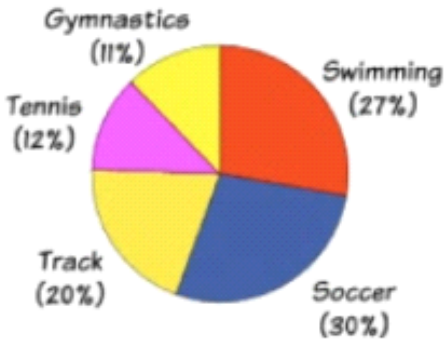
1.1 - Constructing & Interpreting Graphs

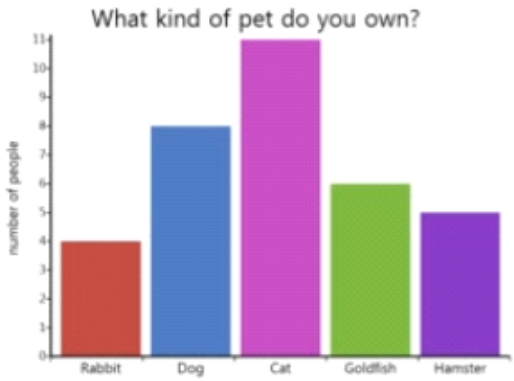
Friday, February 10, 2017 9:58 AM

Categorical Data

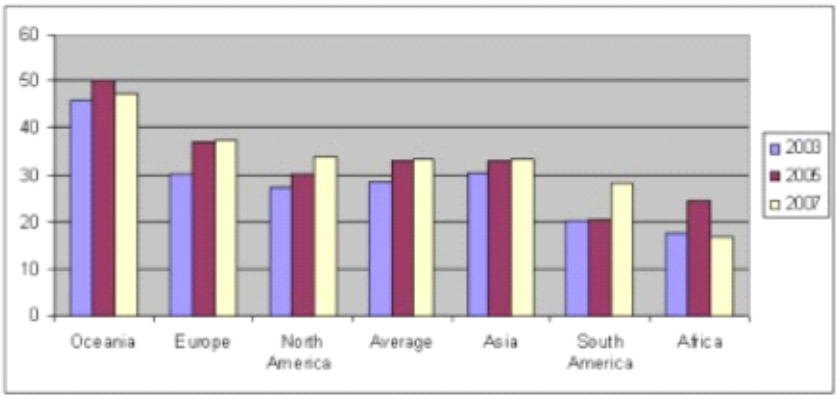
- Pie Charts



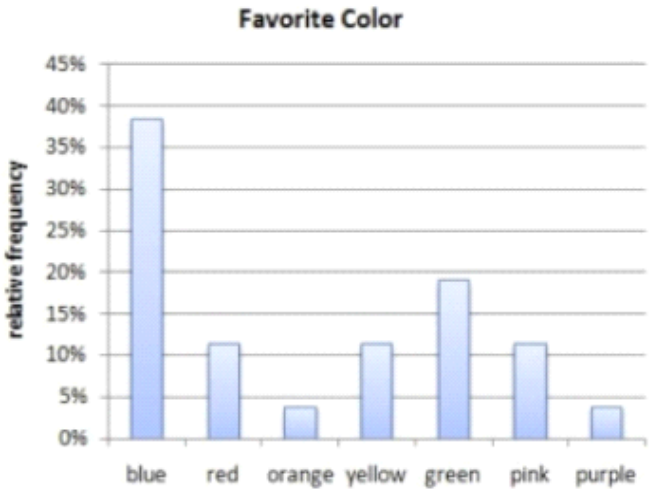
- Bar Graphs



- Comparative Bar Graphs

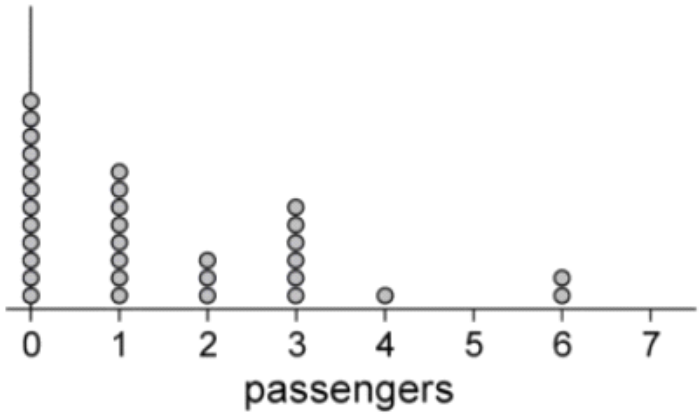


- Relative Frequency Bar Graphs



Numerical Data (Discrete)

- Dot Plots

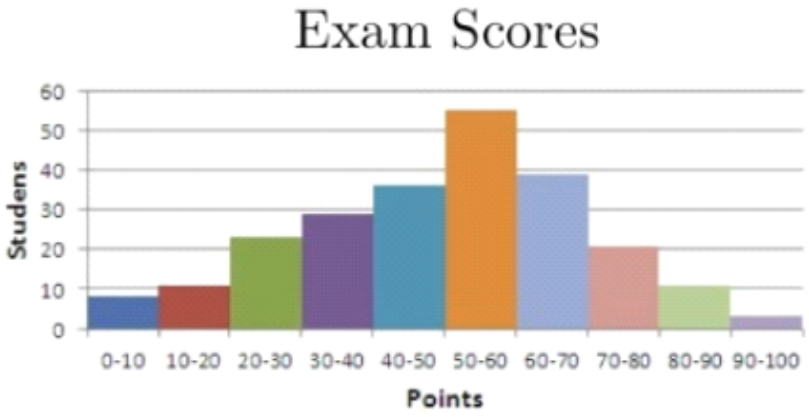


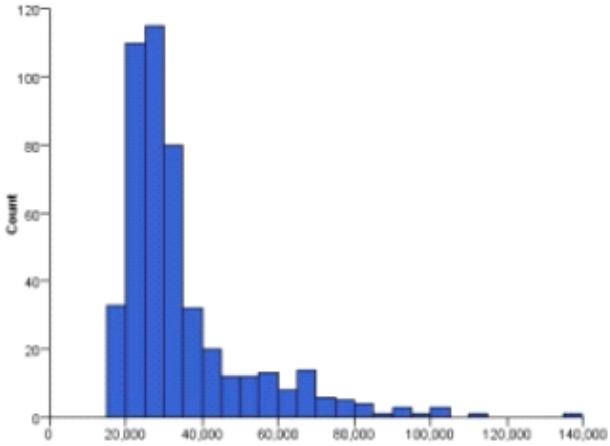
- Stem (and leaf) plots

stem	leaf
9	0 4
8	4 7 9 5 3
7	9 2 0 2
6	8 9 5
5	3

Numerical Data (Continuous)

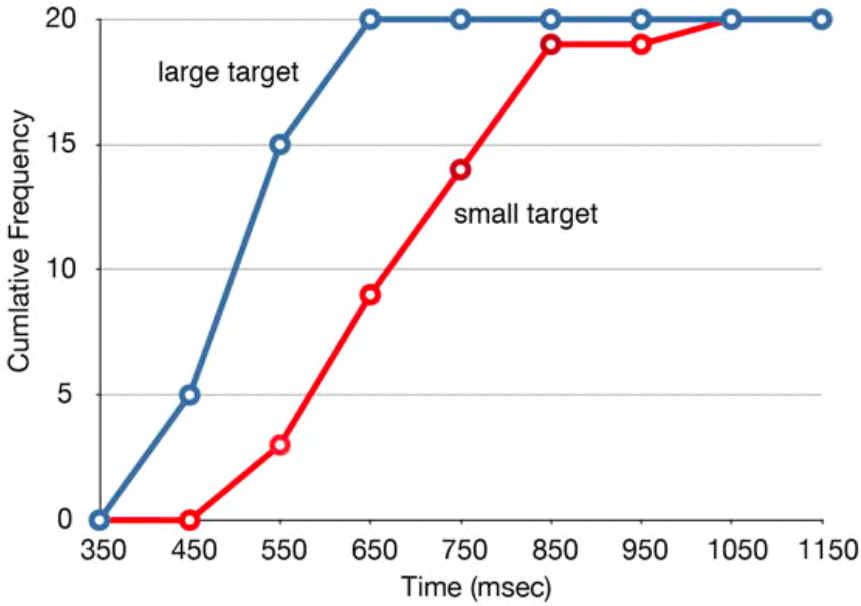
- Histogram



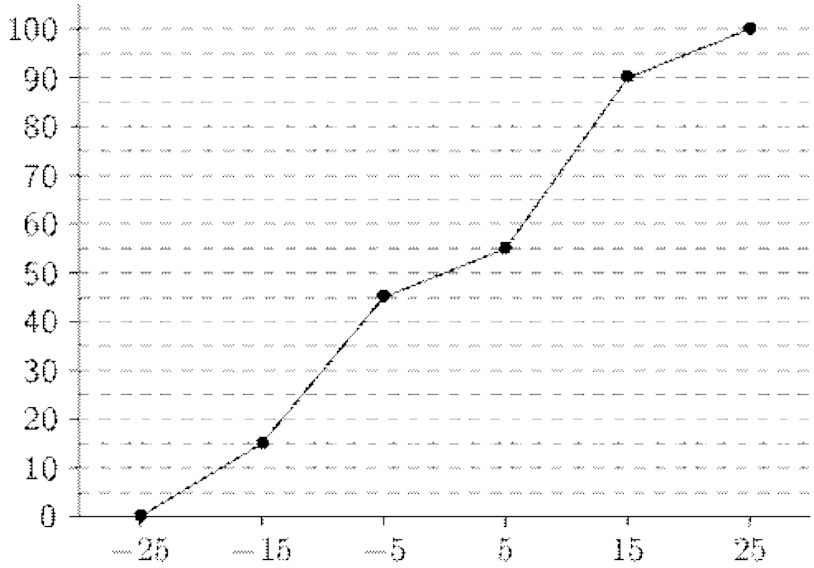


Numerical Data (Cumulative Frequency Plots)

- Frequency Polygon



- Ogive Plot

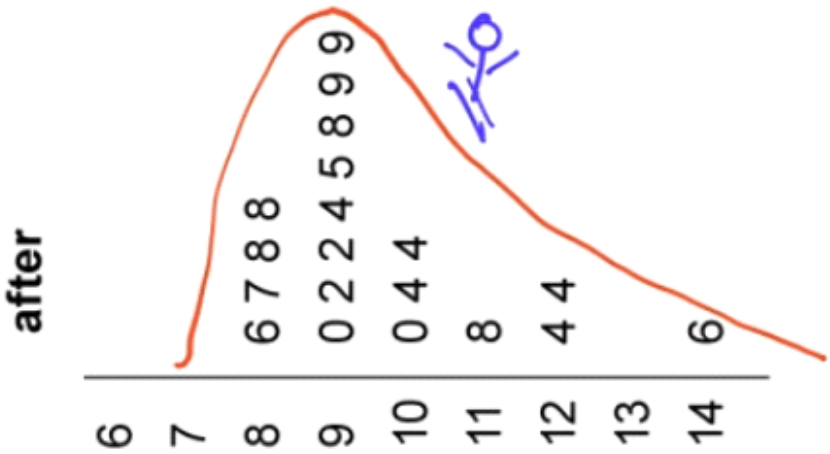


Stem Plots to Compare Two Groups of Data

pulse rate

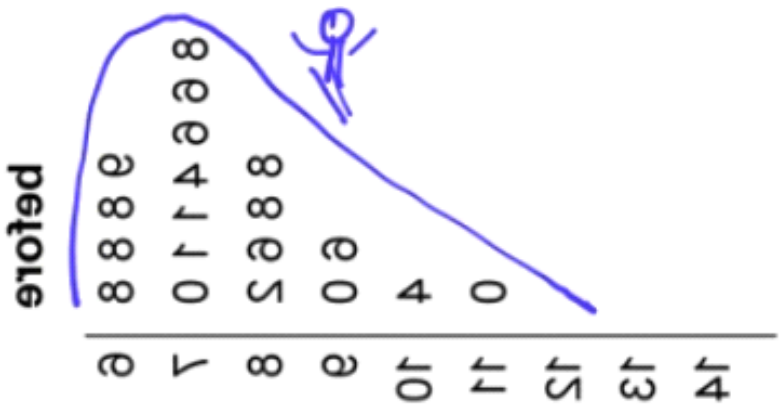
before		after
9 8 8 8	6	
8 6 6 4 1 1 0	7	
8 8 6 2	8	6 7 8 8
6 0	9	0 2 2 4 5 8 9 9
4	10	0 4 4
0	11	8
	12	4 4
	13	
	14	6

- Compare the distribution pulse rate before and after administering a new drug
    - After



- Skewed right
- An outlier at 146
- Centered around 95
- Spread between 86 and 146
- Range of 60

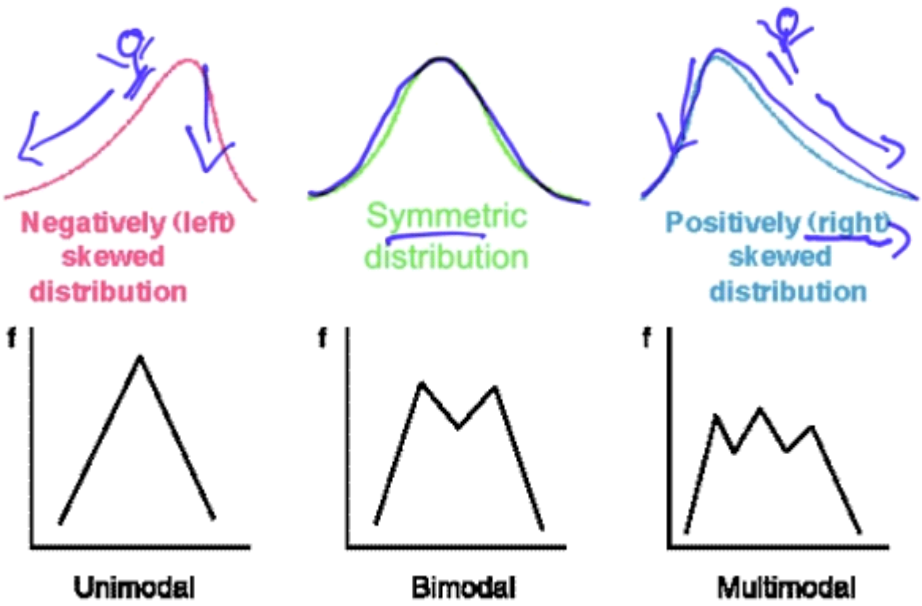
- Before



- Skew right
- No outliers
- Lower center at around 70
- Spread between 68 to 110
- Smaller range of 42

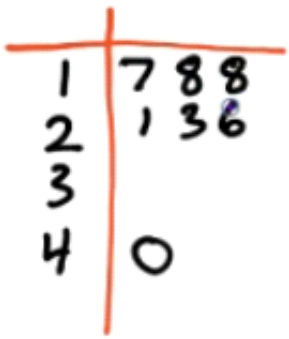
Describe the Distribution

- Shape (only for numerical data)

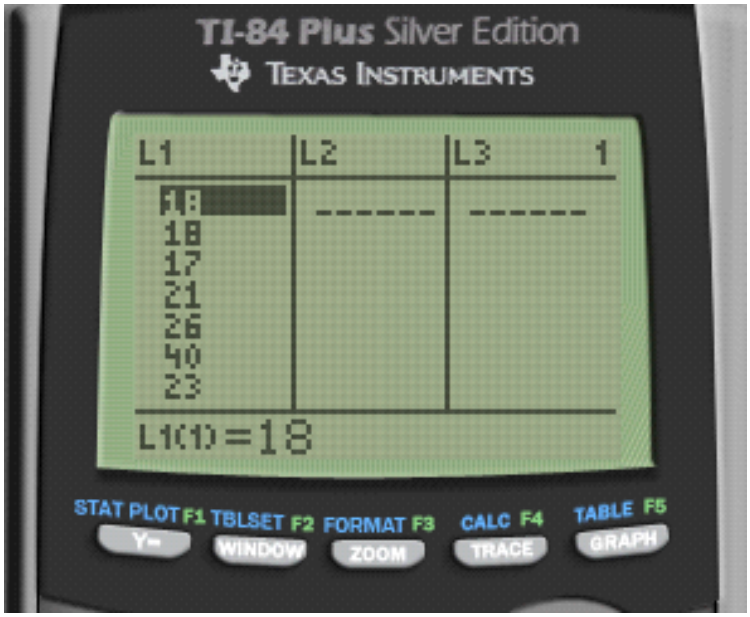


Examples

- Make a stem plot of the ages in a college classroom 18, 18, 17, 21, 26, 40, 23

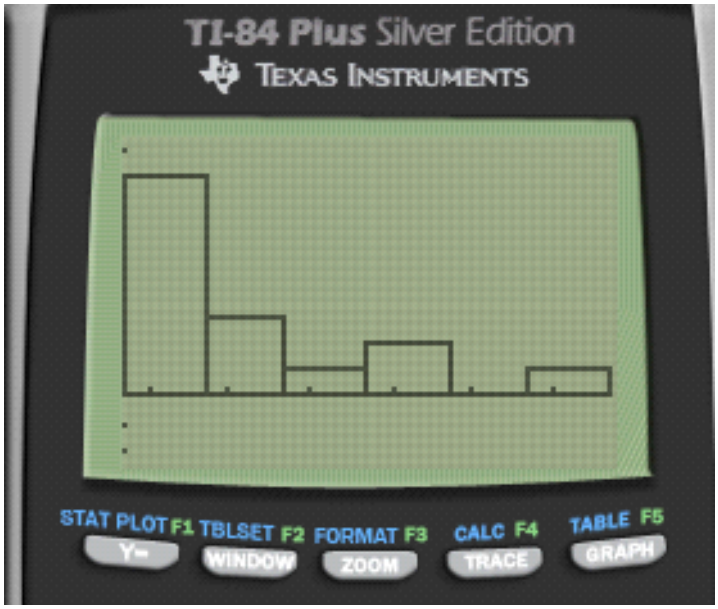


- Use your calculator to make a histogram of ages in a college classroom 18, 18, 17, 21, 26, 40, 23, 27, 22, 19, 20, 21, 18, 35, 32
    - STAT --> EDIT --> Enter & Type in the data

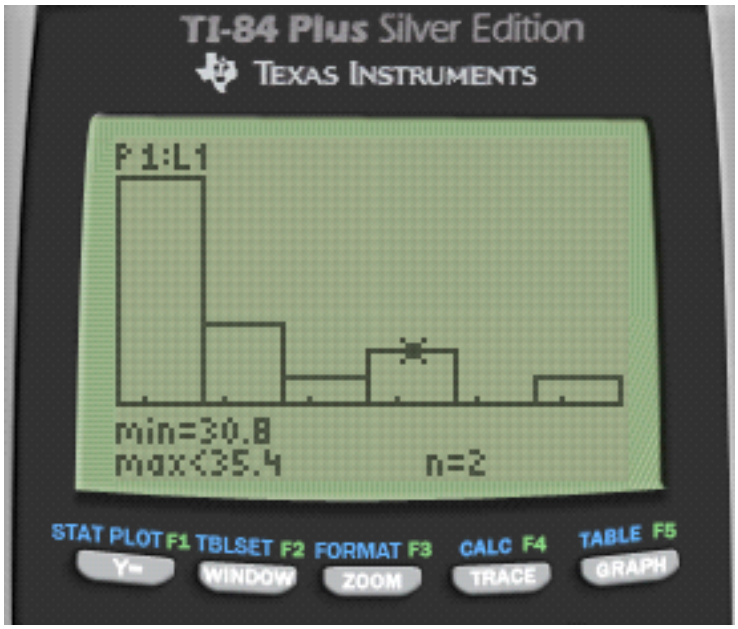


- STAT PLOT (2ND + Y=) Turn on & Select the type

- Zoom + 9



- TRACE



- Describe the distribution of the graph above
  - skew right
- Students took a statistic quiz. The score for the quiz are below. Describe the distribution of the quiz scores

$0|8$

$1|245$

$2|01889$

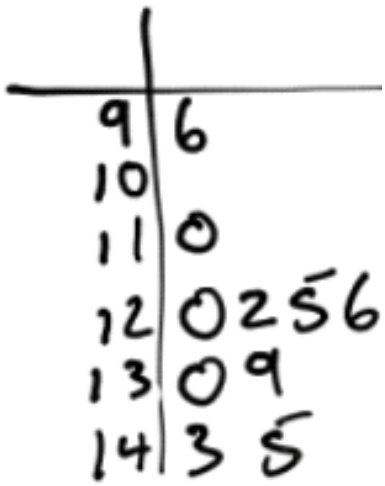
$3|001369$

$4|22446788$

$5|00$

  - Slightly skewed left
  - No outliers
  - Centered around 42
  - Spread of 8 to 50
- Here are the IQ test scores of a few students. Make a stem plot of these scores.

