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*

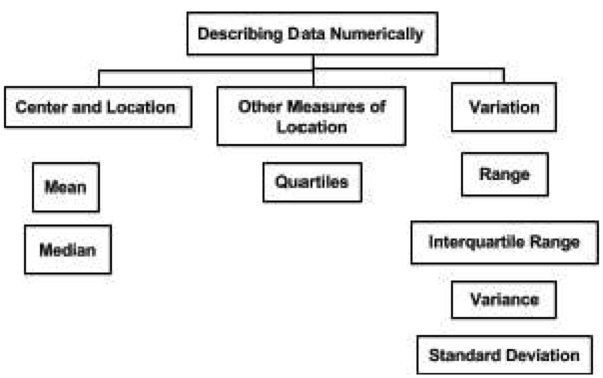
Observational Study - *there can always be lurking variables affecting results*

- ➔ *i.e, strong positive association between shoe size and intelligence for boys*
- ➔ ***should never show causation***

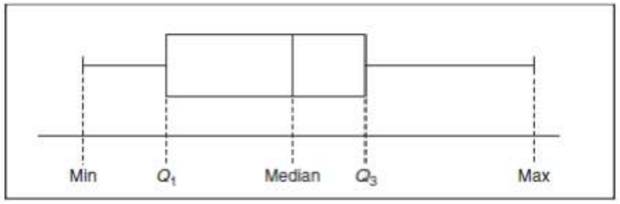
Experimental Study- *lurking variables can be controlled; can give good evidence for causation*

Descriptive Statistics Part I

- ➔ **Summary Measures**



- ➔ **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** = $X_{maximum} - X_{minimum}$
  - ◆ **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*

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

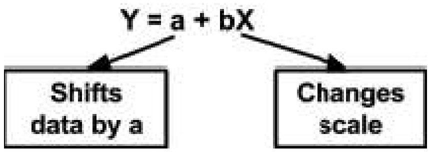
  - ◆ s_x^2 **gets rid of the negative values**
  - ◆ **units are squared**
- ➔ **Standard Deviation** - *shows variation about the mean*

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

  - ◆ **highly affected by outliers**
  - ◆ **has same units as original data**
  - ◆ finance = horrible measure of risk (trampoline example)

Descriptive Statistics Part II

Linear Transformations



- ➔ Linear transformations change the center and spread of data
- ➔ $Var(a + bX) = b^2 Var(X)$
- ➔ *Average(a+bX) = a+b[Average(X)]*

- **Effects of Linear Transformations:**
  - $mean_{new}$ = **a + b*mean**
  - $median_{new}$ = **a + b*median**
  - $stdev_{new} = |b|$ ***stdev**
  - $IQR_{new} = |b|$ ***IQR**
- **Z-score** - *new data set will have mean 0 and variance 1*

$$z = \frac{X-\overline{X}}{S}$$

Empirical Rule

- *Only for mound-shaped data*

Approx. 95% of data is in the interval:

$$(\overline{x} - 2s_x,\ \overline{x} + 2s_x) = \overline{x} + / - 2s_x$$

- *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*
  - $|z| = \left|\frac{X-\overline{X}}{S}\right| \geq 2$
- **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* $|skewness| < 0.8$ **=** *don't need to transform data*

Measurements of Association

- **Covariance**
  - **Covariance > 0 = larger x, larger y**
  - **Covariance < 0 = larger x, smaller y**
  - $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})$
  - **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*
  - $r_{xy} = \frac{covariance_{xy}}{(std.dev._x)(std.dev._y)}$
  - *correlation is between **-1 and 1***
  - *Sign: direction of relationship*
  - *Absolute value: strength of relationship (-0.6 is stronger relationship than +0.4)*

Magnitude of r	Interpretation
.00-.20	Very weak
.20-.40	Weak to moderate
.40-.60	Medium to substantial
.60-.80	Very strong
.80-1.00	Extremely strong

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

Combining Data Sets

- **Mean (Z) =** $\overline{Z} = a\overline{X} + b\overline{Y}$
- **Var (Z) =** $s_z^2 = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$

Portfolios

- **Return on a portfolio:**

$$R_p = w_A\overline{R_A} + w_B\overline{R_B}$$

  - *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)***

Probability Rules

1. *Probabilities range from*
   $0 \le Prob(A) \le 1$
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:**

$P(A \; or \; B) = P(A) + P(B) - P(A \; and \; B)$

Contingency/Joint Table

➔ *To go from contingency to joint table, divide by total # of counts*
➔ *everything inside table adds up to 1*

Conditional Probability

➔ $P(A|B)$
➔ $P(A|B) = \dfrac{P(A \; and \; B)}{P(B)}$
➔ **Given** event B has happened, what is the probability event A will happen?
➔ **Look out for: "given", "if"**

