QuickStudy BUSINESS

ELEMENTARY STATISTICS

IMPORTANT TERMS

- Categorical/qualitative data falls into **groups** or **categories**.
  EX: State, type of pet, and gender
- Quantitative data represents **counts** or **measurements**.
  EX: Profit, number of people in line, and lifetime of a product
- **Population:** The entire group studied.
  EX: All U.S. registered voters
- **Sample:** A subset of a population from which data is collected.
  EX: **1,000** randomly sampled, registered U.S. voters
- **Parameter:** A number that summarizes a population and is typically unknown.
  EX: The **average** price of gas in the U.S.
- **Statistic:** A number that summarizes a sample.
  EX: The **average** price of gas from **1,000** gas stations randomly selected from the U.S.

DATA TABLES

- **Frequencies:** The number in each group.
- **Relative frequencies:** The percentage in each group.

Data Table: Type of Pet		
Type of pet	**Frequency**	**Relative frequency**
Dog	**10**	**.33**
Cat	**15**	**.50**
Other	**5**	**.17**
Total	**30**	**1.00**

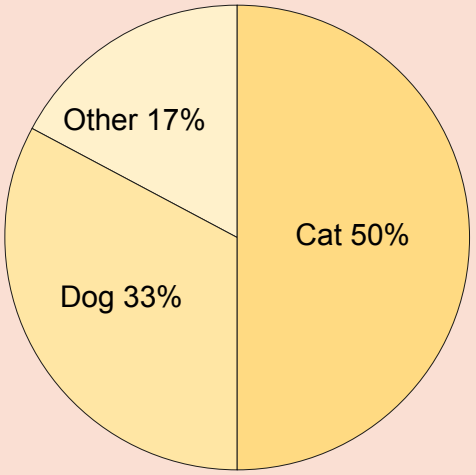
Data Table: Age Group		
Age group	**Frequency**	**Relative frequency**
Under **18**	**100**	**.24**
18–30	**200**	**.47**
Over **30**	**125**	**.29**
Total	**425**	**1.00**

GRAPHS FOR SINGLE VARIABLE

The purpose of a graph is to show a visual of data.

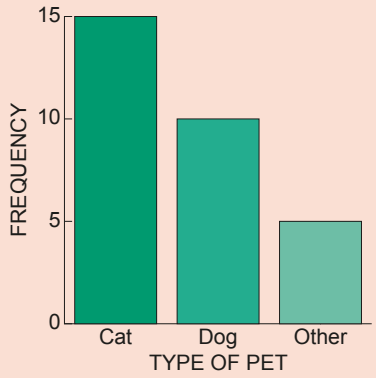
PIE CHART

Pie chart: A graph of categorical data showing frequency or relative frequency in a circle with a slice for each group.



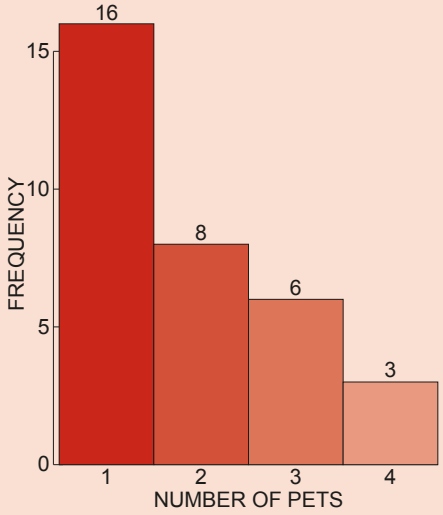
BAR GRAPH

Bar graph: A graph of categorical data showing frequency or relative frequency in a bar for each group.



HISTOGRAM

Histogram: A graph of quantitative data with the variable on the ***X***-axis divided into groups (bars) and the frequency or relative frequency in each group on the ***Y***-axis.



STEM-LEAF PLOT

Stem-leaf plot: A numerical graph of quantitative data with the last digit on the right of the line and the leading digits to the left of the line. You can recover the data set from a stem-leaf plot.

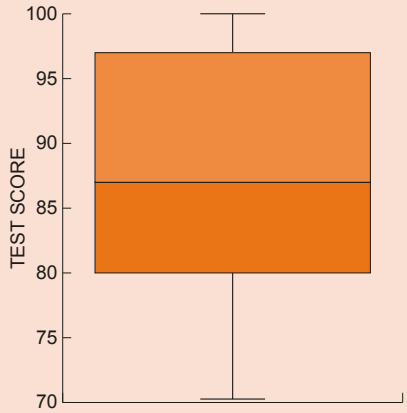
EX: Test scores with **70**, **70**, **71**, **71**, **76**, and **79** as the lowest **6** scores

Test Scores	
Stem	**Leaf**
7	**001169**
8	**00447779**
9	**0123777799**
10	**00**

BOXPLOT

Boxplot: A one-dimensional graph of quantitative data that shows the locations of the **5**-number summary.

- **25**% of the data lies in each section.
- A box contains the **middle 50**% of the data.
- The line in the box indicates the **median**.
- The lines coming out of the box end at the **minimum** and **maximum**.
- In this boxplot, the minimum = **70**, **Q1** = **80**, **Q2** = the median = **87.5**, **Q3** = **97**, and the maximum = **100**.



DESCRIPTIVE STATISTICS FOR SINGLE VARIABLE

Summarize quantitative data to indicate the center, variation, and relative standing.

MEASURES OF CENTER

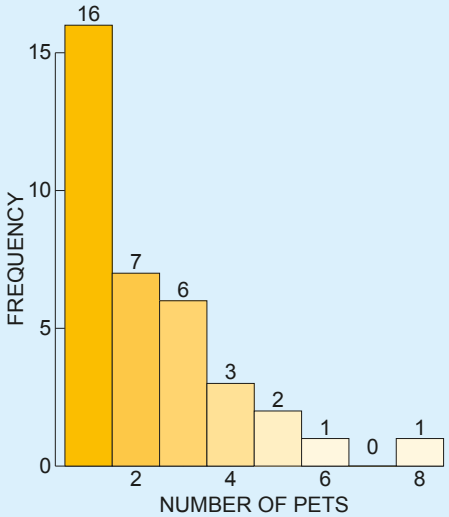
Measures of center indicate where the "middle" of the data is in different ways.