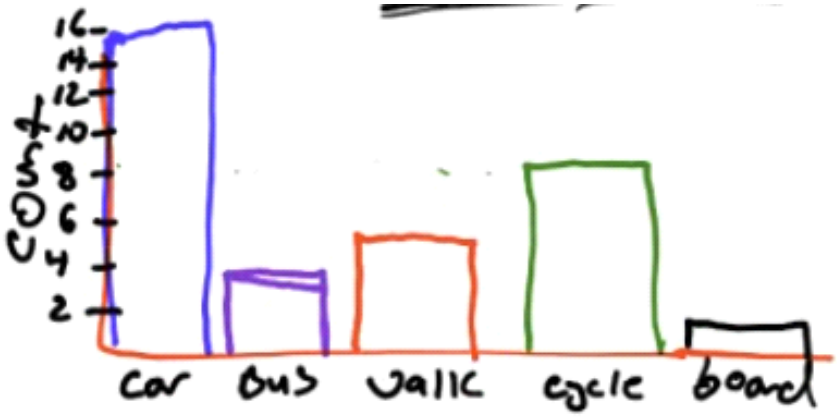
  145 139 126 122 125 130 96 110 120 143



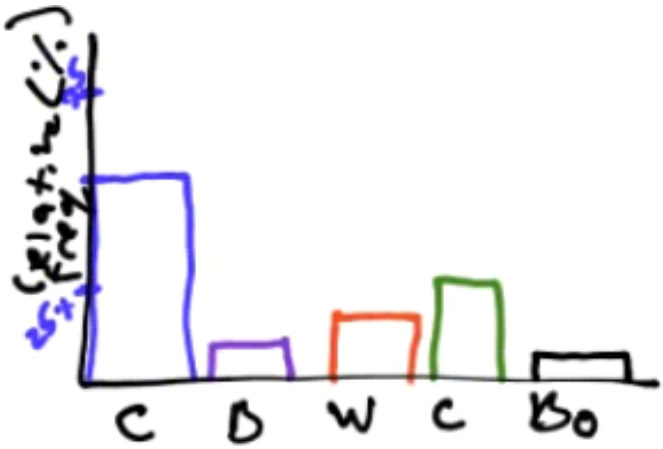
- You are interested in how students in your class get to school in the morning. You take a survey and collect the following data:

  Car 15, Bus 3, Walk 5, Bicycle 8, Skateboard 1

  - Construct a bar graph of your data



- Construct a relative frequency bar graph



1.2 - Summarizing Distributions of Univariate Data

Friday, February 10, 2017 10:38 AM

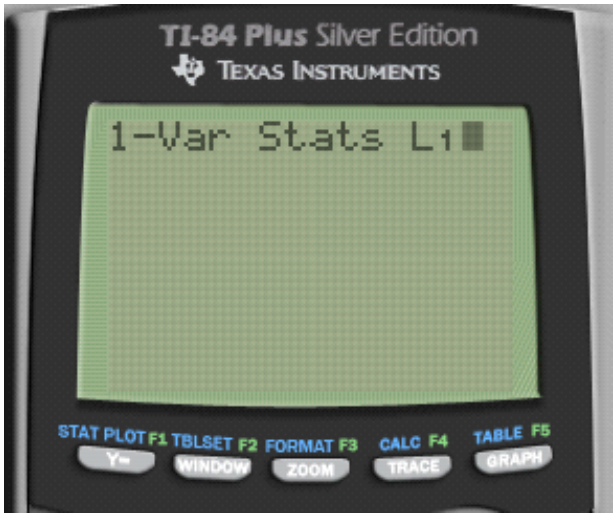
Measuring Center

- Median
  - The middle number
  - The median is **resistant to outliers**
- Mean
  - The average number

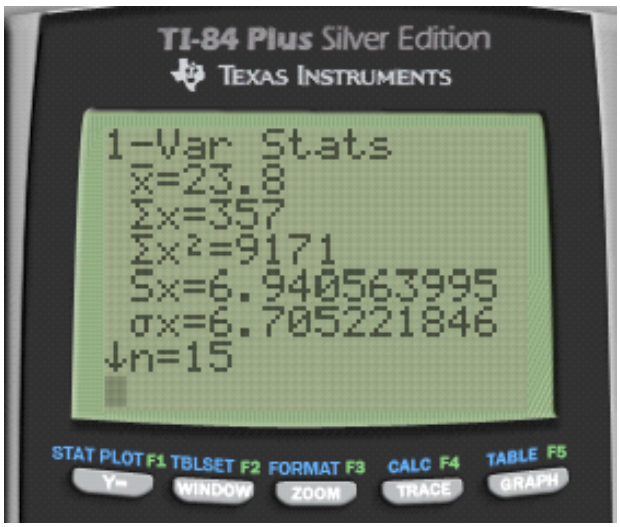
$$\bar{X} = \frac{x_1 + x_2 \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

- Calculation
  - STAT + CALC + 1

  - 2ND + 1 (L1)



- Enter



Measuring Position

- Percentiles
  - Percentage of observations your value is above
  - 30th percentile is the value **below which** 30 percent of the observations may be found
  - Take the average if there are two values
- Quartiles
  - Q1 = first quartile = 25th percentile
  - Q2 = median = 50 th percentile
  - Q3 = third quartile = 75th percentile

Measuring Spread

- Range

- Highest - Lowest

- IQR (Interquartile Range)
  - Q3 - Q1

- Variance / Standard Deviation

$$Variance = \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

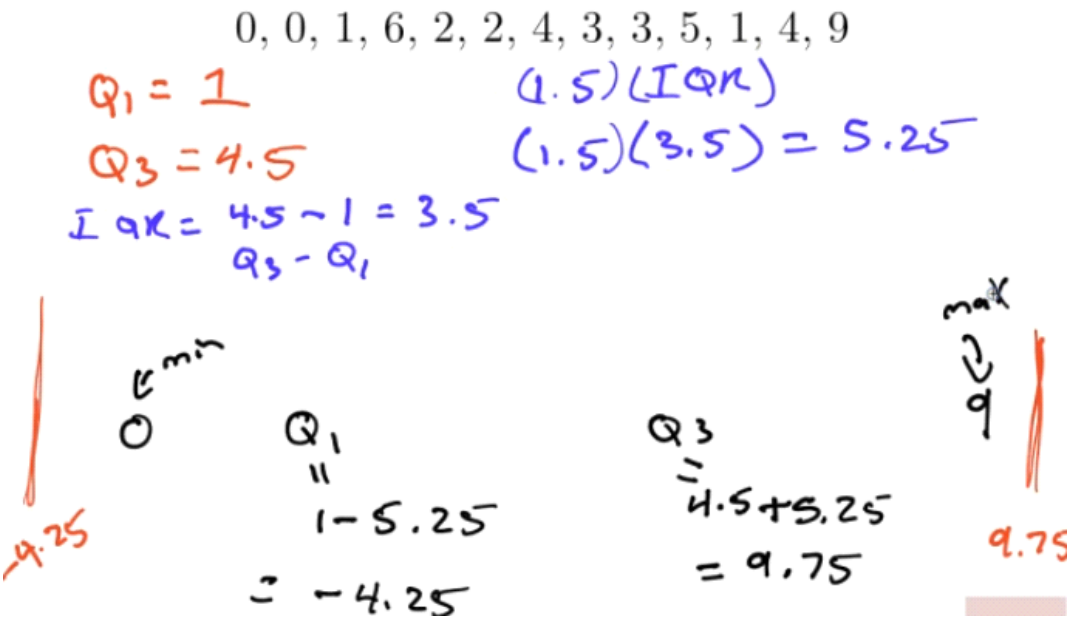
$$SD = \sqrt{Variance} = \sigma$$

  - Use Sx in the TI-84 calculator

Outliers

- 1.5 * IQR rule
  - Q1 - 1.5 * IQR = Bottom fence
  - Q3 + 1.5 * IQR = Upper fence
  - Outside the "fence" = Outlier

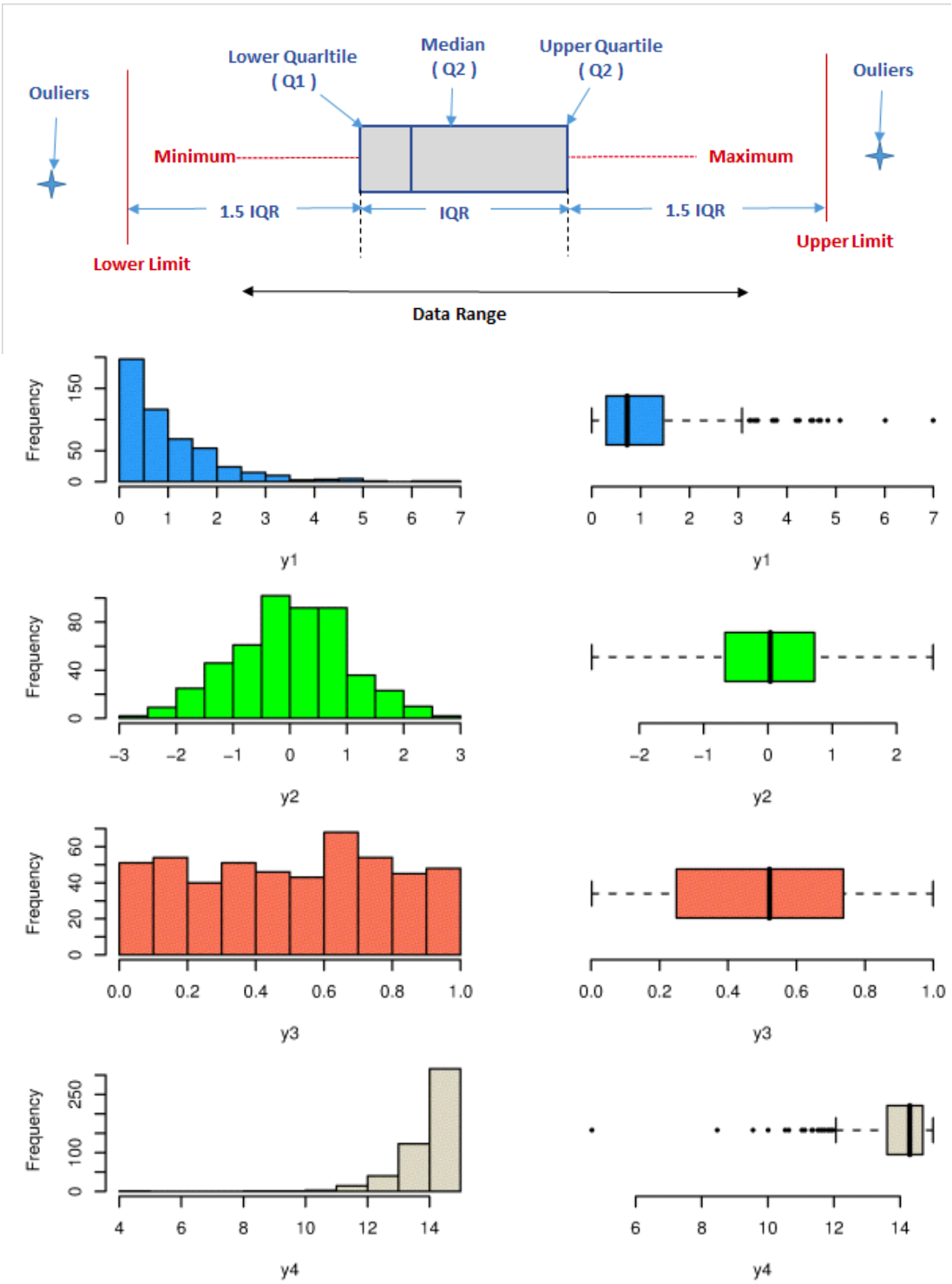
- Example



Boxplots

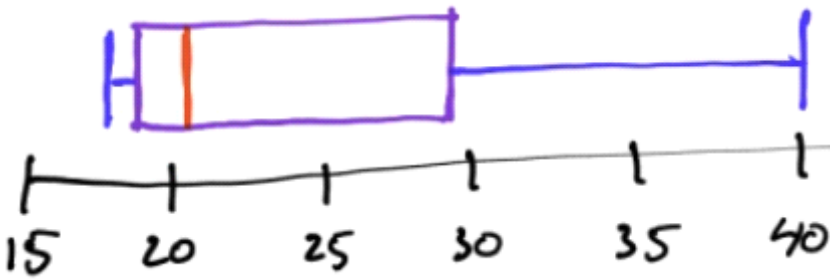
- Graph of the "5-number Summary"

- Min Q1 Median Q3 Max

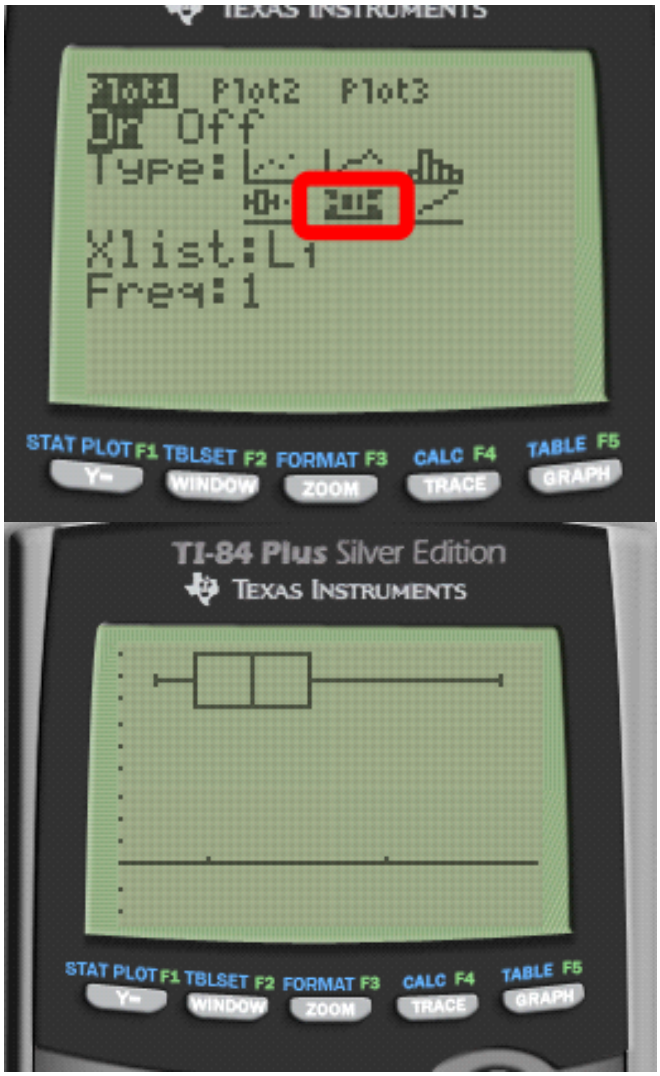


- Example
  - 18, 18, 17, 27, 22, 19, 20, 21, 18, 35, 32, 40
  - Min: 17
  - Q1: 18
  - Median: 20.5

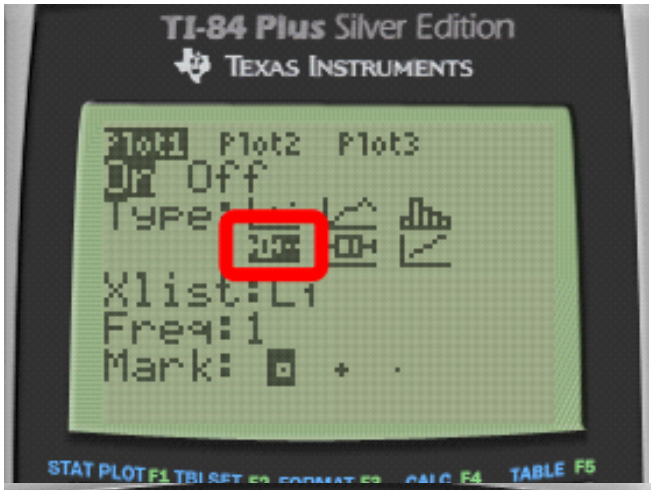
- Q3: 29.5
- Max: 40

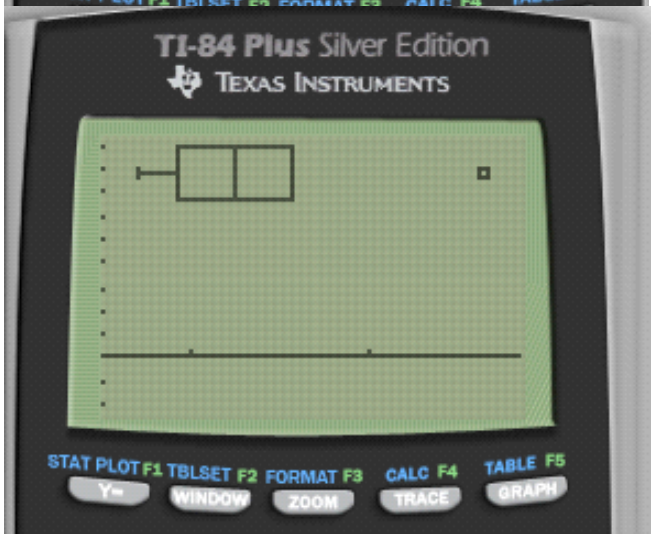


- Calculator
  - Boxplot



  - Modified Boxplot





SOCS

- Shape
  - Skewed left/right
- Outlier
  - Q1 - 1.5 * IQR
  - Q3 + 1.5 * IQR
- Center
  - Mean or Median
- Spread
  - SD or IQR
- Example
  - 18, 18, 17, 21, 26, 40, 23, 27, 22, 19, 20, 21, 18, 35, 32
  - Skewed right