Independence

➔ **Independent if:**
$P(A|B) = P(A)$ **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:**
$P(A \; and \; B) = P(A) \cdot P(B)$

➔ **Another way to find joint probability:**
$P(A \; and \; B) = P(A|B) \cdot P(B)$
$P(A \; and \; B) = P(B|A) \cdot P(A)$

2 x 2 Table

	B	$\bar{B}$
A	$P(A \text{ and } B)$ $= P(B)P(A \mid B)$	$P(A \text{ and } \bar{B})$ $= P(\bar{B})P(A \mid \bar{B})$
$\bar{A}$	$P(\bar{A} \text{ and } B)$ $= P(\bar{A})P(B \mid \bar{A})$	$P(\bar{A} \text{ and } \bar{B})$ $= P(\bar{A})P(\bar{B} \mid \bar{A})$

Called the rule of total probability

$P(A) = P(A \text{ and } B) + P(A \text{ and } \bar{B})$
$= P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})$

Decision Analysis

➔ **Maximax solution =** *optimistic approach. Always think the best is going to happen*
➔ **Maximin solution =** *pessimistic approach.*

Maximin	SIZE OF FIRST STATION	GOOD MARKET ($)	FAIR MARKET ($)	POOR MARKET ($)	Maximax
-10,000	Small	50,000	20,000	–10,000	50,000
-20,000	Medium	80,000	30,000	–20,000	80,000
-40,000	Large	100,000	30,000	–40,000	100,000
-160,000	Very large	300,000	25,000	–160,000	300,000

➔ **Expected Value Solution =**

$EMV = X_1(P_1) + X_2(P_2) ... + X_n(P_n)$

Example: EV (Average factory) = 90(.3) + 120(.5) + (-30)(.2)
= 81

Decision Tree Analysis

➔ **square = your choice**
➔ **circle = uncertain events**

Discrete Random Variables

➔ $P_X(x) = P(X = x)$

Expectation

➔ μ_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

➔ $\mu_s = E(s) = a + b\mu_x$
➔ Var(s)= $b^2 \cdot \sigma^2$

Jointly Distributed Discrete Random Variables

➔ **Independent if:**

$P_{x,y}(X = x \; and \; Y = y) = P_x(x) \cdot P_y(y)$

- **Combining Random Variables**
  - **If X and Y are independent:**

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

  - **If X and Y are dependent:**

$$E(X+Y) = E(X) + E(Y)$$
$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

- **Covariance:**

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

- **If X and Y are independent, Cov(X,Y) = 0**

Calculate the Covariance

- We will use the formula Cov(X,Y)=E(XY)-E(X)E(Y)
- For a die E(X)=E(Y)=3.5
- We need to find E(XY)

Probability	X	Y	XY	Prob × XY
1/6	1	6	6	6/6 = 1
1/6	2	5	10	10/6 = 5/3
1/6	3	4	12	12/6 = 2
1/6	4	3	12	12/6 = 2
1/6	5	2	10	10/6 = 5/3
1/6	6	1	6	6/6 = 1
				$E(XY)$ = sum = $9\frac{1}{3}$ = 9.333

- So 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

$$P(all\ failures) = (1-p)^n$$

2.) All Successes

$$P(all\ successes) = p^n$$

3.) At least one success

$$P(at\ least\ 1\ success) = 1 - (1-p)^n$$

4.) At least one failure

$$P(at\ least\ 1\ failure) = 1 - p^n$$

5.) Binomial Distribution Formula for x=exact value

Binomial Distribution Formula

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

x = number of 'successes' in sample, (x = 0, 1, 2, ..., n)

p = probability of "success" per trial

q = probability of "failure" = (1 – p)

n = number of trials (sample size)

Example: Flip a coin four times, let x = # heads:

n = 4

p = 0.5

q = (1 - .5) = .5

x = 0, 1, 2, 3, 4

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

$$P(x=1) = \frac{20!}{1!(20-1)!}(0.1)^1(0.9)^{19} = 0.27017034353$$
$$P(x=2) = \frac{20!}{2!(20-2)!}(0.1)^2(0.9)^{18} = 0.28517980706$$
$$P(x=3) = \frac{20!}{3!(20-3)!}(0.1)^3(0.9)^{17} = 0.19011987138$$

0.27017034353 + 0.28517980706 + 0.19011987138 = 0.745470022

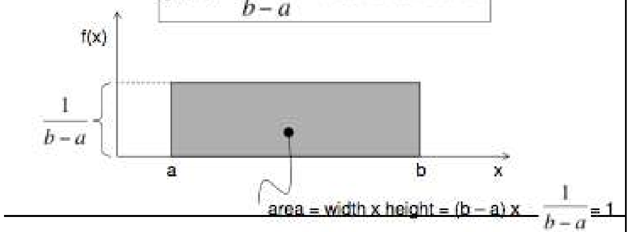
Continuous Probability Distributions

- *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:

$$f(x) = \frac{1}{b-a}, \quad where \quad a \le x \le b$$



- $X \sim Unif(a,b)$

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 two donuts in a month.

