Population - entire collection of objects or individuals about which information is desired.
→ easier to take a sample
  ♦ Sample - part of the population that is selected for analysis
  ♦ Watch out for:
    • Limited sample size that might not be representative of population
  ♦ Simple Random Sampling - Every possible sample of a certain size has the same chance of being selected

Observational Study - there can always be lurking variables affecting results
→ i.e., strong positive association between shoe size and intelligence for boys
→ **should never show causation

Experimental Study - lurking variables can be controlled; can give good evidence for causation

Descriptive Statistics Part I
→ Summary Measures

<table>
<thead>
<tr>
<th>Describing Data Numerically</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center and Location</td>
</tr>
<tr>
<td>Other Measures of Location</td>
</tr>
<tr>
<td>Variation</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Quartiles</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Interquartile Range</td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Standard Deviation</td>
</tr>
</tbody>
</table>

Mean - arithmetic average of data values
→ **Highly susceptible to extreme values (outliers). Goes towards extreme values
→ Mean could never be larger or smaller than max/min value but could be the max/min value

Median - in an ordered array, the median is the middle number
→ **Not affected by extreme values

Quartiles - split the ranked data into 4 equal groups
→ Box and Whisker Plot

Range = $X_{\text{maximum}} - X_{\text{minimum}}$
→ Disadvantages: Ignores the way in which data are distributed; sensitive to outliers

Interquartile Range (IQR) = 3rd quartile - 1st quartile
→ Not used that much
→ Not affected by outliers

Variance - the average distance squared

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
→ $s^2$ gets rid of the negative values
→ units are squared

Standard Deviation - shows variation about the mean

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$
→ highly affected by outliers
→ has same units as original data
→ finance = horrible measure of risk (trampoline example)

Descriptive Statistics Part II
Linear Transformations

$$Y = a + bX$$
→ Linear transformations change the center and spread of data

$$V \alpha(a + bX) = b^2 V \alpha(X)$$
$$Average(a+bX) = a+b[Average(X)]$$
Effects of Linear Transformations:
- \( \text{mean}_{\text{new}} = a + b \cdot \text{mean} \)
- \( \text{median}_{\text{new}} = a + b \cdot \text{median} \)
- \( \text{stddev}_{\text{new}} = |b| \cdot \text{stddev} \)
- \( IQR_{\text{new}} = |b| \cdot IQR \)

**Z-score** - new data set will have mean 0 and variance 1
\[
z = \frac{x - \bar{x}}{s}
\]

**Empirical Rule**
- Only for mound-shaped data
- Approx. 95% of data is in the interval:
  \((\bar{x} - 2s, \bar{x} + 2s) = \bar{x} + / - 2s\)
- Only use if you just have mean and std. dev.

Chebyshev’s Rule
- Use for any set of data and for any number \( k \), greater than 1 (1.2, 1.3, etc.)
- \( 1 - \frac{1}{k^2} \)
- (Ex) for \( k=2 \) (2 standard deviations), 75% of data falls within 2 standard deviations

Detecting Outliers
- Classic Outlier Detection
  - Doesn’t always work
  - \( |z| = \left| \frac{x - \bar{x}}{s} \right| \geq 2 \)
- The Boxplot Rule
  - Value \( X \) is an outlier if:
    \( X < Q1 - 1.5(Q3 - Q1) \)
    or
    \( X > Q3 + 1.5(Q3 - Q1) \)

Skewness
- Measures the degree of asymmetry exhibited by data
  - Negative values = skewed left
  - Positive values = skewed right
  - If \( |\text{skewness}| < 0.8 \) don’t need to transform data

**Measurements of Association**
- **Covariance**
  - Covariance > 0 = larger \( x \), larger \( y \)
  - Covariance < 0 = larger \( x \), smaller \( y \)
  - \( s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \)
  - Units = Units of \( x \) · Units of \( y \)
  - Covariance is only +, -, or 0 (can be any number)

- **Correlation** - measures strength of a linear relationship between two variables
  - \( r_{xy} = \frac{\text{covariance}_{xy}}{(\text{std.dev.}_x) \cdot (\text{std.dev.}_y)} \)
  - Correlation is between -1 and 1
  - Sign: direction of relationship
  - Absolute value: strength of relationship (-0.6 is stronger relationship than +0.4)

- **Correlation doesn’t imply causation**
- The correlation of a variable with itself is one

**Combining Data Sets**
- Mean (\( Z \)) = \( \bar{Z} = a\bar{x} + b\bar{y} \)
- Var (\( Z \)) = \( s_Z^2 = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X,Y) \)