- No outliers
- Centered at a median of 20.5
- A spread of IQR = 11.5

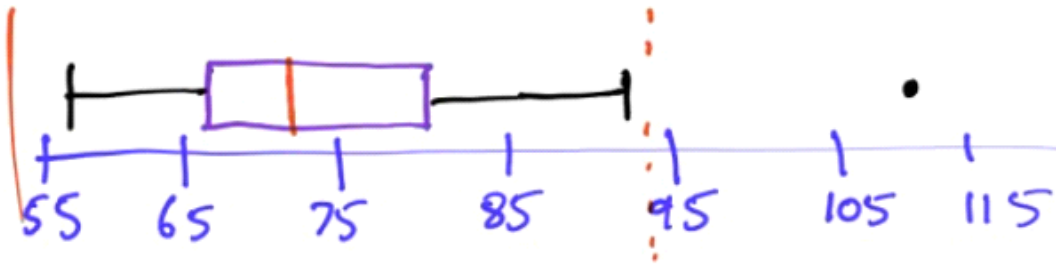
The Effect of Changing Units on Summary Measures

- Linear Transformations
  - $X_{new} = A + B * X_{old}$
  - A: only affect mean and median
  - B: affect all

Examples

72	93	70	59	78	74	65	73	80	57	67
72	57	83	76	74	56	68	67	74	110	

- Construct a modified boxplot for this data set
  - Min = 56
  - Q1 = 66
  - Med = 72
  - Q3 = 77
  - Max = 100
  - IQR = Q3 - Q1 = 11
  - IQR * 1.5= 16.5
  - Bottom fence = Q1 - IQR * 1.5 = 49.5
  - Upper fence = Q3 + IQR * 1.5 = 93.5



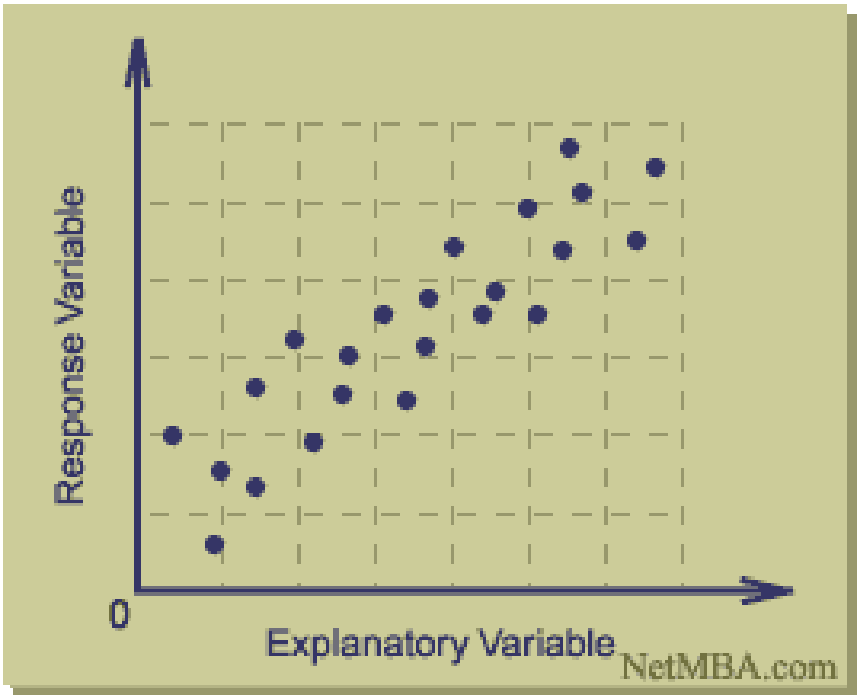
- Describe the distribution of test scores
  - Symmetric
  - One outlier at 110

- Centered at a median of 72
- With a spread of IQR of 11

2.1 - Correlation & Regression

Friday, February 10, 2017 2:11 PM

Scatterplots

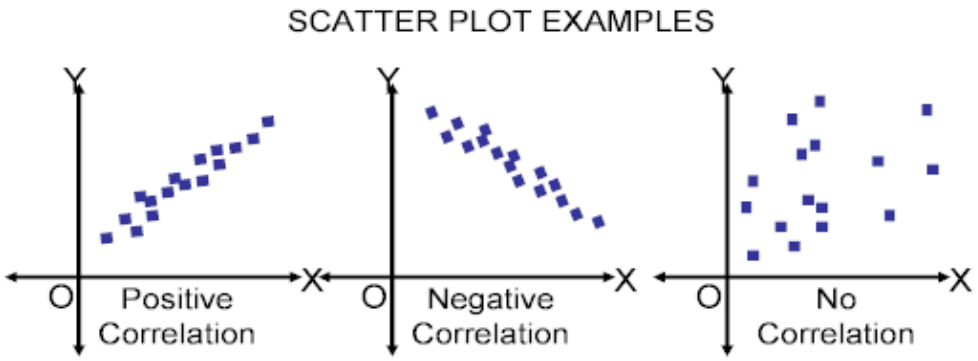


Explanatory **Response**

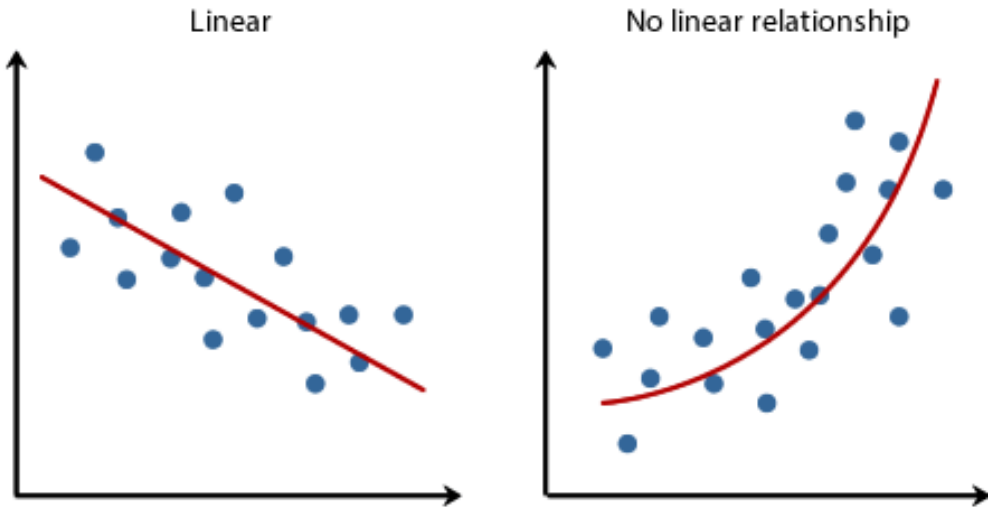
	Age	Distance
Driver 1	**18**	**510**
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

Interpreting Scatterplots

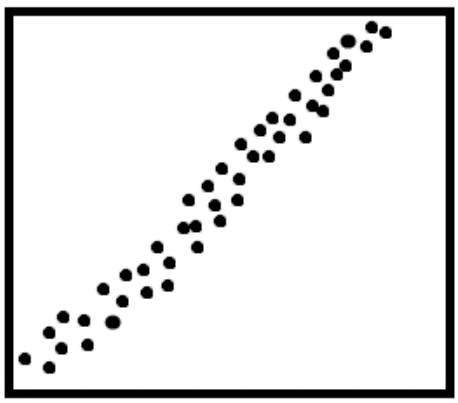
- Direction: Positive or Negative



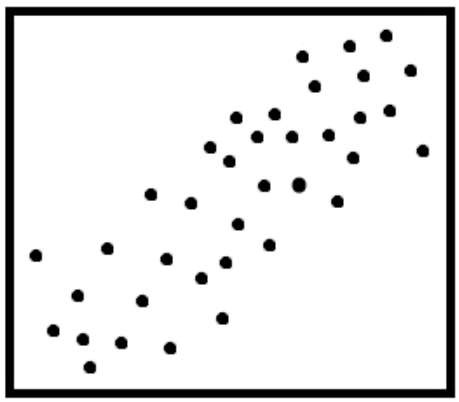
- Form: Linear or Non-linear



- Strength: Weak, Moderate or Strong

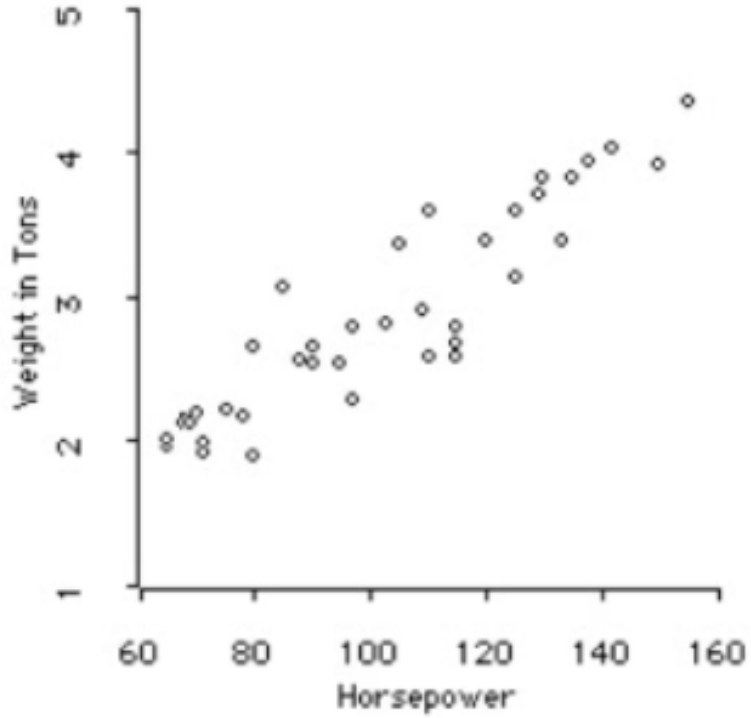


strong positive linear association



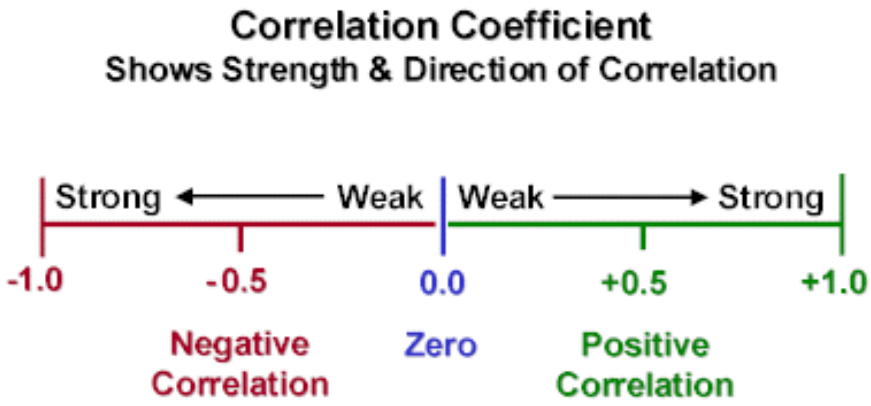
weak positive linear association

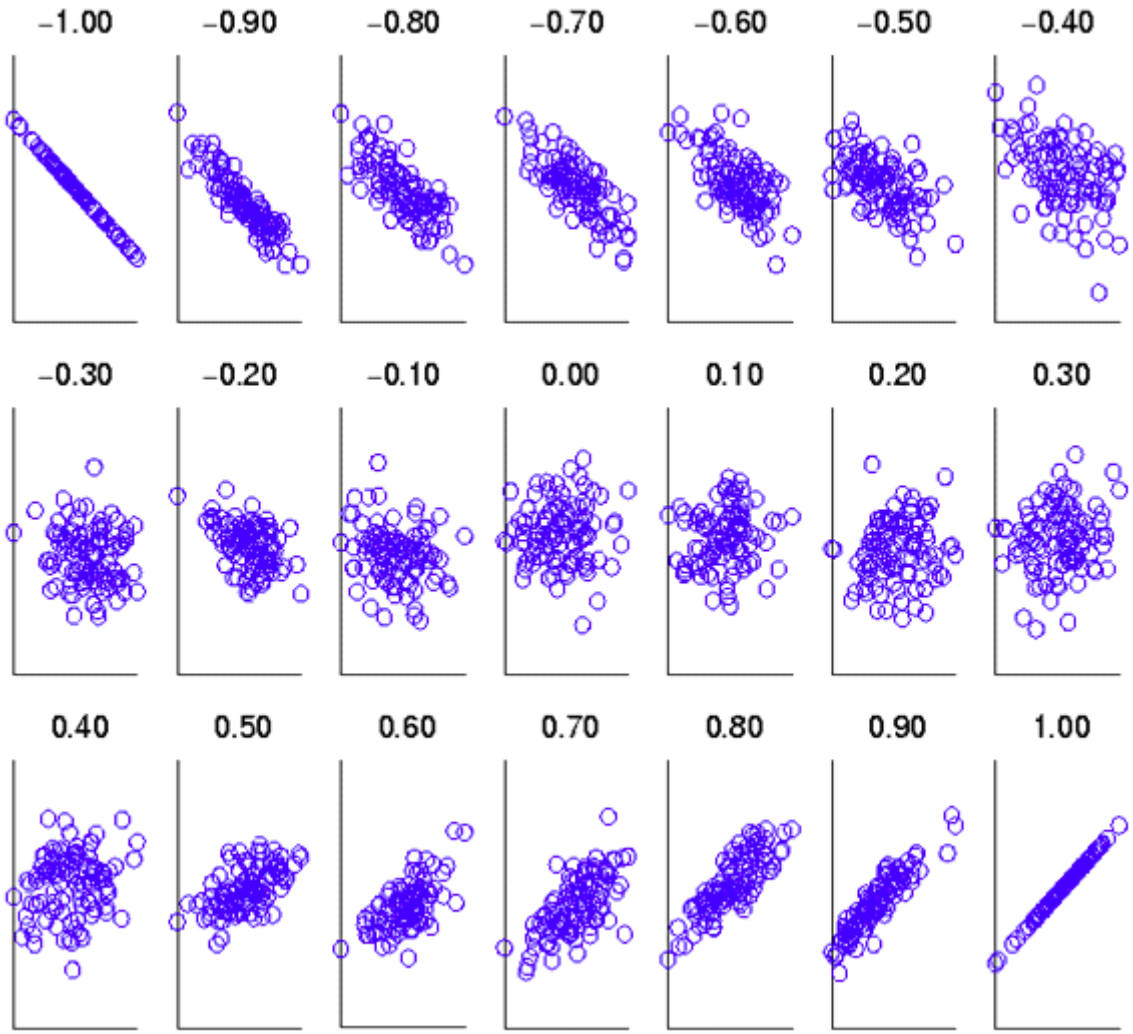
- Example



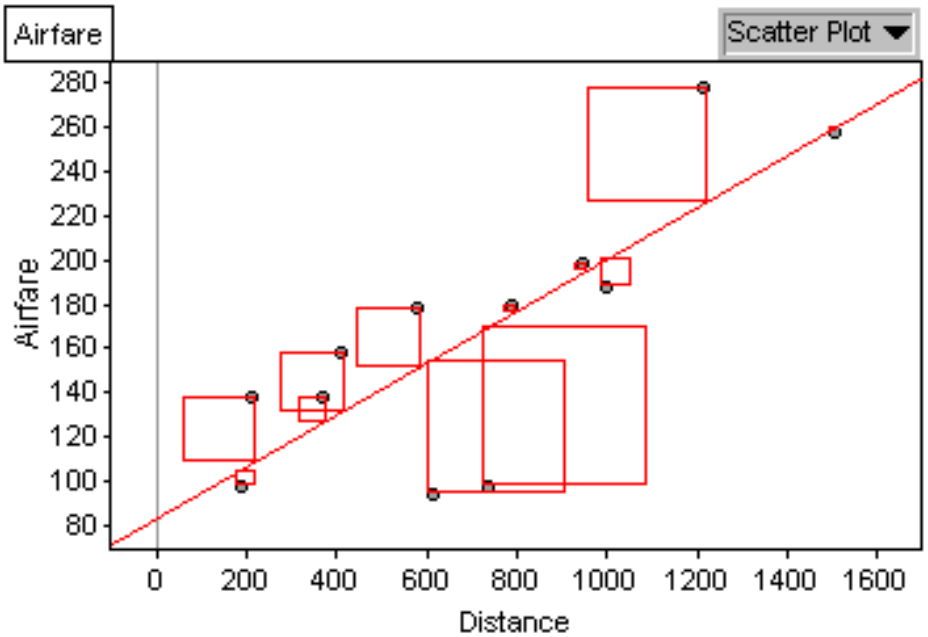
- Positive, linear, strong relationship between horsepower and weight in tons.

