Important Concepts not on the AP Statistics Formula Sheet

Part I:

IQR = $Q_3 - Q_1$
Test for an outlier:
1.5(IQR) above Q_3 or below Q_1
The calculator will run the test for you as long as you choose the boxplot with the oulier on it in STATPLOT

Linear transformation:
Addition: affects center NOT spread
adds to $\bar{x}$, M, Q_1, Q_3, IQR

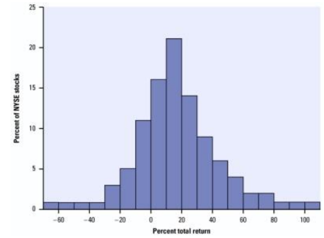
not σ

Multiplication: affects both center and spread
multiplies $\bar{x}$, M, Q_1, Q_3, IQR, σ

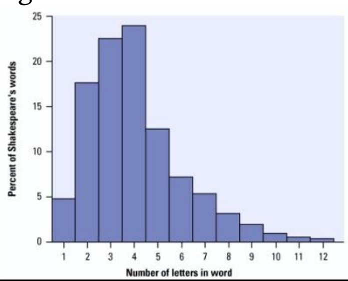
When describing data: describe center, spread, and shape.

Give a 5 number summary or mean and standard deviation when necessary.

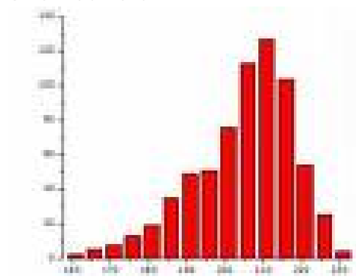
Histogram:
fairly symmetrical
unimodal



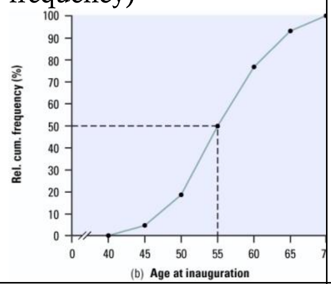
skewed
right



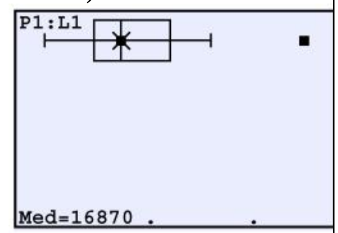
Skewed left



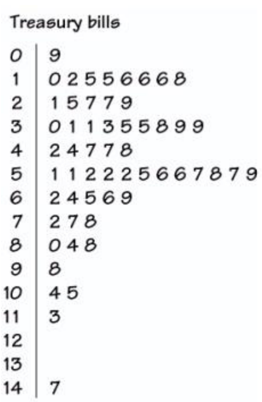
Ogive (cumulative frequency)



Boxplot (with an outlier)



Stem and leaf

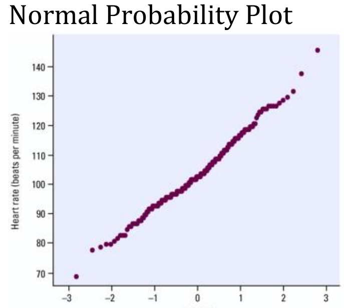


Treasury bills

0	9
1	02556668
2	15779
3	011355899
4	24778
5	1122256678 79
6	24569
7	278
8	048
9	8
10	45
11	3
12	
13	
14	7

(b)

Normal Probability Plot



The 80th percentile means that 80% of the data is below that observation.

$$z = \frac{x - mean}{standard\ dev}$$

or

$$z = \frac{x - \mu}{\sigma}$$

HOW MANY STANDARD DEVIATIONS AN OBSERVATION IS FROM THE MEAN

68-95-99.7 Rule for Normality
N(μ,σ)
N(0,1) Standard Normal

r: correlation coefficient, The strength of the linear relationship of data. Close to 1 or -1 is very close to linear

r^2: coefficient of determination. How well the model fits the data. Close to 1 is a good fit. "Percent of variation in y described by the LSRL on x"

residual = $y - \hat{y}$

residual = observed – predicted

y = a+bx
Slope of LSRL(b): rate of change in y for every unit x

y-intercept of LSRL(a): y when x = 0

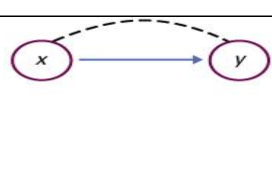
Exponential Model:
$y = ab^x$ take log of y

Power Model:
$y = ax^b$ take log of x and y

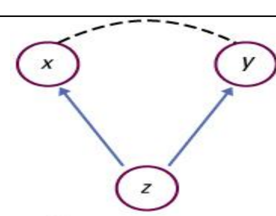
Explanatory variables explain changes in response variables.
EV: x, independent
RV: y, dependent

Lurking Variable: A variable that may influence the relationship bewteen two variables.
LV is not among the EV's

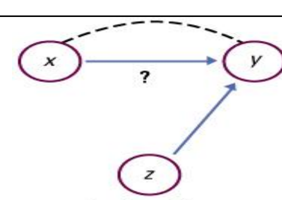
Confounding: two variables are confounded when the effects of an RV cannot be distinguished.



Causation
(a)



Common response
(b)



Confounding
(c)

Part II: Designing Experiments and Collecting Data:

Sampling Methods:

The Bad:
Voluntary sample. A voluntary sample is made up of people who decide for themselves to be in the survey.
Example: Online poll
Convenience sample. A convenience sample is made up of people who are easy to reach.
Example: interview people at the mall, or in the cafeteria because it is an easy place to reach people.