$X \sim Unif(a,b)$ $\qquad f(x) = \frac{1}{b-a}$, where $a \le x \le b$

$X \sim Unif(0.5, 4)$ $\qquad f(x) = \frac{1}{3.5}$, where $0.5 \le x \le 4$

$Probability = Area = Width \times Height$

$Probability = 2 \cdot \frac{1}{3.5}$

$Probability = 0.571428571$

(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.

$P(x \geq 2 \mid x \geq 1.5) = \frac{P(x \geq 2 \text{ and } x \geq 1.5)}{P(x \geq 1.5)}$ **OR** $Probability = Area = Width \times Height$

$P(x \geq 2 \mid x \geq 1.5) = \frac{P(x \geq 2)}{P(x \geq 1.5)}$ $Probability = (4-2) \cdot (\frac{1}{4-1.5})$

$Probability = 2 \cdot \frac{1}{2.5}$

$= 0.8$

$Probability = Area = Width \times Height$

$Probability = 2.5 \cdot \frac{1}{3.5} = 0.714285714$

$P(x \geq 2 \mid x \geq 1.5) = \frac{0.571428571}{0.714285714}$

$Probability = 0.8$

- ➔ **Mean for uniform distribution:**

$$E(X) = \frac{(a+b)}{2}$$

- ➔ **Variance for unif. distribution:**

$$Var(X) = \frac{(b-a)^2}{12}$$

Normal Distribution

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

Standardize Normal Distribution:

$$Z = \frac{X-\mu}{\sigma}$$

- ➔ *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 2.0 minutes, what is the probability that the time required for Ang and her bags to get to the room will be:

$X \sim N(12, 2)$

b) between 10.0 and 14.0 minutes?

$$P(10 \leq x \leq 14) = P(x \leq 14) - P(x \leq 10)$$
$$Z = \frac{10-12}{2} = -1$$
$$Z = \frac{14-12}{2} = 1$$
$$P(Z \leq 1) - P(Z \leq -1) = 0.8413 - 0.1587$$
$$= 0.6826$$

Sums of Normals

- If X_1 and X_2 are each normally distributed

$$X_i \sim N(\mu_i, \sigma_i^2)$$

- 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.

$X = Jill's\ score$

$Y = Jack's\ score$

$S = X - Y$

$Let\ S < 0$

$S \sim N(170-160,\ 20^2 + 15^2)$

$S \sim N(10,\ \sqrt{625})$

$P(x < 0) = P[(x-10) \leq (0-10)]$

$P(x < 0) = P[(\frac{x-10}{\sqrt{625}}) \leq (\frac{0-10}{\sqrt{625}})]$

$P(x < 0) = P(Z \leq -0.4)$

$= 0.3446$

- ➔ **Cov(X,Y) = 0 b/c they're independent**

Central Limit Theorem

- ➔ **as n increases,**
- ➔ $\bar{x}$ **should get closer to** μ **(population mean)**
- ➔ **mean**$(\bar{x}) = \mu$
- ➔ **variance** $(\bar{x}) = \sigma^2/n$
- ➔ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
  - ◆ *if population is normally distributed, n can be any value*
  - ◆ *any population, n needs to be* ≥ 30

- ➔ $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

12) The weight of an adult swan is normally distributed with a mean of 30 pounds and a standard deviation of 9.8 pounds. A farmer randomly selected 36 swans and loaded them into his truck. What is the probability that this flock of swans weighs > 1010 pounds?

$X \sim N(30,\ \frac{9.8}{\sqrt{36}})$

$P(\sum_{i=1}^{36} X_i > 1010)$

$P(\bar{x} > \frac{1010}{36}) = P[(\bar{x} - 30) > (1010/36\ - 30)]$

$P(\bar{x} > \frac{1010}{36}) = P[(\frac{\bar{x}-30}{9.8/\sqrt{36}}) > (\frac{1010/36-30}{9.8/\sqrt{36}})]$

$P(\bar{x} > \frac{1010}{36} = P(Z > \ -1.190$

$= 1 - 0.1170\ = 0.883$

Confidence Intervals = *tells us how good our estimate is*

***Want high confidence, narrow interval*

***As confidence increases ↑, interval also increases ↑*

A. One Sample Proportion

Estimate Population Parameter…	with Sample Statistic
Proportion: π	$\hat{p}$

- ➔ $\hat{p} = \frac{x}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$

$$(\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}})$$

- ➔
- ➔ *We are thus 95% confident that the true population proportion is in the interval…*
- ➔ *We are assuming that n is large, n*$\hat{p}$*>5 and our sample size is less than 10% of the population size.*

Standard Error and Margin of Error

- The confidence interval is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The Standard Error

The Margin of Error

- The standard form of any confidence interval is estimate±(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.

