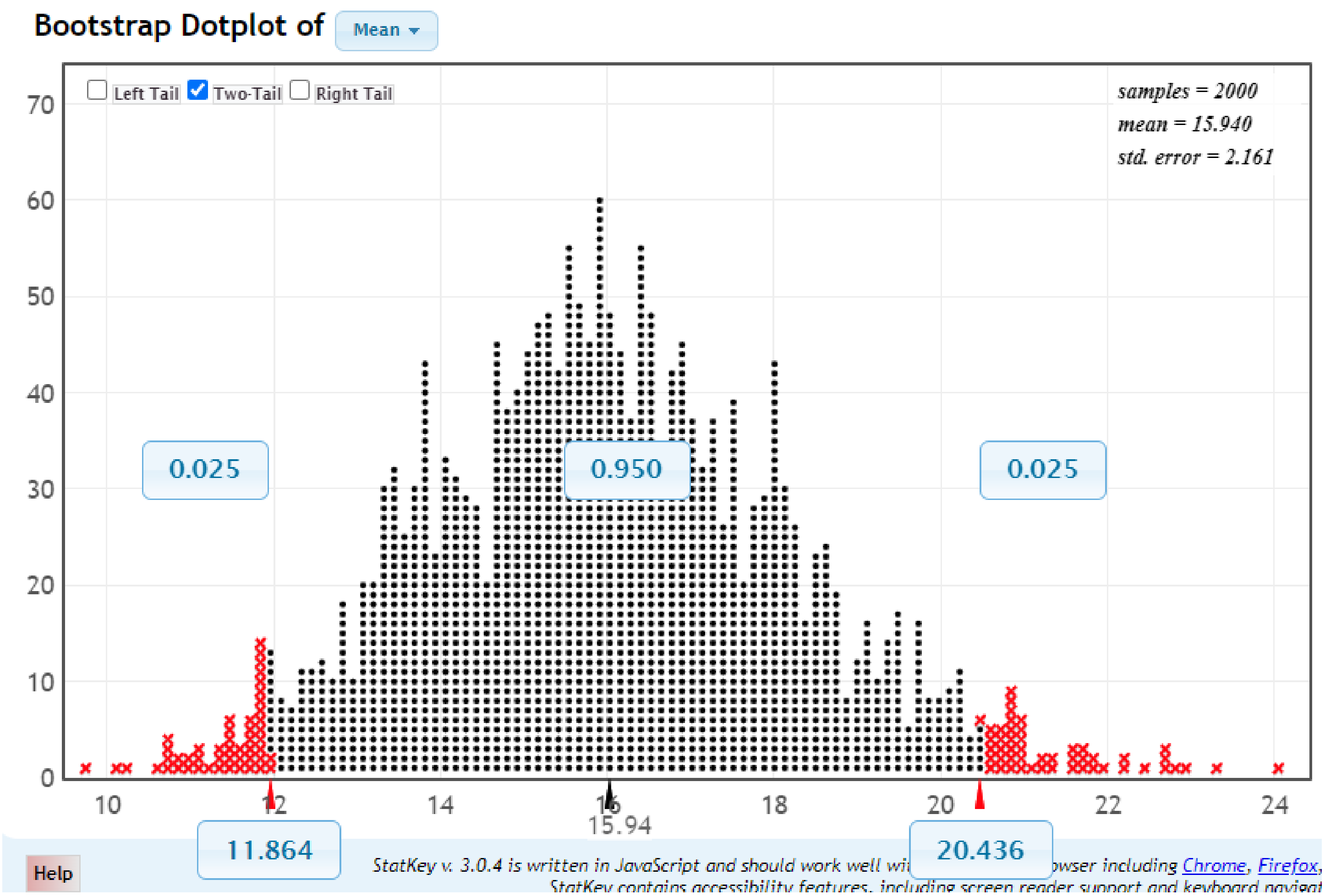
a. Find the sample statistic and the estimated SE from the bootstrap distribution. (Note if we do this again, we will get a different result.)



b. Calculate the 95% bootstrap confidence interval for the mean price of Ford Mustangs using the SE, with units.

Bootstrap Confidence Intervals Using Percentiles

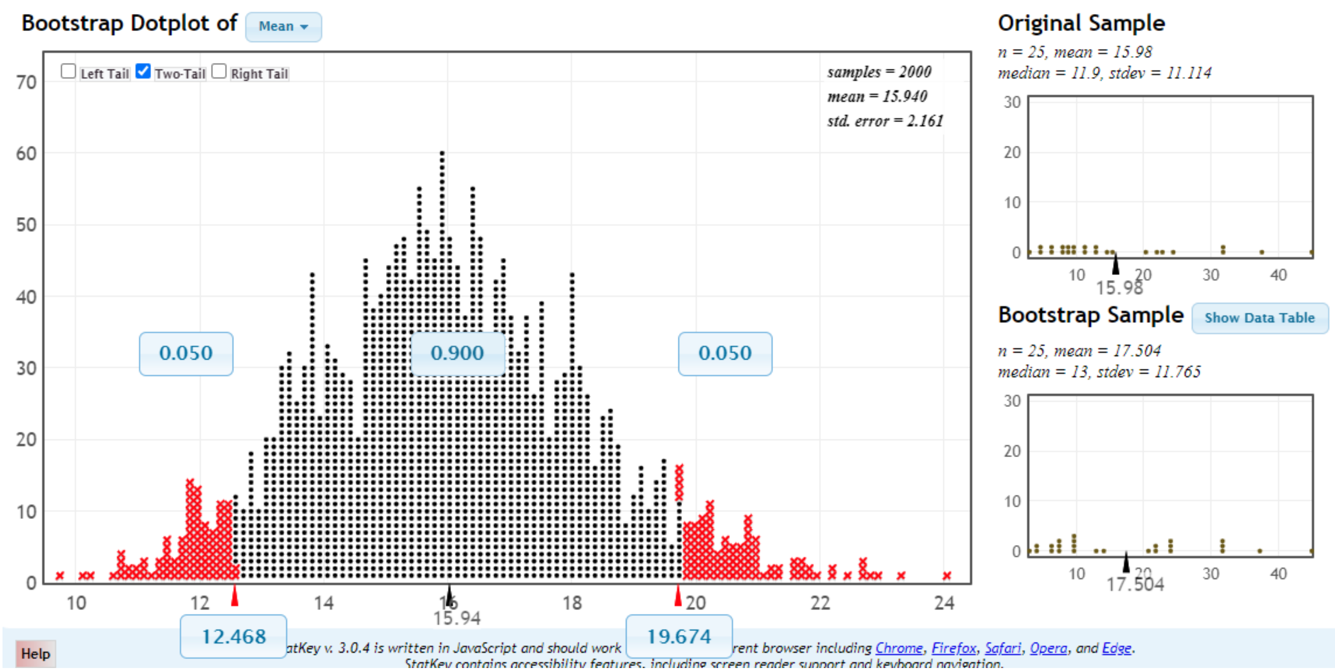
c. Use the Two-Tail feature to get the 95% bootstrap confidence interval. Are these the same? Why or why not?



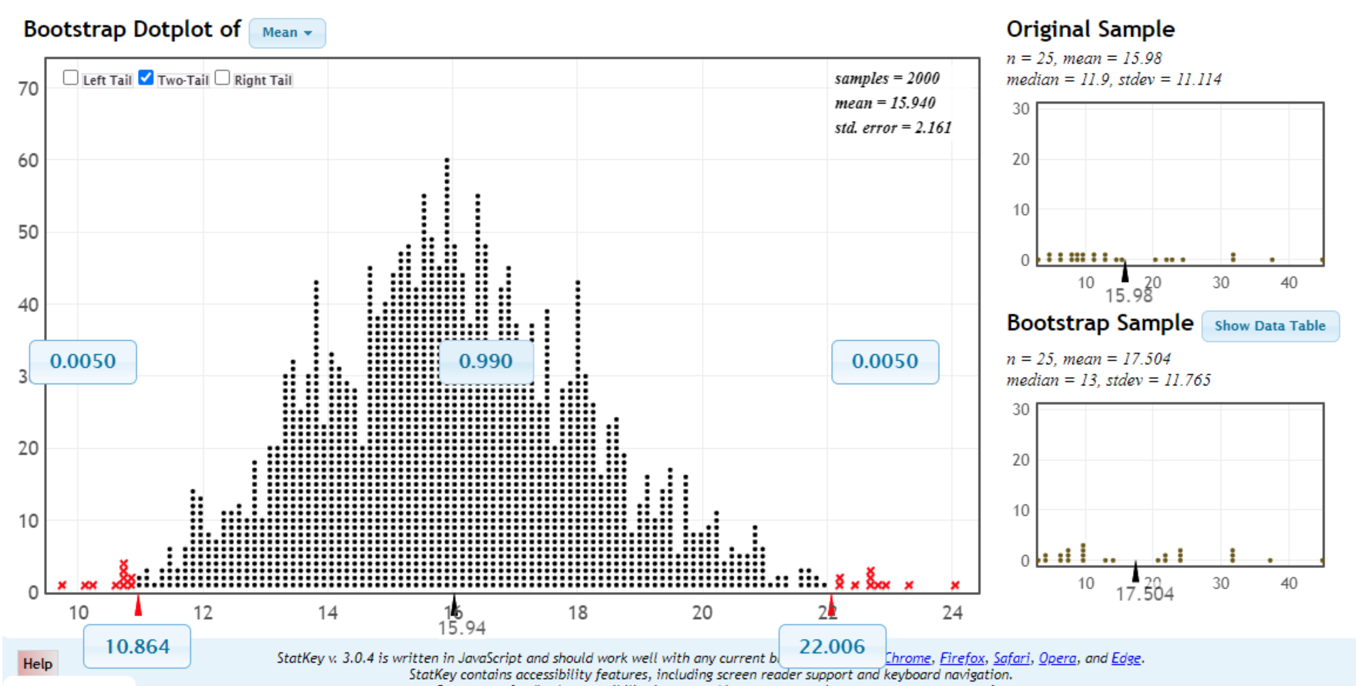
What if we want a confidence level other than 95%? We can get any confidence level we want using our bootstrap simulation.

Make sure Two-Tail is checked and set the middle area box to the desired confidence level in decimal form. Then read the values off the horizontal scale in the boxes.

d. Find a 90% confidence interval using percentiles.



e. Find a 99% confidence interval.



How do you show your work? Use a snipping tool or take a screenshot when guided to.

- Windows – snipping tool, select a rectangle.
- Mac – Command, Shift, 4, then select a rectangle.

Example 2. Let's find bootstrap confidence intervals for a proportion. Open the StatKey Bootstrap Confidence Interval for a Single Proportion. Let's use the Reese's Pieces data set. We want to estimate the proportion of all Reese's Pieces that are orange. They come in three colors, orange, yellow and brown. Do the colors have equal proportions?

The original sample says 72/150 are orange. Generate 2000 bootstrap samples and use your distribution to find each of the following confidence intervals.

a. A 95% confidence interval for the proportion of orange using the bootstrap SE.

b. A 95% confidence interval for the proportion of orange using percentiles.

c. A 90% confidence interval for the proportion of orange using percentiles.

d. An 85% confidence interval for the proportion of orange using percentiles.

e. What do you notice about the width of the confidence interval and the confidence level?

f. Write the interpretation for the 85% confidence interval.

What's the difference between a sampling distribution and a bootstrap distribution?

Sampling distribution – resampling from the population

Bootstrap distribution – resampling from a sample

Effect of Sample Size on Confidence Intervals

With Reese's Pieces example we saw that a higher confidence level led to a larger interval. If we leave the confidence level the same, what do you think will happen if we change the sample size? Let's continue with Reese's Pieces example.

a. The original sample was 72/150 orange Reese's Pieces. Generate 2000 bootstrap samples and use your distribution to find a 95% confidence interval for the proportion of orange using percentiles. Also write down the standard error of the bootstrap distribution.