**Portfolios**
- Return on a portfolio:
  \( R_p = w_A \bar{R}_A + w_B \bar{R}_B \)
  - Weights add up to 1
  - Return = mean
  - Risk = std. deviation

- Variance of return of portfolio
  \( s_p^2 = w_A^2 s_A^2 + w_B^2 s_B^2 + 2w_A w_B (s_{A,B}) \)
  - Risk(variance) is reduced when stocks are negatively correlated. (When there’s a negative covariance)

**Probability**
- Measure of uncertainty
- All outcomes have to be exhaustive (all options possible) and mutually exhaustive (no 2 outcomes can occur at the same time)

<table>
<thead>
<tr>
<th>Magnitude of ( r )</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00–0.20</td>
<td>Very weak</td>
</tr>
<tr>
<td>0.20–0.40</td>
<td>Weak to moderate</td>
</tr>
<tr>
<td>0.40–0.60</td>
<td>Medium to substantial</td>
</tr>
<tr>
<td>0.60–0.80</td>
<td>Very strong</td>
</tr>
<tr>
<td>0.80–1.00</td>
<td>Extremely strong</td>
</tr>
</tbody>
</table>
### Probability Rules
1. Probabilities range from $0 \leq P(\text{prob}(A)) \leq 1$
2. The probabilities of all outcomes must add up to 1
3. The complement rule = A happens or A doesn’t happen
   \[ P(\overline{A}) = 1 - P(A) \]
   \[ P(A) + P(\overline{A}) = 1 \]
4. Addition Rule:
   \[ P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \]

### Contingency/Joint Table
- To go from contingency to joint table, divide by total # of counts
- everything inside table adds up to 1

### Conditional Probability
- \[ P(A|B) \]
- \[ P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \]
- Given event B has happened, what is the probability event A will happen?
- Look out for: "given", "if"

### Independence
- Independent if:
  \[ P(A|B) = P(A) \text{ or } P(B|A) = P(B) \]
- If probabilities change, then A and B are dependent
- **hard to prove independence, need to check every value

### Multiplication Rules
- If A and B are INDEPENDENT:
  \[ P(A \text{ and } B) = P(A) \cdot P(B) \]

### Decision Tree Analysis
- square = your choice
- circle = uncertain events

### Discrete Random Variables
- \[ P_X(x) = P(X = x) \]

### Expectation
- \[ \mu_x = E(x) = \sum x_i P(X = x_i) \]
- Example: \( (2)(0.1) + (3)(0.5) = 1.7 \)

### Variance
- \[ \sigma^2 = E(x^2) - \mu_x^2 \]
- Example: \( (2)^2(0.1) + (3)^2(0.5) - (1.7)^2 = 2.01 \)

### Rules for Expectation and Variance
- \[ \mu_x = E(s) = a + b\mu_x \]
- Var(s) = \( b^2 \cdot \sigma^2 \)

### Jointly Distributed Discrete Random Variables
- Independent if:
  \[ P_{X,Y}(X = x \text{ and } Y = y) = P_X(x) \cdot P_Y(y) \]
Combining Random Variables
- If \( X \) and \( Y \) are independent:
  \[
  E(X + Y) = E(X) + E(Y) \\
  \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)
  \]
- If \( X \) and \( Y \) are dependent:
  \[
  E(X + Y) = E(X) + E(Y) \\
  \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)
  \]

Covariance:
- \( \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \)
- If \( X \) and \( Y \) are independent, \( \text{Cov}(X, Y) = 0 \)

Calculate the Covariance
- We will use the formula \( \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \)
- For a die \( E(X) = E(Y) = 3.5 \)
- We need to find \( E(XY) \)

<table>
<thead>
<tr>
<th>Probability</th>
<th>( X )</th>
<th>( Y )</th>
<th>( XY )</th>
<th>( X \times Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/6</td>
<td>1</td>
<td>5</td>
<td>5/6</td>
<td>5/6</td>
</tr>
<tr>
<td>1/6</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>1/6</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>1/6</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>1/6</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1/6</td>
<td>6</td>
<td>5</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

So \( \text{Cov}(X, Y) = 9.33 - (3.5)(3.5) = -2.91 \)

The covariance is negative because larger values of \( X \) are associated with smaller values of \( Y \).