$n = 1012$

$\hat{p} = \frac{x}{n} = \frac{901}{1012}$ = **0.890316206**

$\hat{p} - 1.96\sqrt{\frac{\hat{p}q}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$

$0.890316206 - 1.96\sqrt{\frac{(0.890316206)(0.109683794)}{1012}}, 0.890316206 + 1.96\sqrt{\frac{(0.890316206)(0.109683}{1012}}$

= **(0.871062728, 0.909569683)**

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?

$(\hat{p} - 1.96\sqrt{\frac{\hat{p}q}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}})$

$(0.22, 0.24) = 0.1\%\ Margin\ of\ Error$

$\hat{p} - 0.01 = 0.22$

$\hat{p} = 0.23$

$n = \frac{(1.96)^2(0.23)(1-0.23)}{(.01)^2}$

= 6804 people should be used in the sample size

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 σ 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

s

```
. summarize bodytemp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    bodytemp |       106        98.2    .6228963       96.5       99.6

. ci bodytemp

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    bodytemp |        106        98.2     .060501        98.08004    98.31996
```

$s/\sqrt{n}$

This tells us how variable the sample mean is

For samples n < 30

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

The t distribution

Looks like the normal but fatter tails

T Distribution used when:

➔ *σ is not known, n < 30, and data is normally distributed*

Replace the 1.96 value with a **t value** to get:

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

All we are doing is pumping up the volume

where "**t**" comes from **Student's t distribution**, and depends on the sample size through the **degrees of freedom "n-1"**.

****Stata always uses the t-distribution when computing confidence intervals***

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, 2006 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. ($\hat{p}$ = 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: The approval rating = 40%

H_a: The approval rating ≠ 40%

$n = 500$

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.34 - 1.96\sqrt{\frac{(0.34)(1-0.34)}{500}},\ 0.34 + 1.96\sqrt{\frac{(0.34)(1-0.34)}{500}}$$

= (0.298477595, 0.381522405)

```
. cii 500 170, wald

                                                  — Binomial Wald —
    Variable |        Obs        Mean    Std. Err.     [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        500         .34     .0211849     .2984784    .3815216
```

Based off our confidence of (0.2984784, 0.3815216), the null hypothesis of the approval rating = 40% is rejected. There is sufficient evidence to conclude that the approval rating is now different from 40%.

2. Test Statistic Approach (Population Mean)

The Test Statistic $\quad t_{stat} = \dfrac{\bar{X} - \mu_0}{s / \sqrt{n}}$

$H_0 : \mu = \mu_o$ $\quad$ If $|t_{stat}| > 1.96$ $\quad$ reject H_o

$H_a : \mu \neq \mu_o$

$H_0 : \mu \geq \mu_o$ $\quad$ If $t_{stat} < -1.64$ $\quad$ reject H_o

$H_a : \mu < \mu_o$

$H_0 : \mu \leq \mu_o$ $\quad$ If $t_{stat} > 1.64$ $\quad$ reject H_o

$H_a : \mu > \mu_o$

2) A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms. Test the hypothesis that μ=8 kilograms against the alternative that μ ≠ 8 kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms with a standard deviation of 0.5 kilograms. Be sure to clearly state your conclusion.

$n = 50$
$\bar{x} = 7.8$
$s = 0.5$

**If $|t_{stat}| > 1.96$, reject H_0

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t_{stat} = \frac{7.8 - 8}{0.5/\sqrt{50}}$$

$= |-2.828427125| > 1.96$

2.828 > 1.96, therefore we reject the null hypothesis. At the 5% level of significance, we did find sufficient evidence to conclude that the average breaking strength of the fishing line is different than 8 kg.

$$t_{stat} = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} = \frac{3061.56 - 3417}{263.9/\sqrt{137}}$$

```
. ttest det_food=3417

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
det_food |     137    3061.562    22.54632    263.8979    3016.975    3106.149
------------------------------------------------------------------------------
    mean = mean(det_food)                                     t = -15.7648
Ho: mean = 3417                              degrees of freedom =      136

   Ha: mean < 3417            Ha: mean != 3417             Ha: mean > 3417
 Pr(T < t) = 0.0000       Pr(|T| > |t|) = 0.0000         Pr(T > t) = 1.0000
```