- **Mean:** The average of the data set. Sample mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where:
  - x_1 = **1st** data value
  - x_2 = **2nd** data value
  - x_n = last data value (n = number of values)

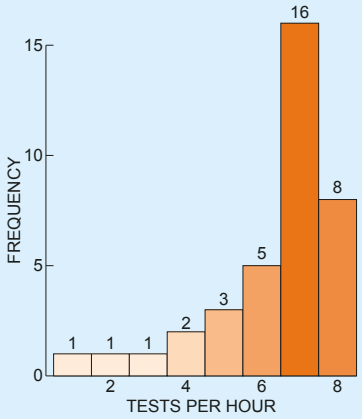
  EX: Given **2**, **4**, and **5**, the average is $\frac{2+4+5}{3} = 3.67$.
- **Median:** Divides the ordered set in half.
  - If the data has an **odd** sample size, the median is the middle value.
    EX: Given **1**, **2**, **3**, **4**, and **5**, the median is **3**.
  - If the data has an **even** sample size, the median is the average of the two middle values.
    EX: Given **1**, **2**, **3**, and **4**, the median is $\frac{2+3}{2} = 2.5$.
- **Mode:** The data value that occurs most often. This is not a good measure of center.

PROPERTIES

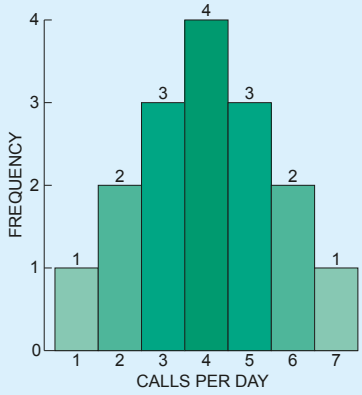
- Mean is affected by outliers; median is not.
- If mean > median, data are skewed right.
  EX: In this graph, mean = **2.33** and median = **2**.



- If mean < median, data are skewed left.
  EX: In this graph, mean = **6.35** and median = **7**.



- If mean = median (or close), data are fairly symmetric.
  EX: In this graph, mean = median = **4**.



QUARTILES & PERCENTILES

Quartiles: Divide data into **4** parts with **25%** in each part. Each cutoff point is a quartile.

Percentiles: A certain percentage of the data is less than this value.

- **Q1** = **1st** quartile = **25th** percentile
  EX: Given **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, and **10**, **Q1** = **3** (**25%** of the data is less than **3**).
- **Q2** = **2nd** quartile = **50th** percentile = median
  EX: Given **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, and **10**, **Q2** = **5.5** (splits the data in half).
- **Q3** = **3rd** quartile = **75th** percentile
  EX: Given **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, and **10**, **Q3** = **8** (**75%** of the data is less than **8**).

FIVE-NUMBER SUMMARY

The five-number summary consists of the minimum (min.), **Q1**, median, **Q3**, and the maximum (max.).

EX: Given **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, and **10**, the five-number summary is **1**, **3**, **5.5**, **8**, and **10**.

MEASURES OF VARIATION

Measures of variation indicate the concentration of data either from the data to the mean (on average) or between two values (e.g., min. and max.).

- **Variance:** Squared deviations from the mean, roughly averaged.

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

- **Standard deviation:** Square root of variance that is roughly the average distance from the mean.

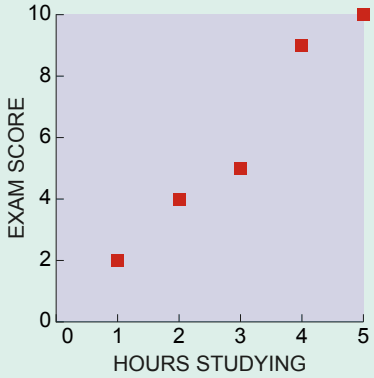
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - Standard deviation is:
    - Always greater than or equal to zero
    - Equal to **0** when all data are the same
    - Measured in the same units as the original data
    - Affected by outliers and skewness
- **Range:** Max–min. It is:
  - Affected by outliers in data
  - A crude measure of variation
- **Interquartile range:** **Q3**–**Q1**. It is:
  - The range of the middle **50%** of data
  - Not affected by outliers

CORRELATION & REGRESSION

The goal of correlation and regression is to explore and build models for linear relationships between two quantitative variables.

- When graphing the relationship between two quantitative variables ***X*** and ***Y***:
  - ***X*** = independent variable (e.g., hours studying)
  - ***Y*** = dependent variable (e.g., exam score)
- **Scatterplot:** Shows the strength, direction, and pattern of the (***X***, ***Y***) data.
  EX: This scatter plot shows (**1**, **2**), (**2**, **4**), (**3**, **5**), (**4**, **9**), and (**5**, **10**).
  - **Interpretation:** In this scatterplot, there is a strong uphill (positive) linear relationship.



CORRELATION

Correlation: Measures the strength and direction of the linear relationship between two quantitative variables ***X*** and ***Y***.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$, where:

- n = number of data points
- $\bar{x}$ = mean of the ***X*** values
- s_x = standard deviation of the ***X*** values
- $\bar{y}$ = mean of the ***Y*** values
- s_y = standard deviation of the ***Y*** values

PROPERTIES OF CORRELATION

Correlation:

- Is only between two quantitative variables
- Is always between **–1** and **+1**
  - r = **–1** when the points make a perfect downhill line.
  - r = **+1** when the points make a perfect uphill line.
  - r = **0** if there is no **linear relationship** (there may be some other relationship).
    - r = ±**0.7** = strong linear relationship
    - r = ±**0.5** = moderate linear relationship
    - r = ±**0.3** = weak linear relationship
- Has no units (it has a universal interpretation)
- Does not change if you switch ***X*** and ***Y***
  EX: If the correlation between the height and weight is **.6**, the correlation between the weight and height is also **.6**.
- Is affected by outliers and skewness, as the formula depends on the means and standard deviations of ***X*** and ***Y*** (all affected by outliers and skewness)

SIMPLE LINEAR REGRESSION

Simple linear regression fits the best line to the data, with the lowest **SSE** (**sum of squares for error**, where "error" means the distance from the line to the data point).