Binomial Distribution
- doing something \( n \) times
- only 2 outcomes: success or failure
- trials are independent of each other
- probability remains constant

1.) All Failures
\[ P(\text{all failures}) = (1 - p)^n \]

2.) All Successes
\[ P(\text{all successes}) = p^n \]

3.) At least one success
\[ P(\text{at least 1 success}) = 1 - (1 - p)^n \]

4.) At least one failure
\[ P(\text{at least 1 failure}) = 1 - p^n \]

5.) Binomial Distribution Formula for \( x \) = exact value

\[
P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}
\]

6.) Mean (Expectation)
\[ \mu = E(x) = np \]

7.) Variance and Standard Dev.
\[ \sigma^2 = npq \]
\[ \sigma = \sqrt{npq} \]
\[ q = 1 - p \]

Binomial Example
3) During the semester a professor cycles to school on 5 days of the week. On any given day, the probability that he arrives at school after 9am is 0.1. For a period of 4 weeks (20 days), calculate the probability that he arrives after 9am
b) On at least 1 day but no more than 3 days
\[ P(x = 1) = \frac{20!}{5!(15)!} (0.1)^1 (0.9)^{15} = 0.27017034353 \]
\[ P(x = 2) = \frac{20!}{5!(15)!} (0.1)^2 (0.9)^{14} = 0.28517986706 \]
\[ P(x = 3) = \frac{20!}{5!(15)!} (0.1)^3 (0.9)^{13} = 0.19011987138 \]
\[ 0.27017034353 + 0.28517986706 + 0.19011987138 = 0.745473022 \]

Continuous Probability Distributions
- the probability that a continuous random variable \( X \) will assume any particular value is 0

Density Curves
- Area under the curve is the probability that any range of values will occur.
- Total area = 1

Uniform Distribution
- It is described by the function:
\[
f(x) = \frac{1}{b-a}, \quad \text{where } a \leq x \leq b
\]

Uniform Example
5) Suppose the number of donuts a nine-year old child eats per month is uniformly distributed from 0.5 to 4 donuts, inclusive
a) Find the probability that a randomly selected nine-year old child eats more than 2 donuts in a month.
\[ X = \text{Unif}(0.5, 4) \]
\[ f(x) = \frac{1}{4-0.5}, \quad \text{where } 0.5 \leq x \leq 4 \]
\[ f(2) = \frac{1}{3.5}, \quad \text{where } 0.5 \leq x \leq 4 \]

Probability = Area = Width \times Height
\[ \text{Probability} = 2 \cdot \frac{1}{3.5} = 0.571428571 \]

(Example cont’d next page)
b) Find the probability that a different nine-year old child eats more than two donuts given that his or her amount is more than 1.5 donuts.

\[
P(x \geq 2 | x \geq 1.5) = \dfrac{P(x = 2) + P(x = 3)}{P(x \geq 1.5)}
\]

**Probability = Area = Width x Height**

\[
P(x \geq 2 | x \geq 1.5) = \dfrac{P(x \geq 2)}{P(x \geq 1.5)}
\]

\[
\begin{align*}
\text{Probability} &= \text{Area} = \text{Width} \times \text{Height} \\
P(x \geq 2 | x \geq 1.5) &= \dfrac{0.5(2)}{0.71} = 0.71 \\
\text{Probability} &= 0.8
\end{align*}
\]

**Mean for uniform distribution:**

\[
E(X) = \dfrac{(a+b)}{2}
\]

**Variance for unif. distribution:**

\[
Var(X) = \dfrac{(b-a)^2}{12}
\]

**Normal Distribution**

- governed by 2 parameters: 
  - \( \mu \) (the mean) and \( \sigma \) (the standard deviation)
- \( X \sim N(\mu, \sigma^2) \)

**Standardize Normal Distribution:**

\[
Z = \dfrac{X - \mu}{\sigma}
\]

- Z-score is the number of standard deviations the related \( X \) is from its mean
- **Z< some value, will just be the probability found on table**
- **Z> some value, will be (1-probability) found on table**

**Normal Distribution Example**

8) It has been reported that the average hotel check-in time, from curbside to delivery of bags into the room, is 12.0 minutes. Ang has just left the cab that brought her to her hotel. Assuming a normal distribution with a standard deviation of 2.5 minutes, what is the probability that the time required for Ang and her bags to get to the room will be:

\[
X \sim N(12, 2)
\]

**Sums of Normals**

- If \( X_1 \) and \( X_2 \) are each normally distributed 
  \[ X_1 \sim N(\mu_1, \sigma_1^2) \]
- Then the sum is normally distributed 
  \[ aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}) \]

**Sums of Normals Example:**

11) Jill's bowling scores are normally distributed with mean 170 and standard deviation 20, whereas Jack's scores are normally distributed with mean 160 and standard deviation 15. If Jack and Jill each bowl one game, find the probability that Jack's score is higher.