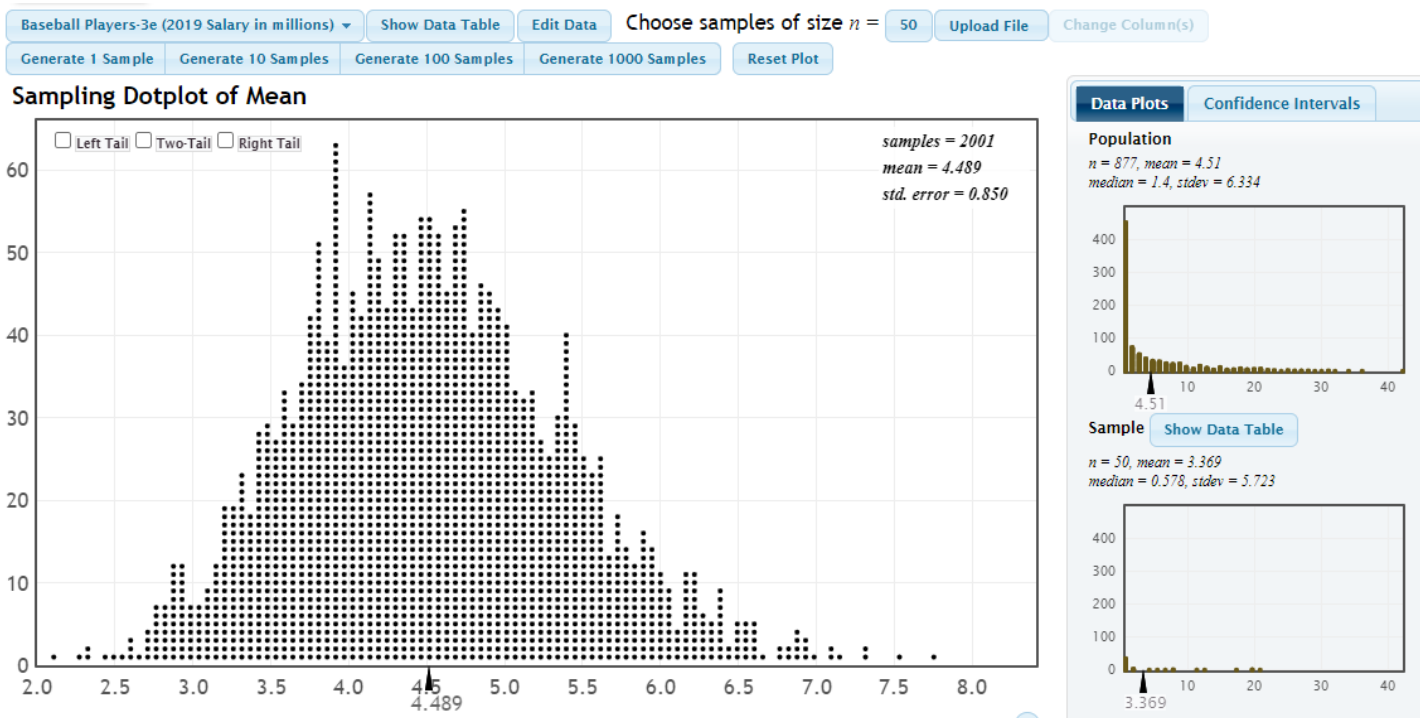
b. Now let's say we have the same proportion, 0.48, but the sample was 48/100 orange. Would you expect the 95% confidence interval to be wider or narrower? Go ahead and find it with a bootstrap simulation and make a note of the standard error.

c. Now let's say our sample was 720/1500 orange. What do you think will happen? Find the 95% bootstrap confidence interval using percentiles and write down the standard error.

Let's summarize our findings here:

Comparing a Sampling Distribution with a Bootstrap Distribution

We have been using the baseball salaries as a population, so here's a sampling distribution for the mean with 2000 random samples from the population.



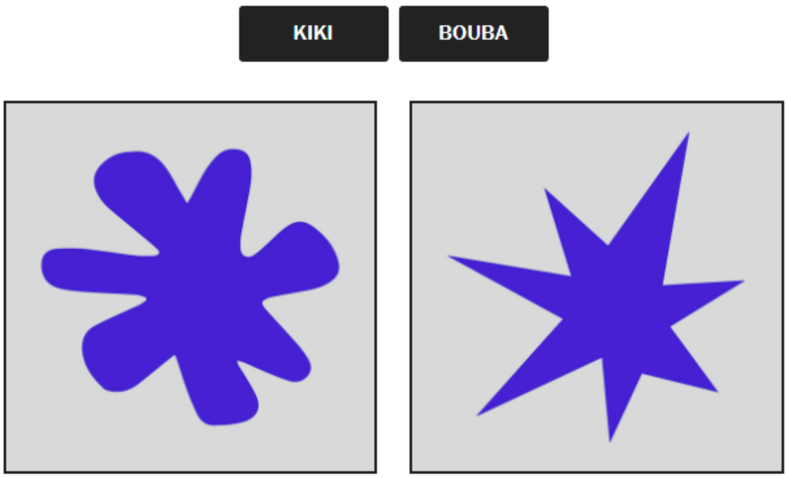
The SE for the sampling distribution is 0.850 million.

Now, I'm going to take one sample of size n=50 from this population and use that for a bootstrap distribution. I imported that sample into the StatKey Bootstrap app.

I found the SE for the bootstrap distribution. Let's repeat this a few times:

Hypothesis Test Example with Randomization Distribution

Let's return to our Kiki/Bouba test. We did a lot of the steps and now we'll learn how to write out the hypotheses and make it a formal test.



Null hypothesis: There is no association between these words and shapes. (It's a 50-50 guess).

Alternate hypothesis: There is an association between the rounded shape and the word bouba. (It's more likely to match the words and shapes in one way than the other)

In symbols, we write them together like this:

$H_0: p = 0.50$

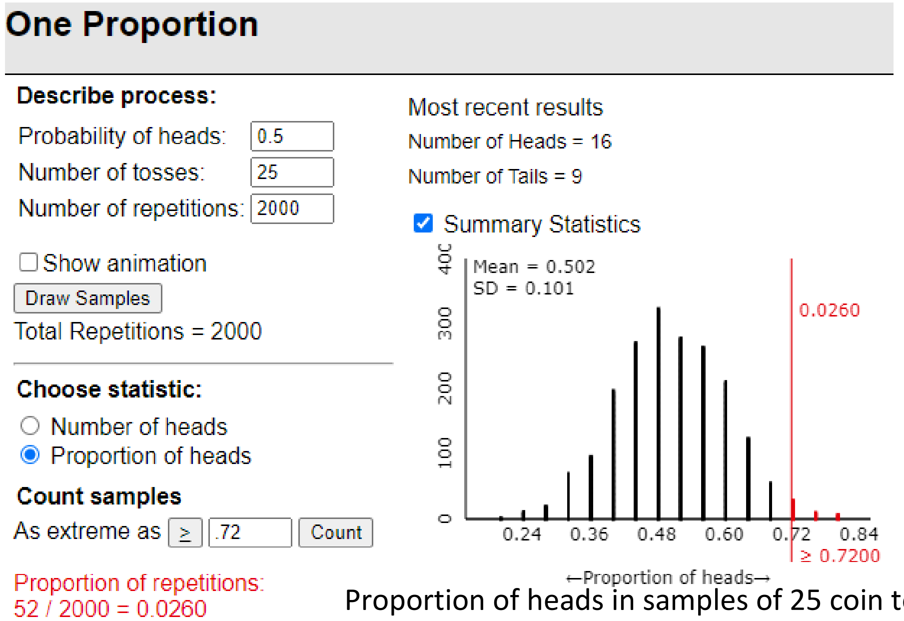
$H_A: p > 0.50$

We will build a randomization distribution for the null hypothesis and our sample size of n=25. Let's try this with both the Rossman/Chance One Proportion Applet and the StatKey Randomization Test for a Proportion Applet.

I counted from a previous class how many students matched the rounded shape with bouba, and that's our observed statistic.

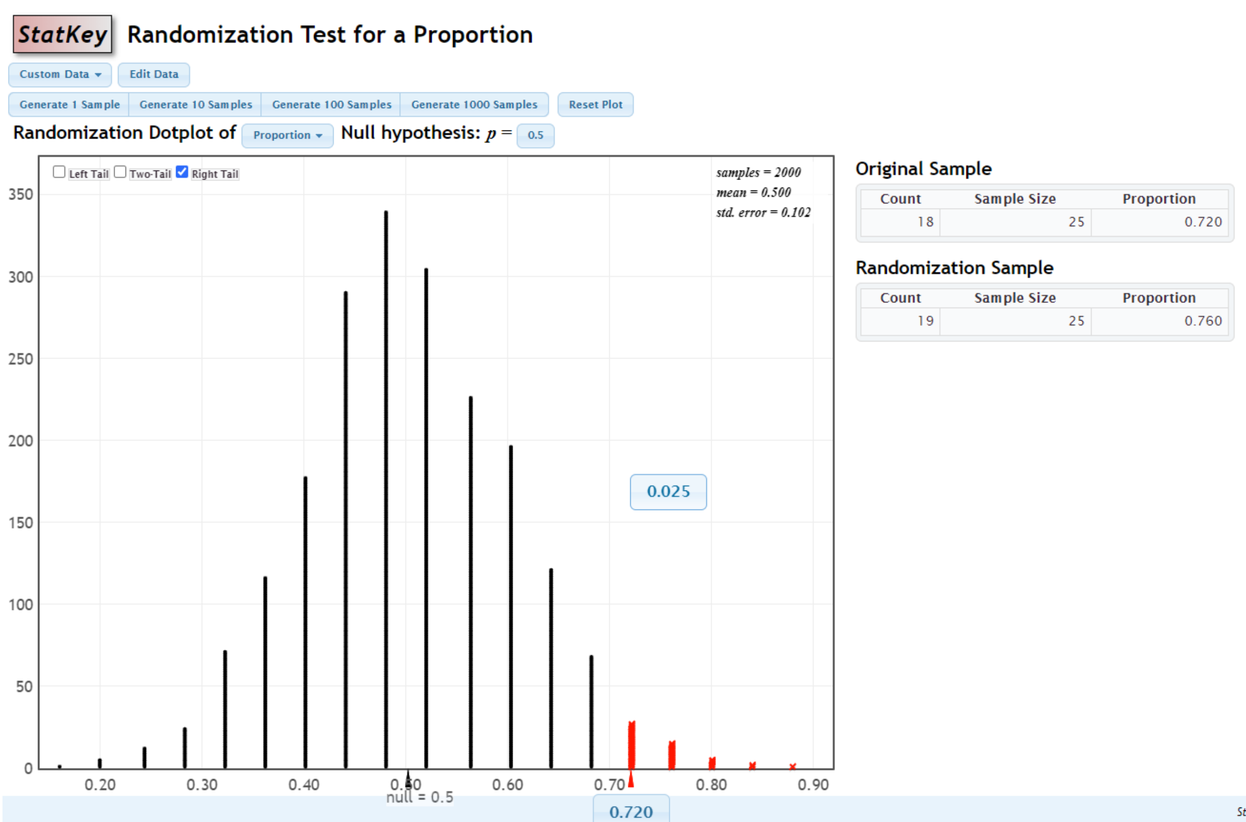
$\hat{p} = \frac{18}{25} = 0.72.$

We used the Rossman/Chance applet and got a probability of 0.026 of getting this value or toward the tail, this is called a p-value



Proportion of heads in samples of 25 coin tosses

I'll also show you on the StatKey Randomization Test for a Proportion. We click on right-tail because this is a right-tail test. Then enter our observed statistic along the horizontal axis. This shows us a p-value of 0.025.



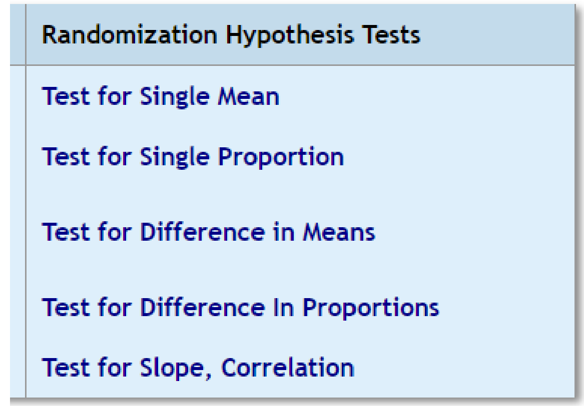
Proportion of heads in samples of 25 coin tosses

Conclusion: It is unlikely to get a sample proportion of 0.72 under the null hypothesis due to random chance, so we reject the null hypothesis. Our conclusion is we have evidence for the claim that there is a significant association between the word bouba and the rounded shape.

Setting Up Hypothesis Tests of One Parameter with Randomization Distributions

We will be testing hypotheses in 5 scenarios:

StatKey Applets



One parameter

- Process probabilities (proportion)
- Proportions
- Means

Two parameters

- Difference of two proportions (difference of two groups, not matched pairs)
- Difference of two means (difference of two groups, not matched pairs)

Writing the Hypotheses

The null hypothesis, H_0, pronounced, H-null or H-naught, is the accepted value, status quo, or no effect. "The null is dull."