Equation of the best fitting regression line: $\hat{y} = b_0 + b_1 x$ (e.g., $\hat{y} = 2 + 3x$), where:

- ***X*** = independent variable
- ***Y*** = dependent variable
- $\hat{y}$ = predicted value of ***Y***
- b_1 = slope of the regression line
  - b_1 is in units of ***Y*** per **1**-unit change in ***X***.
  - The formula is $r\frac{s_y}{s_x}$, where:
    - s_y = standard deviation of the ***y***-values
    - s_x = standard deviation of the ***x***-values
- b_0 = ***y***-intercept of the regression line in units of ***Y***
  - The formula is $\bar{y} - b_1\bar{x}$, where:
    - $\bar{x}$ = mean of the ***x***-values
    - $\bar{y}$ = mean of the ***y***-values

Caution: If you switch ***X*** and ***Y***, the slope and ***y***-intercept will change; you will get a different regression line.

SIMPLE LINEAR REGRESSION ANALYSIS

Simple linear regression analysis: Finding the best-fitting regression line.

EX: If you have (**1**, **2**), (**2**, **4**), (**3**, **5**), (**4**, **9**), and (**5**, **10**) from the given scatterplot, the slope and ***Y***-intercept of the line can be found in the output in column **2** of the table.

Parameter Estimates

Parameter	Estimate	Statistical error	Alternative	DF	*T*-stat	*P*-value
Intercept	–0.3	0.8346656	≠ 0	3	–0.35942538	0.7431
Slope	2.1	0.25166115	≠ 0	3	8.3445538	0.0036

The ***y***-intercept of the line (b_0) is **–.3**. The slope of the line (b_1) is **2.1**. Put these together to get $\hat{y} = b_0 + b_1 x$ or $\hat{y} = -.3 + 2.1x$.

INTERPRETATION & PREDICTION

- To interpret the slope, $\frac{b_1}{1}$ means change in ***Y*** per **1**-unit change in ***X***.
  EX: In $\hat{y} = -.3 + 2.1x$, the slope is $\frac{2.1}{1}$, so ***Y*** increases by **2.1** per every **1**-unit increase in ***X***.
- To interpret the ***y***-intercept, there must be data near the ***y***-intercept and the value must make sense.
  EX: In $\hat{y} = -.3 + 2.1x$, if ***y*** is in years, the ***y***-intercept can't be interpreted since it's negative. If ***x*** is salary, there is no data in the area where $x = 0$.
- To predict ***y*** using ***x***, plug the value of ***x*** into the equation and solve $\hat{y} = b_0 + b_1 x$.
  EX: In $\hat{y} = -.3 + 2.1x$, to predict ***Y*** when $x = 1$, you plug in **1** to get $\hat{y} = -.3 + 2.1(1) = 1.8$.

LINE FIT

Measuring: How well the regression line fits.

EX: Computer output for the simple linear regression results:

Dependent variable: ***Y***

Independent variable: ***X***

r (correlation coefficient) = **0.97913005**

The **correlation** is **.9791** and the points are close to a line, indicating a strong positive linear relationship.

R^2

- **Coefficient of determination:** The percentage of the variation in ***Y*** that is explained by its relationship with ***X***.
- R^2 is always between **0** and **1**.
- R^2 equals the correlation, r, squared.
- The larger the value of R^2, the better the fit of the line. In the example above,
- $R^2 = r^2 = (.9791)^2 = .9587$, which is very strong.

RESIDUAL

Residual = $y - \hat{y}$ = observed ***y*** − predicted ***y*** at a certain value of ***x***.

EX: Suppose (**1**, **2**) is a point in the data and the best-fitting line is $\hat{y} = -.3 + 2.1x$. The observed ***y*** is **2**. The predicted value of ***Y*** at $X = 1$ is $-.3 + 2.1(1) = 1.8$. Subtract the observed minus the predicted to get $2 - 1.8 = .2$.

- **Positive residual:** Line underestimated the value of ***y***.
- **Negative residual:** Line overestimated the value of ***y***.
- **Zero residual:** Line estimated the value of ***y*** exactly.

RESIDUAL PLOT

- An ***X***, ***Y*** plot showing the ***X*** values on the ***X***-axis and the residuals on the ***Y***-axis. The horizontal line at $Y = 0$ is typically highlighted.
- If line fits well, the residual plot shows randomly scattered points close to the line $y = 0$ with no patterns or systematic changes and few outliers.
- The residuals can be standardized, where you subtract their mean and divide by their standard deviation. If the line fits well, most residuals should fall within **–3** and **3**.

PROBABILITY

- S = entire sample space of interest
  EX: S = {**HH**, **HT**, **TH**, **TT**} if you flip a coin twice
  Note: Each outcome has an equal probability of ¼.
- **Event:** Subset of S in which you are interested.
  EX: A = getting the same outcome on both flips (**HH**, **TT**); B = getting at least one head (**HH**, **HT**, **TH**)
  - **Marginal probability:** $P(A)$; the probability of a single event or characteristic.
    EX: $P(A) = P(\text{HH, TT}) = ¼ + ¼ = ½$
- **Conditional probability:** $P(A|B)$; the probability of A occurring given that B has occurred.
  EX: $P(A|B) = P(\text{same outcome on both flips}|\text{at least one head}) = P(\text{HH given HH, TH, HT}) = ⅓$
- **Joint probability:** $P(A \text{ and } B)$; the probability of the intersection of A and B.
  EX: $P(A \text{ and } B) = P(\text{same outcome and at least one head}) = P(\text{HH}) = ¼$
- **"Or" probability:** $P(A \text{ or } B)$; the probability of the union of A and B; a.k.a. A and/or B.
  EX: A or B = **HH**, **HT**, **TH**, **TT**, so $P(A \text{ or } B) = ¼ = 1$
- **Independence:** Events A and B (non-zero) are independent if and only if:
  - $P(A|B) = P(A)$
    EX: $P(A|B) = P(\text{same outcome}|\text{at least one head}) = ⅓$ and $P(A) = P(\text{same outcome}) = P(\text{HH, TT}) = ½$
    Since $P(A|B) \neq P(A)$, then A and B are **not** independent.
  - $P(A|B) = P(A|B^c)$
    EX: $P(A|B) = P(\text{same outcome}|\text{at least one head}) = ⅓$ and $P(A|B^c) = P(\text{same outcome}|\text{no heads}) = P(\text{TT}) = ¼$
    Since $P(A|B) \neq P(A|B^c)$, then A and B are **not** independent.
  - $P(A \text{ and } B) = P(A)P(B)$
    EX: If $P(A \text{ and } B) = P(\text{same outcome and at least one head}) = P(\text{HH}) = ¼$, $P(A) = P(\text{same outcome}) = P(\text{HH, TT}) = ½$, and $P(B) = P(\text{at least one head}) = P(\text{HH, HT, TH}) = ¾$, then $P(A \text{ and } B) = ¼ \neq P(A)P(B) = ½(¾) = ⅜$
    This means A and B are **not** independent.
- **Disjoint events:** A.k.a. **mutually exclusive events**; A and B are disjoint if $P(A \text{ and } B) = 0$.
  EX: If A = (**HH**) and B = (**TT**), then $P(\text{HH and TT}) = 0$, so A and B are disjoint
- **Remember:** Independent does not mean disjoint. Independent events coexist without affecting each other's probabilities, while disjoint events exclude each other.