**Confidence Intervals**

- tells us how good our estimate is
- **Want high confidence, narrow interval**
- **As confidence increases ↑, interval also increases ↑**

### A. One Sample Proportion

<table>
<thead>
<tr>
<th>Estimate Population Parameter...</th>
<th>with Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion: ( \pi )</td>
<td>( \hat{p} )</td>
</tr>
</tbody>
</table>

**Proportion:**

\[
\hat{p} = \dfrac{\bar{x}}{n} = \dfrac{\text{number of successes in sample}}{\text{sample size}}
\]

\[
(\hat{p} - 1.96\sqrt{\dfrac{\hat{p}\hat{q}}{n}}, \hat{p} + 1.96\sqrt{\dfrac{\hat{p}\hat{q}}{n}})
\]

**Central Limit Theorem**

- as \( n \) increases, 
- \( \bar{x} \) should get closer to \( \mu \) (population mean)
- mean(\( \bar{x} \)) = \( \mu \)
- variance(\( \bar{x} \)) = \( \dfrac{\sigma^2}{n} \)
- \( \bar{x} \sim N(\mu, \dfrac{\sigma^2}{n}) \)
  - if population is normally distributed, \( n \) can be any value
  - any population, \( n \) needs to be \( \geq 30 \)

12) The weight of an adult swan is normally distributed with a mean of 30 pounds and a standard deviation of 8 pounds. A farmer randomly selected 39 swans and loaded them into his truck. What is the probability that this flock of swans weighs > 1010 pounds?

\[
X \sim N(30, \dfrac{8}{\sqrt{39}})
\]

**Confidence Intervals**

**Confidence Intervals**

- tells us how good our estimate is
- **Want high confidence, narrow interval**
- **As confidence increases ↑, interval also increases ↑**

**A. One Sample Proportion**
**Standard Error and Margin of Error**

- The confidence interval is given by
  \[ \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

- The margin of error

- The standard form of any confidence interval is \( \pm \text{margin of error} \).

**Example of Sample Proportion Problem**

2) A recent Gallup poll consisted of 1012 randomly selected adults who were asked whether “cloning of humans should or should not be allowed.” Results showed that 901 of those surveyed indicated that cloning should not be allowed. Construct a 95% confidence interval estimate of the proportion of adults believing that cloning of humans should not be allowed.

\[
\begin{align*}
\hat{p} &= \frac{901}{1012} = 0.89031626 \\
\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{1012}} &= 0.89031626 - 1.96 \sqrt{\frac{0.89031626(1-0.89031626)}{1012}} \\
&= 0.87106278, 0.90956968
\end{align*}
\]

**Determining Sample Size**

\[
n = \frac{(1.96)^2 \hat{p}(1-\hat{p})}{e^2}
\]

- If given a confidence interval, \( \hat{p} \) is the middle number of the interval
- No confidence interval; use worst case scenario
- \( \hat{p} = 0.5 \)

5) Obesity is defined as a body mass index (BMI) of 30 kg/m² or more. A 95% confidence interval for the percentage of U.S. adults aged 20 years and over who were obese was found to be 22% to 24%. What was the sample size?

\[
\begin{align*}
\hat{p} &= 0.23 \\
\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.23 - 1.96 \sqrt{\frac{0.23(0.77)}{n}} \\
n &= 6804
\end{align*}
\]

**B. One Sample Mean For samples \( n > 30 \)**

**Confidence Interval:**

\[
(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})
\]

- If \( n > 30 \), we can substitute \( s \) for \( \sigma \) so that we get:

\[
\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}
\]

**Review of Stata Output**

```
This tells us how variable the sample is
```

<table>
<thead>
<tr>
<th>variable</th>
<th>obs</th>
<th>mean</th>
<th>std. dev.</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>bodytemp</td>
<td>306</td>
<td>98.2</td>
<td>0.6228963</td>
<td>98.5</td>
<td>99.6</td>
</tr>
<tr>
<td>ct bodytemp</td>
<td>306</td>
<td>98.2</td>
<td>0.660501</td>
<td>98.08044</td>
<td>98.31996</td>
</tr>
</tbody>
</table>

\[
s \bigg/ \sqrt{n}
\]

**For samples \( n < 30 \)**

\[\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}\]

**T Distribution used when:**

- \( \mu \) is not known, \( n < 30 \), and data is normally distributed
- Replace the 1.96 value with a t value to get:

\[
\bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right)
\]

where \( t \) comes from Student’s t distribution, and depends on the sample size through the degrees of freedom “\( n-1 \).”

**Hypothesis Testing**

- **Null Hypothesis:**
  - \( H_0 \), a statement of no change and is assumed true until evidence indicates otherwise.