Correlation Coefficient (r)





Least Squares Regression Line



Airfare = 0.117Distance + 83; r^2 = 0.63;
Sum of squares = 14310

- A regression line is a line that describes **how y changes as x changes**
- Can be used to **predict** the value of y for a given value of x
- Called the **Least Squares** regression line because it make the **smallest sum of squares**
- LSRL will always run through the point (mean of x, mean of y)
- Formulas (hat = predicted)

$$\hat{y} = b_0 + b_1 x$$

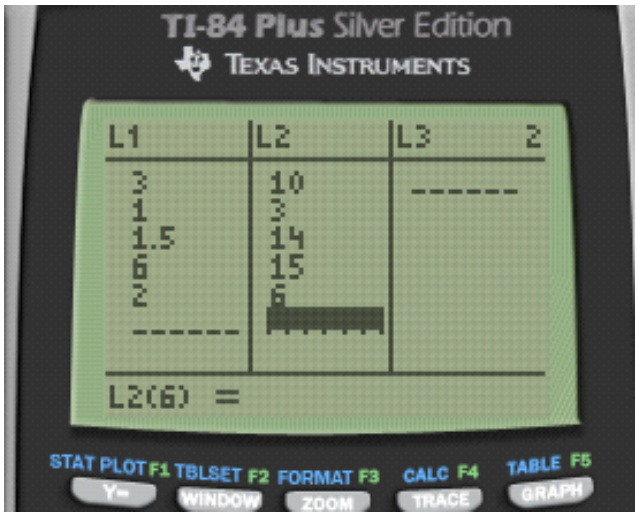
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

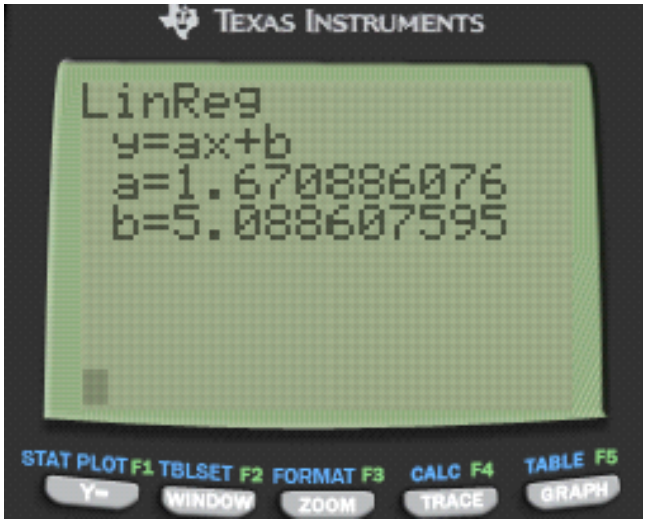
$$b_1 = r \frac{s_y}{s_x}$$

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- **Remember to note what x and y are**
- Calculation
  - Input data

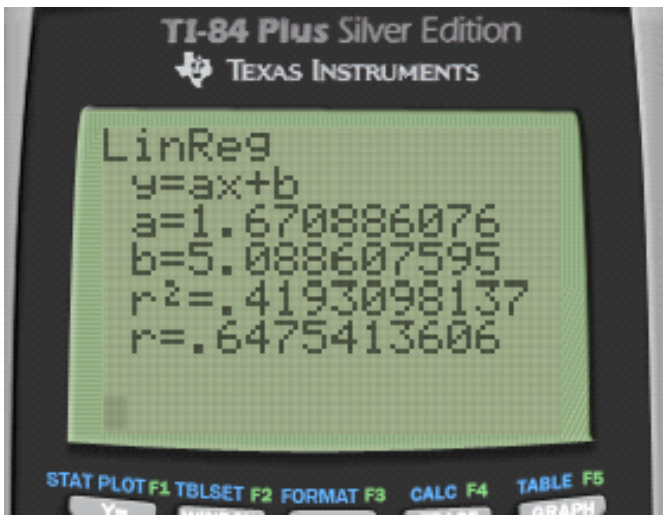


  - STAT-->CALC--> 4:LinReg(ax+b)
  - LinReg(ax+b) L1, L2



- Catalog (2ND + 0) --> DiagnosticOn

- Do LinReg again to display r



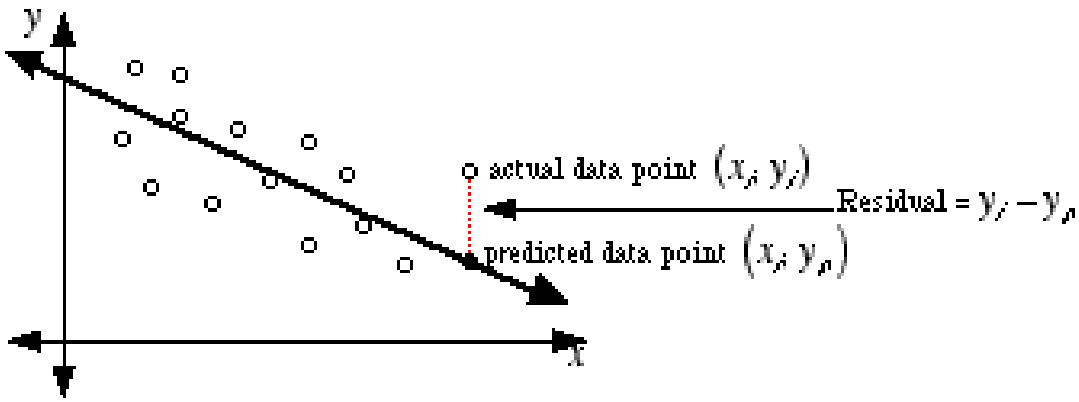
Coefficient of Determination

- R^2=r^2

- Coefficient of Determination = (Correlation Coefficient)^2
- **Percent** of the **change in y** that is explained by the change by the **change in x** / **least squares regression line**
- From the previous example, 41.9% of the change in y can be explained by the change in x

Residuals (≈error)

- Residuals = observed/actual y - predicted y
- Resid = y - y hat
- Resid < 0: Overpredicted
- Resid > 0: Underpredicted



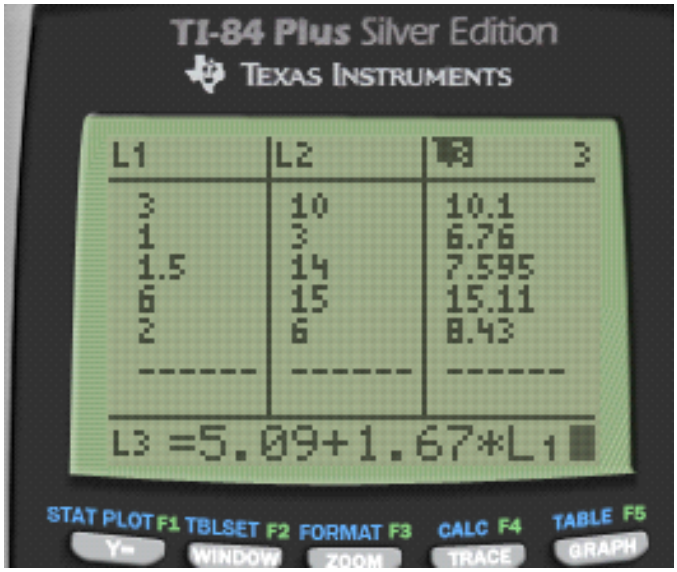
- Residual Plot

  - no pattern = good fit

- Example

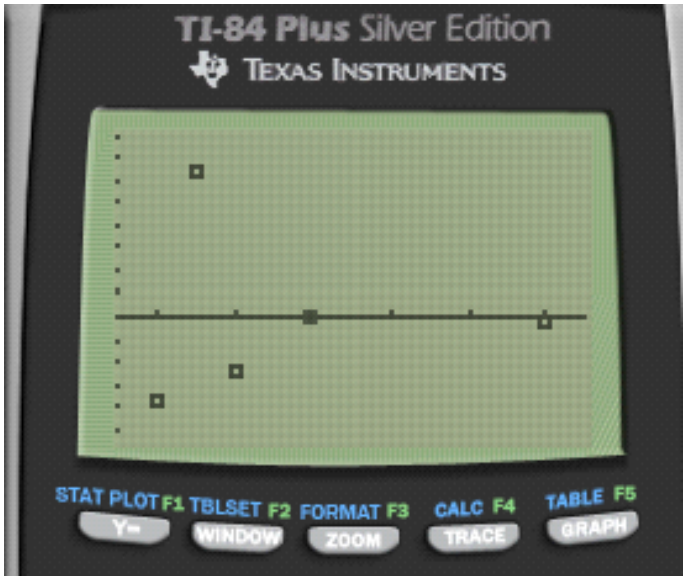
X	3	1	1.5	6	2
Y	10	3	14	15	6
Y hat	10.1	6.76	7.60	15.11	8.43
Residuals	-0.1	-3.76	6.4	-0.11	-2.43

- Calculator
  - Type the regression equation in L3 (y hat)



  - L4 = L2 - L3

  - Graph L1, L4 (Residuals)



Examples

- At the summer school, one of Sarah's teachers told her that you can determine air temperature from the number of cricket chirps

1. What is the explanatory variable, and what it the response variable
    - Explanatory/independent variable: cricket chirps
    - Response/dependent variable: air temperature
2. To determine a formula, Sarah collected data on temperature and number of chirps per minute on 14 occasions. She entered the data into her calculator and did 2-Var Stats. Here are some results. Use this information to find the equation of the least-squares regression line

Xbar	165.8
Sx	32.0
Ybar	76.83
Sy	9.23
r	0.361

    - b= r * Sy / Sx = 0.104
    - Ybar = a + b*Xbar
    - a = Ybar - b*Xbar = 59.57
    - Yhat = 59.57 + 0.14 * x
    - Where y = air temperature, and x = cricket chirps
3. One of Sarah's data points was recorded on a particularly hot day (95F). She

counted 2432 cricket chirps in one minute. What is the residual for this data point?

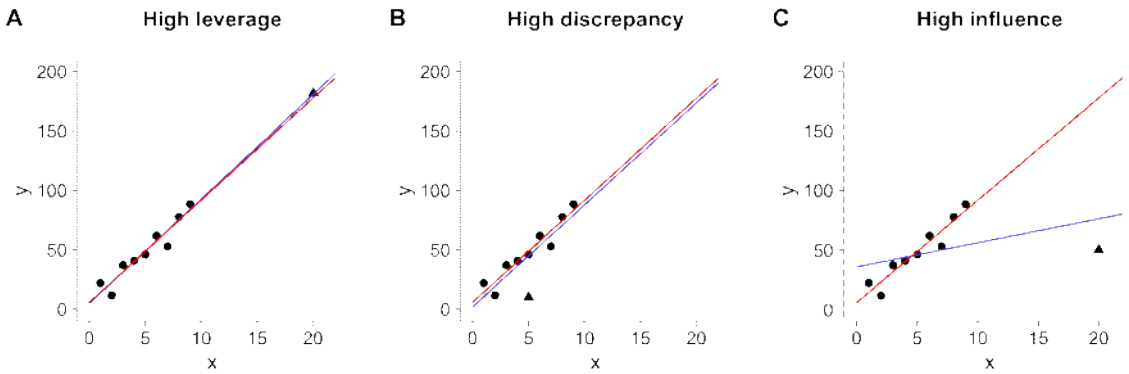
- Residual = Y - Yhat = 95 - (59.57 + 0.104 * 2432) = -217.498

2.2 - Regression, Part II

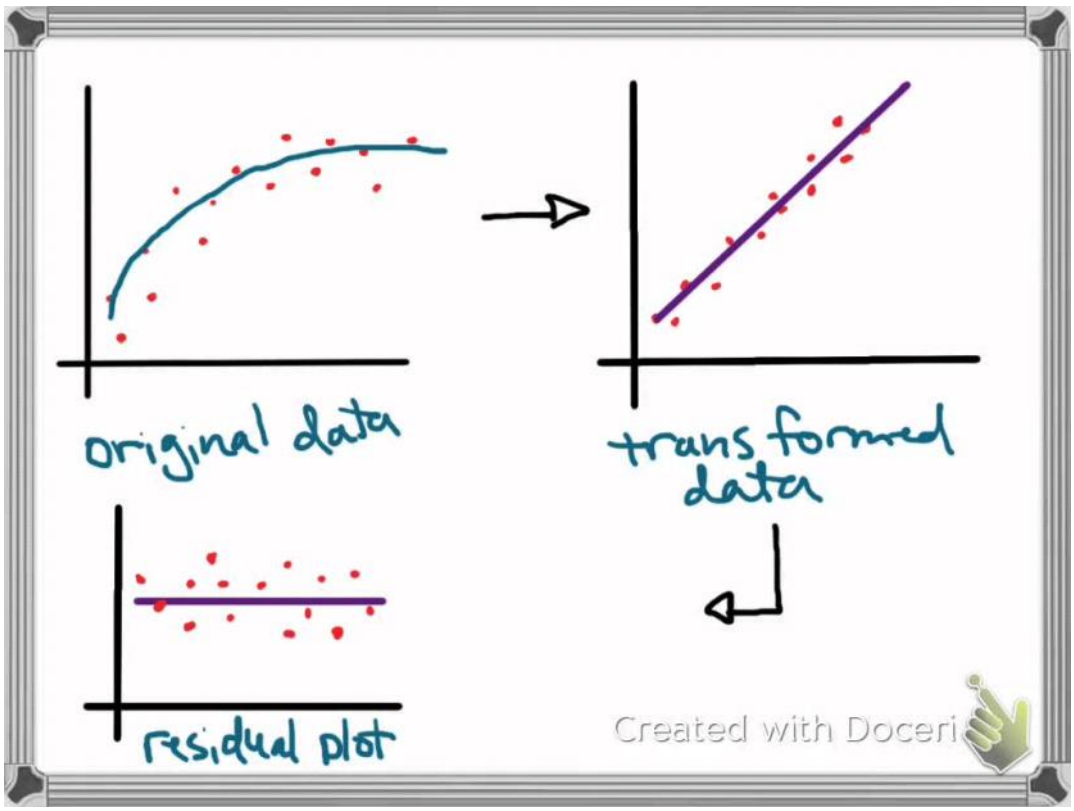
Friday, February 10, 2017 4:49 PM

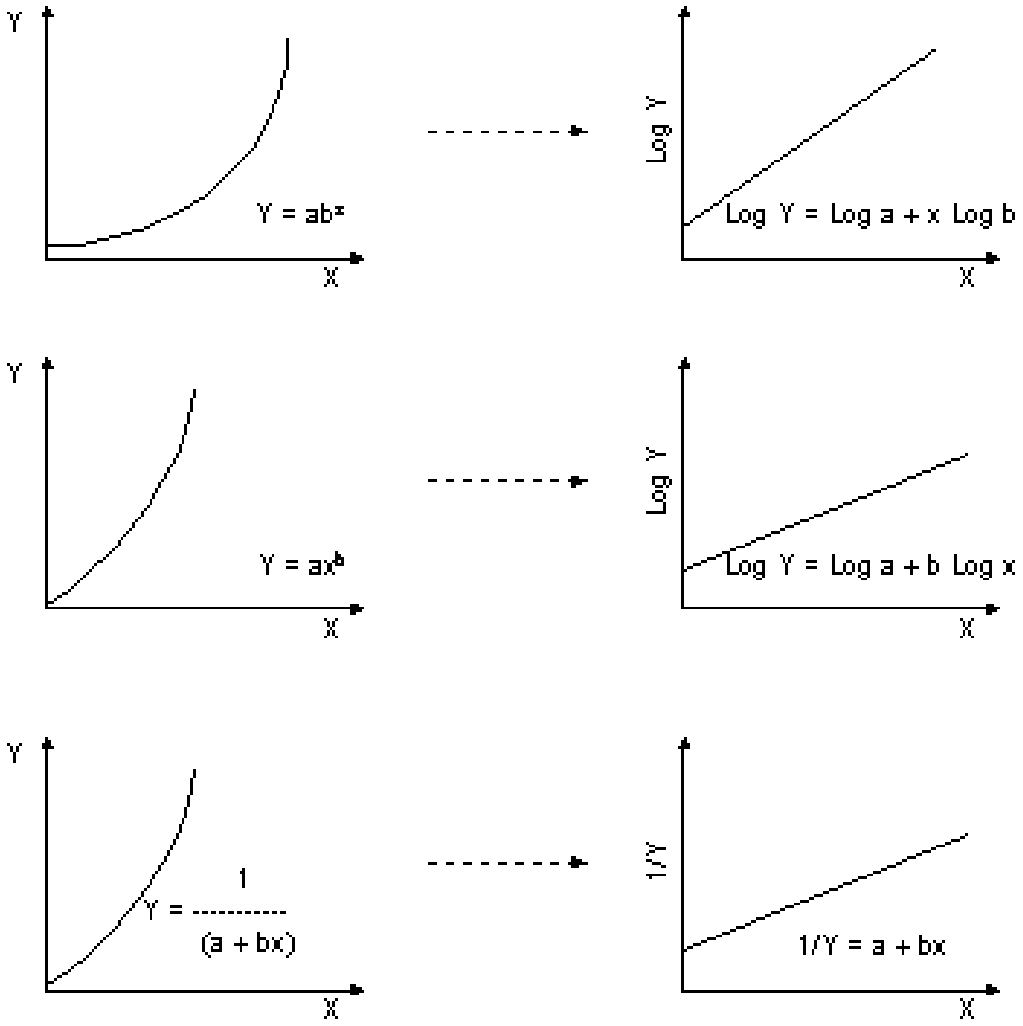
Outliers and Influential Points

- An **outlier** is an observation that lies outside the overall pattern
    - Outliers in the Y direction will lead to large residuals
- **Influential observations** are points that would greatly change the result if removed
- Points with high **leverage** have values far from the mean (of X or Y)

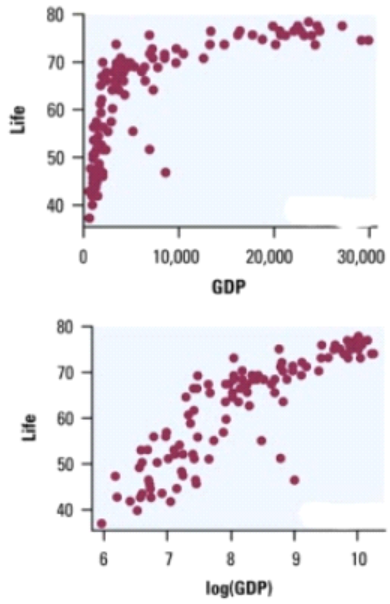


Transformations to Achieve Linearity





- Example

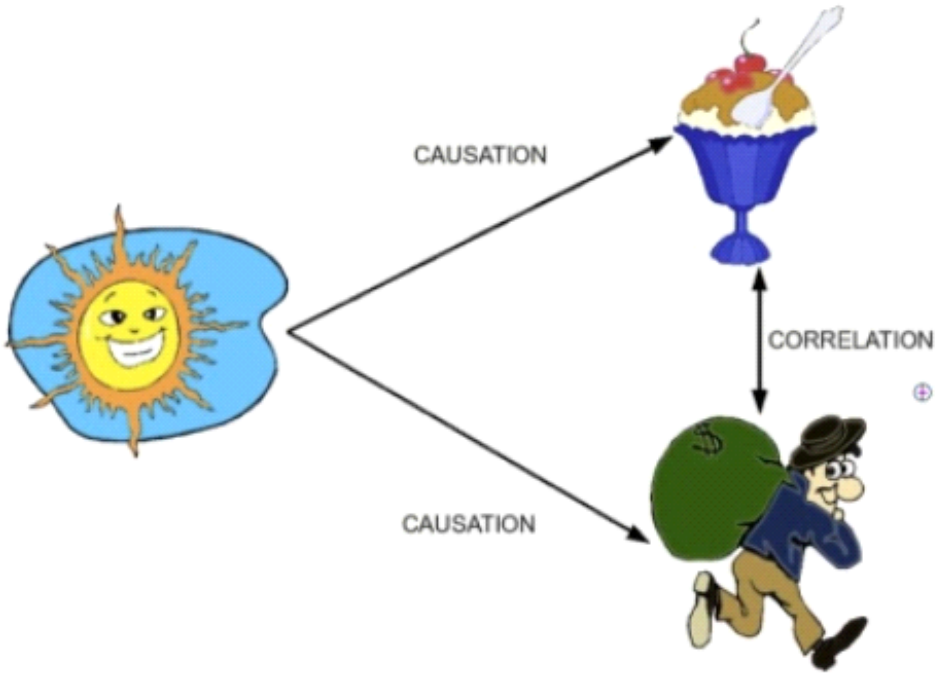


$$\text{Life} = 10 + .008(\text{GDP})$$

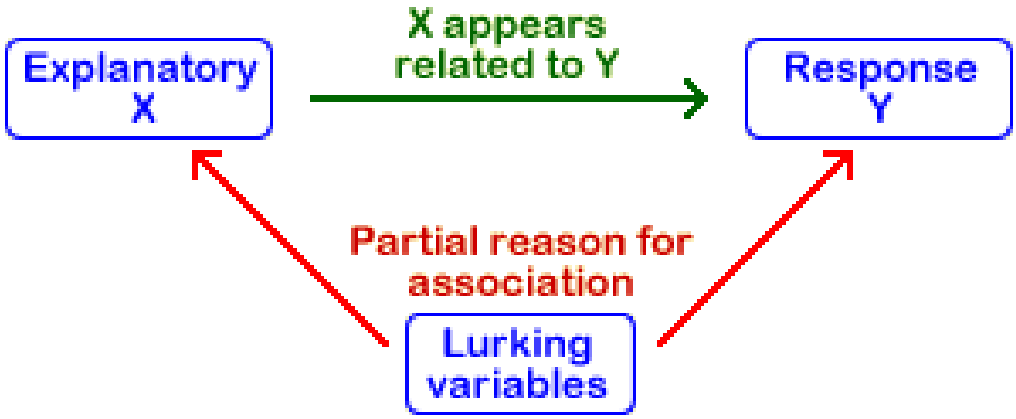
$$\text{Life} = 30 + 4.3(\text{Ln}(\text{GDP}))$$