- **Probability rules:** Let A = student has a backpack, A^c = student doesn't have a backpack, B = student owns a computer, and B^c = student doesn't own a computer. Suppose $P(A) = .7$, $P(B|A) = .57$, and $P(B|A^c) = .67$.
  - **Complement rule:** $P(A^c) = 1 - P(A)$
    EX: $P(\text{no backpack}) = P(A^c) = 1 - P(A) = 1 - .7 = .3$
  - **Multiplication rule:** $P(A \text{ and } B) = P(A)P(B|A)$
    EX: $P(\text{backpack and computer}) = P(\text{backpack})P(\text{computer}|\text{backpack}) = P(A)P(B|A) = .7(.57) = .4$
    $P(\text{computer and no backpack}) = P(A^c \text{ and } B) = P(A^c)P(B|A^c) = (1 - .7)(.67) = .2$
    - **Note:** If A and B are independent, $P(A \text{ and } B) = P(A)P(B)$.
  - **Law of total probability:** $P(B) = P(A \text{ and } B) + P(A^c \text{ and } B) = P(A)P(B|A) + P(A^c)P(B|A^c)$
    EX: $P(\text{computer}) = P(B) = .7(.57) + .3(.67) = .6$
  - **Definition of conditional probability:** $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)}$
    EX: $P(\text{backpack}|\text{computer})$ using the above results $= \dfrac{P(\text{backpack and computer})}{P(\text{computer})} = \dfrac{.4}{.6} = .67$
  - **Bayes Rule:** $P(A|B)$ from scratch $= \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$
    EX: $P(\text{backpack}|\text{computer}) = \dfrac{P(\text{backpack and computer})}{P(\text{computer})} = \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{.7(.57)}{.7(.57) + .3(.67)} = \dfrac{.4}{.6} = .67$
  - **Addition rule:** $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
    EX: $P(\text{backpack or computer}) = P(A \text{ or } B) = .7 + .6 - .4 = .9$
    - If A and B disjoint, $P(A \text{ or } B) = P(A) + P(B)$.

Two-Way Probability Distribution of "and" & "Marginal Probabilities"			
	B = computer	B^c = no computer	Total
A = backpack	$P(A \text{ and } B) = .7(.57) = .4$	$P(A \text{ and } B^c) = .7 - .4 = .3$	$P(A) = .7$
A^c = no backpack	$P(A^c \text{ and } B) = .3(.67) = .2$	$P(A^c \text{ and } B^c) = .3 - .2 = .1$	$P(A^c) = 1 - .7 = .3$
Total	$P(B) = .4 + .2 = .6$	$P(B^c) = 1 - .6 = .4$	1

RANDOM VARIABLES

A random variable is a characteristic, measure, or count that takes on certain values randomly according to certain probabilities.

PROBABILITY DISTRIBUTION

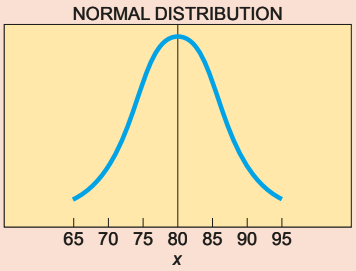
Probability distribution: A list of all possible X values and all their probabilities.

X	1	2	3
$P(x)$	.5	.3	.2

$P(X = 1) = .5$
$P(X > 1) = P(X = 2) + P(X = 3) = .3 + .2 = .5$
$P(X < 2) = P(X = 1) = .5$
Mean of X $= \mu_X = \sum_{\text{all } x} xp(x)$
$\mu_x = 1(.5) + 2(.3) + 3(.2) = 1.7$
Note: The answer is not **2** since you do not average the X values themselves; you multiply them by their weights (probabilities).
Variance of X $= \sigma_X^2 = \sum_{\text{all } x} (x - \mu_x)^2 p(x)$
$\sigma_X^2 = (1 - 1.7)^2(.5) + (2 - 1.7)^2(.3) + (3 - 1.7)^2(.2) = .61$
Standard deviation of X $= \sigma_X = \sqrt{\sum_{\text{all } x} (x - \mu_x)^2 p(x)}$
$\sigma_X = \sqrt{.61} = .78$

NORMAL DISTRIBUTION

- X has a bell-shaped curve according to the function $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.
- **Mean** $= \mu_X$
- **Standard deviation** $= \sigma_X$
- This graph shows a normal distribution with a **mean** = **80** and a **standard deviation** = **5**.



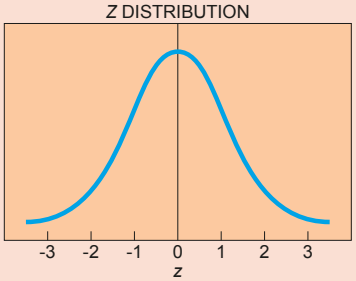
68-95-99.7 RULE

A.k.a. the **empirical rule**, if X has a normal distribution, then:

- About **68%** of the values lie within **1 standard deviation** of the mean (in the section of the graph that is between **75 and 85**).
- About **95%** of the values lie within **2 standard deviations** of the mean (in the section of the graph that is between **70 and 90**).
- About **99.7%** of the values lie within **3 standard deviations** of the mean (in the section of the graph that is between **65 and 95**).