- **Alternative Hypothesis:** \( H_a \) is a statement that we are trying to find evidence to support.

- **Type I error:** reject the null hypothesis when the null hypothesis is true.
  - (considered the worst error)

- **Type II error:** do not reject the null hypothesis when the alternative hypothesis is true.

**Example of Type I and Type II errors**

- According to a study published in March, 2008 the mean length of a phone call on a cellular telephone was 3.25 minutes. A researcher believes that the mean length of a call has increased since then.

- A Type I error occurs if the sample evidence leads the researcher to conclude that \( \mu > 3.25 \) when, in fact, the actual mean call length on a cellular phone is still 3.25 minutes.

- A Type II error occurs if the researcher fails to reject the hypothesis that the mean length of a phone call on a cellular phone is 3.25 minutes when, in fact, it is longer than 3.25 minutes.

**Methods of Hypothesis Testing**

1. **Confidence Intervals**
2. **Test statistic**
3. **P-values**
   - C.I and P-values always safe to do because don’t need to worry about size of \( n \) (can be bigger or smaller than 30)
One Sample Hypothesis Tests

1. Confidence Interval (can be used only for two-sided tests)

11) You want to test whether your candidate’s approval rating has changed from the previous dismal 40% after a major policy announcement. You run a survey and 170 out of a random sample of 500 voters approve of your candidate, (β = 34%). Construct a hypothesis test using a two-sided confidence interval to test if the approval rating is now different from 40%. Clearly state your conclusion.

\[ H_0 : \text{The approval rating} = 40\% \]
\[ H_1 : \text{The approval rating} \neq 40\% \]
\[ n = 500 \]
\[ \bar{x} = 34\% \]
\[ s = 0.5 \]
\[ t_{\text{stat}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \]
\[ t_{\text{stat}} = \frac{34\% - 40\%}{0.5/\sqrt{500}} \]
\[ t_{\text{stat}} = \frac{-6}{0.0577} \]
\[ t_{\text{stat}} = -10.5 \]

Since 1.767766953 isn't greater than 1.96, we can't reject the null hypothesis. Therefore, at the 5% level of significance, we did not find sufficient evidence to conclude that the percent of employees that are qualified for promotion is different from 40%.

2. Test Statistic Approach (Population Mean)

The Test Statistic
\[ t_{\text{stat}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \]

\[ H_0 : \mu = \mu_0 \quad \text{If} \quad |t_{\text{stat}}| > 1.96 \quad \text{reject} \ H_0 \]
\[ H_1 : \mu \neq \mu_0 \]

\[ H_0 : \mu \geq \mu_0 \quad \text{If} \quad t_{\text{stat}} < -1.64 \quad \text{reject} \ H_0 \]
\[ H_1 : \mu < \mu_0 \]

\[ H_0 : \mu \leq \mu_0 \quad \text{If} \quad t_{\text{stat}} > 1.64 \quad \text{reject} \ H_0 \]
\[ H_1 : \mu > \mu_0 \]

3. Test Statistic Approach (Population Proportion)

The Test Statistic
\[ t_{\text{stat}} = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}/n} \]

\[ H_0 : \pi = \pi_0 \quad \text{If} \quad |t_{\text{stat}}| > 1.96 \quad \text{reject} \ H_0 \]
\[ H_1 : \pi \neq \pi_0 \]

\[ H_0 : \pi \geq \pi_0 \quad \text{If} \quad t_{\text{stat}} < -1.64 \quad \text{reject} \ H_0 \]
\[ H_1 : \pi < \pi_0 \]

\[ H_0 : \pi \leq \pi_0 \quad \text{If} \quad t_{\text{stat}} > 1.64 \quad \text{reject} \ H_0 \]
\[ H_1 : \pi > \pi_0 \]

4. P-Values

\[ \text{a number between 0 and 1} \]
\[ \text{the larger the p-value, the more consistent the data is with the null} \]
\[ \text{the smaller the p-value, the more consistent the data is with the alternative} \]

\[ **\text{If P is low (less than 0.05),} \]
\[ H_0 \text{ must go - reject the null hypothesis} \]

5) The Francis Company is evaluating the promotability of its employees—that is, determining the proportion of employees whose ability, training, and supervisory experience qualify them for promotion to the next level of management. The human resources director of Francis Company tells the president that 80 percent of the employees in the company are "promotable." However, a special committee appointed by the president finds that only 75 percent of the 200 employees who have been interviewed are qualified for promotion. Test \( H_0 : p = 0.8 \quad H_1 : p < 0.8 \) using whatever method you want. Clearly explain your conclusion.

\[ H_0 : x \leq 0.8 \]
\[ H_1 : x > 0.8 \]
\[ *x \geq 1.96 \quad \text{reject} \ H_0 \]
\[ t_{\text{stat}} = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} \]
\[ t_{\text{stat}} = \frac{0.75 - 0.8}{0.0577} \]
\[ t_{\text{stat}} = -10.5 \]

Since 1.767766953 isn't greater than 1.96, we can't reject the null hypothesis. Therefore, at the 5% level of significance, we did not find sufficient evidence to conclude that the percent of employees that are qualified for promotion is different from 80%.