The Good:
Simple random sampling. Simple random sampling refers to a method in which all possible samples of n objects are equally likely to occur.
Example: assign a number 1-100 to all members of a population of size 100. One number is selected at a time from a list of random digits or using a random number generator. The first 10 selected are the sample.
Stratified sampling. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a SRS is taken. In stratified sampling, the groups are called **strata**.
Example: For a national survey we divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.
Cluster sampling. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen using a SRS. Only individuals within sampled clusters are surveyed.
Example: Randomly choose high schools in the country and only survey people in those schools.
<u>Difference</u> between cluster sampling and stratified sampling. With stratified sampling, the sample includes subjects from each stratum. With cluster sampling the sample includes subjects only from sampled clusters.
Multistage sampling. With multistage sampling, we select a sample by using combinations of different sampling methods.
Example: Stage 1, use cluster sampling to choose clusters from a population. Then, in Stage 2, we use simple random sampling to select a subset of subjects from each chosen cluster for the final sample.
Systematic random sampling. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* subjects on the population list. Thereafter, we select every *kth* subject on the list.
Example: Select every 5^{th} person on a list of the population.

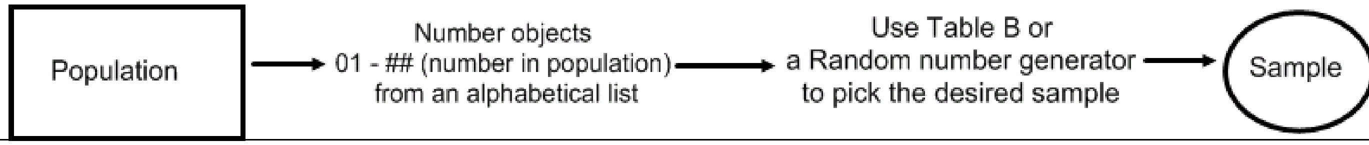
<u>Experimental Design:</u>

A well-designed experiment includes design features that allow researchers to eliminate extraneous variables as an explanation for the observed relationship between the independent variable(s) and the dependent variable.
Experimental Unit or Subject: The individuals on which the experiment is done. If they are people then we call them subjects
Factor: The explanatory variables in the study
Level: The degree or value of each factor.
Treatment: The condition applied to the subjects. When there is one factor, the treatments and the levels are the same.
Control. Control refers to steps taken to reduce the effects of other variables (i.e., variables other than the independent variable and the dependent variable). These variables are called **lurking variables**.
Control involves making the experiment as similar as possible for subjects in each treatment condition. Three control strategies are control groups, placebos, and blinding.
Control group. A control group is a group that receives no treatment
Placebo. A fake or dummy treatment.
Blinding: Not telling subjects whether they receive the placebo or the treatment
Double blinding: neither the researchers or the subjects know who gets the treatment or placebo
Randomization. Randomization refers to the practice of using chance methods (random number tables, flipping a coin, etc.) to assign subjects to treatments.
Replication. Replication refers to the practice of assigning each treatment to many experimental subjects.
Bias: when a method systematically favors one outcome over another.

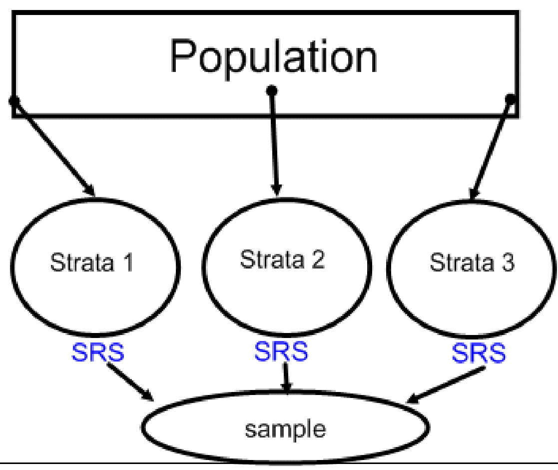
Types of design:
Completely randomized design With this design, subjects are randomly assigned to treatments.
Randomized block design, the experimenter divides subjects into subgroups called **blocks**. Then, subjects within each block are randomly assigned to treatment conditions. Because this design reduces variability and potential confounding, it produces a better estimate of treatment effects.
Matched pairs design is a special case of the randomized block design. It is used when the experiment has only two treatment conditions; and subjects can be grouped into pairs, based on some blocking variable. Then, within each pair, subjects are randomly assigned to different treatments.

Part II in Pictures: Sampling Methods

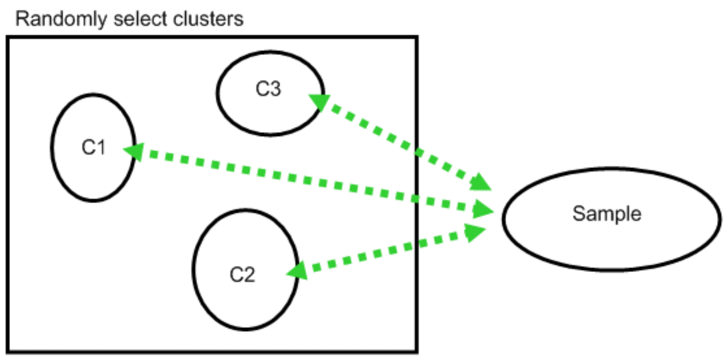
Simple Random Sample: Every group of n objects has an equal chance of being selected.



Stratified Random Sampling:
Break population into strata (groups) then take an SRS of each group.



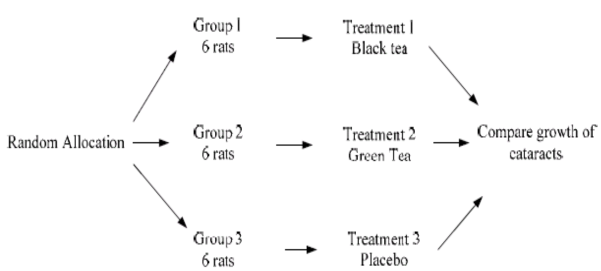
Cluster Sampling:
Randomly select clusters then take all Members in the cluster as the sample.



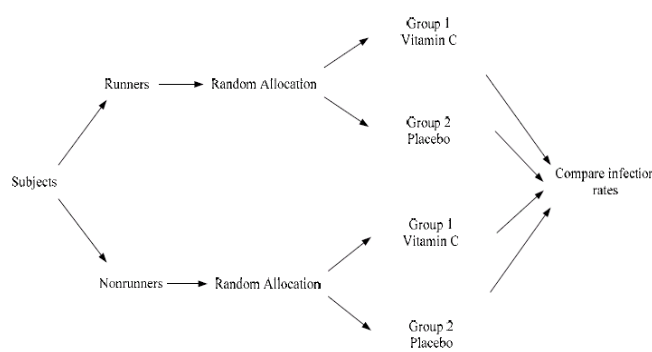
Systematic Random Sampling:
Select a sample using a system, like selecting every third subject.