H_0: parameter = null value

The alternate hypothesis, H_A, pronounced, H-A, is the claim that we are looking to show evidence of. It can also be written as H_1.

Options for the alternate hypothesis:

H_A: parameter > null value (right-tail test)

H_A: parameter < null value (left-tail test)

H_A: parameter $\neq$ null value (two-tail test)

Steps for a Hypothesis Test with a Randomization Distribution

a. Write the null and alternate hypotheses.

b. Simulate the null hypothesis with a randomization distribution for the given sample size.

c. Write your observed statistic, find the approximate p-value and insert an image of your randomization distribution with p-value.

d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

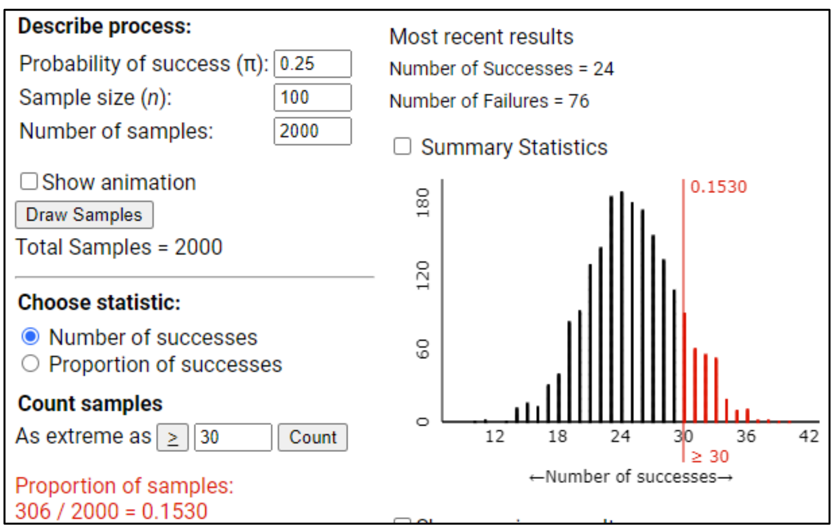
Hypothesis Test Example 1 – Flat Tire

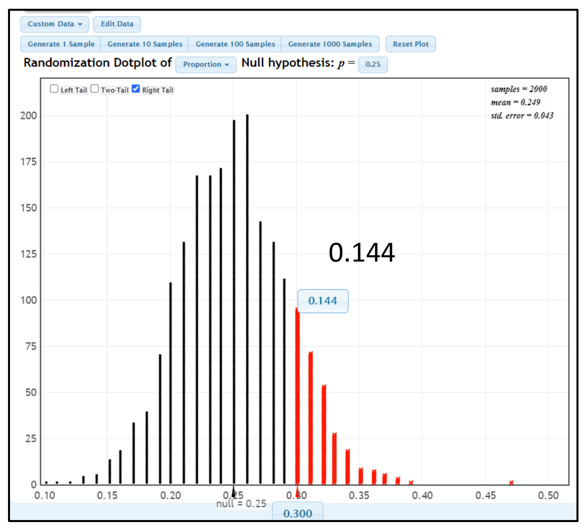
Example 1. A legendary story on college campuses concerns two students who miss a chemistry exam because of excessive partying but blame their absence on a flat tire. The professor allowed them to take a make-up exam, and he sent them to separate rooms to take it. The first question, worth 5 points, was quite easy. The second question, worth 95 points asked, "Which tire was it?" It turns out there is a tire most commonly chosen which one do you think it is? In a survey of 100 students, 30 said the right front tire. Test the claim that it is more common to choose the right front tire at the 5% significance level.

a. What is the parameter of interest and what is the direction of the test? Use this to write the hypotheses.

b. Using an appropriate randomization applet, enter the observed statistic and simulate the randomization distribution for the sample size used in the study.

c. Use the randomization distribution to find an approximate p-value and insert an image.





Random samples of 100 with p=0.25

d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

P-value, Significance Level and Conclusion

We set up our randomization distribution assuming the null hypothesis is true, with a given sample size. This is the accepted value, status quo or no effect. The p-value shows how strong our evidence is under this assumption.

What is a p-value?

The p-value is the probability of getting a result at least as extreme as the observed result (sample statistic), given that the null is true. The smaller the p-value, the stronger the evidence against the null hypothesis. We compare the p-value with the significance level.

Significance Level, alpha, α

The most common significance level is 0.05. If there is less than a 5% chance of getting the observed result, then that is statistically significant. For medical tests we may use $\alpha = 0.01$. In the social sciences we may use $\alpha = 0.10$. Unless otherwise stated, use $\alpha = 0.05$.

Stating your Conclusion

If the **p-value** ≥ **α,** we fail to reject the null hypothesis. There is not enough evidence to suggest that the alternate hypothesis is true. (The result is fairly likely under the null, so it is likely due to chance or random variation, and the result is not statistically significant.)

If the **p-value** < **α,** we reject the null hypothesis. There is evidence to suggest that the alternate hypothesis is true. (It is unlikely that the result would occur under the null hypothesis, so it is unlikely due to chance or random variation alone. This is evidence for the claim, and we say the result is statistically significant.)

Why do we Fail to Reject?

You might be wondering why we don't "accept" the null. We can only fail to reject it. Think about an example using the court system. A defendant is innocent until proven guilty. The burden of proof lies with the alternative hypothesis.

H_0: The defendant is innocent.

H_A: The defendant is guilty.

If there is sufficient evidence, then the defendant may be proven guilty (reject the null hypothesis). Otherwise, we fail to reject the null, and they are proven not guilty. It is impossible to prove that someone is innocent, even though they very well may be.

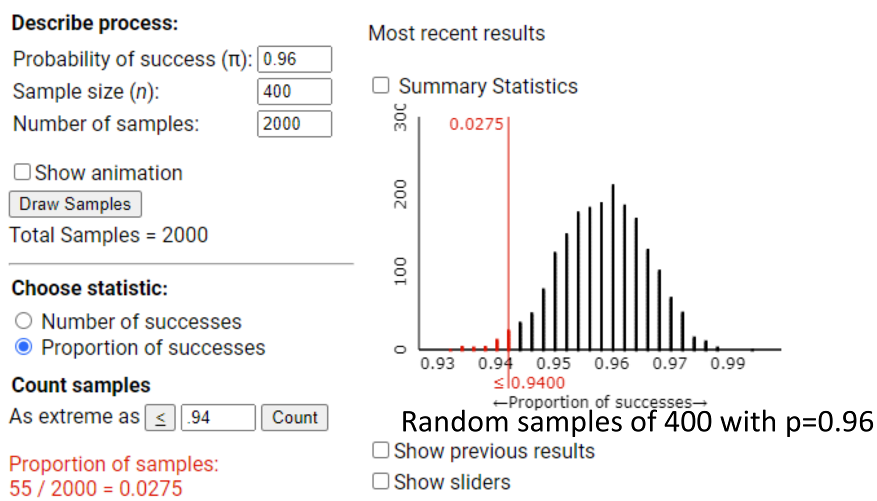
Hypothesis Test Example 2 – Smartphone Training

Example 2. A company develops what it hopes will be better instructions for its customers to set up their smartphones. The goal is to have 96% of customers succeed. The company tests the new system on 400 people, of whom 376 were successful. Is this strong evidence that the new system fails to meet the company goal?

a. What is the parameter of interest and what is the direction of the test? Use this to write the hypotheses.

b. Using an appropriate randomization applet, enter the observed statistic and simulate the randomization distribution for the sample size used in the study.

c. Use the randomization distribution to find an approximate p-value and insert an image.

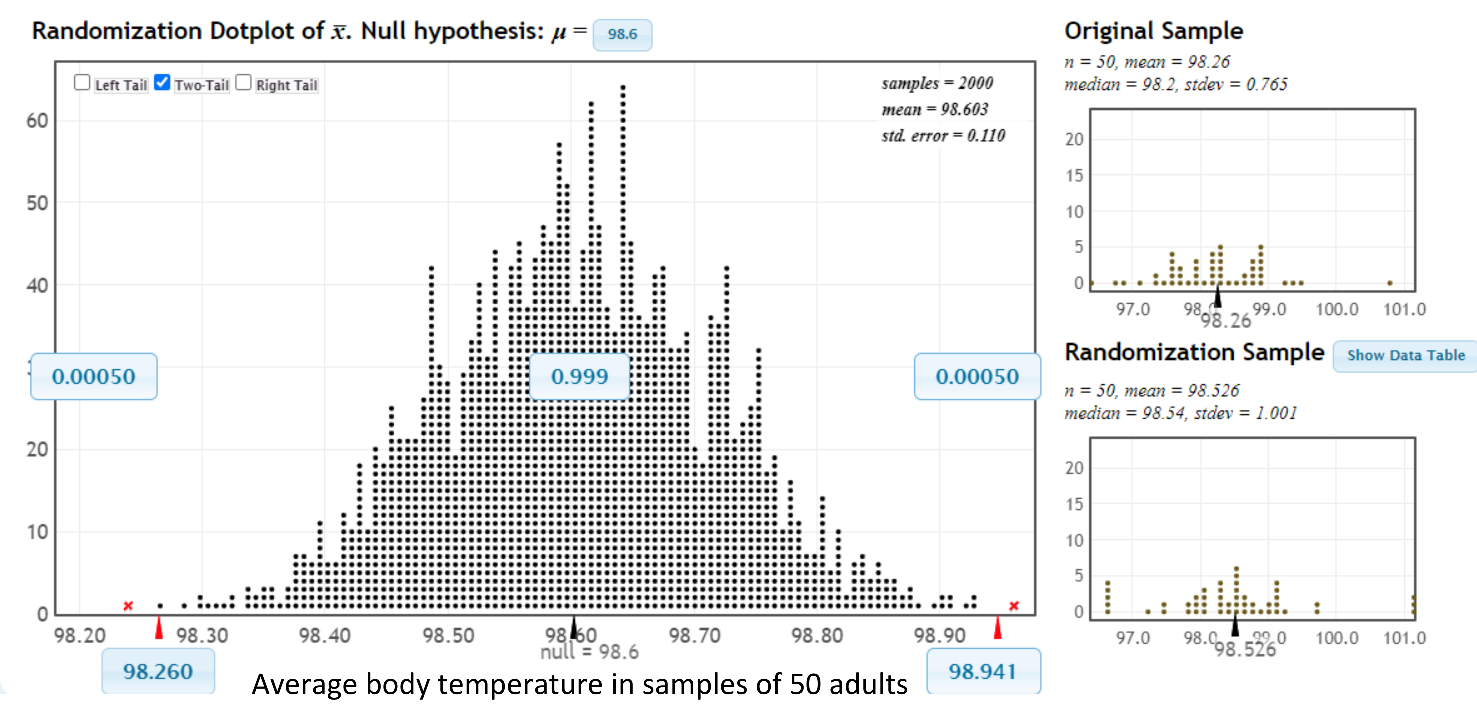


d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

Hypothesis Test Example 3 – Body Temperature

Example 3. The regular body temperature for healthy humans is said to be 98.6 degrees Fahrenheit. Is this really true, or has it changed? Allen Shoemaker presented some data derived from a study of healthy adults where the sample mean of 50 adults was 98.26°F. Do these data provide significant evidence at a 5% level that the average body temperature is really different from the standard 98.6°F?