3. A state environmental study concerning the number of scrap-tires accumulated per tire dealership during the past year was conducted. The null hypothesis is \( H_0 : \mu = 2500 \) and the alternative hypothesis is \( H_1 : \mu \neq 2500 \), where \( \mu \) represents the mean number of scrap-tires per dealership in the state. For a random sample of 88 dealerships, the mean is 2725 and the standard deviation is 955.

The p-value for this hypothesis test is 0.0227. Since it is smaller than 0.05, we can reject the null hypothesis and conclude that the average number of accumulated scrap tires is different than 2500.
Two Sample Hypothesis Tests

1. Comparing Two Proportions (Independent Groups)

**Calculate Confidence Interval**

\[ (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \]

B) Many doctors believe that early prenatal care is very important to the health of a baby and its mother. A study has recently been focused on teen mothers. A random sample of 52 teenagers who gave birth revealed that 32 of them began prenatal care in the first trimester of their pregnancy. A random sample of 209 women in their twenties who gave birth revealed that 163 of them began prenatal care in the first trimester of their pregnancy.

a. Construct a 95% confidence interval for the difference between the proportion of teen mothers who get early prenatal care and the proportion of mothers in their twenties who get early prenatal care. (you may do this by hand or Stata, but it would be good practice to do it by hand).

\[
\hat{p}_1 = 32/52 = 0.615384615 \\
\hat{p}_2 = 163/209 = 0.779403506 \\
\]

\[
\left(\hat{p}_1 - \hat{p}_2\right) \pm 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\
\left(0.615384615 - 0.779403506\right) \pm 1.96 \sqrt{\frac{0.615384615(1-0.615384615)}{52} + \frac{0.779403506(1-0.779403506)}{209}} \\
= (-0.164519061) \pm 0.069916938 \\
\]

The confidence interval is (-0.164519061, 0.069916938)

2. Comparing Two Means (large independent samples \(n>30\))

**Calculating Confidence Interval**

\[
\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
\]

**Test Statistic for Two Means**

\[
T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\
\]

\[H_0: \mu_1 = \mu_2 \quad \text{If } |T| < 1.96 \text{ reject } H_0\]

\[H_a: \mu_1 \neq \mu_2 \quad \text{If } |T| > 1.96 \text{ reject } H_0\]

**Test Statistic for Two Proportions**

\[
T = \frac{(\hat{p}_1 - \hat{p}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \\
\text{where } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \\
\]

\[H_0: \pi_1 = \pi_2 \quad \text{If } |T| > 1.96 \text{ reject } H_0\]

\[H_a: \pi_1 \neq \pi_2 \quad \text{If } |T| < -1.96 \text{ reject } H_0\]

11) Are male high school graduates equally likely to attend college the following fall as female high school graduates? A random sample of 1354 males who graduated high school in 2007 found that 860 of them were enrolled in college in 2007. A sample of 1415 females who graduated high school in 2007 found that 996 of them were enrolled in college in 2007. At the 0.05 level of significance, test the null hypothesis that the proportion of male graduates that go on to college is the same proportion of female graduates that go on to college. (you may do this by hand or Stata)

\[n_1 = 1354, \hat{p}_1 = 0.635155606 \\
n_2 = 1415, \hat{p}_2 = 0.705002212 \\
\]

\[
\hat{p} = \frac{860 + 996}{1354 + 1415} = 0.669916938 \\
\]

\[T = \frac{(0.635155606 - 0.705002212)}{0.069916938} = -3.805163016 \\
\]

**Matched Pairs**

**Two samples are DEPENDENT**

Example:

a) Using Stata, construct a 95% confidence interval for the mean of the differences between the scores before the concert and the scores after the concert.