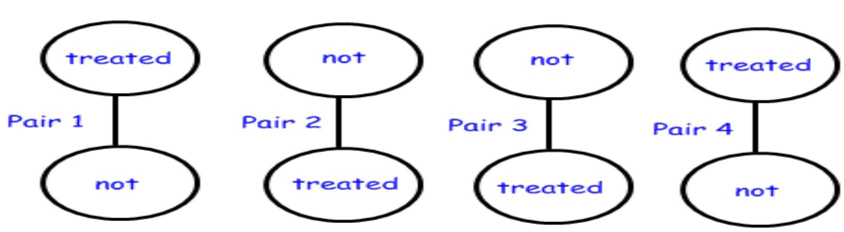
Experimental Design:

Completely Randomized Design:



Randomized Block Design:





Matched Pairs Design:

Part III: Probability and Random Variables:

Counting Principle:
Trial 1: a ways
Trial 2: b ways
Trial 3: c ways …
The there are a x b x c ways to do all three.

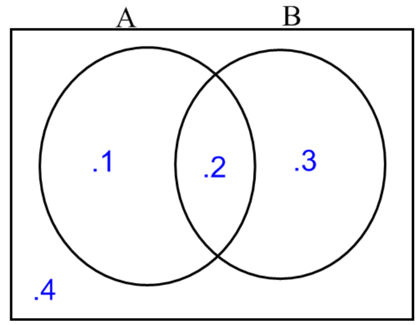
$0 \le P(A) \le 1$

$1 - P(A) = P(A^c)$

A and B are **disjoint** or **mutually exclusive** if they have no events in common.
Roll two die: DISJOINT
rolling a 9
rolling doubles
Roll two die: not disjoint
rolling a 4
rolling doubles

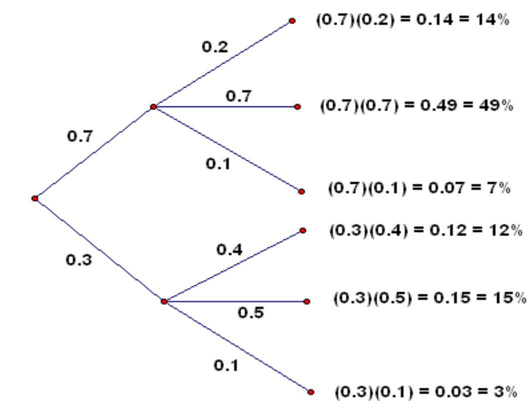
A and B are **independent** if the outcome of one does not affect the other.

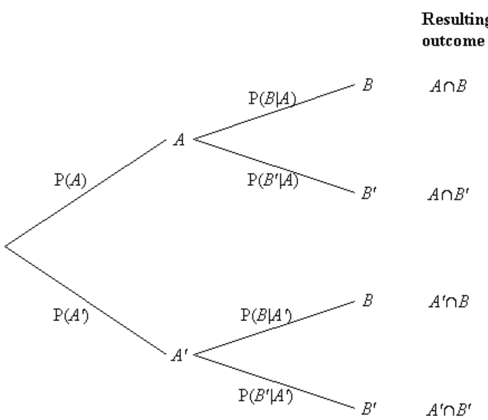
Mutually Exclusive events <u>CANNOT</u> BE Independent



P(A) = 0.3
P(B) = 0.5
P(A ∩B) = 0.2
P(A U B) = 0.3 + 0.5 - 0.2 = 0.6
P(A|B) = 0.2/0.5 = 2/5
P(B|A) = 0.2 /0.3 = 2/3

For Conditional Probability use a TREE DIAGRAM:





For Binomial Probability:
Look for x out of n trials
1. Success or failure
2. Fixed n
3. Independent observations
4. p is the same for all observations

P(X=3) Exactly 3
use binompdf(n,p,3)
P(X≤ 3) at most 3
use binomcdf(n,p,3) (Does 3,2,1,0)
P(X≥3) at least 3 is 1 - P(X≤2)
use 1 - binomcdf(n,p,2)

Normal Approximation of Binomial:
for np ≥ 10 and n(1-p) ≥ 10
the X is approx N(np,$\sqrt{np(1-p)}$)

Discrete Random Variable: has a countable number of possible events (Heads or tails, each .5)
Continuous Random Variable: Takes all values in an interval: (EX: normal curve is continuous)
Law of large numbers. As n becomes very large $\bar{x} \to \mu$

Linear Combinations:

$\mu_{a+bx} = a + b\mu_x$

$\mu_{X+Y} = \mu_x + \mu_Y$

$\sigma^2_{a+bx} = b^2\sigma^2_X$

$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ $\quad\quad$ $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

Geometric Probability:
Look for # trial until first success
1. Success or Failure
2. X is trials until first success
3. Independent observations
4. p is same for all observations

P(X=n) = $p(1-p)^{n-1}$
μ is the expected number of trails until the first success or $\frac{1}{p}$

$\sigma^2 = \frac{1-p}{p^2}$

P(X > n) = $(1 - p)^n$ = 1 - P(X ≤ n)

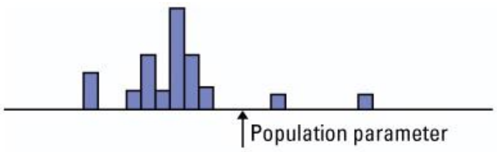
Sampling distribution: The distribution of all values of the statistic in all possible samples of the same size from the population.