3. Test Statistic Approach (Population Proportion)

$$t_{stat} = \frac{(\hat{p} - \pi_o)}{\sqrt{\pi_o(1-\pi_o)/n}}$$

$H_0 : \pi = \pi_o$ $\quad$ If $|t_{stat}| > 1.96$ $\quad$ reject H_o

$H_a : \pi \neq \pi_o$

$H_0 : \pi = \pi_o$ $\quad$ If $t_{stat} < -1.64$ $\quad$ reject H_o

$H_a : \pi < \pi_o$

$H_0 : \pi = \pi_o$ $\quad$ If $t_{stat} > 1.64$ $\quad$ reject H_o

$H_a : \pi > \pi_o$

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 $H_o : p = 0.8$ $H_a : p \neq 0.8$ using whatever method you want. Clearly explain your conclusion.

$H_0 : \pi = 0.8$

$H_a : \pi \neq 0.8$

**If $|t_{stat}| > 1.96$, reject H_0

$$t_{stat} = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

$$t_{stat} = \frac{.75 - .8}{\sqrt{.8(1-.8)/200}}$$

= -1.767766953

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%.

4. P-Values

- a number between 0 and 1
- the larger the p-value, the more consistent the data is with the null
- the smaller the p-value, the more consistent the data is with the alternative
- ***If P is low (less than 0.05), H_0 must go - reject the null hypothesis***

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 Ho: μ = 2500 and the alternative hypothesis is Ha : μ ≠ 2500, where μ represents the mean number of scraptires per dealership in the state. For a random sample of 85 dealerships, the mean is 2725 and the standard deviation is 955.

```
One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      85        2725    103.5843         955    2519.011    2930.989
------------------------------------------------------------------------------
    mean = mean(x)                                            t =   2.1721
Ho: mean = 2500                              degrees of freedom =       84

   Ha: mean < 2500            Ha: mean != 2500             Ha: mean > 2500
 Pr(T < t) = 0.9837       Pr(|T| > |t|) = 0.0327         Pr(T > t) = 0.0163
```

The p-value for this hypothesis test is 0.0327. 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**

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

8) Many doctors believe that early prenatal care is very important to the health of a baby and its mother. Efforts have 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).

$n_1 = 52,\ \hat{p}_1 = \frac{32}{52} = 0.615384615$

$n_2 = 209,\ \hat{p} = \frac{163}{209} = 0.779904306$

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.615384615 - 0.779904306) \pm 1.96\sqrt{\frac{0.615384615(1-0.615384615)}{52} + \frac{0.779904306(1-0.779904306)}{209}}$$

$$(-0.164519691) \pm 1.96\sqrt{0.004551661 + 0.000821309}$$

= (-0.308188761, -0.020850621)

```
. prtesti 52 32 209 163,count

Two-sample test of proportions                  x: Number of obs =       52
                                                y: Number of obs =      209

    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .6153846    .067466                      .4831537    .7476155
           y |   .7799043   .0286585                      .7237347    .8360739
-------------+----------------------------------------------------------------
        diff |  -.1645197   .0733005                      -.3081861   -.0208533
             |  under Ho:   .0673588    -2.44    0.015
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =  -2.4424
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0073         Pr(|Z| < |z|) = 0.0146          Pr(Z > z) = 0.9927
```

Since our 95% confidence interval is (-0.308188761, -0.020850621), all of our values are negative. This means that the proportion of teenage mothers that 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.

➔ **Test Statistic for Two Proportions**

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},\ where\ \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$H_0: p_1 = p_2$ If $|T| > 1.96$ $reject H_o$

$H_a: p_1 \neq p_2$

$H_0: p_1 = p_2$ If $T < -1.64$ $reject H_o$

$H_a: p_1 < p_2$

$H_0: p_1 = p_2$ If $T > 1.64$ $reject H_o$

$H_a: p_1 > p_2$

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 October 2007. A sample of 1415 females who graduated high school in 2007 found that 995 of them were enrolled in college in October 1997. 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 as the proportion of female graduates that go on to college against the two sided alternative. You may do this by hand or Stata. Clearly state your conclusion.

$n_1 = 1354,\ \hat{p}_1 = \frac{860}{1354} = 0.635155096$

$n_2 = 1415,\ \hat{p} = \frac{995}{1415} = 0.703180212$

$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{(1354)(0.635155096)+(1415)(0.703180212)}{1354+1415} = 0.669916938$

***If $|T| > 1.96$, reject H_0*

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

$$T = \frac{(0.635155096 - 0.703180212)}{\sqrt{0.669916938(1-0.669916938)(\frac{1}{1354}+\frac{1}{1415})}} = -3.80516301$$

```
. prtesti 1354 860 1415 995,count

Two-sample test of proportions                  x: Number of obs =     1354
                                                y: Number of obs =     1415

    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .6351551   .0130823                      .6095142     .660796
           y |   .7031802   .0121451                      .6793762    .7269842
-------------+----------------------------------------------------------------
        diff |  -.0680251   .0178508                      -.103012   -.0330382
             |  under Ho:   .0178771    -3.81   0.000
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =  -3.8052
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0001         Pr(|Z| < |z|) = 0.0001          Pr(Z > z) = 0.9999
```

Since our test statistic value of 3.80516301 is greater than 1.96, we can reject the null hypothesis. Looking at our p-value, 0.0001 is less than 0.05, so we can reject the null hypothesis. Therefore, at the 5% level of significance, we find sufficient evidence to conclude that the proportion of male graduates that go on to college is different from the proportion of female graduates.