Average body temperature in samples of 50 adults with mean=98.6 degrees Fahrenheit

Setting Up Hypothesis Tests for Two Parameters by Simulation

We will be testing hypotheses in 5 scenarios:

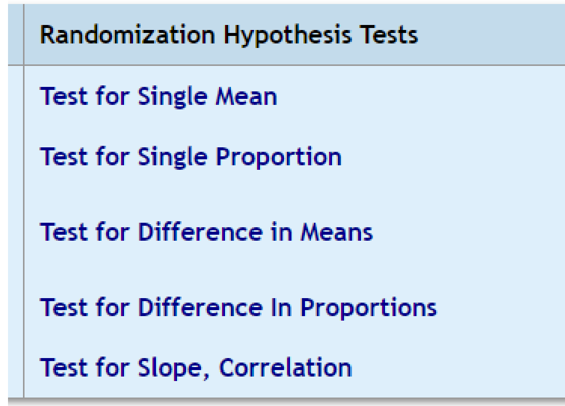
One parameter

- Process probabilities (proportion)
- Proportions
- Means

Two parameters

- Difference of two proportions (difference of two groups, not matched pairs)
- Difference of two means (difference of two groups, not matched pairs)

StatKey Applets



Hypothesis Test for Two Parameters – Beer and Mosquitos

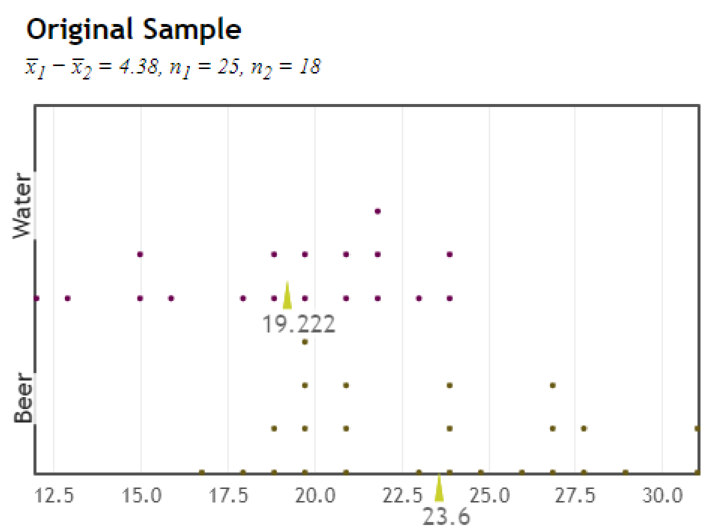
Example 4. Does drinking beer attract mosquitos? A study done in Burkino Faso, Africa[1], about the spread of malaria investigated the connection between beer consumption and mosquito attraction. In the experiment, 25 volunteers consumed a liter of beer while 18 volunteers consumed a liter of water. The volunteers were assigned to the two groups randomly. Mosquitoes were released and caught in traps as they approached the volunteers. For each group, the number of mosquitos caught per person is listed below. Test the claim that the beer drinkers attracted more mosquitoes than the water group after drinking.

Beer Drinkers: 27 20 21 26 27 31 24 19 23 24 28 19 24 29 20 17 31 20 25 28 21 27 21 18 20 mosquitos

Water Drinkers: 21 22 15 12 21 16 19 15 24 19 23 13 22 20 24 18 20 22 mosquitos

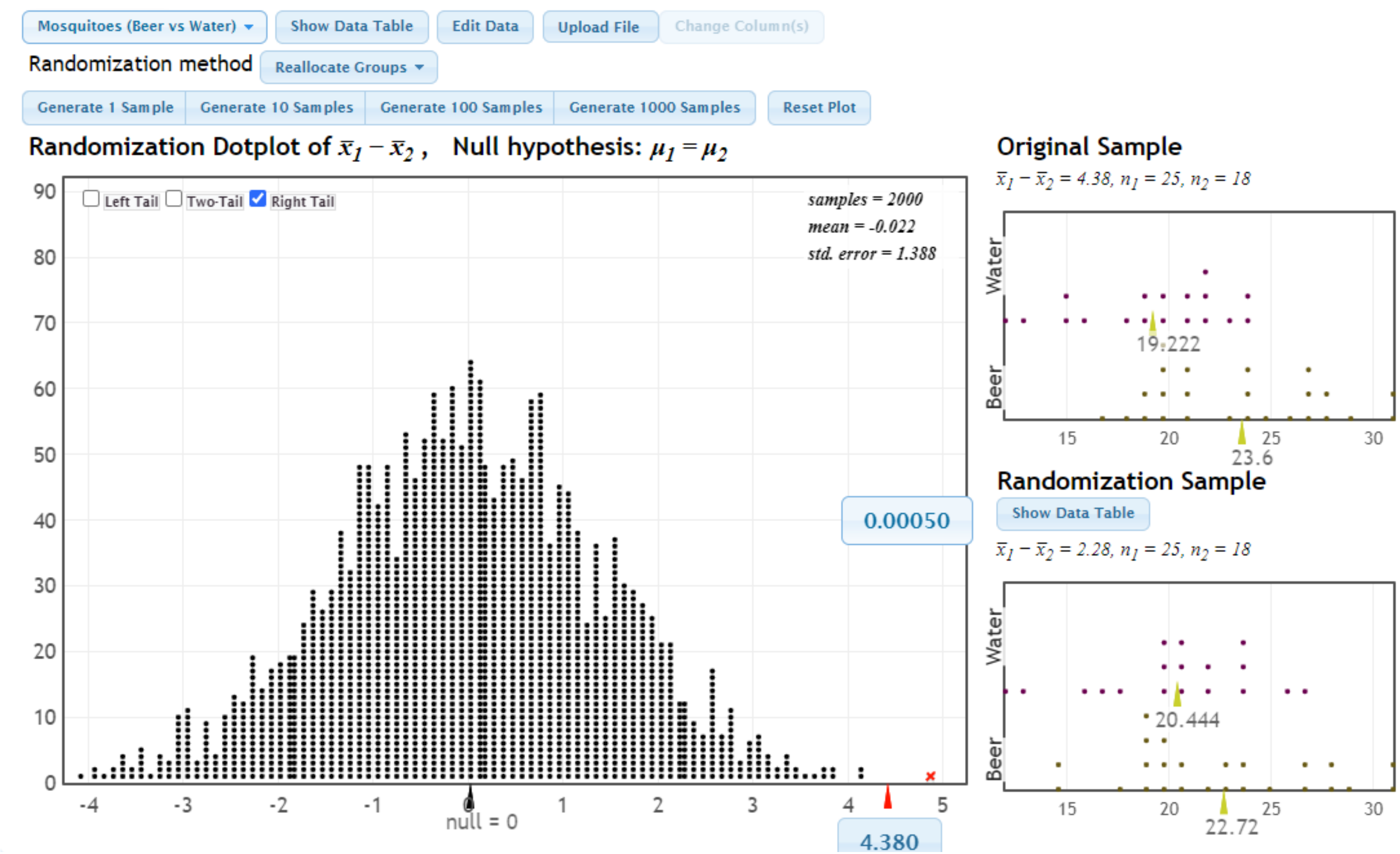
a. What is the parameter of interest and what is the direction of the test? Use this to write the hypotheses.

b. Using an appropriate randomization test applet, select the data set and simulate the randomization distribution.



[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832015/

c. Use the randomization distribution to find an approximate p-value and insert an image.



d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

Hypothesis Test for Two Parameters – Treatments for Cocaine Addiction

Example 5. Recall the cocaine treatment experiment we looked at earlier. We did an independence test and noticed a difference in the success rates of the different treatment groups. We can also look at these data using a hypothesis test.

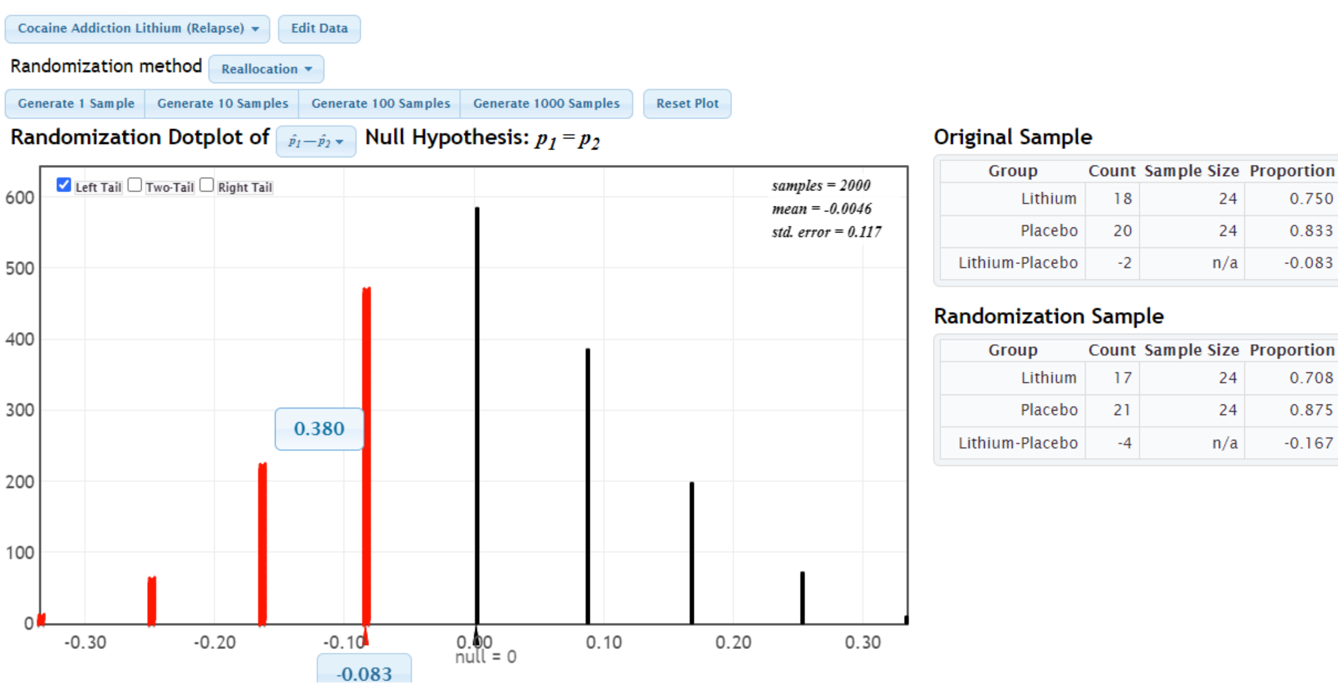
	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
Total	48	24	72

In a hypothesis test we will test two proportions at a time. Test the hypothesis that taking the drug lithium reduces the proportion of participants that have a relapse compared with the placebo.

a. What is the parameter of interest and what is the direction of the test? Use this to write the hypotheses.

b. Using an appropriate randomization applet, select the data set or enter the observed statistic(s) and simulate the randomization distribution for the sample size used in the study.

c. Use the randomization distribution to find an approximate p-value and insert an image.



d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

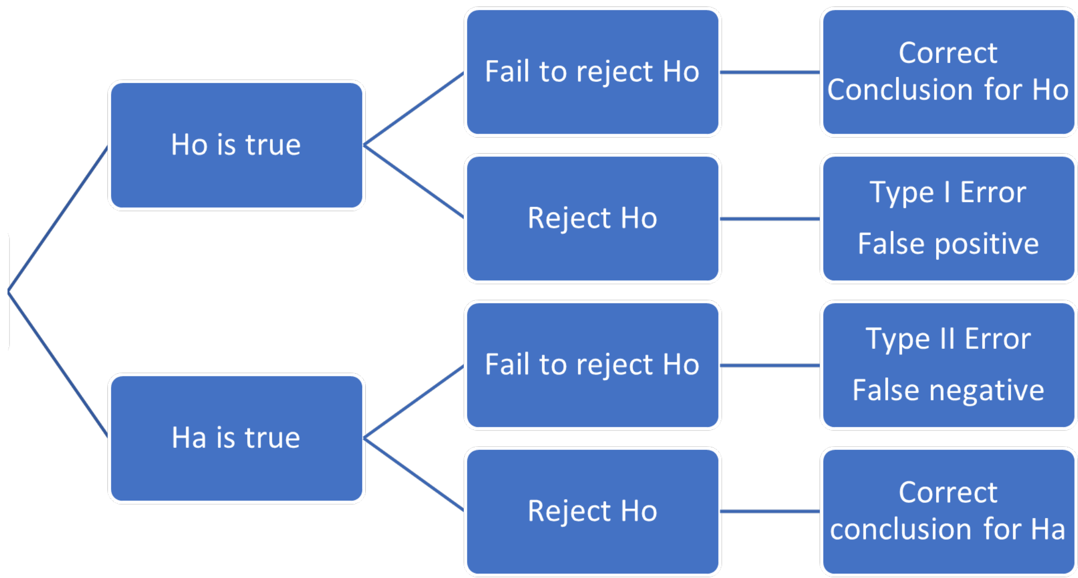
Decision Errors in Hypothesis Testing

A **Type I Error** is when we conclude for the alternate hypothesis when it's not actually true. The probability of doing this in the long run is alpha or the significance level. This is known as a false positive.

If making a Type I error is more dangerous or costly, we can make alpha smaller to reduce the chance we reject the null hypothesis.

A **Type II Error** is when we conclude for the null when it's not actually true. This is called a false negative.

If making a Type II error is more dangerous or costly, we can choose a higher significance level to be cautious about failing to reject the null hypothesis. There is more about power and sample size that you may learn in a future course.



Example 1. A new drug is being tested to see if it has a more significant effect than the current one.

a. Write out the Hypotheses in words:

b. What is a Type 1 Error in this case and what are the consequences?

c. Wha is a Type II Error in this case and what are the consequences?