STANDARD NORMAL (Z) DISTRIBUTION

Standard distribution: A special normal distribution with a **mean** = **0** and a **standard deviation** = **1**.



- Formula for transforming from X to Z: $Z = \dfrac{X - \mu}{\sigma}$
  EX: If X is normal with a mean $\mu_x = 80$ and a standard deviation $\sigma_x = 5$, then $X = 90$ transforms into $Z = \dfrac{90 - 80}{5} = 2$. You can see on the **Normal Distribution** and **Z-Distribution** graphs how **80** on the X-distribution corresponds to **2** on the Z-distribution.
- This table shows the first **12** lines for $P(Z \leq z)$, typically for values of Z from **–3** to **3**.

Z-Table										
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	*.8413*	.8438	.8461	.8485	.8508	.8531	.8554	.8577	*.8599*	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	*.8997*	.9015

FINDING & CALCULATING PROBABILITIES & CRITICAL VALUES WITH THE Z-TABLE

- **Less than:** $P(Z < 1.00)$ is in row **1.0** and column **.00**. You see **.8413**. To find $P(Z < 1.08)$, look at row **1.0** and column **.08** to get **.8599**.
- **Greater than:** $P(Z > 1.00) = 1 - P(Z < 1.00) = 1 - .8413 = .1587$
- **Between:** $P(1.00 < Z < 1.08) = P(Z < 1.08) - P(Z < 1.00) = .8599 - .8413 = .0186$
- **Critical values:** If you have an **80%** confidence interval, $\dfrac{1 - .80}{2} = .10$ of the probability lies above Z and $1 - .10 = .90$ of the probability lies below Z. Find the number closest to **.90** in the body of the table (**.8997**) and work backward. It lies in row **1.2** and column **.08**, so $Z = 1.28$.

FIND A PERCENTILE FOR X (REVERSE NORMAL)

- To find the **84.13th** percentile for X, where $\mu = 80$ and $\sigma = 5$, find **.8413** in the body of the table and work backward. It is located in row **1.0** and column **.00**.
- Put the row and column together to get $Z = 1.00$. This is the **84.13th** percentile for Z.
- Change to X using $Z = \dfrac{X - \mu}{\sigma} \rightarrow X = Z\sigma + \mu = 1.00(5) + 80 = 85$.
- **85** is the **84.13th** percentile of X.

THE *t* DISTRIBUTION

- X has a mound-shaped symmetric curve that is flatter than a normal distribution.
- A family of t-distributions exists, each with a different amount of flatness in the curvature. Each t-distribution is denoted by its **degrees of freedom**, $n-1$.
- As the degrees of freedom increase, the t-distribution approaches a Z-distribution.
- The **mean** = **0** and the **standard deviation** differs with each degree of freedom.
- ***t*-table:** Shows selected values of t for certain degrees of freedom and certain **right-tail probabilities** (i.e., the area to the right of the value of t).
- The below table shows t-values with certain right-tail probabilities for **1–10** degrees of freedom.

t-table							
df \area	.20	.15	.10	.05	.025	.01	.005
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.978	1.250	1.638	2.353	3.183	4.541	5.841
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.906	1.134	1.440	1.943	2.447	3.154	3.707
7	0.896	1.119	1.415	1.895	2.365	2.998	3.450
8	0.889	1.108	1.397	1.856	2.306	2.896	3.355
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.879	1.093	*1.372*	1.813	*2.228*	2.764	3.167

FINDING & CALCULATING PROBABILITIES & CRITICAL VALUES WITH THE *t*-TABLE

- If $t = 1.372$ with **10** degrees of freedom, $P(t > 1.372) = .10$. (Find **1.372** in the table and look at the top of the column it is in to find the probability.)
- If $t = -1.372$ with **10** degrees of freedom, you know that $P(t < -1.372) = .10$ by symmetry.
- If $|t| = 1.372$ with **10** degrees of freedom, $P(|t| > 1.372) = 2(.10) = .20$.
- **Critical values:** If you have a **95%** confidence interval with **10** degrees of freedom, t = **2.228** (row **10**, column **.025**) since $\frac{1-.95}{2} = .025$ of the probability lies above, and this table shows the above values only.

BINOMIAL DISTRIBUTION

The only outcomes of a binomial are success (yeses) and failures (noes).

- **Characteristics of a binomial:**
  - X = number of successes (yeses) in n independent trials
  - n = a fixed value
  - p = probability of success (yes)
    - p stays the same on every trial.
- **Mean of binomial** = np
- **Variance** = $np(1-p)$
- **Standard deviation** = $\sqrt{np(1-p)}$
- **Formula for calculating binomial probabilities:**

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where:
  - $n! = n(n-1)(n-2)\ldots(2)(1)$
  - $2! = 2(1) = 2$
  - $1! = 1$
  - $0! = 1$
- A binomial table shows the probabilities $P(X \le x)$ with X = **0**, **1**, …, n (number of yeses/successes) for selected values of n and p. Each n has its own mini-table. Below is the mini-table for n = **3** with certain values of p from **.10** to **.50**.

Binomial Table (n = 3)									
$X \setminus p$	.10	.15	.20	.25	.30	.35	.40	.45	.50
0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
1	.9720	.9393	.8960	.8438	.7840	.7183	.6480	.5748	.5000
2	.9990	.9966	.9920	.9844	.9730	.9571	.9360	.9089	.8750
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

FINDING & CALCULATING PROBABILITIES WITH A BINOMIAL TABLE

Rewrite the probability as something involving ≤.

Let n = **3** and p = **.4** in the binomial table.