Central Limit Theorem: As n becomes very large the sampling distribution for $\bar{x}$ is approximately NORMAL

Use ($n \geq 30$) for CLT

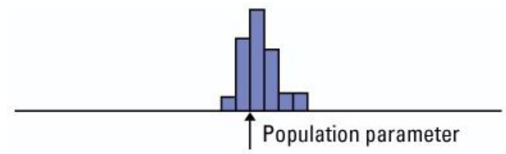
Low Bias: Predicts the center well
Low Variability: Not spread out

High Bias: Does not predict center well
High Variability: Is very spread out



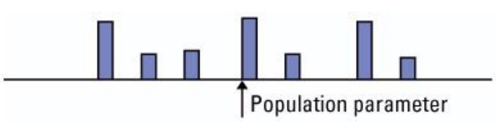
(a)

High bias, High Variability



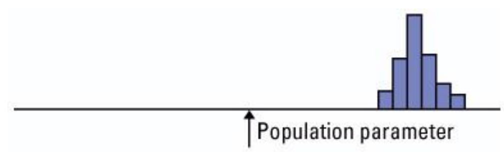
(b)

Low Bias, Low Variability



(c)

Low Bias, High Variability



(d)

High Bias, Low Variability

See other sheets for Part IV

ART is my BFF

Type I Error: Reject the null hypothesis when it is actually True

Type II Error: Fail to reject the null hypothesis when it is False.

ESTIMATE – DO A CONFIDENCE INTERVAL **EVIDENCE - DO A TEST**

<table>
<tr>
<td>Paired Procedures
Must be from a matched pairs design:
Sample from one population where each subject receives two treatments, and the observations are subtracted. OR
Subjects are matched in pairs because they are similar in some way, each subject receives one of two treatments and the observations are subtracted</td>
<td>Two Sample Procedures
Two independent samples from two different populations OR
Two groups from a randomized experiment (each group would receive a different treatment) Both groups may be from the same population in this case but will randomly receive a different treatment.</td>
</tr>
</table>

Major Concepts in Probability

For the expected value (mean, μ_X) and the σ_X or σ_X^2 of a probability distribution use the formula sheet

Binomial Probability

Fixed Number of Trials
Probability of success is the same for all trials
Trials are independent

If X is B(n,p) then (ON FORMULA SHEET)
Mean $\mu_X = np$

Standard Deviation $\sigma_X = \sqrt{np(1-p)}$
For Binomial probability use $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ or use:

Exactly: $P(X = x)$ = binompdf(n , p, x)

At Most: $P(X \le x)$ = binomcdf(n , p, x)

At least: $P(X \ge x)$ = 1 - binomcdf(n , p, x-1)

More than: $P(X > x)$ = 1 - binomcdf(n , p, x)

Less Than: $P(X < x)$ = binomcdf(n , p, x-1)

You may use the normal approximation of the binomial distribution when $np \ge 10$ and $n(1-p) \ge 10$. Use then mean and standard deviation of the binomial situation to find the Z score.

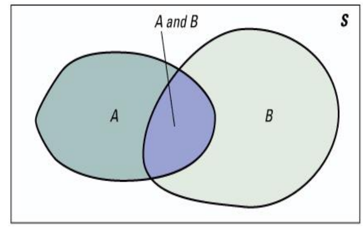
Simple Probability (and, or, not):

Finding the probability of multiple simple events.
Addition Rule: P(A or B) = P(A) + P(B) – P(A and B)
Multiplication Rule: P(A and B) = P(A)P(B|A)

Mutually Exclusive events CANNOT be independent
A and B are **independent** if the outcome of one does not affect the other.
A and B are **disjoint** or **mutually exclusive** if they have no events in common.

Roll two die: DISJOINT
rolling a 9
rolling doubles

Roll two die: NOT disjoint
rolling a 4
rolling doubles



Independent: P(B) = P(B|A)
Mutually Exclusive: P(A and B) = 0

Geometric Probability

You are interested in the amount of trials it takes UNTIL you achieve a success.
Probability of success is the same for each trial
Trials are independent

Use simple probability rules for Geometric Probabilities.

$P(X=n) = p(1-p)^{n-1}$ $\quad$ $P(X > n) = (1 - p)^n = 1 - P(X \le n)$
μ_X is the expected number of trails until the first success or $\frac{1}{p}$

Conditional Probability

Finding the probability of an event given that another even has already occurred.

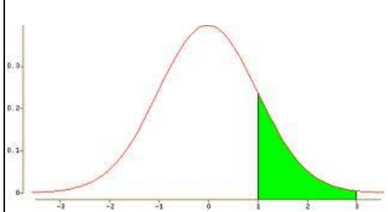
Conditional Probability: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$

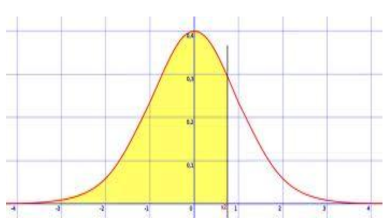
Use a two way table or a Tree Diagram for Conditional Problems.
Events are Independent if P(B|A) = P(B)

Normal Probability

For a single observation from a normal population

$$P(X > x) = P(z > \frac{x-\mu}{\sigma}) \qquad P(X < x) = P(z < \frac{x-\mu}{\sigma})$$





To find $P(x < X < y)$ Find two Z scores and subtract the probabilities (upper – lower)

Use the table to find the probability or use normalcdf(min,max,0,1) after finding the z-score

<u>For the **mean** of a random sample of size n from a population.</u>

When $n > 30$ the sampling distribution of the sample mean $\bar{x}$ is approximately Normal with:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

If $n < 30$ then the population should be Normally distributed to begin with to use the z-distribution.

$$P(\bar{X} > x) = P(z > \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}) \qquad P(\bar{X} < x) = P(z < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}})$$

To find $P(x < X < y)$ Find two Z scores and subtract the probabilities (upper – lower)
Use the table to find the probability or use normalcdf(min,max,0,1) after finding the z-score

Mutually Exclusive vs. Independence

You just heard that Dan and Annie who have been a couple for three years broke up.
This presents a problem, because you're having a big party at your house this Friday night and you have invited them both. Now you're afraid there might be an ugly scene if they both show up.
When you see Annie, you talk to her about the issue, asking her if she remembers about your party.
She assures you she's coming. You say that Dan is invited, too, and you wait for her reaction.
If she says, "That jerk! If he shows up I'm not coming. I want nothing to do with him!", they're **mutually exclusive**.
If she says, "Whatever. Let him come, or not. He's nothing to me now.", they're **independent**.