Confounding

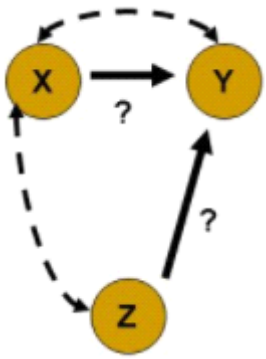
- Correlation does not prove causation



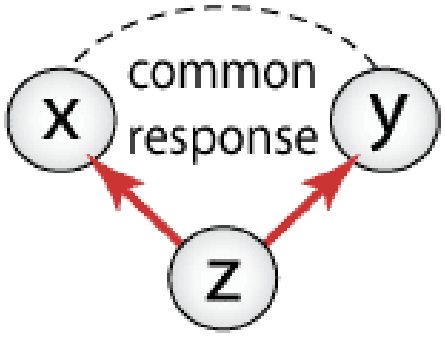
- A **Lurking Variables** is one that is not the explanatory (x) or response (y) variable in the study, but can influence how we interpret the relationship between them
- What looks like an association between x and y but is a **Common Response**

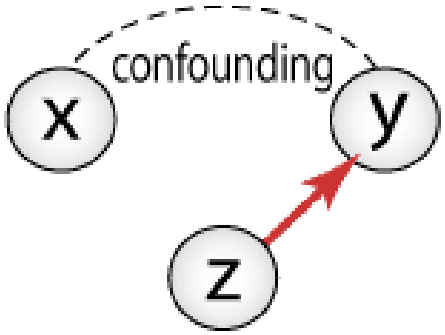


- Confounding
  - The response is at least partially due to a third factor

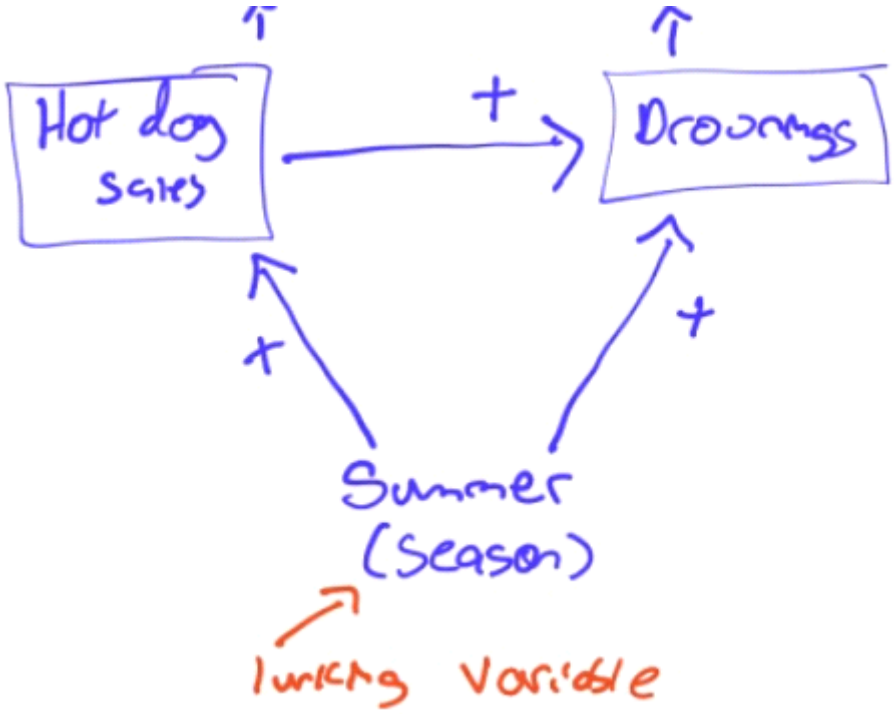


- Comparison

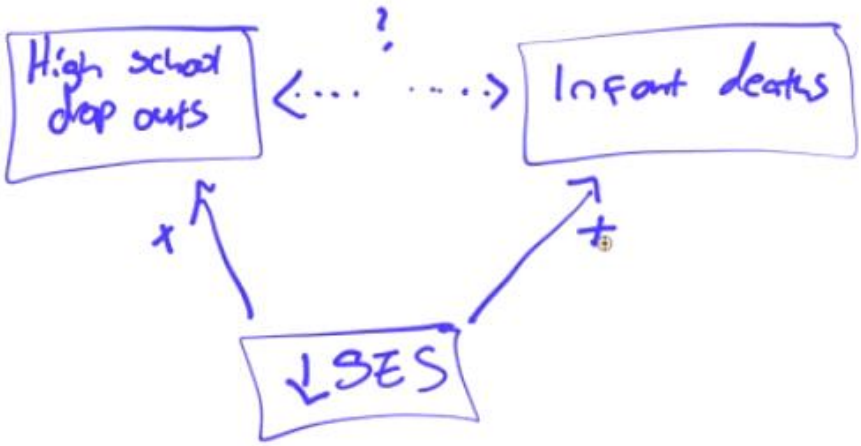




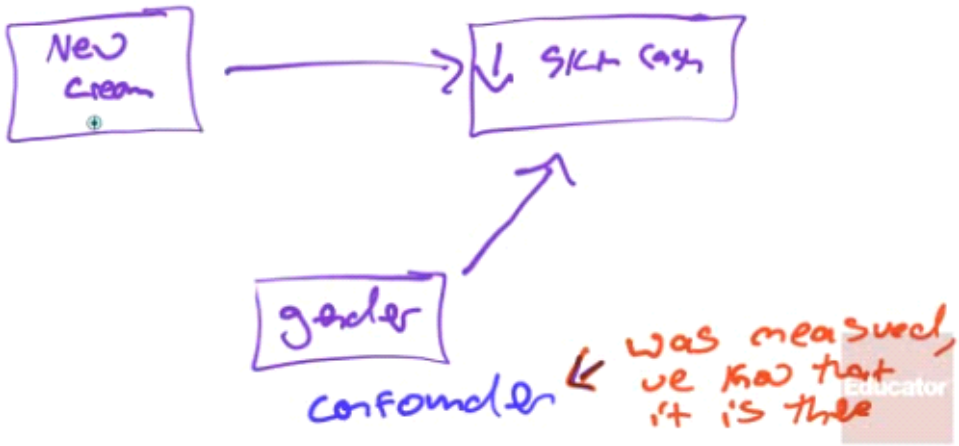
- Example
    - There is a positive association between the number of drowning and hot dog sales. Is the association between two variables most likely due to causation, cofounding, or common response? Justify your answer.
        - Answer: Common Response



- According to the 19990 census, those states with an above-average number X of people who fail to complete high school tend to have an above average number Y of infant deaths. In other words, there is a positive association between X and Y. The most plausible explanation for this is?
  - Answer: Common Response



- A drug company is testing a new cream to relieve skin rashes. They try it on 20 people and a placebo on 20 people and find that it works better. Later someone realizes that the new cream was tested on mostly all men and the placebo was tested on mostly all women. Is it possible the difference seen could have been due to confounding? What about common response?

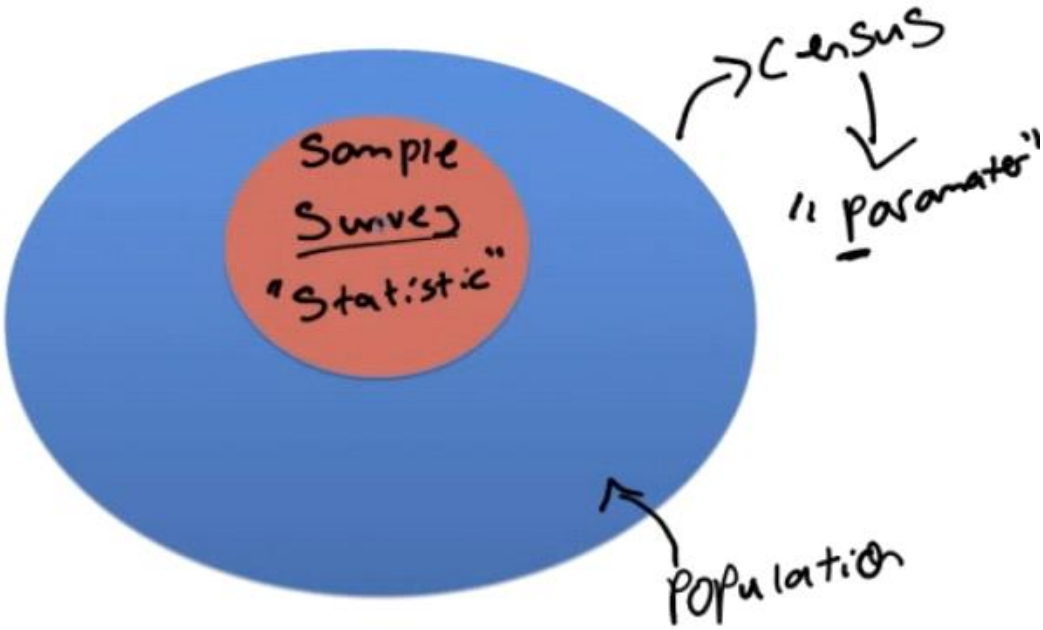

New Cream
Skin cancer
gender
confounder
was measured, we know that it is there

3.1 - Planning & Conducting Surveys

2017年2月10日 星期五　　下午 5:54

Census vs. Survey

- A **parameter** is a numerical description of a **population** characteristic
- A **statistic** is a numerical description of a **sample** characteristic



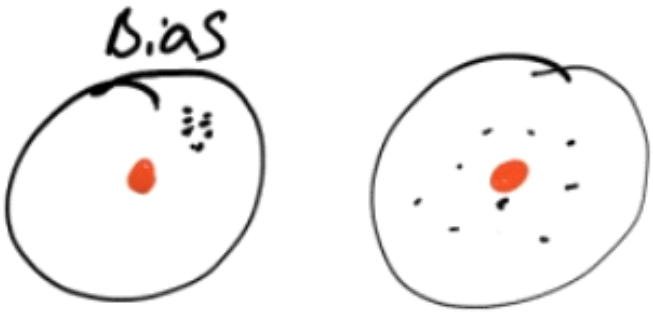
Census	Population	Parameter
Survey	Sample	Statistic

Characteristics of a Well-Designed and Well-Conducted Survey

- Representative sample
    - Must represent the population that you want to **draw inference** about
- Random sample
- Does not introduce bias

Bias

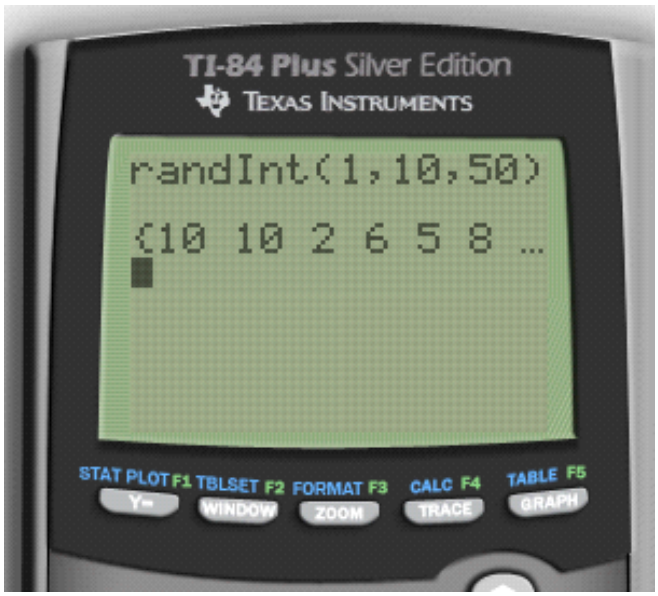
- What is it?



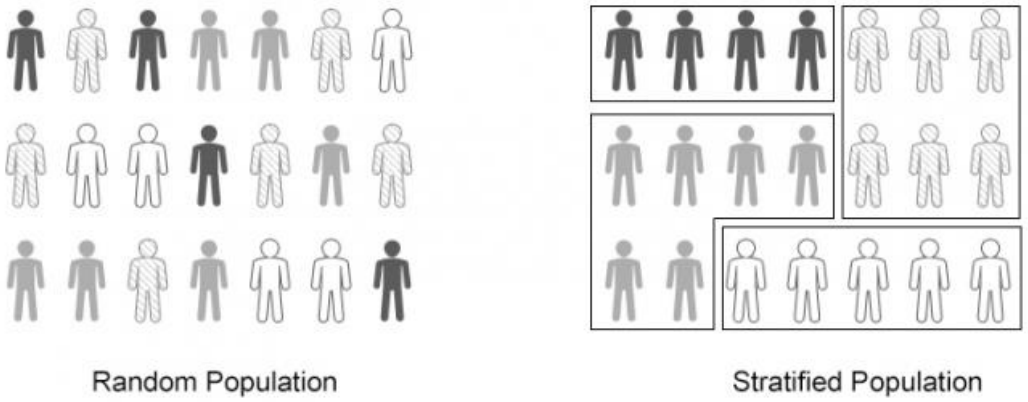
- How might it occur?
    - Wording of the question
    - Sample technique
        - Under-coverage: miss a certain group of people
        - Non-response: someone in your sample choose not to respond

Random Sampling

- Simple Random Sample (SRS)
    - Random digit table
    - Calculator
        - MATH --> PRB --> randInt(lower, upper, number of numbers)