Since these errors can occur, the replication of studies is important. It is good practice to make sure a study is conducted in a way that is transparent and reproduceable by other researchers.

Connecting the Types of Simulated Distributions

A **sampling distribution** is a distribution of sample statistics from random samples from a population and is centered at the true value of the population parameter.

A **bootstrap distribution** is a distribution of sample statistics from re-samples of an original sample. This is to simulate a sampling distribution, but it will be centered at the original sample statistic.

A **randomization distribution** is a distribution of sample statistics assuming the null hypothesis is true. It will be centered at the value of the null parameter.

The Problem with p-values

P-value methods have been criticized and even banned from some journals. The significance level of 0.05 is arbitrary and makes a high-stakes binary decision for publication of studies. There are many ways to influence a study to get significant results. The p-value also measures evidence against the null hypothesis. It doesn't measure evidence for the alternative.

Instead of a binary, reject or fail to reject, some researchers suggest a scale:

- p-value > 0.10 — not much evidence against the null; the null is plausible
- $0.05 <$ p-value ≤ 0.10 — moderate evidence against the null hypothesis
- $0.01 <$ p-value ≤ 0.05 — strong evidence against the null hypothesis
- p-value ≤ 0.01 — very strong evidence against the null hypothesis

There have been many letters and articles written about the problem with p-values. Some are linked in D2L.

Relating Confidence Intervals and Hypothesis Testing

Comparison of a Hypothesis Test with a Confidence Interval

Instead of or in addition to doing a hypothesis test we could take our sample statistic and form a 95% confidence interval. This would be similar to doing a two-tailed test at a 5% significance level.

The main difference in the simulation methods is a confidence interval is calculated with a bootstrap distribution while a hypothesis test is performed with a randomization simulation of the null hypothesis.

Example 2. A national survey of 50,347 households in December 2011 found that 30.4% of American households have a pet cat (2012 U.S. Pet Ownership and Demographics Sourcebook, American Veterinary Medical Association).

a. Conduct a significance test to determine whether the sample data provide evidence that the population proportion of households who have a pet cat differs from one-third. State the hypotheses, report the p-value, and draw a conclusion in the context of this study.

b. Produce a 95% confidence interval for the population proportion of households who have a pet cat. Interpret this interval.

c. Are the test decision and confidence interval consistent with each other? Explain how you can tell.

d. Do the sample data provide very strong evidence that the population proportion of households who own a pet cat is different from one-third? Explain whether the p-value or the confidence interval helps you decide.

e. Do the sample data provide strong evidence that the population proportion of households who own a cat is very different than one-third? Explain whether the p-value or the confidence interval helps you decide.

If the null value falls within the confidence interval, that supports the null hypothesis and we fail to reject H_0.

If the null value falls outside the confidence interval, that is evidence against the null hypothesis and we reject H_0.

Practical vs. Statistical Significance

Statistical Significance vs. Practical Significance

If a very large sample size is used, then very small differences can be statistically significant. The difference may not be meaningful. In later courses, you may learn how to choose the sample size so that the statistical significance reflects a meaningful or practical difference.

Size of the Effect

Confidence intervals should accompany significance tests to estimate the size of an effect or difference.

More Practice with Hypothesis Testing and Stapplet

Example 3. An experiment compared the ability of three groups of participants to remember briefly presented chess positions. The data are shown below. The numbers represent the average number of pieces correctly remembered from three chess positions. You can use either StatKey or Stapplet One Quantitative Variable, Multiple Groups. I find Stapplet easier when I need to enter a data set that is not in StatKey.

Setup and do a randomization hypothesis test using this data to test whether the beginners' ability to remember the chess positions are significantly different than the non-players.

Non-players (pieces)	Beginners (pieces)	Tournament Players (pieces)
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

The Normal Model and Probabilities

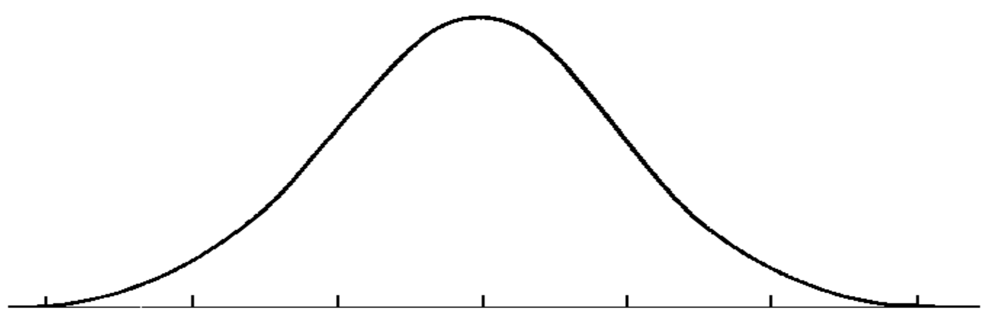
We've seen the unimodal, symmetric, bell shape many times in our simulations. In this module we'll learn about the theoretical normal model and use it to find theoretical probabilities. This is the basis for many theoretical confidence intervals and hypothesis tests.

Notation for a normal model: $X \sim N(\mu, \sigma)$. The inputs are the mean, μ, and the standard deviation, σ. Write out a definition statement for each new model you use. This is how statisticians define their models.

If you're curious, here's the formula for the normal curve: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. We will be using online normal calculators to find our probabilities. Stapplet.com/normal is one.

Example 1. The mean annual rainfall in Portland is approximately normally distributed with a mean of 40 inches and a standard deviation of 8 inches, rounded to the nearest inch.

a. Define, draw and label the normal distribution model for this situation. If you are typing your notes you can copy/paste from Stapplet.



Using an online normal calculator, write a probability statement and find the probability that the mean annual rainfall in Portland is

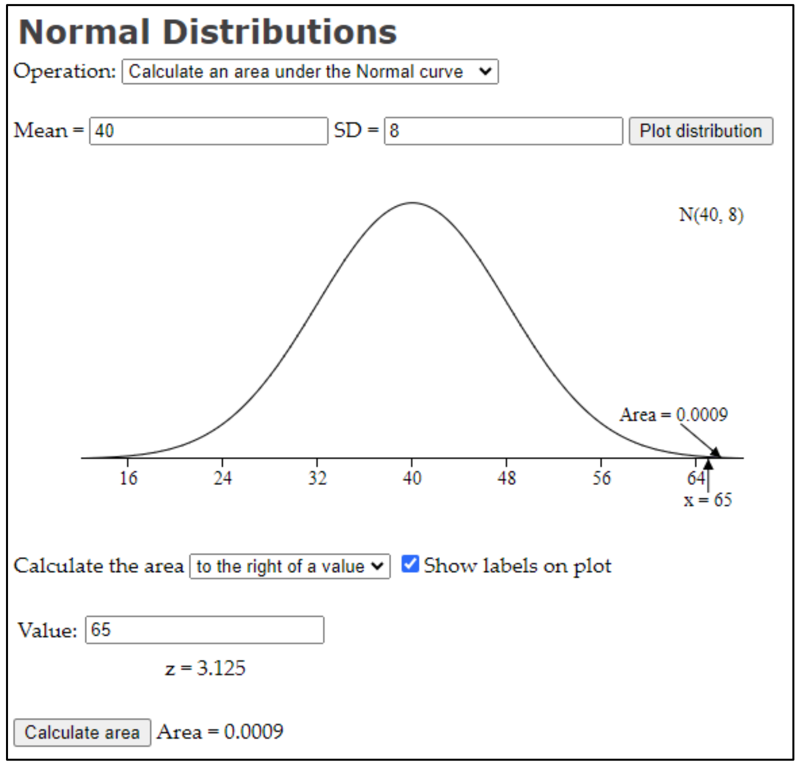
b. between 30-50 inches.

c. less than 22 inches.

d. greater than 65 inches.

e. greater than or equal to 65 inches.

Finding Normal Probabilities:
https://www.stapplet.com/normal.html



In the Stapplet Normal Calculator, make sure the Operation Menu is set to:
Calculate an area under the Normal Curve.

Type in the values for the mean, μ, and the standard deviation, σ, and click Plot Distribution.

Then select the type of probability you want: between two values, to the left, or right and enter your value(s) and click on Calculate area. Notice that Stapplet will give you the Z-score of the value and label the point if you select show labels on plot.

Example 2. In a medical study the population of children in Wisconsin were found to have serum cholesterol levels that were normally distributed with a mean μ = 1.75 mg/ml and a standard deviation σ = 0.30 mg/ml.

a. Define and draw or paste and label the normal model for children's cholesterol in Wisconsin.

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

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

Z-Scores and the Standard Normal Model

Remember earlier we computed Z-scores to compare an individual to the group. We can also use them to compare unlike events.

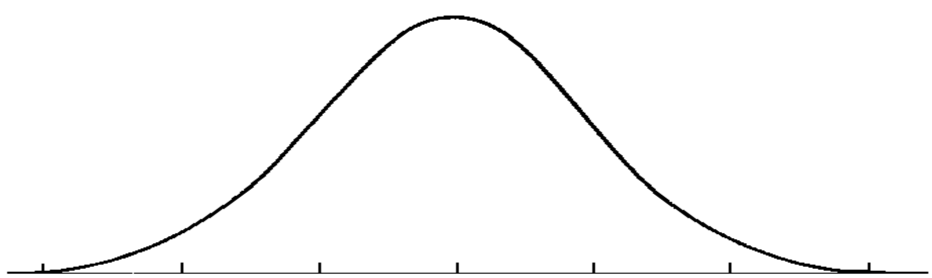
Example 3. Assume the average annual rainfall in Portland is 40 inches per year with a standard deviation of 8 inches. Also assume that the average wind speed in Chicago is 10 mph with a standard deviation of 2 mph. Suppose that one year Portland's annual rainfall was only 24 inches and Chicago's average wind speed was 13 mph. Which of these events was more extraordinary?

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

a. Find the Z-score for 24 inches of rain in Portland.

b. Find the Z-score for a wind speed of 13 mph in Chicago. Which of these events is more extraordinary?

The normal model of Z-scores is called the **Standard Normal Model.** It has a mean of 0 and a standard deviation of 1. We denote this with $Z \sim N(0,1)$

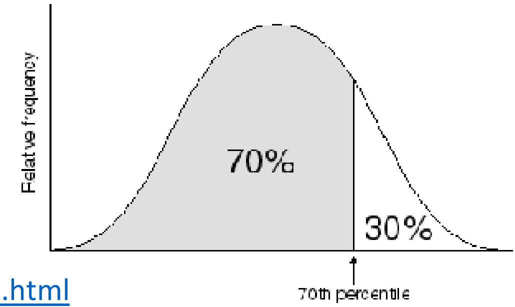


Example 4. Find these following probabilities.

a. $P(Z < 1.5)$

b. $P(-0.3 \leq Z \leq 0.5)$

Finding Percentiles and Cutoff Values

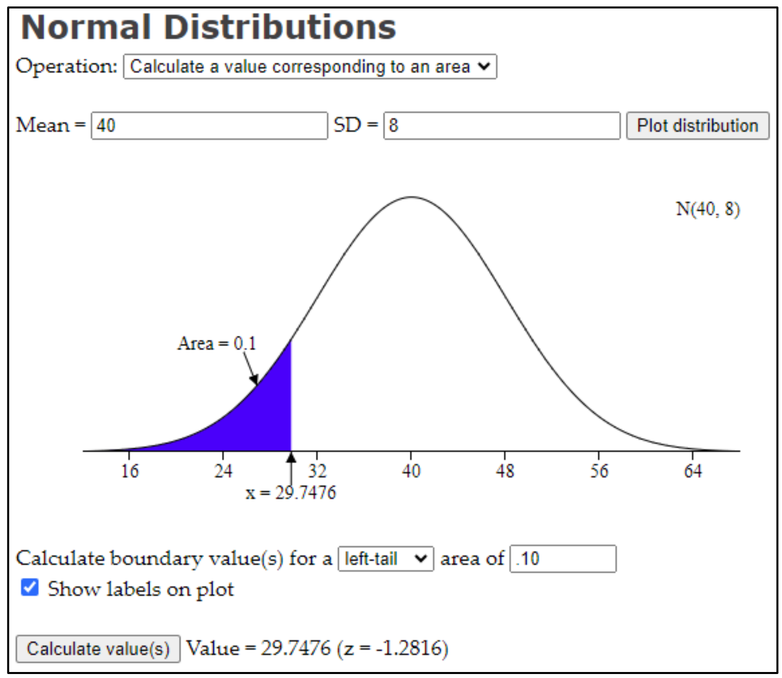


For any percentage of data, there is a corresponding **percentile** or **cutoff value.** That is the value that leaves the given percentage of data below it. We may be given a percentage and need to find the cutoff value or cut score.