Mutually Exclusive and Independence are two very different ideas

Mutually Exclusive (disjoint): ***P(A and B) = 0***	**Independence:** ***P(B) = P(B|A)***
Events A and B are mutually exclusive if they have no outcomes in common. That is A and B cannot happen at the same time. Example of **mutually exclusive (disjoint)**: A: roll an odd on a die B: roll an even on a die Odd and even **share no outcomes** P(odd and even) = 0 Therefore, they are mutually exclusive. Example of **not mutually exclusive (joint):** A: draw a king B: draw a face card King and face card **do share outcomes**. All of the kings are face cards. P(king and face card) = 4/52 Therefore, they are not mutually exclusive.	Events A and B are independent if knowing one outcome does not change the probability of the other. That is knowing A does not change the probability of B. Examples of **independent** events: A: draw an ace B: draw a spade P(Spade) =13/52 = 1/4 P(Spade | Ace) = 1/4 Knowing that the drawn card is an ace **does not change** the probability of drawing a spade Examples that are **dependent** (not independent): A: roll a number greater than 3 B: roll an even P(even) = 3/6 = 1/2 P(even | greater than 3) = 2/3 Knowing the number is greater than three **changes** the probability of rolling an even number.

Mutually Exclusive events cannot be independent	Independent events cannot be Mutually Exclusive	Dependent Events may or may not be mutually exclusive
Mutually exclusive and dependent A: Roll an even B: Roll an odd They share no outcomes and knowing that it is odd changes the probability of it being even.	**Independent and not mutually exclusive** A: draw a black card B: draw a king Knowing it is a black card does not change the probability of it being a king and they do share outcomes.	**Dependent and mutually exclusive** A: draw a queen B: draw a king Knowing it is a queen changes the probability of it being a king and they do not share outcomes. **Dependent and not mutually exclusive** A: Face Card B: King Knowing it is a face card changes the probability of it being a king and they do share outcomes.

If events are mutually exclusive then: P(A or B) = P(A) + P(B) If events are not mutually exclusive use the general rule: P(A or B) = P(A) + P(B) – P(A and B)	If events are independent then: P(A and B) = P(A)P(B) If events are not independent then use the general rule: P(A and B) = P(A)P(B|A)

Interpretations from Inference

Interpretation for a Confidence Interval: I am C% confident that the true parameter (mean μ or proportion p) lies between # and #. **INTERPRET IN CONTEXT!!**
Interpretation of C% Confident: Using my method, If I sampled over and over again, C% of my intervals would contain the true parameter (mean μ or proportion p). **NOT**: The parameter lies in my interval C% of the time. It either does or does not!!
If $p < \alpha$ I **reject** the null hypothesis H_0 and I have **sufficient/strong** evidence to support the alternative hypothesis H_a **INTERPRET IN CONTEXT in terms of the alternative.**
If $p > \alpha$ I **fail to reject** the null hypothesis H_0 and I have **insufficient/poor** evidence to support the alternative hypothesis H_a **INTERPRET IN CONTEXT in terms of the alternative.**
Evidence Against H_o **P-Value** "Some" $0.05 < \text{P-Value} < 0.10$ "Moderate or Good" $0.01 < \text{P-Value} < 0.05$ "Strong" $\text{P-Value} < 0.01$
Interpretation of a p-value: The probability, assuming the null hypothesis is true, that an observed outcome would be as extreme or more extreme than what was actually observed.
Duality: Confidence intervals and significance tests. If the hypothesized parameter lies **outside** the C% confidence interval for the parameter I can REJECT H_0 If the hypothesized parameter lies **inside** the C% confidence interval for the parameter I FAIL TO REJECT H_0
Power of test: The probability that at a fixed level α test will reject the null hypothesis when and alternative value is true.

Confidence Intervals

<table>
<tr>
<td>

<u>One Sample Z Interval</u>

Use when estimating a single population mean and σ is known
Conditions:
-SRS
-Normality: CLT, stated, or plots
-Independence and N≥10n
Interval:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

CALCULATOR:
7: Z-Interval

</td>
<td>

<u>One Sample t Interval</u>

Use when estimating a single mean and σ is NOT known
[Also used in a <u>matched paired design</u> for the mean of the difference: *PAIRED t PROCEDURE*]
Conditions:
-SRS
-Normality: CLT, stated, or plots
-Independence and N≥10n
Interval:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

df = n-1
CALCULATOR:
8: T-Interval

</td>
<td>

<u>One Proportion Z Interval</u>

Use when estimating a single proportion
Conditions:
-SRS
-Normality: $n\hat{p} \geq 10,\ n(1-\hat{p}) \geq 10$
-Independence and N≥10n
Interval:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

CALCULATOR:
A: 1-PropZInt

</td>
</tr>
<tr>
<td>

<u>Two Sample Z Interval</u>
<u>σ known (RARELY USED)</u>

Use when estimating the difference between two population means and σ is known
Conditions:
-SRS for both populations
-Normality: CLT, stated, or plots for -both populations
-Independence and N≥10n for both populations.
Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

CALCULATOR:
9: 2-SampZInt

</td>
<td>

<u>Two Sample t Interval</u>
<u>σ unknown</u>

Use when estimating the difference between two population means and σ is NOT known
Conditions:
-SRS for both populations
-Normality: CLT, stated, or plots for both populations
-Independence and N≥10n for both populations.
Interval:

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

Use lower n for df (df = n-1) or use calculator
CALCULATOR:
0: 2-SampTInt

</td>
<td>

<u>Two Proportion Z Interval</u>

Use when estimating the difference between two population proportions.
Conditions:
-SRS for both populations
-Normality:
$n_1\hat{p}_1 \geq 10\ \ n_1(1-\hat{p}_1) \geq 10$
$n_2\hat{p}_2 \geq 10\ \ n_2(1-\hat{p}_2) \geq 10$
-Independence and N≥10n for both populations.
Interval:

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

CALCULATOR:
B: 2-PropZInt

</td>
</tr>
<tr>
<td colspan="3">

<u>Confidence interval for Regression Slope:</u>

Use when estimating the slope of the true regression line
Conditions:
1. Observations are independent
2. Linear Relationship (look at residual plot)
3. Standard deviation of y is the same(look at residual plot)
4. y varies normally (look at histogram of residuals)
Interval:

$$b \pm t^* SE_b$$

df = n - 2
CALCULATOR: **LinRegTInt** Use technology readout or calculator for this confidence interval.