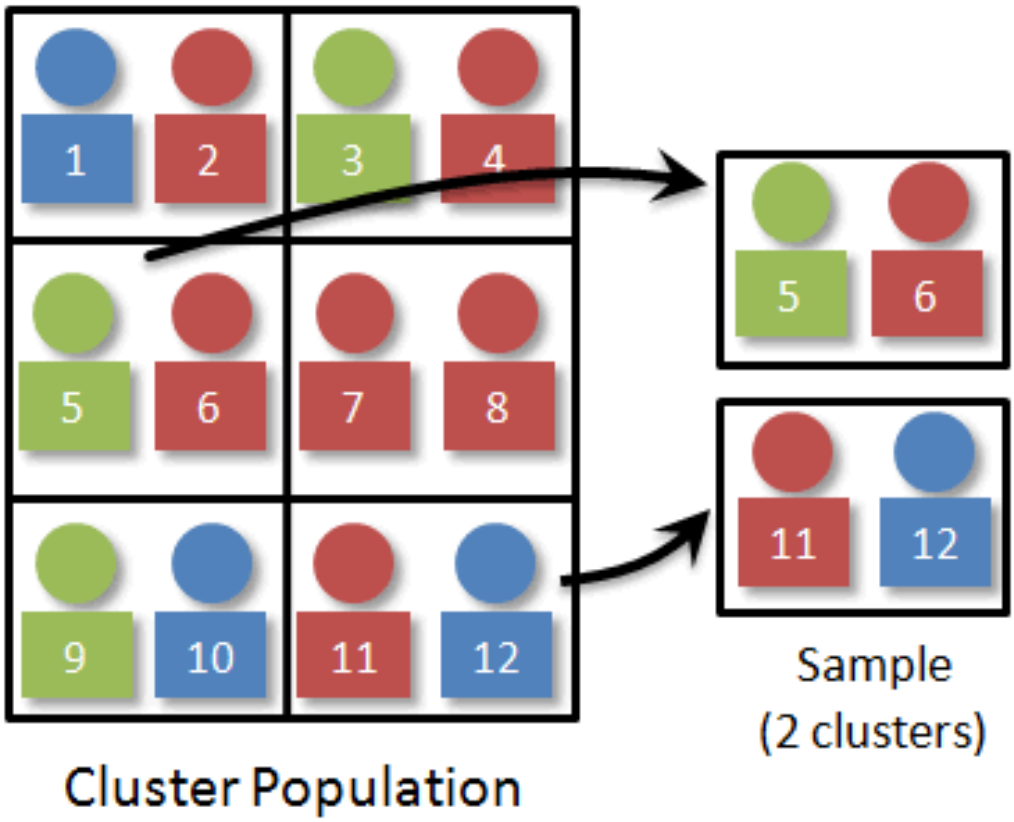


- Stratified Random Sampling

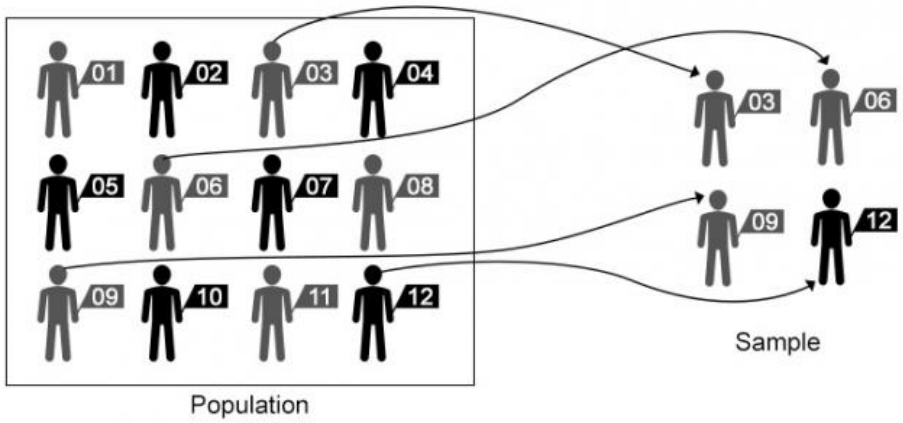


Stratified Random Sampling

- Cluster Sample



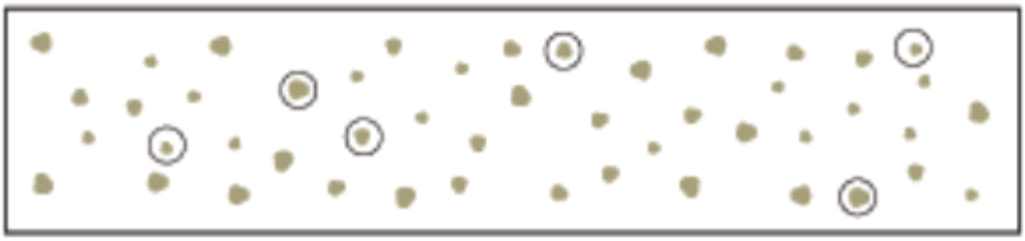
- Systematic Random Sample



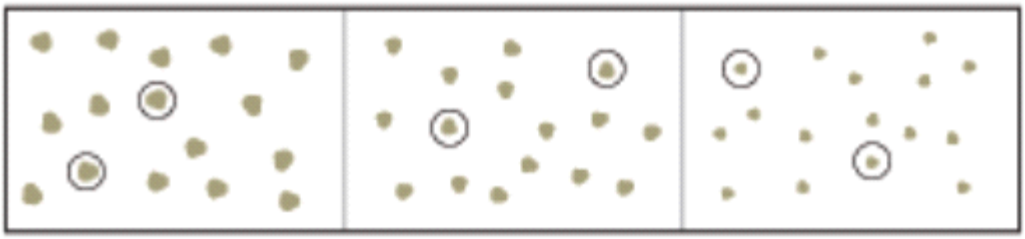
Systematic Random Sampling

- Comparison

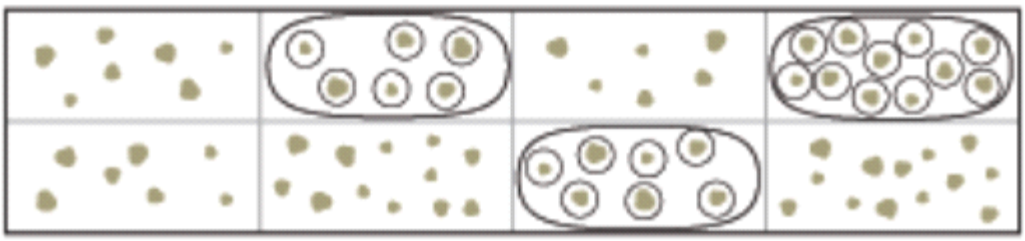
Simple Random Sampling



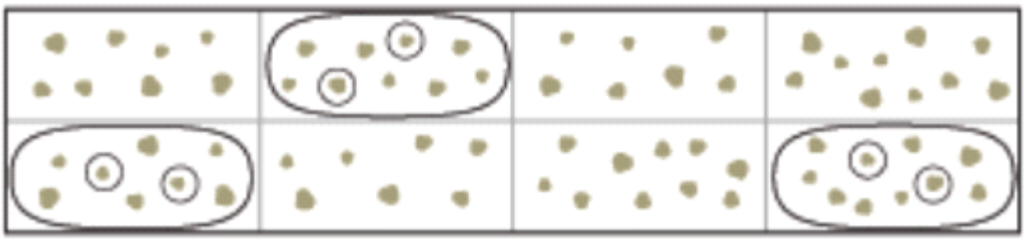
Stratified Random Sampling



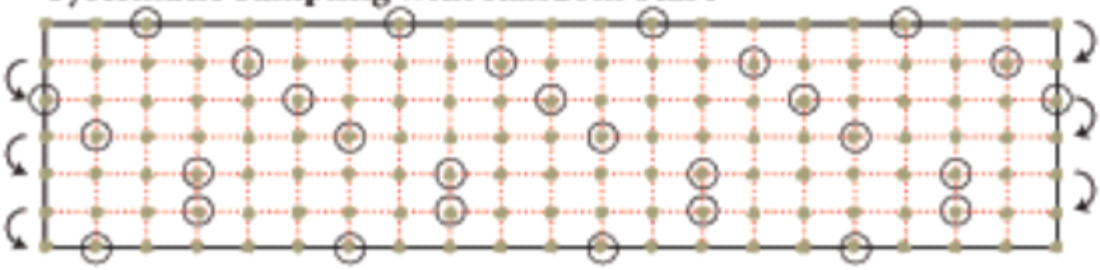
Cluster Sampling



Two-Stage Cluster Sampling



Systematic Sampling with Random Start



- Example
    a. If you wanted to select a sample of school children and you did this by selecting an SRS of **classrooms** this would be an example of what type of sampling?

        - Cluster

    b. If you wanted to select a sample of school children and you did this by randomly selecting 5 children from each **grade** this would be an example of what type of sampling?

        - Stratified

Non-Random Sampling

- Convenience Sample
    - Selecting those that are easy of convenient
- Voluntary Response Sample

3.2 - Planning & Conducting Experiments

Saturday, February 11, 2017 12:11 AM

Experiments vs. Observational Studies

- Ovservational Study
  - Observe individuals
  - Measure variables
  - **Do NOT influence the response**
    - Has global warming effected penguin mating behavior
- Experiment
  - **Do something** to your individuals
  - Observe/measure response
    - Does housing penguins in warmer environments effect mating behavior?

Placebo Effect

- The phenomenon where **patients get better** because they **expect** the treatment to work.
- Many statistical studies involve testing the effectiveness of drugs. A placebo looks identical to the actual drug but contains no active ingredient and so has no real physical effect.
- Humans want to be helped by the medication that is administered to them. If they think they are receiving a drug to help their condition, they tend to improve even if it turns out that the drug is a placebo.

Characteristics of a Well-Designed and Well-Conducted Experiment

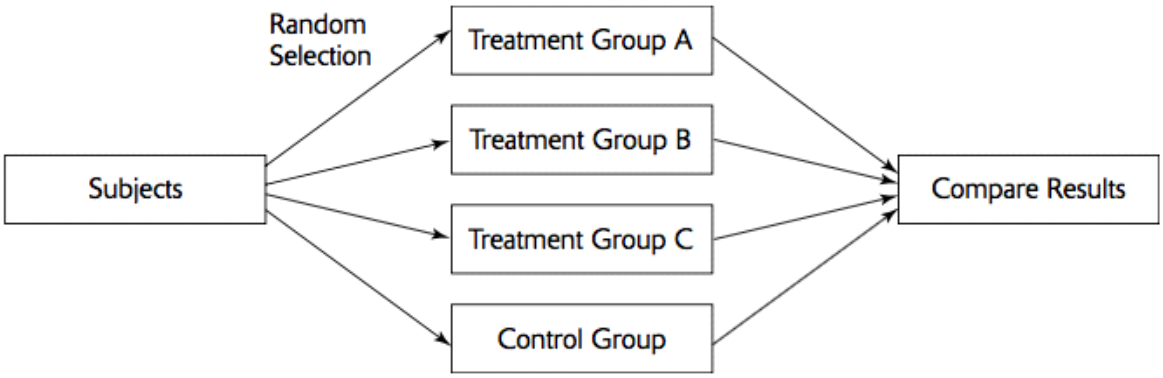
- Control
  - The effect of **lurking variables**, most often by comparing treatments
  - Example: a "Control group" in a drug study to eliminate the "confounding effects" of environment or the placebo effect
- Replicate

- Each treatment on many units to **reduce chance variation**
- Example: do the mouse study many times

- Randomize
  - Use probability (chance) to assign experimental units to treatments
  - May be the most **important**!!
  - Because it allows us to say the different treatment groups **start out similar**

Completely Randomized Design

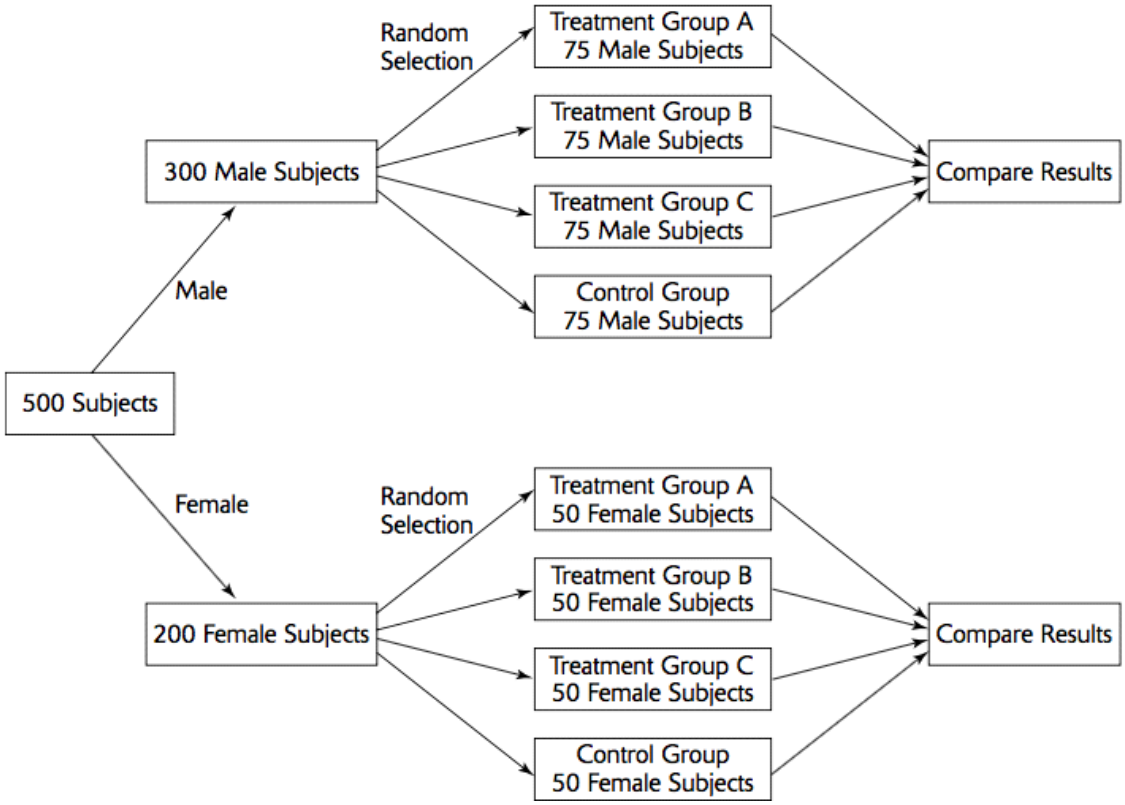
- If all the experimental units (subjects of the experiment) are **randomly** assigned to either the **control group** or to the **treatment group**, then the experiment has a completely randomized design.



- Randomize by assigning each subject a number and then generating it to choose treatment groups

Block Randomization

- Placing subjects into **groups of similar individuals**. The random assignments into treatment groups is carried out separately within each block (think stratified random sample)

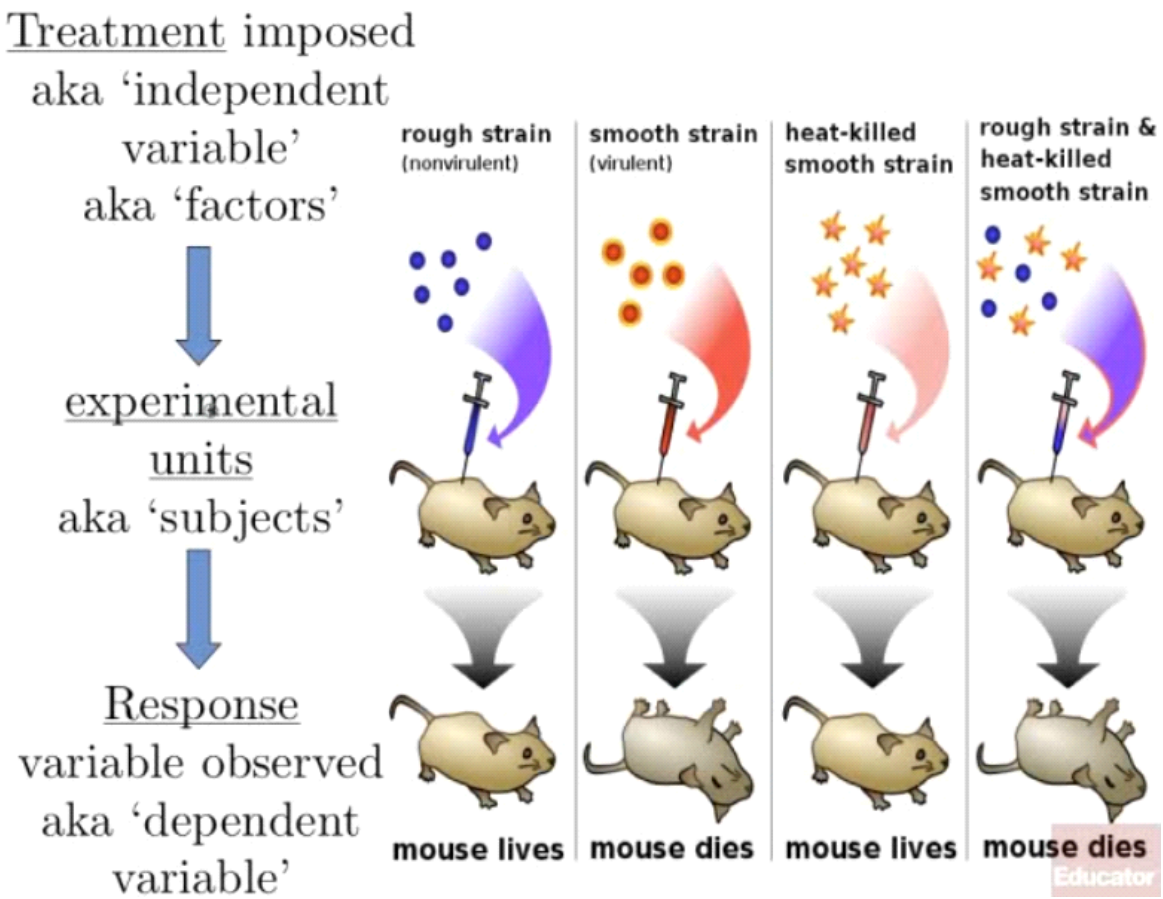


Matched Pairs Design

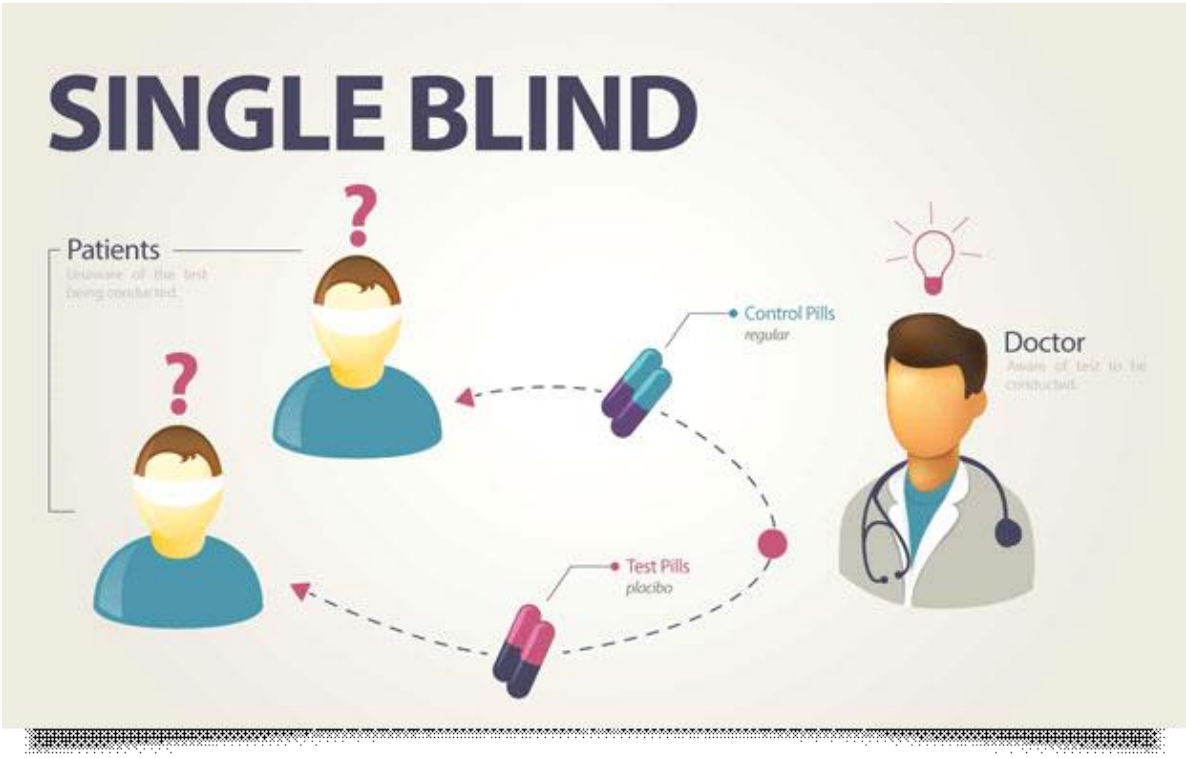
- Subjects are **matched into pairs** and get **different treatments**
- Matched pairs are **more similar** than random **unmatched subjects**
- Randomizing the **rest of the experiment** is still important!!!

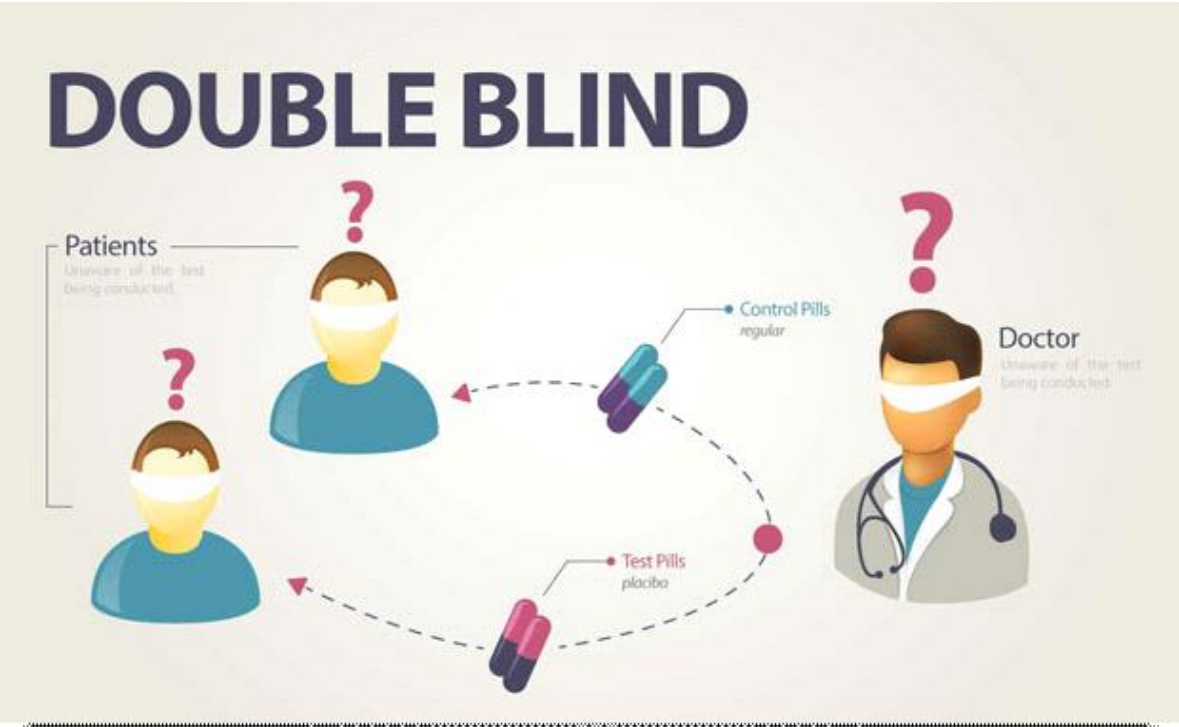
Experimental Set Up

- Treatment Imposed = Independent Variable = Factors
- Experimental Units = Subjects
- Response Variable Observed = Dependent Variable



Double-Blind Experiment





- In a double-blind experiment, neither the subjects nor the researchers know to which group, treatment, or control, subjects have been assigned. If a researcher knows that a subject is in the control group, they do not expect a treatment effect, and their measurement of a response might be understated. If a researcher knows that a subject is in the treatment group, they might overstate a response simply because they expect it.
- An experiment might also be single-blind. In this case, only one of the participants, either the subjects or the researchers, knows to which group the subjects have been assigned.
- **Avoids unconscious bias**

Generalizability of Results

- To determine if our data is "statically significant"
    - i.e. is an observed effect so large that it would rarely occur by chance
- If we designed and conducted our experiment well, we can generalize these results to the population!