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

➔ **Calculating Confidence Interval**

$$(\overline{x}_1 - \overline{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

➔ **Test Statistic for Two Means**

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ the test statistic

$H_0: \mu_1 = \mu_2$ If $|T| > 1.96$ $reject H_o$

$H_a: \mu_1 \neq \mu_2$

$H_0: \mu_1 = \mu_2$ If $T < -1.64$ $reject H_o$

$H_a: \mu_1 < \mu_2$

$H_0: \mu_1 = \mu_2$ If $T > 1.64$ $reject H_o$

$H_a: \mu_1 > \mu_2$

Assuming both sample sizes > 30

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

	before	after	diff
1	9	8	1
2	10	8	2
3	9	9	0
4	8	6	2
5	8	6	2
6	9	7	2
7	9	10	-1
8	9	8	1
9	8	5	3
10	10	9	1
11	9	9	0
12	10	8	2
13	8	8	0
14	8	9	-1
15	9	9	0
16	9	7	2
17	9	6	3
18	9	6	3

. summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
before	18	8.888889	.6763995	8	10
after	18	7.666667	1.414214	5	10
diff	18	1.222222	1.308594	-1	3

. ttesti 18 1.222222 1.308594 0

One-sample t test

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	18	1.222222	.3084386	1.308594	.5714735	1.87297

mean = mean(x) t = 3.9626
Ho: mean = 0 degrees of freedom = 17

Ha: mean < 0 Pr(T < t) = 0.9995
Ha: mean != 0 Pr(|T| > |t|) = 0.0010
Ha: mean > 0 Pr(T > t) = 0.0005

Simple Linear Regression

- *used to predict the value of one variable (dependent variable) on the basis of other variables (independent variables)*
- $\hat{Y} = b_0 + b_1 X$
- **Residual:** $e = Y - \hat{Y}_{fitted}$
- **Fitting error:**

  $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_i X_i$
  - **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) $b_1 = r\frac{s_y}{s_x}$

  $b_0 = \overline{Y} - b_1 \overline{X}$

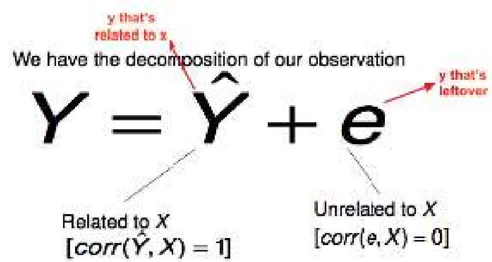
. reg temp chirps

	temp	Coef.	Std. Err.
slope b1	chirps	.8917975	.0247
b0	_cons	40.02525	.7441376

- **Interpretation of slope -** *for each additional x value (e.x. 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.x. y=3.25-0.0614xSkippedClass, 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*

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** $\overline{Y} = \overline{\hat{Y}}$



3.
4. **Correlation Matrix**

Correlation matrix:

$corr(\hat{Y}, e)$

. corr
(obs=100)

	price	odometer	yhat	resid
price	1.0000			
odometer	-0.8063	1.0000		
yhat	0.8063	-1.0000	1.0000	
resid	0.5915	-0.0000	-0.0000	1.0000

$corr(\hat{Y}, X) = 1$ $corr(e, X) = 0$

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

A Measure of Fit: R^2

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

$$\underbrace{SST}_{\text{Total Variation in Y}} = \underbrace{SSR}_{\text{amt. of variation in Y explained by X}} + \underbrace{SSE}_{\text{amt. of variation in Y not explained by X}}$$

- **Good fit: if SSR is big, SEE 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.x. 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

SOURCE	DF	SS	MS
Regression	k	SSR	SSR/k
Error	n-k-1	SSE	SSE/(n-k-1)
Total	n-1	SST	

Assumptions of Simple Linear Regression

1. *We model the AVERAGE of something rather than something itself*
2. $$E(Y|X) = \beta_0 + \beta_1 X$$

where E(Y|X) is the expected value (average) of Y for a given X value.