</td>
</tr>
</table>

Hypothesis Testing

One Sample Z Test

Use when testing a mean from a single sample and **σ** is known

H_o: $\mu = \#$

H_a: $\mu \neq \#$ **-or-** $\mu < \#$ **-or-** $\mu > \#$

Conditions:

-SRS

-Normality: Stated, CLT, or plots

-Independence and N≥10n

Test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

CALCULATOR: **1: Z-Test**

One Sample t Test

Use when testing a mean from a single sample and σ is NOT known

[Also used in a matched paired design for the mean of the difference: *PAIRED t PROCEDURE*]

H_o: $\mu = \#$

H_a: $\mu \neq \#$ **-or-** $\mu < \#$ **-or-** $\mu > \#$

Conditions:

-SRS

-Normality: Stated, CLT, or plots

-Independence and N≥10n

Test statistic: df = n-1

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

CALCULATOR: **2: T-Test**

One Proportion Z test

Use when testing a proportion from a single sample

H_o: $p = \#$

H_a: $p \neq \#$ **-or-** $p < \#$ **-or-** $p > \#$

Conditions:

-SRS

-Normality: $np_0 \geq 10$, $n(1-p_0) \geq 10$

-Independence and N≥10n

Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

CALCULATOR: **5: 1-PropZTest**

Two Sample Z test σ known (RARELY USED)

Use when testing two means from two samples and σ is known

H_o: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$ **-or-** $\mu_1 < \mu_2$ **-or-** $\mu_1 > \mu_2$

Conditions:

-SRS for both populations

-Normality: Stated, CLT, or plots for both populations

-Independence and N≥10n for both populations.

Test statistic:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

CALCULATOR: **3: 2-SampZTest**

Two Sample t Test σ unknown

Use when testing two means from two samples and σ is NOT known

H_o: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$ **-or-** $\mu_1 < \mu_2$ **-or-** $\mu_1 > \mu_2$

Conditions:

-SRS for both populations

-Normality: Stated, CLT, or plots for both populations

-Independence and N≥10n for both populations.

Test statistic: use lower n for df (df = n-1) or use calculator.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

CALCULATOR: **4: 2-SampTTest**

Two Proportion Z test

Use when testing two proportions from two samples

H_o: $p_1 = p_2$

H_a: $p_1 \neq p_2$ **-or-** $p_1 < p_2$ **-or-** $p_1 > p_2$

Conditions:

-SRS for both populations

-Normality:

$n_1\hat{p}_c \geq 10$ $n_1(1-\hat{p}_c) \geq 10$

$n_2\hat{p}_c \geq 10$ $n_2(1-\hat{p}_c) \geq 10$

Independence and N≥10n for both populations. Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

CALCULATOR: **6: 2-PropZTest**

X^2 – Goodness of Fit L1 and L2

Use for categorical variables to see if one distribution is similar to another

H_o: The distribution of two populations are not different

H_a: The distribution of two populations are different

Conditions:

-Independent Random Sample

-All expected counts ≥ 1

-No more than 20% of Expected Counts <5

Test statistic: df = n-1

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

CALCULATOR:**D: X^2GOF-Test(on 84)**

X^2 – Homogeneity of Populations r x c table

Use for categorical variables to see if multiple distributions are similar to one another

H_o: The distribution of multiple populations are not different

H_a: The distribution of multiple populations are different

Conditions:

-Independent Random Sample

-All expected counts ≥ 1

-No more than 20% of Expected Counts <5

Test statistic: df = (r-1)(c-1)

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

CALCULATOR: **C: X^2-Test**

X^2 – Independence/Association r x c table

Use for categorical variables to see if two variables are related

H_o: Two variables have no association (independent)

H_a: Two variables have an association (dependent)

Conditions:

-Independent Random Sample

-All expected counts ≥ 1

-No more than 20% of Expected Counts <5

Test statistic: df = (r-1)(c-1)

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

CALCULATOR: **C: X^2-Test**

Significance Test for Regression Slope:

H_o: $\beta = 0$ H_a: $\beta \neq \#$ **-or-** $\beta < \#$ **-or-** $\beta > \#$

Conditions:
1. Observations are independent
2. Linear Relationship (look at residual plot)
3. Standard deviation of y is the same(look at residual plot)
4. y varies normally (look at histogram of residuals)

CALCULATOR: **LinRegTTest** Use technology readout or the calculator for this significance test $t = \frac{b}{SE_b}$ df = n-2