- **Less than or equal to:** $P(X \le 2) = .9360$ (done)
- **Less than:** $P(X < 2) = P(X \le 1) = .6480$
- **Greater than or equal to:** $P(X \ge 2) = 1 - P(X \le 1) = 1 - .6480 = .3520$
- **Greater than:** $P(X > 2) = 1 - P(X \le 2) = 1 - .9360 = .0640$
- **Between:** $P(1 < X < 4) = P(X \le 3) - P(X \le 1) = 1.000 - .6480 = .3520$
- **Equal to:** $P(X = 1) = P(X \le 1) - P(X \le 0) = .6480 - .2160 = .4320$

Note: If p only goes up to **.5** on your binomial table and you need a probability for p beyond **.5**, count the number of failures/noes instead and use $1 - p$ as your new p.

EX: Suppose $n = 10$ and $p = .9$. Finding $P(X = 2$ yeses) is the same as finding $P(X = 10 - 2 = 8$ noes) using $p = 1 - .9 = .1$.

SPECIAL PROBABILITIES

Suppose X is a binomial with n = **3**.

- $P(X$ is **at least 1**$) = P(X \ge 1) = P(X = 1) + \ldots + P(X = 3)$
- $P(X$ is **at most 1**$) = P(X \le 1) = P(X = 0) + P(X = 1)$

NORMAL APPROXIMATION TO A BINOMIAL

Let X be a binomial with n and p given.

$$P(X \le x) \approx P\left(Z \le \frac{X - np}{\sqrt{np(1-p)}}\right)$$

Conditions: Use when $np \ge 10$ and $n(1-p) \ge 10$.

SAMPLING DISTRIBUTION OF $\bar{X}$

Let X be a random variable with **mean** μ_X and **standard deviation** σ_X.

- $\bar{X}$ is a random variable that represents the average value from **any** sample of size n.
  Note: $\bar{x}$ is the average value from a **particular** sample of size n.
- **Mean** of $\bar{X}$ is denoted $\mu_{\bar{x}} = \mu_X$.
- **Standard error** (deviation) of $\bar{X}$ is denoted $\sigma_{\bar{x}}$ and is equal to $\frac{\sigma_X}{\sqrt{n}}$.
- If X has a normal distribution, then $\bar{X}$ has a normal distribution.

$$P(X \le x) = P\left(Z \le \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = P\left(Z \le \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}\right)$$

- If X has any distribution besides normal, then $\bar{X}$ has an **approximate** normal distribution as long as n is large enough (typically $n > 30$). This result is due to the **central limit theorem**.

$$P(X \le x) \approx P\left(Z \le \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = P\left(Z \le \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}\right)$$

SAMPLING DISTRIBUTION OF $\hat{p}$

Let X = a binomial random variable with n trials and p = the probability of success on each trial.

- $\hat{p}$ is a random variable that represents the proportion of successes in **any** sample of size n. It is calculated by taking the number of successes X, divided by the number of trials, n.
- **Mean** of $\hat{p}$ is denoted $\mu_{\hat{p}} = p$.
- **Standard error** (deviation) of $\hat{p}$ is denoted $\sigma_{\hat{p}}$ and is equal to $\frac{p(1-p)}{n}$.
- If $np \ge 10$ and $n(1-p) \ge 10$, use the normal distribution to find approximate probabilities for $\hat{p}$.

EX: A ≤ probability would be $P\left(Z \le \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P\left(Z \le \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$.

(1 – α)% CONFIDENCE INTERVALS

FOR ONE POPULATION MEAN μ_X

One population mean: To estimate, or give a range of values for, the population mean.

WITH σ_X KNOWN & (1 – α) THE CONFIDENCE LEVEL

$$\bar{X} \pm Z\left(\sigma_X / \sqrt{n}\right)$$

- **Typical values of Z:** If $(1-\alpha)\%$ = **99%** → **2.575**, **95%** → **1.96**, **90%** → **1.645**, and **80%** → **1.28**
- **Margin of error (MOE):** $Z\left(\sigma_X / \sqrt{n}\right)$
  - As the confidence level $(1-\alpha)\%$ ↑, MOE ↑
  - As the sample size n ↑, MOE ↓
  - As population standard deviation σ_X ↑, MOE ↑
- Sample size needed to achieve a margin of error ≤ MOE:

$$n \ge \left\lceil \left(\frac{Z^* \sigma_X}{\text{MOE}} \right)^2 \right\rceil$$, where ⌈ ⌉ indicates round up to the nearest integer

EX: **345.2** rounds up to **346**.

- To cut MOE in half, you need **4×** as many values in your sample.
- To cut MOE down by a factor of $\frac{1}{n}$, you need n^2 values.

WITH σ_X UNKNOWN

$\bar{X} \pm t_{n-1}\left(s / \sqrt{n}\right)$, where:

- t_{n-1} = a value on the t-distribution with $n-1$ degrees of freedom
- s = the standard deviation of the sample data

FOR ONE POPULATION PROPORTION p

$$\hat{p} \pm Z\left(\hat{p}(1-\hat{p}) / \sqrt{n}\right)$$

- **Conditions:** Use when $n\hat{p} \ge 10$ and $n(1-\hat{p}) \ge 10$
- **Typical values of Z:** If $(1-\alpha)\%$ = **99%** → **2.575**, **95%** → **1.96**, **90%** → **1.645**, and **80%** → **1.28**

- **Margin of error (MOE):** $Z\left(\hat{p}(1-\hat{p})/\sqrt{n}\right)$
  - As the confidence level $(1-\alpha)$% ↑, MOE ↑.
  - As the sample size n ↑, MOE ↓.
  - As the sample proportion $\hat{p}$ approaches **0** or **1**, MOE ↓.
  - As the sample proportion $\hat{p}$ approaches **.5**, MOE ↑.
  - MOE is maximized at $\hat{p} = .5$ where the data is as unpredictable as it can get.
- A **conservative** sample size to achieve a margin of error ≤ MOE assumes $\hat{p} = .5$ and $(1-\alpha)$% = **95**% confidence level.
  - $n \geq \left\lceil \left(\frac{1}{\text{MOE}}\right)^2 \right\rceil$, where $\lceil\ \rceil$ indicates round up to the nearest integer.

EX: To achieve a MOE of ±.**02** in a survey, you need $n \geq \left\lceil \left(\frac{1}{.02}\right)^2 \right\rceil$ = **2,500** people.