Practice Questions

- Control groups are used in experiments in order to
    a. Control the effects of outside variables on the outcome

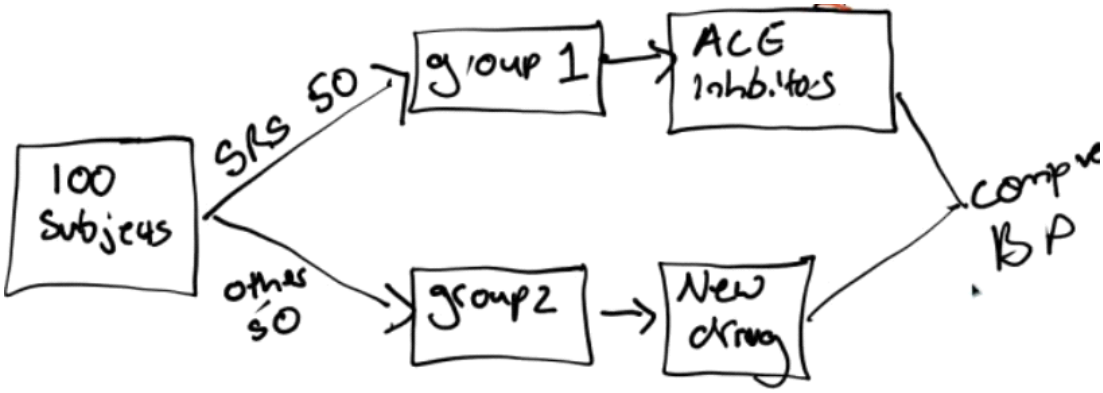
b. Control the subjects of a study to ensure that all participate equally

c. Guarantee that someone other than the investigators, who have a vested interest in the outcome, controls how the experiment is conducted

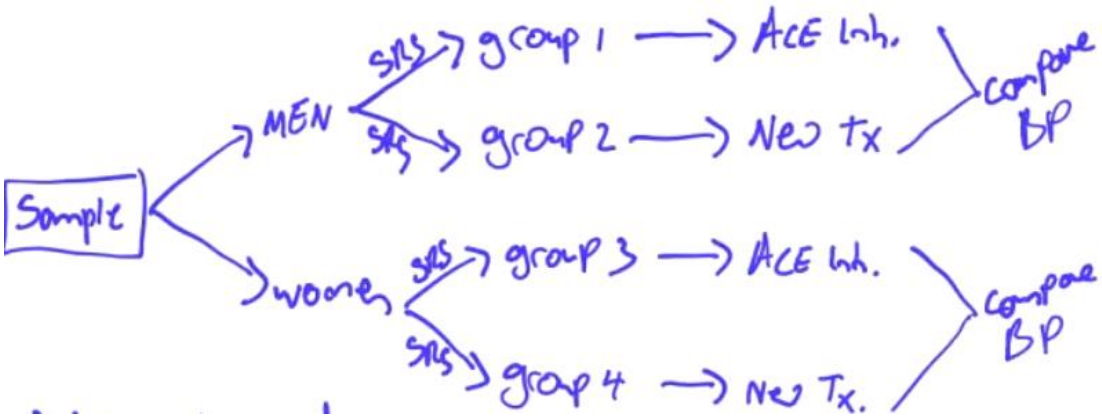
d. Achieve a proper and uniform level of randomization

Answer: a

- Angiotensin-converting enzyme (ACE) inhibitors are used to treat high blood pressure. We want to conduct an experiment to see if a new blood pressure drug works even better than ACE inhibitors. Design a completely randomized experiment to test this.



- In conducting an experiment to see if a new blood pressure drug works even better than ACE inhibitors. We learn that men and women may react differently to common cardiovascular drug treatments. Design a randomized experiment to test this with your new information on gender.
    - We will conduct a randomized blocked experiment, blocking on gender.



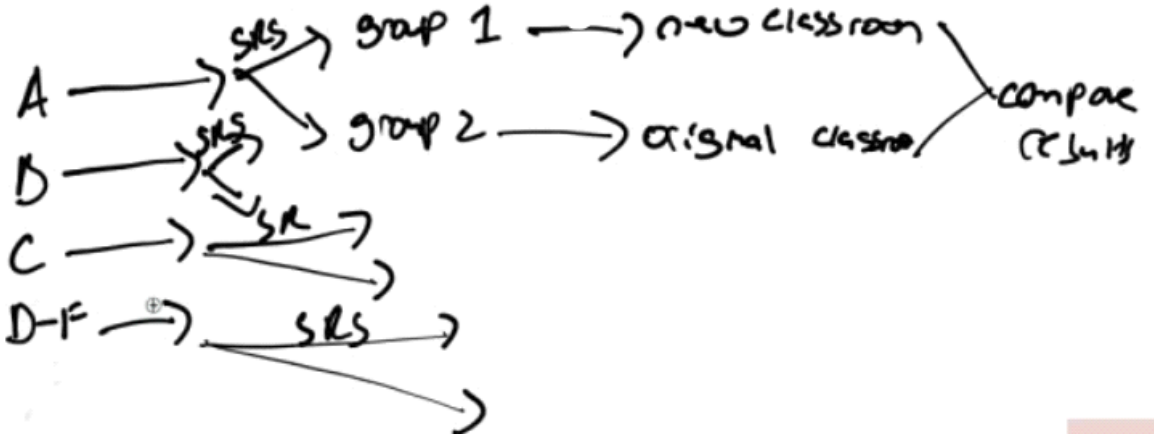
- The Community Intervention Trial for Smoking Cessation (COMMIT) asked whether a community-wide advertising campaign would reduce smoking. The researchers located 11 **pairs** of communities that were **similar** in location, size, economic status, and so on. One community in each pair participated in the advertising campaign and the other did not. This is

a. an observational study
b. a matched pairs experiment
c. a completely randomized experiment
d. a block design

Answer: b

- A study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same sex, age, and race who did not have brain cancer, then asked about the use of cell phones. This is
  a. an observational study
  b. an uncontrolled experiment
  c. a randomized comparative experiment
  d. a matched pairs experiment
  e. a survey

  Answer: a
- A fitness instructor wants to test the effectiveness of a performance-enhancing herbal supplement. Design an experiment to test this supplement
  - Double blind, (placebo controlled), matched pairs experiment:
  - Match subjects based on performance in a fitness test and gender
  - Randomize who in the pair gets the new supplement and who gets the old supplement / placebo.
  - Give both in the same packaging, making sure the subject doesn't know the group and person measuring fitness doesn't know the group either.
- A researcher believes that students may do better on a test when taken in the same classroom where the material was learned. To test this theory she plans to present a lecture and then give students a multiple choice quiz on the material. She knows there is a lot of variability in the students' academic ability. Design a study to test her hypothesis.
  - Block on academic ability


A
B
C
D-F
SRS
group 1
group 2
new classroom
original classroom
compare

4.1 - Probability Overview

Monday, February 13, 2017 3:39 PM

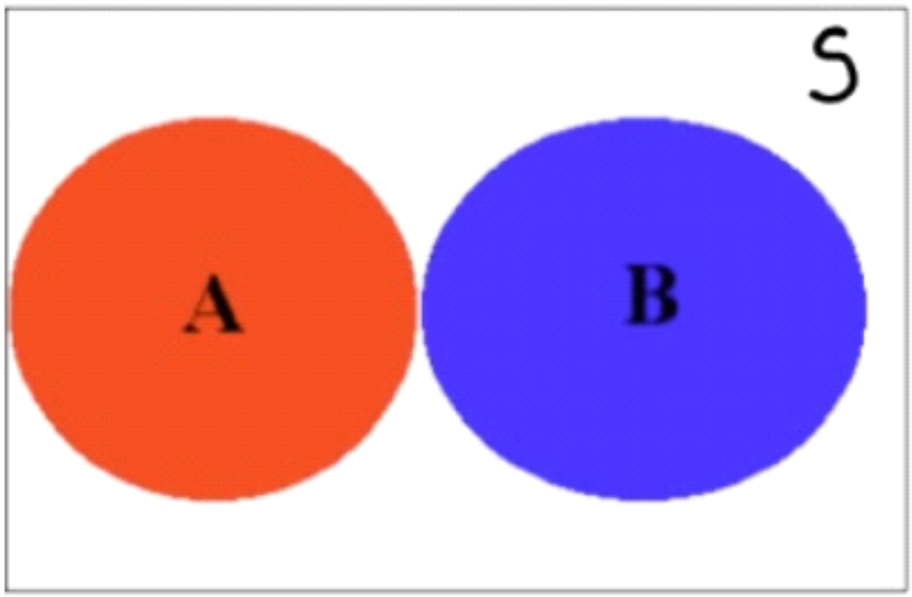
Probability Model

- A mathematical description of a random phenomenon.
- The Probability Model consists of
  - **Sample Space** (S) = the set of all possible outcomes
  - **Event** within the sample space = an outcome or set of outcomes in S
  - **Probabilities** associated with each event

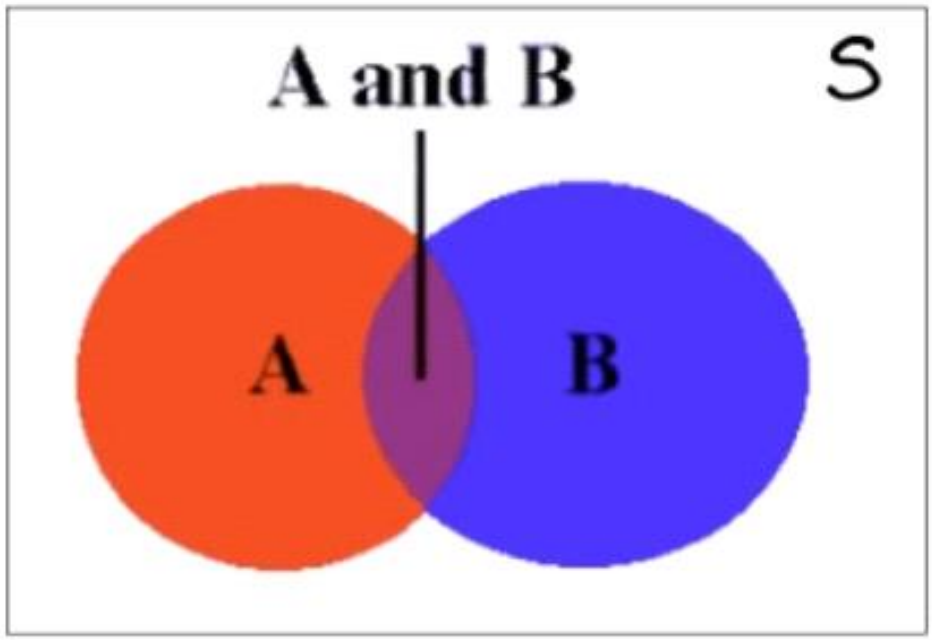
X	H	T
P(x)	1/2	1/2

Disjoint Events (aka Mutually Exclusive)

- Disjoint / mutually exclusive example



- Not disjoint / not mutually exclusive example



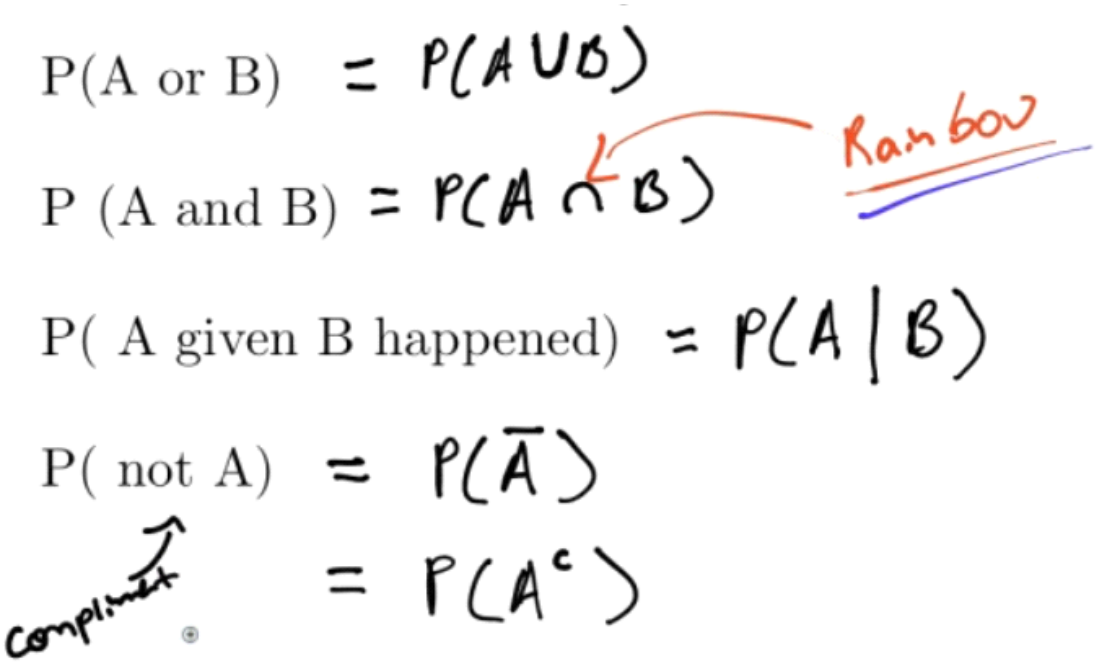
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independence vs. Dependence

- A conditional probability is the probability of some event occurring, given that some other event has already occurred. The conditional probability of event X occurring, given that some other event Y has already occurred, is written as P(X|Y).
- For example, P(M|N) would be the probability of the occurrence of event M given that event N has already occurred. It would be read as "the probability of M, given N."
- As stated earlier, two events are considered independent if the occurrence of one of the events does not change the probability of the other event from what it would have been had the first event not occurred. Thus, two events, X and Y, are independent if P(X|Y) = P(X) or P(Y|X) = P(Y)
- Actually, these two conditional relationships are related. If one is true, the other must be true. If one is false, the other must be false.
- If **P(X|Y) = P(X)**, then **P(Y|X) = P(Y)**, and the events are **independent**.
- If **P(X|Y) ≠ P(X),** then **P(Y|X) ≠ P(Y)**, and the events are **dependent**.

Probability Rules

- Notation



- Rules

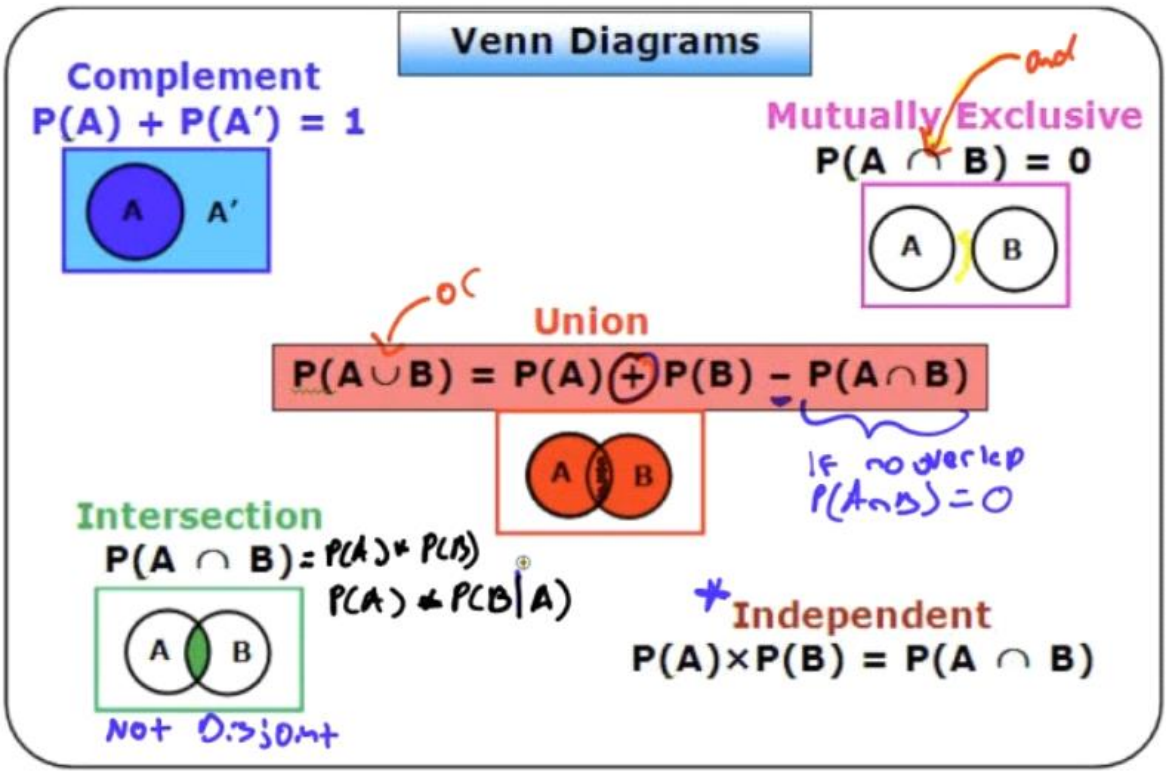
The Addition Rule: $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$.

Special case of *The Addition Rule*: If A and B are *mutually exclusive*, $P(\text{A and B}) = 0$, so $P(\text{A or B}) = P(\text{A}) + P(\text{B})$.

The Multiplication Rule: $P(\text{A and B}) = P(\text{A}) \bullet P(\text{B|A})$.

Special case of *The Multiplication Rule:* If A and B are *independent*, $P(\text{B|A}) = P(\text{B})$, so $P(\text{A and B}) = P(\text{A}) \bullet P(\text{B})$.