Difference = Sound score Before - Sound score After

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
<th>diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>-3</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Since our 95% confidence interval is (-0.615818761, -0.020850021), all of our values are negative. This means that the proportion of teen mothers who started prenatal care in their first trimester of pregnancy is smaller than the proportion of mothers in their twenties that started prenatal care since \(\hat{p}_1 - \hat{p}_2\) is negative. 0 isn't in the interval, therefore, we are 95% confident that the two proportions aren't equal.
Simple Linear Regression

- used to predict the value of one variable (dependent variable) on the basis of other variables (independent variables)
- $\hat{Y} = b_0 + b_1X$
- Residual: $e = Y - \hat{Y}_{fitted}$
- Fitting error:
  - $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1X_i$
  - $e$ is the part of $Y$ not related to $X$
- Values of $b_0$ and $b_1$ which minimize the residual sum of squares are:
  - (slope) $b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$
  - $b_0 = \overline{Y} - b_1\overline{X}$
- reg temp chirps

Properties of the Residuals and Fitted Values

1. Mean of the residuals = 0; Sum of the residuals = 0
2. Mean of original values is the same as mean of fitted values: $\overline{Y} = \overline{\hat{Y}}$

- Interpretation of slope - for each additional x value (e.g., mile on odometer), the y value decreases/increases by an average of $b_1$ value
- Interpretation of y-intercept - plug in 0 for x and the value you get for $\hat{Y}$ is the y-intercept (e.g., $y = 3.25 - 0.0614X$ if a student who skips no classes has a GPA of 3.25.)
- **danger of extrapolation** - if an x value is outside of our data set, we can't confidently predict the fitted y value

- corr($\hat{\hat{Y}}, e$) = 0

A Measure of Fit: $R^2$

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

$$SST = SSR + SSE$$

- Good fit: if SSR is big, SSE is small
- SST = SSR, perfect fit
- $R^2$ : coefficient of determination
  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- $R^2$ is between 0 and 1, the closer $R^2$ is to 1, the better the fit
- Interpretation of $R^2$ : (e.g., 65% of the variation in the selling price is explained by the variation in odometer reading. The rest 35% remains unexplained by this model)
- **$R^2$ doesn’t indicate whether model is adequate**
- As you add more X's to model, $R^2$ goes up
- Guide to finding SSR, SSE, SST

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>k</td>
<td>SSR</td>
<td>SSR/k</td>
</tr>
<tr>
<td>Error</td>
<td>n-k-1</td>
<td>SSE</td>
<td>SSE/(n-k-1)</td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST</td>
<td></td>
</tr>
</tbody>
</table>
**Assumptions of Simple Linear Regression**

1. We model the AVERAGE of something rather than something itself
   \[ E(Y|X) = \beta_0 + \beta_1 X \]
   where \( E(Y|X) \) is the expected value (average) of \( Y \) for a given \( X \) value.

**ASSUMPTIONS** of the Simple Linear Regression Model

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

- \( \beta_0 + \beta_1 X \) the part of \( Y \) related to \( X \)
- \( \varepsilon \) the part of \( Y \) unrelated to \( X \): \( \varepsilon \sim N(0, \sigma^2) \)

Note: the distribution of \( \varepsilon \) does not depend on \( X \)

- As \( \varepsilon \) (noise) gets bigger, it's harder to find the line

**Estimating \( S_e \)**

\[ S_e^2 = \frac{SSE}{n-2} \]

\( S_e \) is our estimate of \( \sigma^2 \)

\[ S_e = \sqrt{S_e^2} \text{ is our estimate of } \sigma \]

95% of the \( Y \) values should lie within the interval \( b_0 + b_1 X \pm 1.96S_e \)

**Example of Prediction Intervals:**

We are roughly 95% confident that the (average) price of an Accord with 50,000 miles is in the interval

\[ 17066 - 0.06(50000) \pm 1.96(303.14) = (13472, 14660) \]

**Standard Errors for \( b_1 \) and \( b_0 \)**

- standard errors ↑ when noise ↑
- \( s_{b_0} \), amount of uncertainty in our estimate of \( \beta_0 \) (small \( s \) good, large \( s \) bad)
- \( s_{b_1} \), amount of uncertainty in our estimate of \( \beta_1 \)

Confidence Intervals for \( b_1 \) and \( b_0 \)

\[ b_1 \pm 1.96(s_{b_1}) \]

\[ Var(b_1) = s_{b_1}^2 = \frac{s_e^2}{(n-1)s_e^2} \]

\[ b_0 \pm 1.96(s_{b_0}) \]

\[ Var(b_0) = s_{b_0}^2 = s_e^2 \left( 1 + \frac{1}{n} \frac{\bar{X}^2}{s_e^2} \right) \]

- \( n \) small → bad
- \( s_{b_1} \) big → bad
- \( s_e^2 \) small → bad (wants \( x \)'s spread out for better guess)