FOR THE DIFFERENCE IN TWO POPULATION MEANS $\mu_X - \mu_Y$

INDEPENDENT SAMPLES WITH σ_X & σ_Y KNOWN

Conditions: **2** independent samples and **2** normal distributions

$\bar{X} - \bar{Y} \pm Z\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$, where for the two populations, respectively:

- n_1 and n_2 = independent samples
- σ_X^2 and σ_Y^2 = population variances

INDEPENDENT SAMPLES WITH σ_X & σ_Y UNKNOWN & EQUAL

$\bar{X} - \bar{Y} \pm t s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where:

- n_1 and n_2 = independent samples from the two populations, respectively
- s_p = pooled sample standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_x^2 + (n_2-1)s_Y^2}{n_1 + n_2 - 2}}$$

- Degrees of freedom for $t = n_1 + n_2 - 2$, so use $t_{n_1+n_2-2}$

INDEPENDENT SAMPLES WITH σ_X & σ_Y UNKNOWN & UNEQUAL

$\bar{X} - \bar{Y} \pm t_{df} \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$, where:

- n_1 and n_2 = independent samples from the two populations, respectively
- s_X and s_Y = sample standard deviations from the two populations, respectively
- Degrees of freedom for $t =$ $df = \dfrac{\left(s_X^2/n_1 + s_Y^2/n_2\right)^2}{\dfrac{\left(s_X^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_Y^2/n_2\right)^2}{n_2 - 1}}$

FOR THE DIFFERENCE IN TWO POPULATION PROPORTIONS $p_1 - p_2$

$\hat{p}_1 - \hat{p}_2 \pm Z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where for the first and second populations, respectively:

- $\hat{p}_1$ and $\hat{p}_2$ = sample proportions
- n_1 and n_2 = sample sizes

Conditions: Use when $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.

α-LEVEL HYPOTHESIS TEST

FOR ONE POPULATION MEAN μ_X

WITH σ_X KNOWN

STEP 1

Set up your hypothesis.

- **Null hypothesis:** The claim being made about the population mean H_o is $\mu_X = \mu_o$, where:
  - μ_X = real value of the population mean (not known)
  - μ_o = claimed value of the population mean (given)
- **Alternative hypothesis:** What you would believe if you had enough evidence against H_o.
  - Choose an H_a of $\mu_X < \mu_o$, $\mu_X \neq \mu_o$, **or** $\mu_X > \mu_o$.

  Note: If the H_a contains > or <, it is called a **one-sided** or **right-tailed/left-tailed test**, respectively. If H_a contains ≠, the test is called a **two-sided** or **two-tailed test**.

STEP 2

Find the test statistic based on the data.

$Z = \dfrac{\bar{X} - \mu_o}{\sigma_X/\sqrt{n}}$, where σ_X = population standard deviation (given)

STEP 3

Find the p-value for the test statistic.

- If H_a is $\mu_X < \mu_o$ **or** $\mu_X > \mu_o$, the p-value is the **probability** of being at or **beyond** the test statistic in the same direction as H_a.
- If H_a is $\mu_X \neq \mu_o$, the p-value is **twice the probability** of being at or **beyond** the test statistic.
- Use the Z-table to find these probabilities.

STEP 4

Make your decision regarding H_o. Are you going to reject it or fail to reject it?

Note: Never **accept** H_o. This is incorrect. Either reject H_o or don't.

- If the p-value is ≤ α, where α is the pre-set cutoff value (significance level), **reject** H_o. You have enough evidence against it.
- If the p-value is > α, where α is the pre-set cutoff value (significance level), **fail to reject** H_o. You do not have enough evidence against it.
- Typical values of α, the significance level of your test: **.001**, **.01**, **.05**, and **.10**. Choose in advance.

STEP 5

Draw your conclusion in the context of the problem.

EX: "We have/don't have enough evidence to say 'H_a' is true," where instead of H_o, you use words in the context of the problem that pertain to H_a.

Type I & II Errors

- If you conduct a hypothesis test and **reject** H_o, you could be wrong. In this case H_o is actually **true**, but you received an abnormal value by chance. This is called a **type I error**, which is a **false alarm**. The chance of a type I error is α (the significance level of the test).
- If you conduct a hypothesis test and **fail to reject** H_o, you could be wrong. In this case H_o is actually **false**, but you did not find enough evidence against it. This is called a **type II error**, which is a **missed opportunity**. The chance of a type II error is β and is related to the sample size (n). Its exact calculation is beyond our scope.
- α and β are inversely related. As α ↑, β ↓ and vice versa.

WITH σ_X UNKNOWN

This type of test is called a ***t*-test**. It does best when X resembles a normal distribution. n does not have to be > **30**.

STEP 1

Set up the hypothesis.

H_o: $\mu_X = \mu_o$

H_a: Choose $\mu_X < \mu_o$, $\mu_X \neq \mu_o$, **or** $\mu_X > \mu_o$

STEP 2

Find the test statistic based on the data.

$t_{n-1} = \dfrac{\bar{X} - \mu_o}{s/\sqrt{n}}$, where s is the sample standard deviation from the data

Here $n - 1$ represents the degrees of freedom for t, where n is the sample size.

STEP 3

Find the p-value for the test statistic.

- If H_a is $\mu_X < \mu_o$ **or** $\mu_X > \mu_o$, the p-value is the probability of being at or **beyond** the test statistic in the same direction as H_a.
- If H_a is $\mu_X \neq \mu_o$, the p-value is twice the probability of being at or **beyond** the test statistic.
- Use the t-table to find the probabilities (a.k.a. **tail probabilities** on some t-tables).

STEPS 4 & 5

See **With σ_X Known** in this section under **For One Population Mean μ_X**.

Connection between Confidence Intervals & Hypothesis Tests

- A $(1-\alpha)$% confidence interval coincides with the hypothesis test with an H_a of $\mu_X \neq \mu_o$ as the **alternative hypothesis** (a.k.a. a **two-tailed test**). The hypothesis test has a significance level of α.

  EX: A **95**% confidence interval coincides with a **.05** two-tailed hypothesis test.
- If the value in H_o **is** within the confidence interval, **fail to reject** H_o in the coinciding two-tailed hypothesis test.
- If the value in H_o is **not** within the confidence interval, **reject** H_o from the coinciding two-tailed hypothesis test.