Note that a percen**tile** is a cutoff value, not a percent**age**. This is called the *Inverse Normal* function because the input and output are reversed.

Find inverse normal values using Stapplet: https://www.stapplet.com/normal.html

In the Stapplet Normal Calculator, set the Operation Menu to:
Calculate a value corresponding to an area.

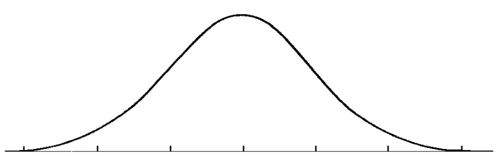


Type in the values for the mean, μ, and the standard deviation, σ, and click Plot Distribution.

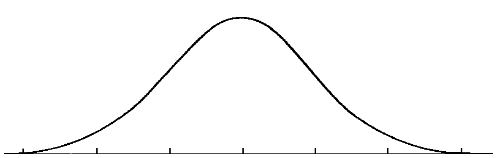
Then select the type of area you have: a left-tail, right-tail or central area, enter your value and click on Calculate value(s). Notice that Stapplet will give you the Z-score of the value and label the boundary value if you select show labels on plot.

Example 5. Let's continue the rainfall example where the mean annual rainfall in Portland is 40 inches with a standard deviation of 8 inches. Shade and find the cutoff values for:

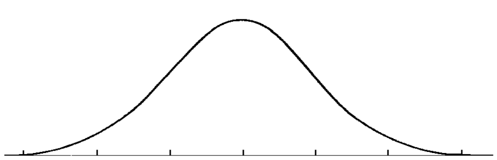
a. The lowest 10% of rainfall (the 10th percentile).



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



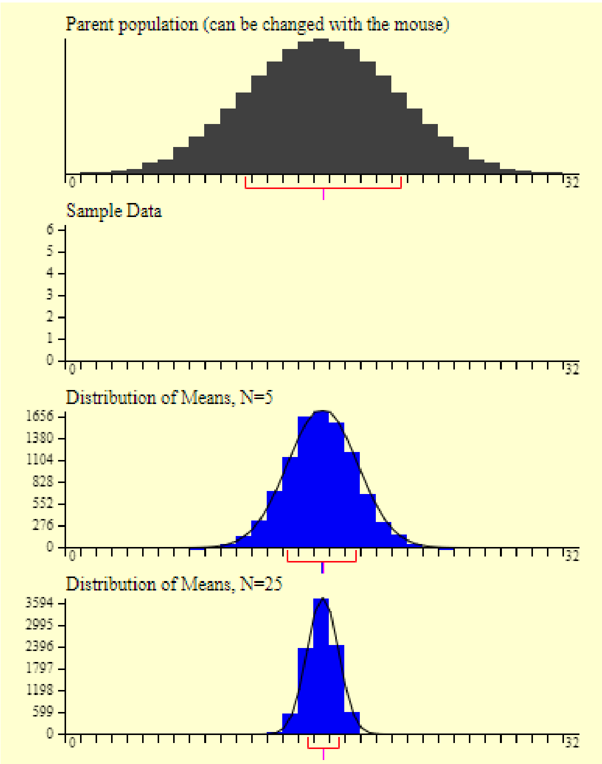
c. The amounts of rainfall that define the middle 50%.



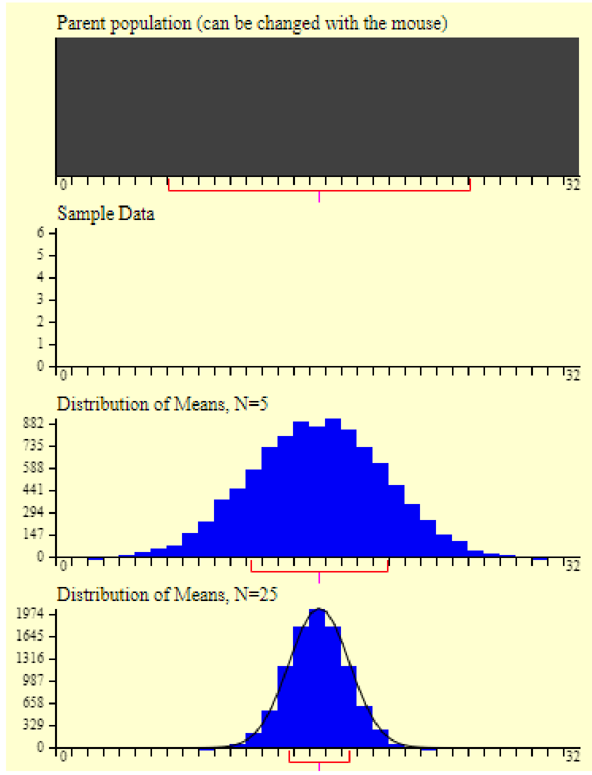
Central Limit Theorem and Sampling Distribution Models

Now we'll return to the Central Limit Theorem and learn about theoretical sampling distributions.

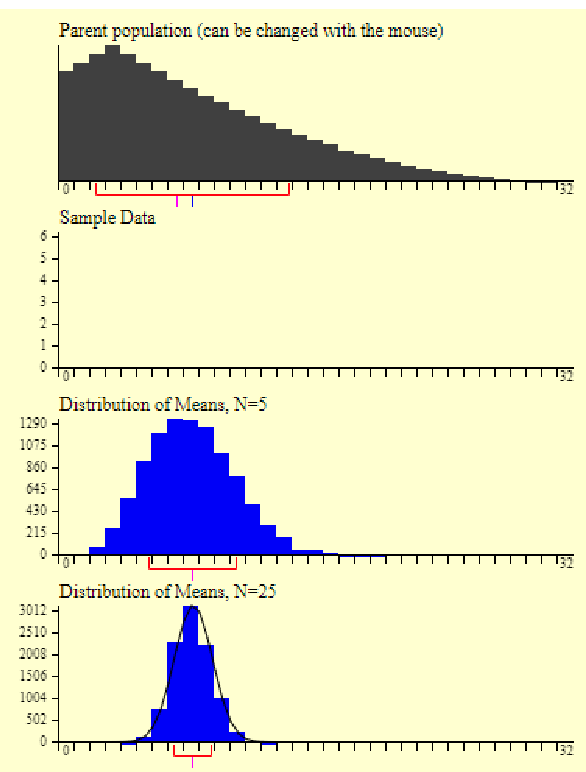
Normal Population



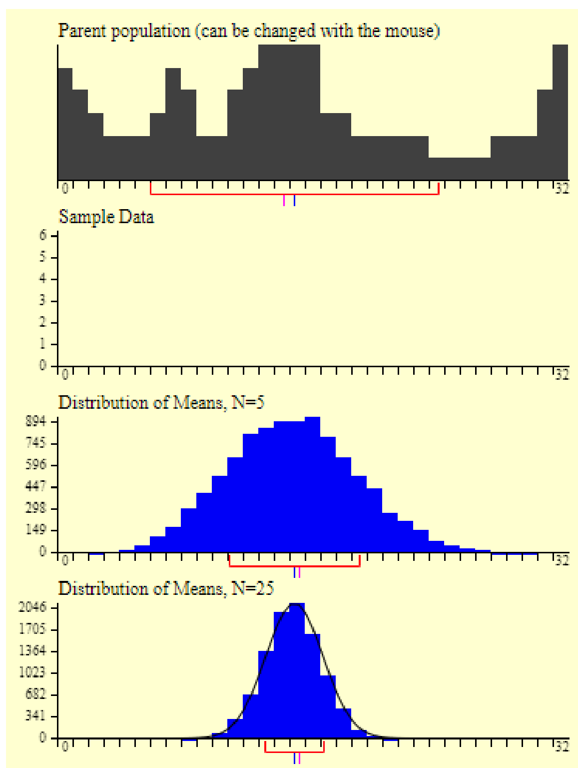
Uniform Population



Skewed Population



Multimodal Population



Sampling Distribution Conditions and Models

We have seen in our simulations that the distribution of a statistic for many random samples of the same size from the same population is close to a normal model. We've also seen that the standard error gets smaller the larger the sample size. Now we'll learn the theoretical models for the sampling distributions of means and proportions.

The Central Limit Theorem Conditions and Models

When using a theoretical model, we need to check the conditions or assumptions to ensure the model applies to the situation.

Conditions: These conditions are common to the models for means and proportions:

1. Independence: The individuals or items must be independent of each other regarding the variable measured.
2. Randomization: The samples need to be randomly chosen, or it's not safe to assume independence.
3. Large Population: If sampling without replacement, we must be sampling from a large population to consider the individuals independent of each other. Sometimes this is called the 10% Condition or said that the population must be 10 or 20 times our sample size.

For a Mean, $\bar{X}$

4. Sample Size:

   If the population is normally distributed, even small sample sizes will have a normal shape.
   If the population is not normally distributed, the sample size, n, should be 30 or larger.

   When the conditions are met, the sampling distribution for a mean, $\bar{X}$, is modeled by a normal distribution with the following parameters:

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

For a Proportion, $\hat{p}$

4. Success/Failure Condition:
   You should expect to have at least 10 successes and 10 failures in your data to ensure a normal distribution. Check that $np \geq 10$ and $nq \geq 10$.

   When the conditions are met, the sampling distribution for a proportion, $\hat{p}$, is modeled by a normal distribution with the following parameters:

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Large Population Condition

Small vs. Large Populations

Imagine you have one fun size bag of m&m's as your population. If you take out one m&m, what happens to the probability of selecting each color?

If you have a huge box of m&m's as your population, what happens to the probability of selecting each color when you take out one?

0.53 ounce bag 25 pound box

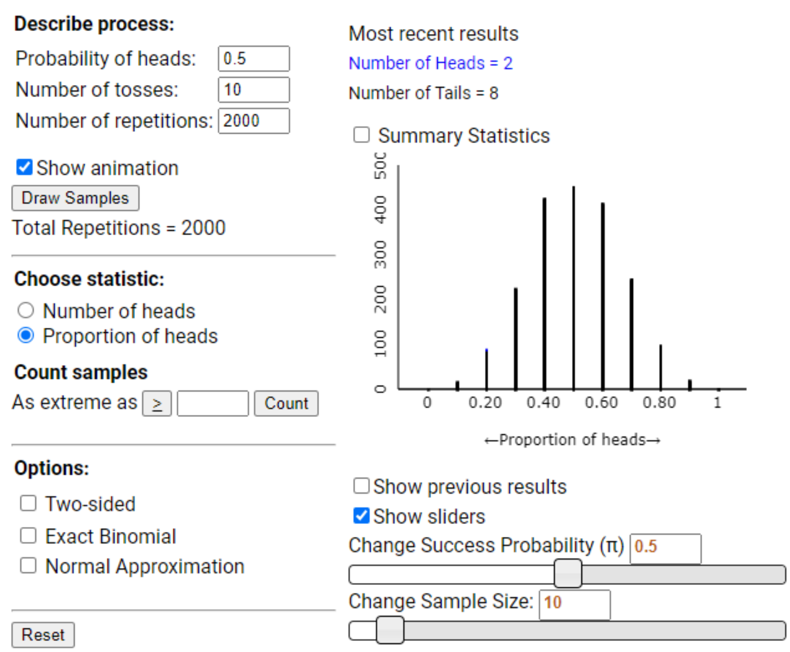
As long as we sample less than 10% of the population, we can say the individuals will be independent and will meet this condition. Some books say the population should be 10-20 times the sample size or more.

There are methods for small samples and small populations that are beyond the scope of this course.

Why the Success/Failure Condition?

The sample size condition for proportions depends on both n and p, rather than just n in the case of means. A good place to explore this is the [Rossman/Chance One Proportion Inference] Applet.

Simulate 2000 repetitions of size n=10 with a 0.50 probability of heads. Then click on show sliders. Move the slider for success probability and notice what happens.



Move the slider for n and notice what happens.

Practice with Theoretical Sampling Distributions

Example 1. Information on a packet of seeds claims that the germination rate is 92%. There are approximately 160 seeds in each packet.

a. Write about each of the conditions, in context, required to use a sampling distribution for the proportion of seeds that will germinate/sprout.

b. Define, draw, and label the sampling distribution model for the proportion of seeds that will sprout.

c. What is the probability that more than 95% of the seeds in a packet will germinate?

d. If you buy a packet of these seeds, would it be unusual for only 140 or fewer of the 160 seeds to germinate? Explain.