**Regression Hypothesis Testing**

*always a two-sided test*

- want to test whether slope (\( \beta_1 \)) is needed in our model
- \( H_0 : \beta_1 = 0 \) (don’t need \( x \))
- \( H_a : \beta_1 \neq 0 \) (need \( x \))
- Need \( X \) in the model if:
  a. 0 isn’t in the confidence interval
  b. \( t > 1.96 \)
  c. P-value < 0.05

**Test Statistic for Slope/Y-intercept**

- can only be used if \( n > 30 \)
- if \( n < 30 \), use p-values

\[ T = \frac{b_1 - \beta_1}{s_{b_1}} \]
Multiple Regression

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$
- Variable Importance:
  - Higher t-value, lower p-value = variable is more important
  - Lower t-value, higher p-value = variable is less important (or not needed)

Adjusted R-squared
- $k = \# \text{ of } X's$
- $R^2 = 1 - \frac{\sum_i^n e_i^2}{\sum_i^n (Y_i - \bar{Y})^2} = 1 - \frac{1}{n-1} \frac{SSE}{SST}$
- Adj. R-squared will ↓ as you add junk x variables
- Adj. R-squared will only ↑ if the x you add in is very useful
- **Want Adj. R-squared to go up and Se low for better model

The Overall F Test
- $f = \frac{(SSR) / k}{SSE / (n-k-1)}$
- Always want to reject F test (reject null hypothesis)
- Look at p-value (if $< 0.05$ reject null)
- $H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ (don’t need any X’s)
- $H_a: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k \neq 0$ (need at least 1 X)
- If no x variables needed, then SSR=0 and SST=SSE

Modeling Regression
Backward Stepwise Regression
- 1. Start will all variables in the model
- 2. at each step, delete the least important variable based on largest p-value above 0.05
- 3. stop when you can’t delete anymore
- **Will see Adj. R-squared ↑ and Se ↓

Dummy Variables
- An indicator variable that takes on a value of 0 or 1, allow intercepts to change
- b) We can also run the two sample test using regression. Run the regression
  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  $R^2 = 0.89$
  $\text{adj } R^2 = 0.88$
  $\text{Se} = 3.4$
  $p = 0.001$
  $R^2$ is very high, these variables are strong predictors

Interaction Terms
- allow the slopes to change
- interaction between 2 or more x variables that will affect Y variable

How to Create Dummy Variables (Nominal Variables)
- If C is the number of categories, create (C-1) dummy variables for describing the variable
  - One category is always the “baseline”, which is included in the intercept
  - $\hat{Y} = 30 - 4 \text{Female} + 5 \text{Black} - 2 \text{Other} + 0.3 \text{Edu}$
  1. Women’s self-esteem is 4 points lower than men’s.
  2. Blacks’ self-esteem is 5 points higher than whites’.
  3. Others’ self-esteem is 2 points lower than whites’ and consequently 7 points lower than black.
  4. Each year of education improves self-esteem by 0.3 units.

Recoding Dummy Variables
Example: How many hockey sticks sold in the summer (original equation)
$hockey = 100 + 10 W tr - 20 Spr + 30 F all$
Write equation for how many hockey sticks sold in the winter
$hockey = 110 + 20 F all - 30 Spri - 10 Summer$
- **Always need to get same exact values from the original equation
Regression Diagnostics

Standardize Residuals
\[ r_i = \frac{e_i}{S_e} \sim N(0,1) \]

Check Model Assumptions

Plot residuals versus Yhat

Outliers
- Regression likes to move towards outliers (shows up as \( R^2 \) being really high)
- Want to remove outlier that is extreme in both X and Y

Nonlinearity (ovtest)
- Plotting residuals vs. fitted values will show a relationship if data is nonlinear (\( R^2 \) also high)

Log transformation - accommodates non-linearity, reduces right skewness in the Y, eliminates heteroskedasticity
- **Only take log of X variable

so that we can compare models. Can’t compare models if you take log of Y.

Transformations cheatsheet

- ovtest: a significant test statistic indicates that polynomial terms should be added
  - \( H_0: \) data = no transformation
  - \( H_a: \) data \# no transformation

Multicollinearity
- when X variables are highly correlated with each other.
- \( R^2 > 0.9 \)
- pairwise correlation > 0.9
- correlate all x variables, include y variable, drop the x variable that is less correlated to y

Summary of Regression Output

Guide to Regression Output

- Homoskedastic: band around the values
- Heteroskedastic: as x goes up, the noise goes up (no more band, fan-shaped)
- If heteroskedastic, fix it by logging the Y variable
- If heteroskedastic, fix it by making standard errors robust