FOR ONE POPULATION PROPORTION p

Conditions to be met: Use when $np_o \geq 10$ and $n(1-p_o) \geq 10$.

STEP 1

Set up the hypothesis.

- H_o: $p = p_o$, where p_o is the claimed value for p.
- H_a: Choose $p < p_o$, $p \neq p_o$, **or** $p > p_o$.

STEP 2

Find the test statistic based on the data.

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

STEPS 3–5

See **With σ_X Known** in this section under **For One Population Mean μ_X**. Use the Z-table to find the p-value.

FOR THE DIFFERENCE OF TWO POPULATION MEANS $\mu_X - \mu_Y$

INDEPENDENT SAMPLES WITH σ_X & σ_Y KNOWN

STEP 1

Set up the hypothesis.

- H_o: $\mu_X - \mu_Y = c$, where c is some constant, often **0** (e.g., if H_o is $\mu_X = \mu_Y$)
- H_a: Choose $\mu_X - \mu_Y > c$, $\mu_X - \mu_Y < c$, **or** $\mu_X - \mu_Y \neq c$
- $\mu_X - \mu_Y$ = difference in the population means
- c = amount of the difference

STEP 2

Find the test statistic based on the data.

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}, \text{ where:}$$

- σ_X and σ_Y = two population standard deviations (given)
- n_1 and n_2 = sample sizes, respectively
- $\mu_X - \mu_Y = c$ for some specified value (often **0**)

STEP 3

Find the *p*-value for the test statistic (*see* **For One Population Mean** μ_X, p. 5).

STEPS 4 & 5

Make your decision and draw your conclusion in the context of the problem.

INDEPENDENT SAMPLES WITH σ_X & σ_Y UNKNOWN & ASSUMED TO BE EQUAL

STEP 1

Set up the hypothesis.

- H_o: $\mu_X - \mu_Y = c$, where c is some constant, often **0** (e.g., if H_0 is $\mu_X = \mu_Y$)
- H_a: Choose $\mu_X - \mu_Y > c$, $\mu_X - \mu_Y < c$, **or** $\mu_X - \mu_Y \neq c$
- $\mu_X - \mu_Y$ = difference in the population means
- c = amount of the difference

STEP 2

Find the test statistic based on the data.

$$t_{df} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where:}$$

- $df = n_1 + n_2 - 2$
- $s_p = \sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}$, where:
  - s_X and s_Y = two sample standard deviations from the data
  - n_1 and n_2 = sample sizes, respectively
  - $\mu_X - \mu_Y = c$ for some specified value (often **0**)

STEP 3

Find the *p*-value for the test statistic (*see* **For One Population Mean** μ_X, p. 5).

STEPS 4 & 5

Make your decision and draw your conclusion in the context of the problem.

INDEPENDENT SAMPLES WITH σ_X & σ_Y UNKNOWN & NOT ASSUMED TO BE EQUAL

STEP 1

Set up the hypothesis.

- H_o: $\mu_X - \mu_Y = c$, where c is some constant, often **0** (e.g., if H_o is $\mu_X = \mu_Y$)
- H_a: Choose $\mu_X - \mu_Y > c$, $\mu_X - \mu_Y < c$, **or** $\mu_X - \mu_Y \neq c$
- $\mu_X - \mu_Y$ = difference in the population means
- c = amount of the difference

STEP 2

Find the test statistic based on the data.

$$t_{df} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}, \text{ where:}$$

- $df = \dfrac{\left(s_X^2/n_1 + s_Y^2/n_2\right)^2}{\frac{\left(s_X^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_Y^2/n_2\right)^2}{n_2 - 1}}$, where:
  - s_X and s_Y = two sample standard deviations from the data
  - n_1 and n_2 = sample sizes, respectively
  - $\mu_X - \mu_Y = c$ for some specified value (often **0**)

STEP 3

Find the *p*-value for the test statistic (*see* **For One Population Mean** μ_X, p. 5).

STEPS 4 & 5

Make your decision and draw your conclusion in the context of the problem.

FOR THE MEAN OF THE PAIRED DISTANCES, μ_d, IN DEPENDENT SAMPLES

STEP 1

Set up the hypothesis.

- H_o: $\mu_d = 0$
- H_a: Choose $\mu_d > 0$, $\mu_d < 0$, **or** $\mu_d \neq 0$

STEP 2

Find the test statistic based on the data.

$$t_{n-1} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}, \text{ where:}$$

- n = number of pairs of data
- $\bar{d}$ = average of the differences in the pairs of data
- s_d = standard deviation of the differences in the pairs of data

STEP 3

Find the *p*-value for the test statistic (*see* **For One Population Mean** μ_X, p. 5).

STEPS 4 & 5

Make your decision and draw your conclusion in the context of the problem.

FOR THE DIFFERENCE IN TWO PROPORTIONS IN INDEPENDENT SAMPLES

STEP 1

Set up the hypothesis.

- H_o: $p_1 - p_2 = 0$ (or $p_1 = p_2$)
- H_a: Choose $p_1 - p_2 > 0$ (or $p_1 > p_2$), $p_1 - p_2 < 0$ (or $p_1 < p_2$), **or** $p_1 - p_2 \neq 0$ (or $p_1 \neq p_2$)
- $p_1 - p_2$ = difference in the population proportions

STEP 2

Find the test statistic based on the data.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where:}$$

- $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$, x_1, and x_2 = number of successes (yeses) in the two samples, respectively
- n_1 and n_2 = two sample sizes, respectively
- $p_1 - p_2 = c$ for some specified value (usually **0**)

Conditions: Use when $n_1\hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1 - \hat{p}_2) \geq 10$.

STEP 3

Find the *p*-value for the test statistic (*see* **For One Population Mean** μ_X, p. 5).

STEPS 4 & 5

Make your decision and draw your conclusion in the context of the problem.

ISBN-13: 978-142324433-2
ISBN-10: 142324433-8
50695
U.S. $6.95

Author: Deborah Rumsey, PhD

9 781423 244332
6 54614 04433 4