Example 2. The composition of pennies was changed in 1983, partly due to the rising cost of copper. The weights of pennies minted beginning in 1983 follow a normal distribution with a mean of 2.5 grams and a standard deviation of 0.03 grams. The earlier pennies made between 1865 and 1982 weigh approximately 3.11 grams.

a. If a single penny is randomly selected, define, draw, and label the distribution for the weight. What's the probability that a single penny weight is less than 2.4 grams?

b. If you have three pennies in your pocket, discuss each of the conditions required to use a sampling distribution for the average weight of the three pennies.

c. Define, draw and label the sampling distribution for the average weight of three pennies relative to your drawing in part a. (Use approximately the same scale.)

d. What's the probability that the mean weight of three pennies is less than 2.4 grams?

Practice with Theoretical Sampling Distributions when the Population Isn't Normal

Example 3. Bills at a given restaurant have an assumed population mean of $32.40 per person and a population standard deviation of $8.16 per person. The bills (per person) are heavily skewed to the right.

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

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

c. If we take a random sample of 50 bills from a month, would all the conditions be met?

d. Define, draw and label the sampling distribution model for the per person average of 50 bills.

e. How likely is it that a random sample of 50 bills will average at least $35 per person?

f. Find the cutoff values for the middle 50% of the mean price per person for 50 bills.

Constructing a Theoretical Confidence Interval for a Proportion

As we have already been practicing, confidence intervals are given by this formula:

$$\boxed{point\ estimate \pm margin\ of\ error}$$

Now we know the formula for the standard error of a sample proportion is $SE = \sqrt{\frac{pq}{n}}$ from the Central Limit Theorem.

As long as the CLT conditions are met, the confidence interval for a population proportion is given by:

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

where the critical z-score is determined by the confidence level.

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

Example 1. A Pew Research Poll[1] found that 59% of US adults think Marijuana should be legalized for both recreational and medical purposes. They surveyed 5,098 adults from Oct. 10-16, 2022. They also broke out their results by subgroups.

a. Calculate the theoretical 95% confidence interval for the population proportion of US adults who think Marijuana should be legal.

b. Calculate a bootstrap 95% confidence interval using the bootstrap estimate of the standard error.

c. Calculate a bootstrap 95% confidence interval using percentiles.

d. How do all three answers compare?

[1] https://www.pewresearch.org/short-reads/2022/11/22/americans-overwhelmingly-say-marijuana-should-be-legal-for-medical-or-recreational-use/

Finding Critical Z-Values Using Stapplet

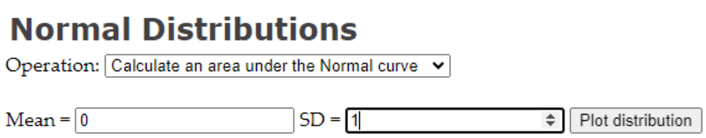
So far, we've used $\pm$ 2SE for our confidence intervals, which approximates the 95% confidence level. We can be more exact with this, and we also want to be able to use other confidence levels.

The critical Z-value for a confidence level is the Z-score that defines the middle area with that percentage.

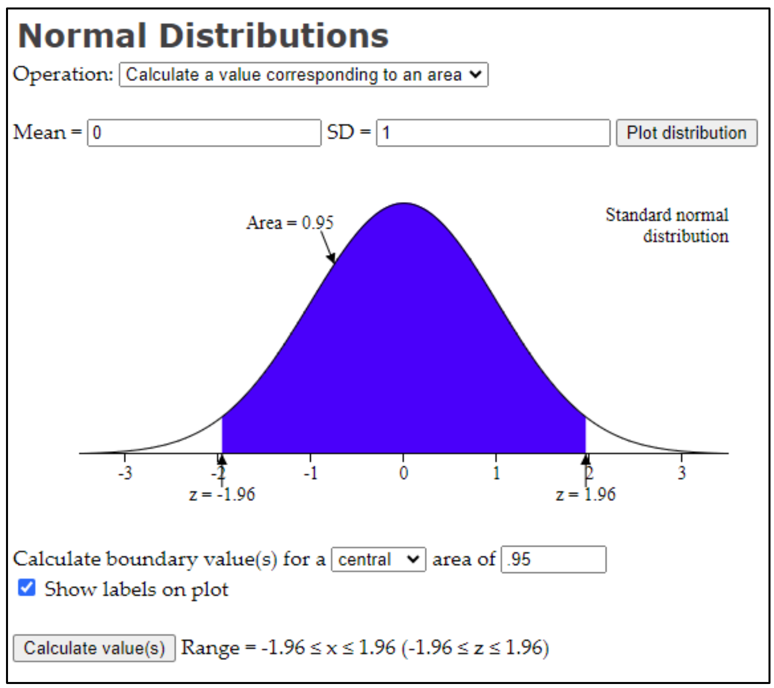
For example, when we go out 2 standard deviations from the mean, we capture approximately the middle 95% of the values in a normal model.

Our Z-score is a cutoff value, so we'll use the inverse normal function on Stapplet.
https://www.stapplet.com/normal.html

Select Calculate a value corresponding to an area. Then enter 0 for the mean and 1 for the standard deviation and click on Plot distribution.



For a 95% confidence interval, we select central and enter 0.95.



Do this for 80%, 90%, 95% and 99%. These are the most commonly used Z-values, but you can find another one if needed.

Confidence Level	Critical Z-value
80%	
90%	
95%	
99%	

Certainty Versus Precision

What do you notice about the width of the area as you are calculating the Z-values? What does this tell you about the widths of the confidence intervals?

Constructing and Interpreting Theoretical Confidence Intervals for a Proportion

Check the conditions:	Construct the confidence interval:	Write the interpretation:
• Independence • Randomization • Large population • Success/Failure $n \cdot \hat{p} \geq 10$ $n \cdot \hat{q} \geq 10$	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$	• Confidence Level • Interval • Wording for population proportion

Example 2.

a. A 2012 poll asked 166 adults whether they were baseball fans; 46% said that they were. Construct and interpret a 95% confidence interval for the true proportion of US adults that are baseball fans. Also state the standard error and the margin of error.

b. A 2016 Gallup poll asked 1021 U.S. adults whether they were satisfied with their current healthcare and 581 people said they were satisfied. Write and interpret a 99% confidence interval for the true proportion of U.S. adults who are satisfied with their healthcare.

Finding an Ideal Sample Size

When given a desired margin of error, ME, we can find the sample size, n, needed.

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

This is found by solving the margin of error formula for n: $ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Sample size must be a whole number. We always round up, even if the decimal is under 0.5. If we were to round down that would make the margin of error larger than the desired value.

Example 3. An article titled, "Tongue Piercing May Speed Tooth Loss, Researchers Say," found that 18 out of 52 participants had receding gums, which can lead to tooth loss.

a. How many people need to be surveyed in order to estimate the proportion of pierced-tongue people with receding gums to within 3% with 95% confidence?

b. Suppose we decide that a margin of error of 8% would be sufficient (again with 95% confidence). What's the necessary sample size?

Example 4. It's believed that 25% of adults over 50 never graduated high school. We wish to see if the same is true among 25- to 30-year-olds.

a. How many of this younger age group must we survey in order to estimate the proportion of non-grads to within 6% with 90% confidence?

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

c. What is the relationship between the number of people sampled and the margin of error?

Theoretical Hypothesis Tests for Proportions

The theoretical hypothesis test also uses the formula for standard error that comes from the Central Limit Theorem.

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

The steps for a hypothesis test using the normal distribution are similar to using a randomization distribution but we find our p-value using the normal distribution.

Steps for a Hypothesis Test with the Normal Distribution (1-proportion Z-test)

a. Write the null and alternate hypotheses.

b. Check the conditions to use the Central Limit Theorem.

c. Calculate the test statistic, $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0 q_0}{n}}}$. Use it to find the p-value and insert an image of your normal distribution with p-value.

d. Compare the p-value with the significance level, α, and determine whether the result is statistically significant.

e. State the conclusion in context, including the p-value and whether we reject the null or fail to reject the null hypothesis.

Since we are using the Normal Model through the Central Limit Theorem Normal, we need our four conditions to be satisfied. Remember we set up a hypothesis test assuming the null hypothesis is true, so we use p_0 when checking conditions and in the test statistic.

- Independence
- Randomization
- Large Population
- Success/Failure condition using p_0.

Comparing Hypothesis Tests with Confidence Intervals

For each example we will also compute a confidence interval to see whether the conclusion would be the same or different in each case. The significance test gives us the strength of the evidence and the confidence interval gives us the size of the effect.

- If the null value falls within the confidence interval, that supports the null hypothesis, and we fail to reject H_0.
- If the null value falls outside the confidence interval, that is evidence against the null hypothesis, and we reject H_0.

Example 1. During the 2013 National Football League (NFL) season, the home team won 153 of 245 regular-season games. Test whether there is a home field advantage at the 5% significance level.

95% confidence interval and result.

Example 2. In 2014, the official poverty rate was 14.8% in the US. A city official wants to test if their county has a different poverty rate than the rest of the US. In a random sample of 2000 county residents, 13.3% were below the poverty level. Is this enough evidence to show that the county's rate is significantly different than the national rate?

95% confidence interval and result

Example 3. A company develops what it hopes will be better instructions for its customers to set up their smartphones. The goal is to have 96% of its customers succeed. The company tests the new system with 400 people, of whom 376 were successful. Is this strong evidence that the new system fails to meet the company goal at the 5% significance level?

95% confidence interval and result.

Two-Variable Numerical Data

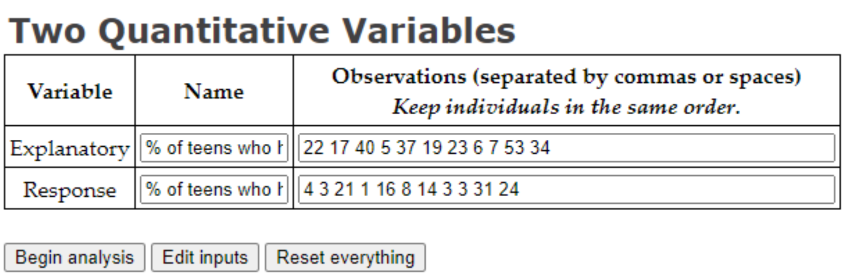
Paired quantitative data (x, y) is usually shown on a **scatterplot**. The pattern of the plotted points is used to determine whether there is a relationship between the two variables.

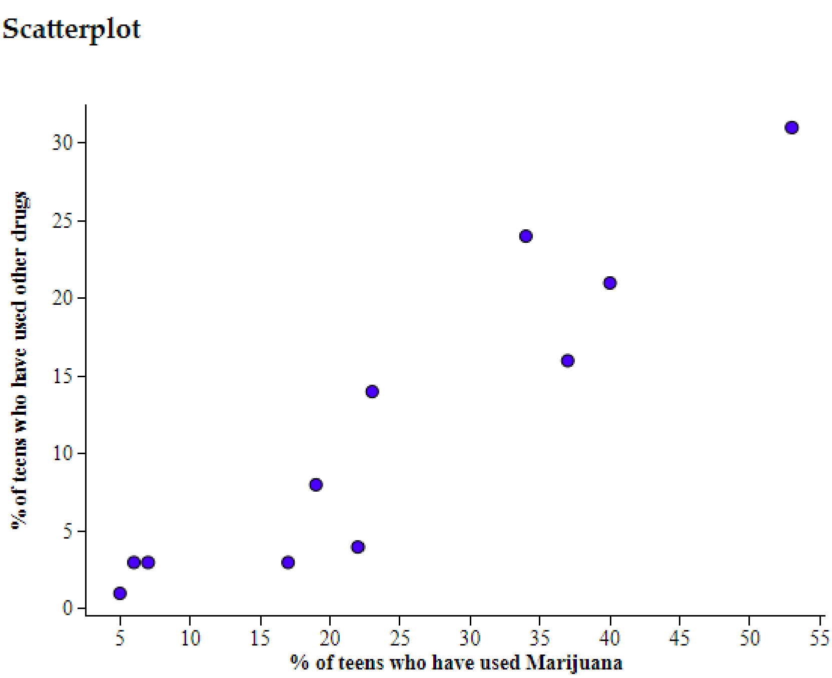
The **response variable** is the dependent variable, y.

The **eXplanatory variable** is the independent variable, x. We think that changes in the explanatory variable might *explain* changes in the response variable.