ASSUMPTIONS of the Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

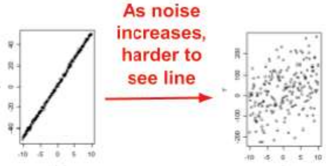
$\beta_0 + \beta_1 X$ the part of *Y* related to *X*

ε the part of *Y* unrelated to *X*: $\varepsilon \sim N(0, \sigma^2)$

Note: the distribution of ε does not depend on *X*

ε is *independent* of *X*.

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



Estimating S_e

- ➔ $S_e^2 = \frac{SSE}{n-2}$
- ➔ S_e^2 *is our estimate of* σ^2
- ➔ $S_e = \sqrt{S_e^2}$ *is our estimate of* σ
- ➔ **95% of the Y values should lie within the interval** $b_0 + b_1 X \pm 1.96 S_e$

S_e

Number of obs	=	100
F(1, 98)	=	182.11
Prob > F	=	0.0000
R-squared	=	0.6501
Adj R-squared	=	0.6466
Root MSE	=	303.14

Example of Prediction Intervals:

```
. scatter price odometer
. regress price odometer
```

Source	SS	df	MS
Model	16734110.9	1	16734110.9
Residual	9005449.88	98	91892.3457
Total	25739560.8	99	259995.563

*Se

Number of obs	=	100
F(1, 98)	=	182.11
Prob > F	=	0.0000
R-squared	=	0.6501
Adj R-squared	=	0.6466
Root MSE	=	303.14

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
odometer	-.0623155	.0046178	-13.49	0.000	-.0714793	-.0531516
_cons	17066.77	169.0246	100.97	0.000	16731.34	17402.19

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 β_0 (small s good, large s bad)**
- ➔ s_{b_1} **amount of uncertainty in our estimate of β_1**

s_{b_0}

```
. scatter price odometer
. regress price odometer
```

Source	SS	df	MS
Model	16734110.9	1	16734110.9
Residual	9005449.88	98	91892.3457
Total	25739560.8	99	259995.563

price	Coef.	Std. Err.	t	P>|t|
odometer	-.0623155	.0046178	-13.49	0.000
_cons	17066.77	169.0246	100.97	0.000

s_{b_1}

Confidence Intervals for b_1 and b_0

- ➔ $b_1 \pm 1.96(s_{b_1})$

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

- ➔
- ➔ $b_0 \pm 1.96(s_{b_0})$

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

- ➔
- ➔ ***n small → bad***

s_e big → bad

s_x^2 small → bad (wants x's spread out for better guess)

Regression Hypothesis Testing

***always a two-sided test**

- ➔ **want to test whether slope (β_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**

$$T = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

$H_0: \beta_1 = \beta_1^*$ If $|T| > 1.96$ *reject H_o*

$H_a: \beta_1 \neq \beta_1^*$

$H_0: \beta_1 \geq \beta_1^*$ If $T < -1.64$ *reject H_o*

$H_a: \beta_1 < \beta_1^*$

$H_0: \beta_1 \leq \beta_1^*$ If $T > 1.64$ *reject H_o*

$H_a: \beta_1 > \beta_1^*$

$H_0: \beta_1 = 0$ $\frac{b_1}{s_{b_1}}$

Source	SS	df	MS
Model	.365486678	1	.365486678
Residual	.449667193	36	.012490755
Total	.815153871	37	.022031186

Number of obs	=	38
F(1, 36)	=	29.26
Prob > F	=	0.0000
R-squared	=	0.4484
Adj R-squared	=	0.4330
Root MSE	=	.11176

anf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sp500	1.611712	.2979518	5.41	0.000	1.007438	2.215987
_cons	.0005632	.0181477	0.03	0.975	-.036242	.0373684

$H_0: \beta_0 = 0$ $\frac{b_0}{s_{b_0}}$

P-values

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 = # of X's**

$$R_a^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n-k-1}SSE}{\frac{1}{n-1}SST}$$

- **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)**
- $H_0: \beta_1 = \beta_2 = \beta_3 ... = \beta_k = 0$ **(don't need any X's)**

  $H_a: \beta_1 = \beta_2 = \beta_3 ... = \beta_k \neq 0$ **(need at least 1 X)**
- **If no x variables needed, then SSR=0 and SST=SSE**

$H_o: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

vs.

H_a: *At least one* $\beta_i \neq 0$

Conclusion?