Notation

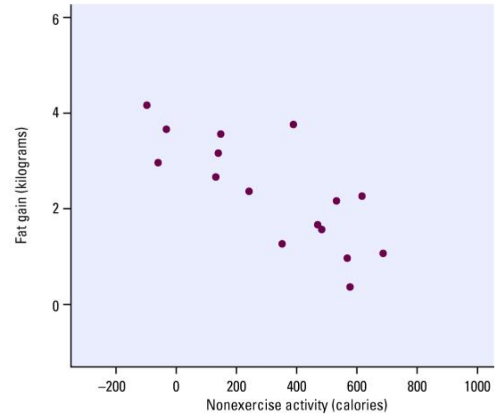
IQR	Inner Quartile Range
$\bar{x}$	Mean of a sample
μ	Mean of a population
s	Standard deviation of a sample
σ	Standard deviation of a population
$\hat{p}$	Sample proportion
p	Population proportion
s^2	Variance of a sample
σ^2	Variance of a population
M	Median
Σ	Summation
Q_1	First Quartile
Q_3	Third Quartile
Z	Standardized value – z test statistic
z^*	Critical value for the standard normal distribution
t	Test statistic for a t test
t^*	Critical value for the t-distribution
N(μ, σ)	Notation for the normal distribution with mean and standard deviation
r	Correlation coefficient – strength of linear relationship
r^2	Coefficient of determination – measure of fit of the model to the data
$\hat{y} = a + bx$	Equation for the Least Squares Regression Line
a	y-intercept of the LSRL
b	Slope of the LSRL
$(\bar{x}, \bar{y})$	Point the LSRL passes through
$y = ax^b$	Power model
$y = ab^x$	Exponential model
SRS	Simple Random Sample
S	Sample Space
P(A)	The probability of event A
A^c	A complement
P(B|A)	Probability of B given A
∩	Intersection (And)
U	Union (Or)
X	Random Variable
μ_X	Mean of a random variable
σ_X	Standard deviation of a random variable
σ_X^2	Variance of a random variable
B(n,p)	Binomial Distribution with observations and probability of success
$\binom{n}{k}$	Combination n taking k
pdf	Probability distribution function
cdf	Cumulative distribution function
n	Sample size
N	Population size
CLT	Central Limit Theorem
$\mu_{\bar{x}}$	Mean of a sampling distribution
$\sigma_{\bar{x}}$	Standard deviation of a sampling distribution
df	Degrees of freedom
SE	Standard error
H_0	Null hypothesis-statement of no change
H_a	Alternative hypothesis- statement of change
p-value	Probability (assuming H_0 is true) of observing a result as large or larger than that observed
α	Significance level of a test. P(Type I) or the y-intercept of the true LSRL
β	P(Type II) or the true slope of the LSRL
χ^2	Chi-square test statistic

Regression in a Nutshell

Given a Set of Data:

Data:

NEA change (cal):	−94	−57	−29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

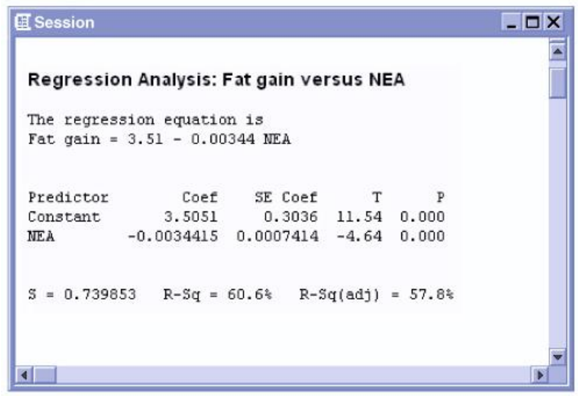


Enter Data into L_1 and L_2 and run **8:Linreg(a+bx)**

The regression equation is:

$$\boxed{\text{predicted fat gain} = 3.5051 - 0.00344(NEA)}$$

Minitab



y-intercept: Fat gain is 3.5051 kilograms when NEA is zero.

slope: Fat gain decreases by .00344 for every unit increase in NEA.

r: correlation coefficient

r = - 0.778

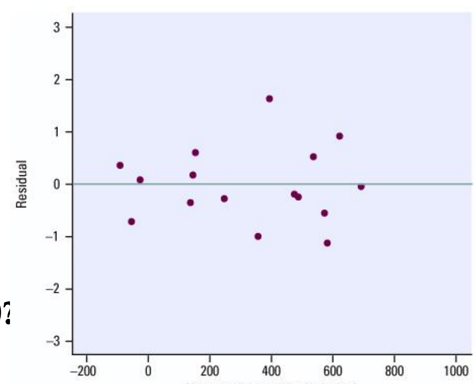
Moderate, negative correlation between NEA and fat gain.

r^2: coefficient of determination

$r^2 = 0.606$

60.6% of the variation in fat gained is explained by the Least Squares Regression line on NEA.
The linear model is a moderate/reasonable fit to the data. It is not strong.

The residual plot shows that the model is a reasonable fit; there is not a bend or curve, There is approximately the same amount of points above and below the line. There is No fan shape to the plot.



Predict the fat gain that corresponds to a NEA of 600.

$$\boxed{\begin{aligned}\text{predicted fat gain} &= 3.5051 - 0.00344\,(600) \\ \text{predicted fat gain} &= 1.4411\end{aligned}}$$

Would you be willing to predict the fat gain of a person with NEA of 1000?

No, this is extrapolation, it is outside the range of our data set.

Residual:

observed y **-** predicted y

Find the residual for an NEA of 473

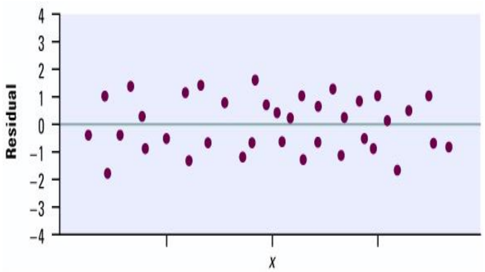
First find the predicted value of 473:

$$\boxed{\begin{aligned}\text{predicted fat gain} &= 3.5051 - 0.00344\,(473) \\ \text{predicted fat gain} &= 1.87798\end{aligned}}$$

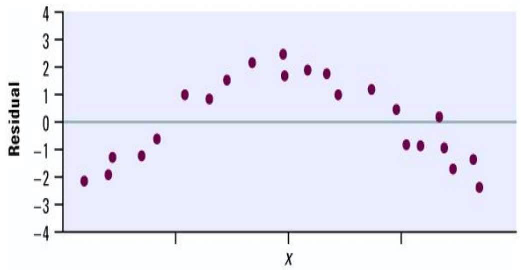
observed – predicted = 1.7 - 1.87798 = -0.17798

Transforming Exponential Data $y = ab^x$	*Transforming Power Data $y = ax^b$*
Take the log or ln of y. The new regression equation is: log(y) = a + bx	Take the log or ln of x and y. The new regression equation is: log(y) = a + b log(x)

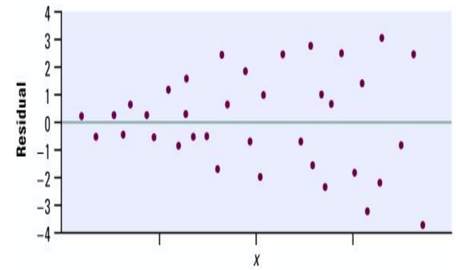
Residual Plot examples:



Linear mode is a Good Fit



Curved Model would be a good fit



Fan shape loses accuracy as x increases

Inference with Regression Output:

```
Regression Analysis
The regression equation is
IQ = 91.3 + 1.49 Crycount

Predictor      Coef      StDev        T         P
Constant     91.268      8.934    10.22     0.000
Crycount     1.4929     0.4870     3.07     0.004

S = 17.50          R-Sq = 20.7%

estimates σ                  SE_b      We usually
                                       ignore
                                       this part.
```

Construct a 95% Confidence interval for the slope of the LSRL of IQ on cry count for the 20 babies in the study.