Example 1. A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the following table. Use the Stapplet Two Quantitative Variables applet to create a scatterplot with labels. Enter the variable names with units and copy each column separately into the applet.

Country	% of teens who have used Marijuana	% of teens who have used other drugs
Czech Republic	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
United States	34	24





a. Does it appear that there might be a relationship between Marijuana use and other drug use?

A Framework to Describe Association

Describe four features of the association between two variables:

- Direction
- Form
- Strength
- Unusual Features (subgroups or outliers)

Direction:

positive negative neither

Form:

linear curved no pattern

Strength:

strong moderate weak

Unusual Features:

groupings outliers

Correlation

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

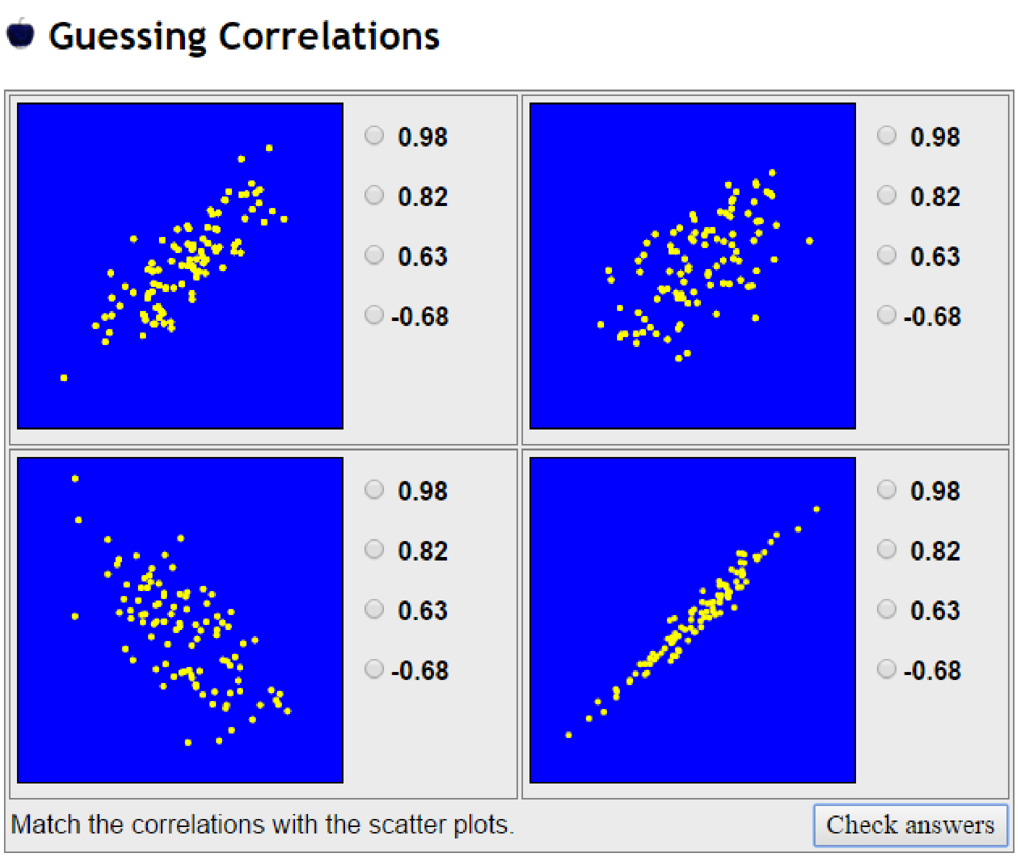
Caution: Even if it appears that y can be "predicted" from x, it does not follow that x **causes** y.

Characteristics of r

- The value of r only makes sense for linear relationships.
- The value of r is between -1 and 1.
- r has no units.
- The sign of r is the direction of the association.

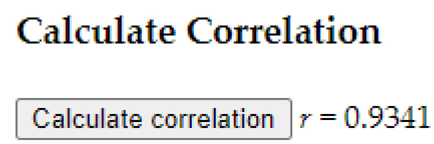
-1 **0** **1**

Matching Correlations Applet. Follow the link and try matching the correlations. Then answer the one in the image. http://www.istics.net/Correlations/



Example 1. Continued:

b. Using Stapplet, click on Calculate Correlation to find the correlation between the percentage of teens who use marijuana and other drugs. (Don't forget to check for a linear pattern.)



c. Write a brief description of the association using the 4-part framework including the correlation coefficient.

<u>The Line of Best Fit</u>

d. On the scatterplot for drug use, use a ruler or straightedge to draw a line that best models this relationship.

e. Draw the vertical distance between each point and the line. These are the residuals.
Residual = observed value – predicted value

- A positive residual means the data point is above the line, so the model underestimates the value for that case.
- A negative residual means the data point is below the line and the model overestimates the value for that case.

The least squares regression line is the line that minimizes the sum of the squared residuals (deviations of y). We will use technology to calculate this for us.

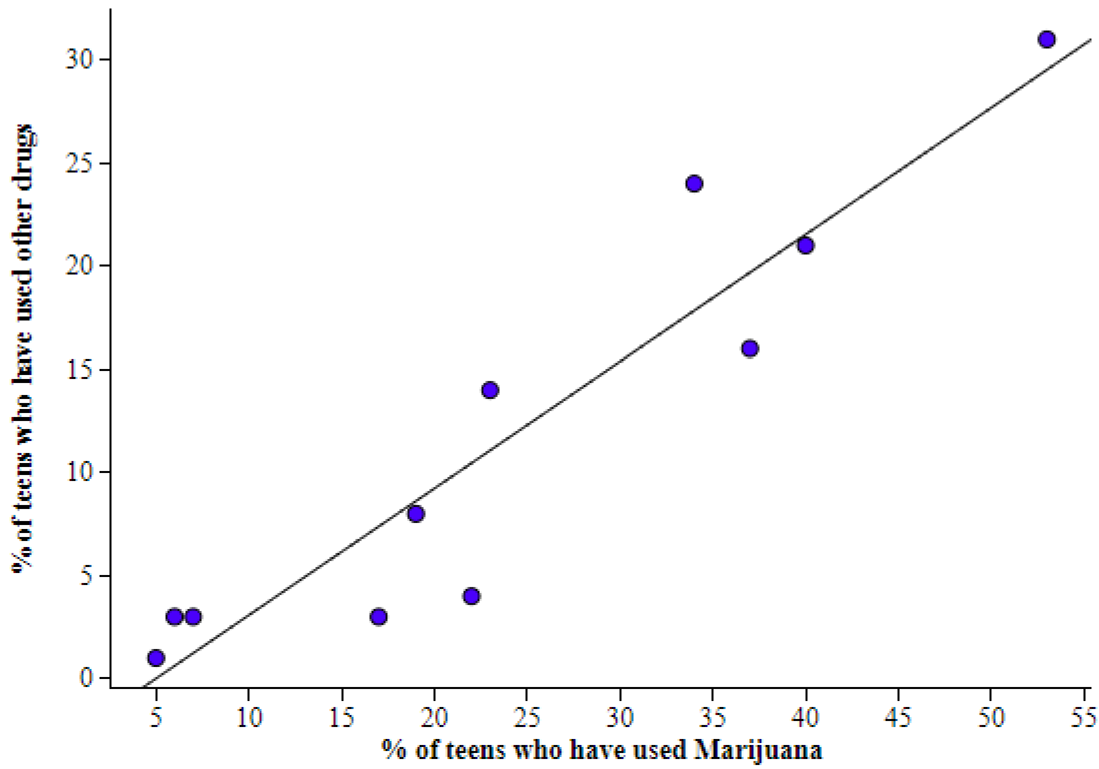
The Least Squares Regression Line, $\boldsymbol{\hat{y} = mx + b}$

Recall the equation of a line from algebra: $y = mx + b$

m represents the ______________________ and b represents the ________________________________.

$\hat{y}$ is read "$y -$ hat," and represents the predicted value of y. The values of m and b are the parameters of the linear model. You may also see the variables written as b_1 and b_0 like this: $\boldsymbol{\hat{y} = b_1x + b_0}$.

Going back to Stapplet, click on Calculate least-squares regression line.



Equation	n	s	r^2
$\hat{y} = -3.0677992 + 0.615003x$	11	3.8535	0.8725

f. Do these results confirm that the increase in marijuana use leads to an increase in other drugs? Explain.

Lurking, Hidden and Confounding Variables

A lurking variable or hidden variable is another variable that is actually responsible for the apparent association. For example, nations with more TV sets have higher life expectancies. Does having a TV make you live longer? No. The wealth of a nation has more to do with having TVs and life expectancy, so it is causing a common response.

A confounding variable is tangled up with the explanatory variable and also affects the response variable. It can be challenging to separate out the effects. For example, when studying the relationship between alcohol consumption and heart disease, whether a person smokes or not is related to alcohol consumption and can also affect heart disease. There are ways to handle this in studies that go beyond the scope of this class.

Example 2. The data below show the cost of the airfare and the distance traveled to each destination from Baltimore, MD.

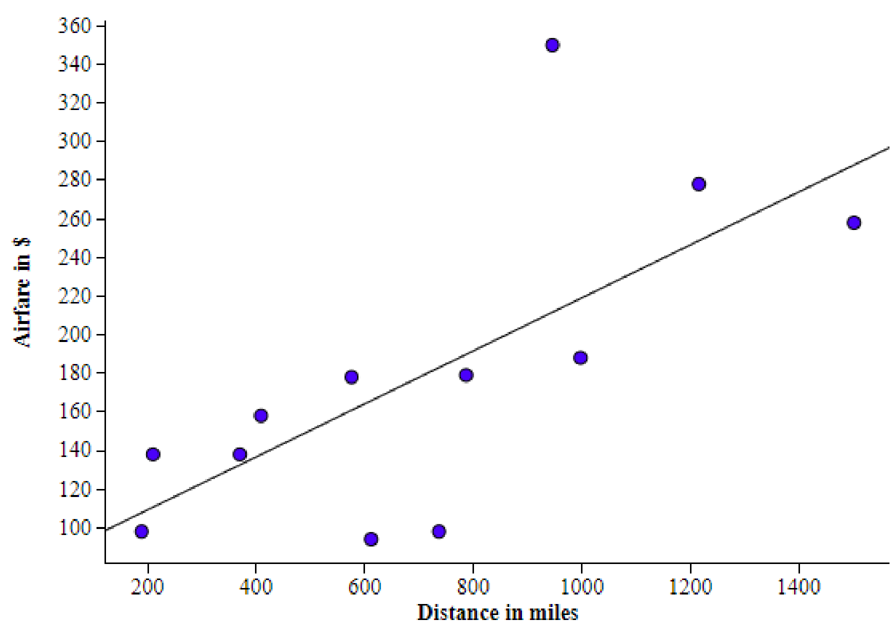
Destination	Distance, miles	Airfare, $
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	350
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

a. Which is the explanatory, and which is the response variable? Create a scatterplot for the data using Stapplet.

b. If the form is linear, what is the regression equation and correlation between the airfare and the distance of the flight?

c. Write a description of the association.

d. Is the residual for Chicago positive or negative? What does that mean?



Making Predictions

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

f. If you wanted to fly to Sydney, Australia from Baltimore (9,782 miles), what would the ticket cost? Do you think this is a good prediction?

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

Interpolation vs Extrapolation

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

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

Linear Regression Practice – Mustang Prices

Here is a data set for Mustang cars for sale online. We have looked at the price variable by itself. Now we will look at three numerical variables (two at a time).

Age	Miles in thousands	Price in thousand $
6	8.5	32
7	33	45
9	82.8	11.9
2	7	24.8
3	23	22
15	111	10
10	136.2	5
9	78.2	9
1	26.1	23
1	1.1	37.9
4	18.2	32.5
14	144.9	3
8	100.8	9
10	51.4	13
5	38.5	14.9
9	61.9	7
6	71.2	16
1	26.4	21
12	117.4	7
14	102	8.2
10	86.4	9.7
13	72.7	8
13	71.8	11.8
12	72.9	12.9
14	115.1	4.9

a. Before making a scatterplot, would each pair of variables have a positive or negative association? Explain why.
   i. Mileage vs. Age
   ii. Price vs. Age
   iii. Price vs. Mileage

b. Use Stapplet to make a scatterplot of price vs. age. If the association is linear, find the correlation coefficient and linear regression model.

c. Describe the association between Mustang price and age.

d. What does a positive residual mean in this model?

e. What does the model predict for the price of a Mustang if the car is 7 years old?

f. Could we use this model to predict the price for a 25-year old Mustang? Why or why not?