For those interested……40.03=1902.47/47.52

```
. regress price size age lotsize                SSR/k

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  3,    11) =   40.03
       Model |  5707.43856     3  1902.47952           Prob > F      =  0.0000
    Residual |  522.797622    11  47.5270565           R-squared     =  0.9161
-------------+------------------------------           Adj R-squared =  0.8932
       Total |  6230.23618    14  445.01687            Root MSE      =   6.894
                          SSE/(n-k-1)
------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        size |   4.146191   .7511855     5.52   0.000     2.492843    5.799539
         age |  -.2360837   .8812207    -0.27   0.794    -2.175637     1.70347
     lotsize |   4.830881    .901075     5.36   0.000     2.847628    6.814134
       _cons |  -16.05802   19.07105    -0.84   0.418    -58.03311    25.91707
------------------------------------------------------------------------------
```

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 t-test using regression. Run the regression $income = \beta_0 + \beta_1(female) + \varepsilon$

```
. regress income female

      Source |       SS       df       MS              Number of obs  =     500
-------------+------------------------------           F(1, 498)      =   57.60
       Model |  4718.27891     1  4718.27891           Prob > F       =  0.0000
    Residual |  40792.3586   498  81.9123666           R-squared      =  0.1037
-------------+------------------------------           Adj R-squared  =  0.1019
       Total |  45510.6375   499  91.2036824           Root MSE       =  9.0505
          b1
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -6.174747   .8135835    -7.59   0.000    -7.773227   -4.576268
       _cons |   27.81111   .6033697    46.09   0.000     26.62565    28.99658
------------------------------------------------------------------------------
       b0
```

i) Interpret the coefficients from this regression

$\beta_0 = 27.81111$ and is referred to the baseline value. This value is the average income for males.

$\beta_1 = -6.174747$ and this value represents the expected difference between a female's income and a male's income.

ii) Show that you obtain the same average values as you did in part(a)

$income = \beta_0 + \beta_1(female) + \varepsilon$

average income for male → $income = \beta_0 + \beta_1(female) + \varepsilon$

$income = (27.81111) + (-6.174747)(0)$

$income = 27.81111$

average income for female → $income = \beta_0 + \beta_1(female) + \varepsilon$

$income = (27.81111) + (-6.174747)(1)$

$income = 21.63636$

Interaction Terms

- **allow the slopes to change**
- **interaction between 2 or more x variables that will affect the 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***

$\hat{Y} = 30 - 4Female + 5Black - 2Other + 0.3Edu$

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 blacks'.
4. Each year of education improves self-esteem by 0.3 units.

*Make sure you get into the habit of saying the slope is the effect of an independent variable **"while holding everything else constant."***

Recoding Dummy Variables

Example: How many hockey sticks sold in the summer (original equation)

$hockey = 100 + 10Wtr - 20Spr + 30Fall$

Write equation for how many hockey sticks sold in the winter

$hockey = 110 + 20Fall - 30Spri - 10Summer$

- ****always need to get same exact values from the original equation**

Regression Diagnostics

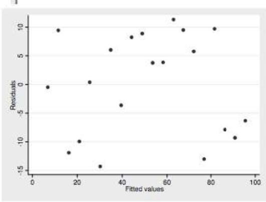
Standardize Residuals

$$r_i = \frac{e_i}{s_e} \approx \frac{\varepsilon_i}{\sigma} \sim N(0,1)$$

Check Model Assumptions

➔ **Plot residuals versus Yhat**

```
. scatter res yhat
```



This is the way a residual plot looks when the model fits the data:

No obvious pattern!!!!!

resids unrelated to X!!!!!!

➔ **Outliers**

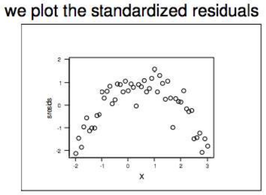
- **Regression likes to move towards outliers (shows up as R^2 being really high)**
- **want to remove outlier that is extreme in both x and y**

➔ **Nonlinearity (ovtest)**

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

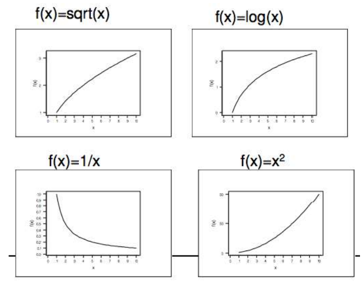
As a *diagnostic*, we plot the standardized residuals versus X:

there should be no relationship between the resids and X!!!!



The nonlinearity is even more evident in the residual plot !! What is wrong with fitting a linear regression to this data?

- **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*

```
. ovtest

Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                 F(3, 6044) =    158.43
                   Prob > F =      0.0000
```

➔ **Normality (sktest)**

- H_0 : *data = normality*

  H_a : *data ≠ normality*
- **don't want to reject the null hypothesis. P-value should be big**

```
. sktest res
```

Skewness/Kurtosis tests for Normality

Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
res	0.869	0.046	4.25	0.1195

➔ **Homoskedasticity (hettest)**

- H_0 : *data = homoskedasticity*
- H_a : *data ≠ homoskedasticity*
- **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**

➔ **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