Formula: df = n – 2 = 20 – 2 = 18

$$b \pm t^* SE_b$$

$$1.4929 \pm (2.101)(0.4870)$$

$$1.4929 \pm 1.0232$$

$$(0.4697, 2.5161)$$

Find the t-test statistic and p-value for the effect cry count has on IQ.

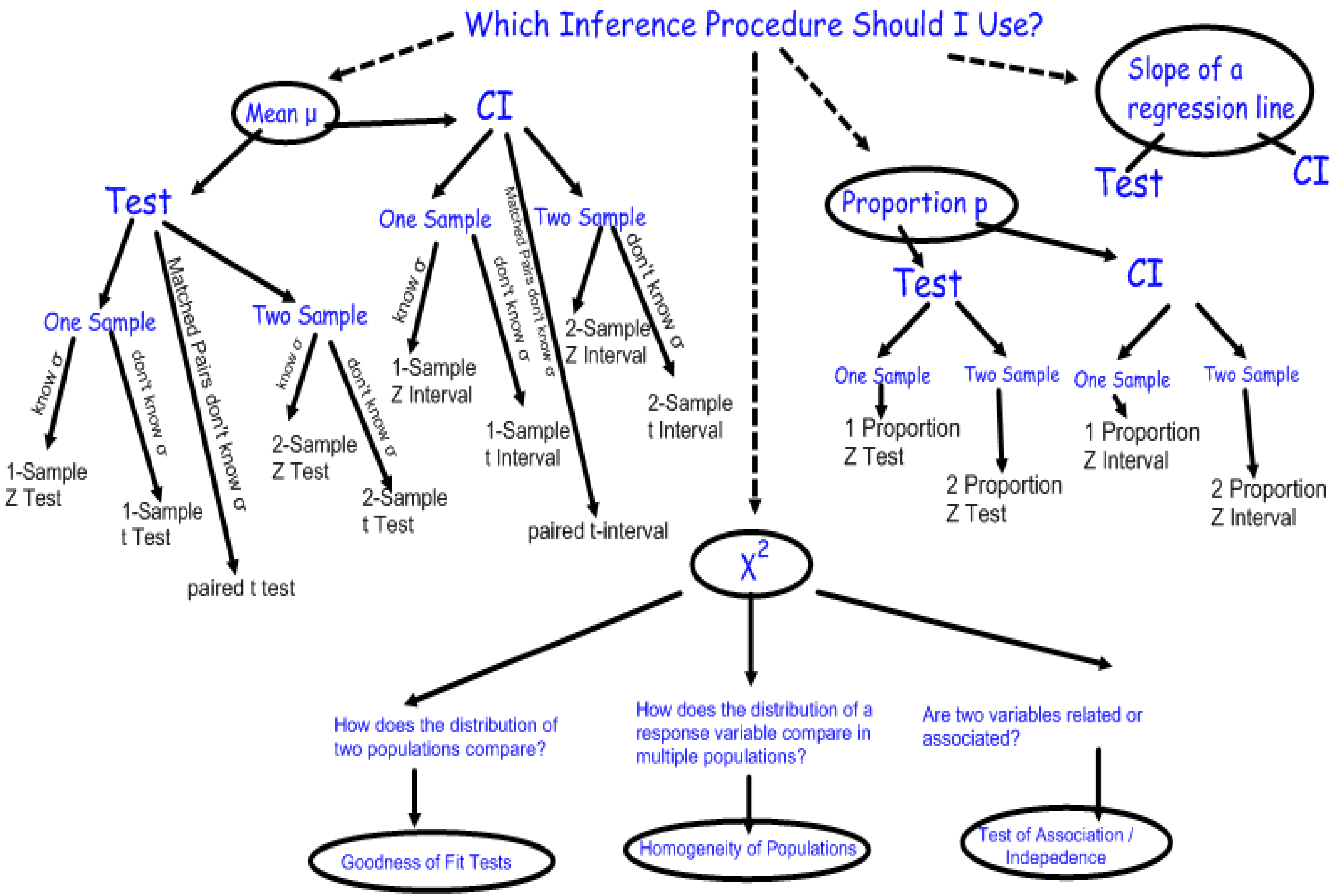
From the regression analysis t = 3.07 and p = 0.004

Or

$$t = \frac{b}{SE_b} = \frac{1.4929}{0.4870} = 3.07$$

s = 17.50

This is the standard deviation of the residuals and is a measure of the average spread of the deviations from the LSRL.



Which Inference Procedure Should I Use?
Mean μ
Test
CI
One Sample
know σ
1-Sample Z Test
don't know σ
1-Sample t Test
Matched Pairs don't know σ
paired t test
Two Sample
know σ
2-Sample Z Test
don't know σ
2-Sample t Test
One Sample
know σ
1-Sample Z Interval
don't know σ
1-Sample t Interval
Matched Pairs don't know σ
paired t-interval
Two Sample
2-Sample Z Interval
don't know σ
2-Sample t Interval
Proportion p
Test
One Sample
1 Proportion Z Test
Two Sample
2 Proportion Z Test
CI
One Sample
1 Proportion Z Interval
Two Sample
2 Proportion Z Interval
Slope of a regression line
Test
CI
χ^2
How does the distribution of two populations compare?
Goodness of Fit Tests
How does the distribution of a response variable compare in multiple populations?
Homogeneity of Populations
Are two variables related or associated?
Test of Association / Indepedence