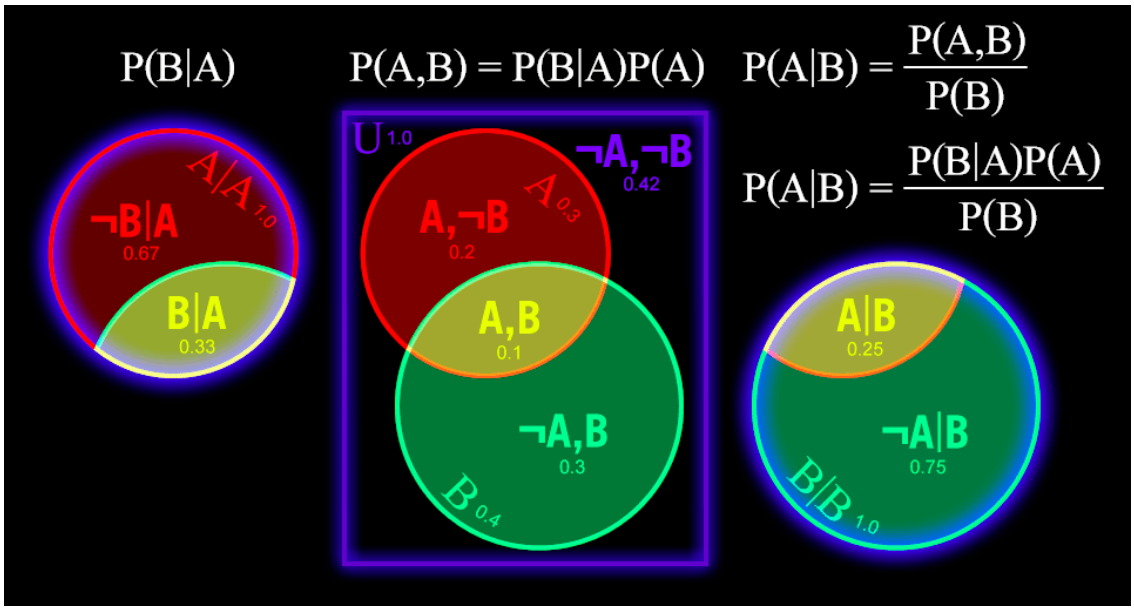
Overview

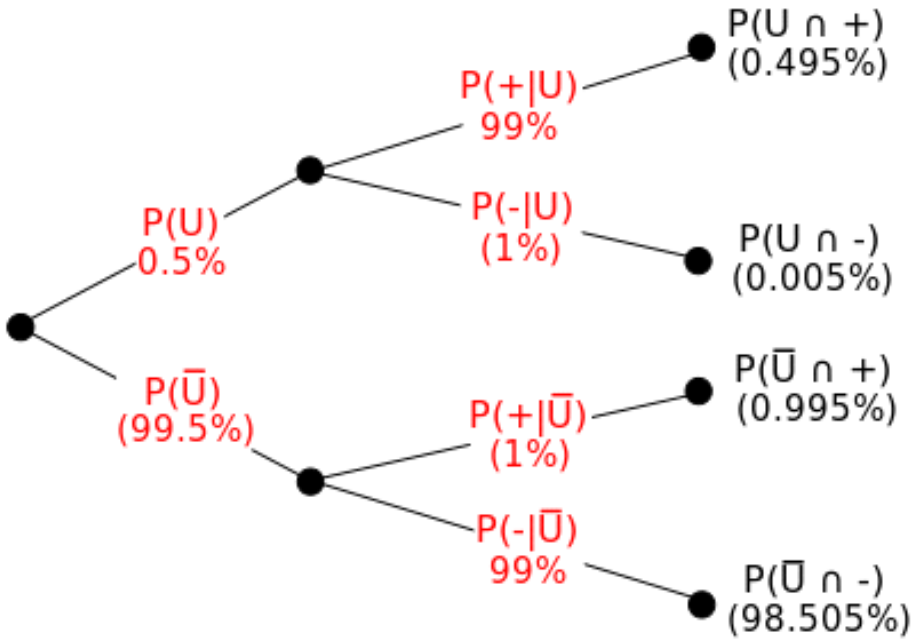


Bayes Rule

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D)P(D)}{P(T)}$$

$$= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)}$$



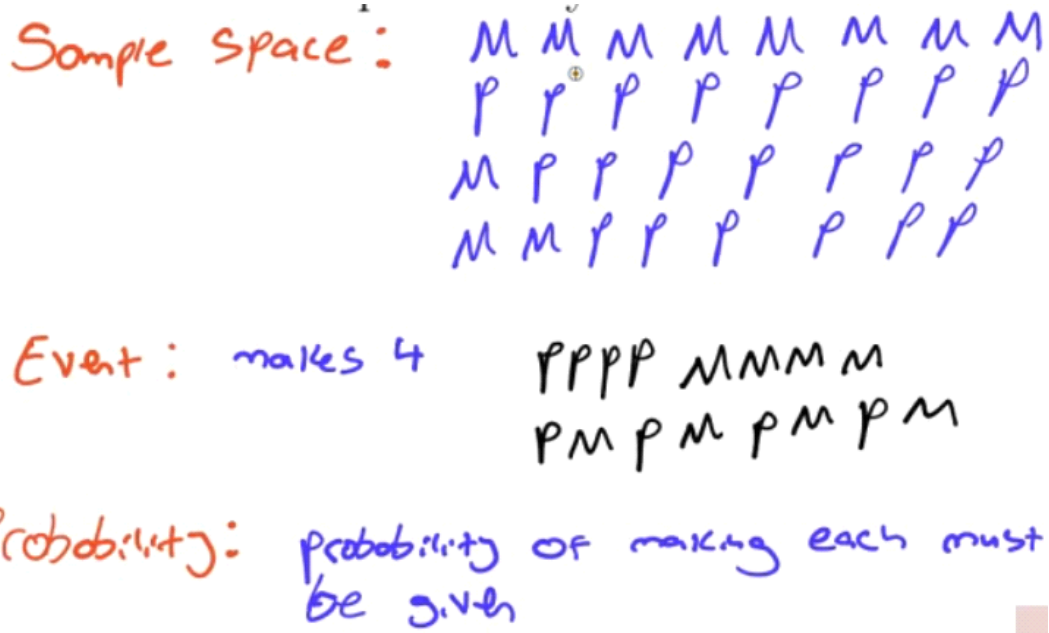


Simulations

- Imitating a real world process
- Follow a few steps
    1. **Describe** the possible outcomes
    2. **Determine** the **probability** of each outcome
    3. **Link** each outcome to one or more **random numbers**
    4. **Choose** a random number for each "trial"
    5. Based on the random number, note the "simulated" outcome
    6. Repeat step 4 and 5 for each "simulation"

Practice Questions

- A basketball player shoots 8 free throws during a game. We are interested in the probability that she makes 4.
    - Describe the probability model in this case



- Let A be the event that a victim of violent death was a woman and B the event that the death was a suicide. The proportion of suicides among violent deaths of woman is expressed in probability notation as
    a. 0.5
    b. 0.126
    c. P(A and B)
    d. P(A|B)
    e. P(B|A)

    Answer: e
- Suppose that for a group of consumers, the probability of eating pretzels is 0.65 and that the probability of drinking Coke is 0.75. The probability of eating pretzels and drinking Coke is 0.45. Determine if those two events are independent.
    - P(P)=0.65
    - P(C)=0.75
    - P(P and C)=0.45
    - P(P and C)≠P(P)*P(C)
    - So P and C are not independent
- What is the probability that the students is either a female of at least 35 years old?

Age	14-17	18-24	25-34	≥35	
Male	30	500	120	6	**656**
Female	50	620	130	12	**812**
	80	**1120**	**250**	**18**	

$P(F \text{ or } \geq 35)$

$P(F \cup \geq 35) = P(F) + P(\geq 35) - P(F \cap \geq 35)$

$\frac{812}{1468} + \frac{18}{1468} - \frac{12}{1468}$

$= .557$

- Generally speaking, 1% of college football players use steroids, but your friend on the football team has tested positive for steroid use. The blood test the league uses correctly comes up positive for 95% of steroid users, but mistakenly comes up positive for 1% of non-users. Given that your friend tested positive, how likely is it that he has been using steroids?

$P(S) = .01$

$P(+|S) = .95$

$P(+|\bar{S}) = .01$

$P(\bar{A}) = 1 - P(A)$

$P(\bar{S}) = 1 - P(S) = .99$

$P(S|+) = \frac{P(S \cap +)}{P(+)} = \frac{P(+|S) \cdot P(S)}{[\text{"}\,\text{"}] +}$

$+ \text{ w/ steroids} = P(+|S) \cdot P(S)$

OR

$+ \text{ w/o steroids} = P(+|\bar{S}) \cdot P(S)$

$= \frac{P(+|S) \cdot P(S)}{[P(+|S) \cdot P(S)] + [P(+|\bar{S}) \cdot P(\bar{S})]}$

$= \frac{(.95) \cdot (.01)}{(.95 \cdot .01) + (.01 \cdot .99)} = .49$

- Of all the soda drinkers in a particular district, 40% prefer brand A and 60% prefer brand B. Of those drinkers who prefer brand A, 30% are females, and of those who prefer brand B, 40% are female. What is the probability that a randomly selected soda drinker prefers brand A, given that the person selected is a female

$P(A) = .40$

$P(\bar{A}) = .60$

$P(F|A) = .30$

$P(F|\bar{A}) = .40$

$$P(A|F) = \frac{P(A \cap F)}{P(F)} = \frac{P(F|A) \cdot P(A)}{P(F)}$$

$F \cup A = P(F|A) \cdot P(A)$

or

$F \cup \bar{A} = P(F|\bar{A}) \cdot P(\bar{A})$

$$P(A|F) = \frac{P(F|A) \cdot P(A)}{[P(F|A) \cdot P(A)] + [P(F|\bar{A}) \cdot P(\bar{A})]}$$

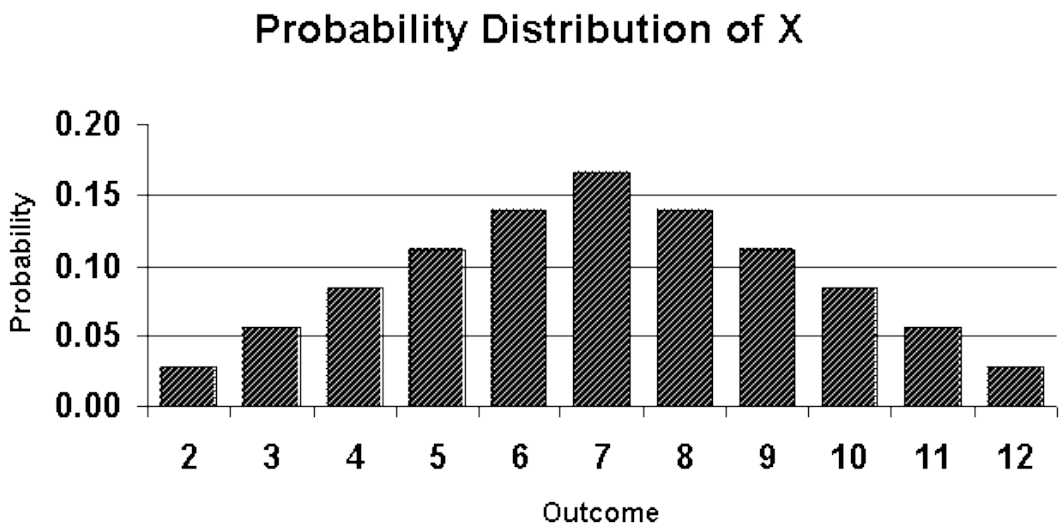
$$= \frac{(.30)(.40)}{[(.30)(.40)] + [(.40)(.60)]} = \frac{.12}{.36} = .33$$

4.2 - Intro to Probability for Discrete Random Variables

Monday, February 13, 2017 5:09 PM

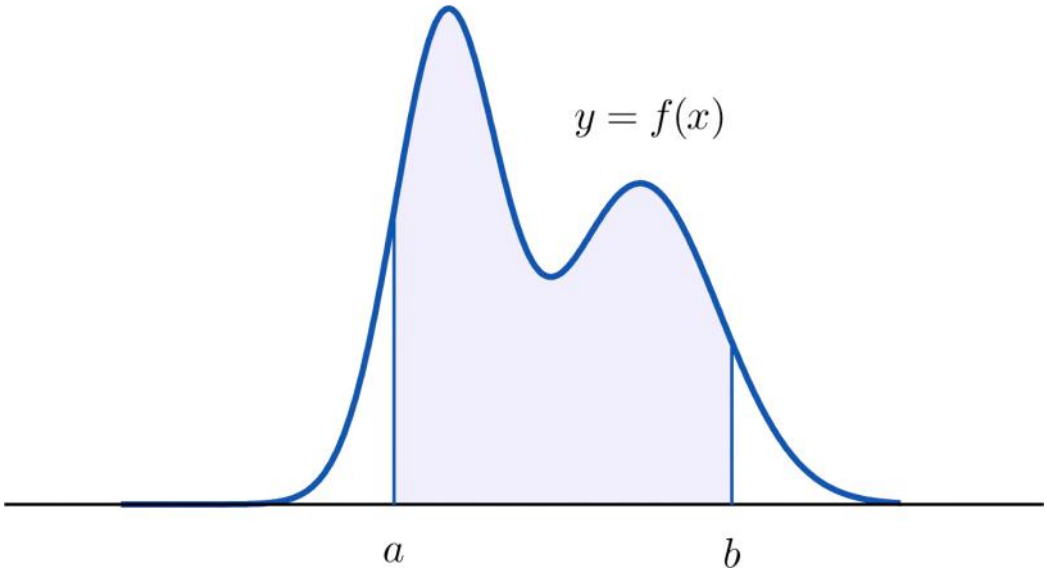
Discrete vs. Continuous RVs (Random Variables)

- A discrete RV has a countable number of possible outcomes
  - number of students present
  - number of red marbles in a jar
  - number of heads when flipping three coins
  - students' grade level



- A continuous RV can take any values in an interval of numbers
  - height of students in class
  - weight of students in class
  - time it takes to get to school
  - distance traveled between classes

$$P(a < X < b) = \text{area of shaded region}$$



Probability Distribution

- The probability distribution of a discrete RV(X) list all the values possible and their probabilities

List of possible values

x	0	1	2
P(X=x)	1/4	1/2	1/4

Probability of each value

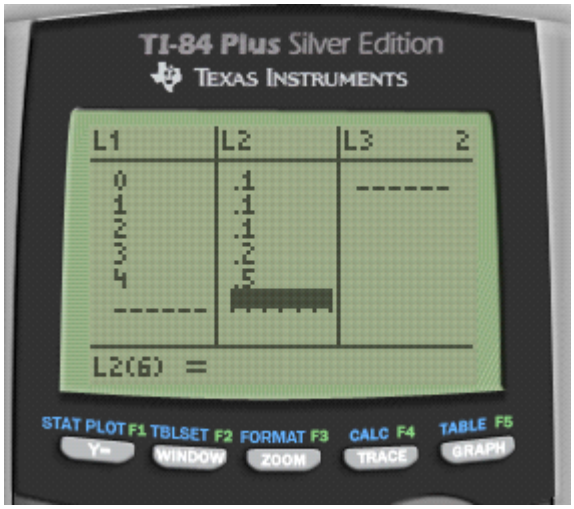
- The probabilities must:
  - All be a number between 0 and 1
  - Together add up to 1

Mean = Expected Value

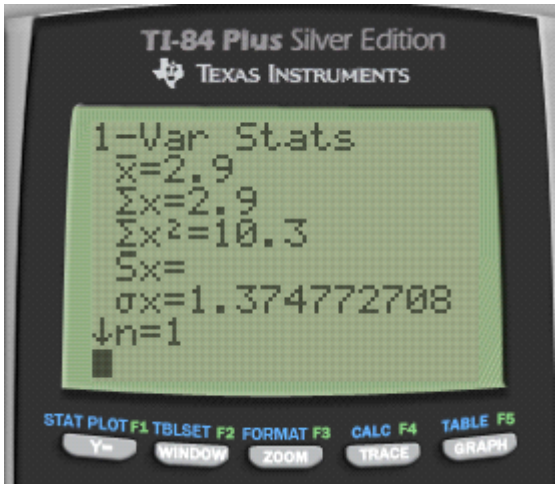
- Formula

$$\mu_x = x_1p_1 + x_2p_2 + \cdots + x_kp_k$$
$$= \sum x_ip_i$$

- Calculator
  - Type in X in L1 and P in L2



- 1-Var Stats L1, L2

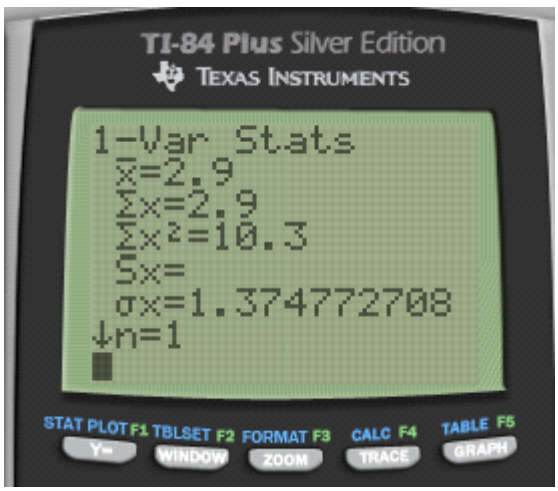


Variance = Standard Deviation

- Formula

$$\sigma_X^2 = \sum (x_i - \mu_x)^2 p_i$$

- Calculator



Practice Questions

- A bin contains ten $1 bills, five $2 bills, three $5 bills, one $10 bill, and one $100 bill. A person is charged $10 to select one bill. Let the random variable (X) be the amount someone wins by playing.
    - Construct a probability distribution for these data.

X	$-\$9$	$-\$8$	$-\$5$	$\$0$	$+\$90$
$P(X=x)$	$\frac{10}{20}$	$\frac{5}{20}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{1}{20}$

    - What is the mean and standard deviation of the amount of money someone can expect to win?

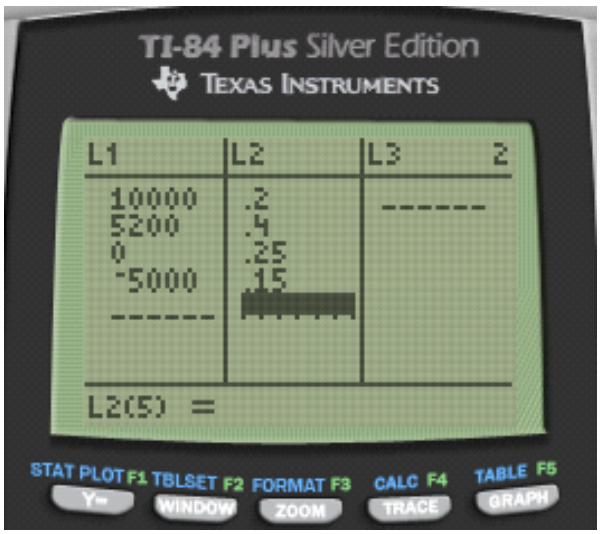
$$\mu_x = \sum_{i=1}^{6}(x_i)(p_i) = (-9)\left(\frac{10}{20}\right) + \cdots + (90)\left(\frac{1}{20}\right)$$

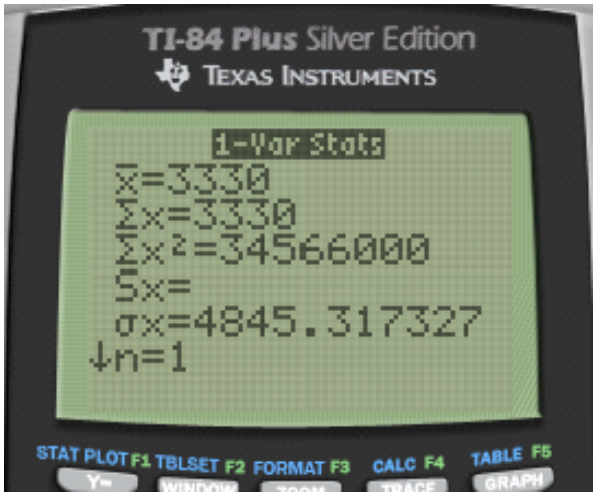
$$-4.5 \quad + \cdots + 4.5$$

$$= -2.75$$

$\sigma_x = 21.39$ $\quad \sigma_x^2 = \sum (x_i - \mu)^2 p_i = [-9-(-2.75)]^2\left(\frac{10}{20}\right) +$

- You work for a company and are tasked with evaluating a proposed venture. The venture stands to make a profit of $10,000 with probability 4/20, to make a profit of $5200 with probability 8/20, to break even with probability 1/4, and to lose $5000 with probability 3/20. The expected profit in dollars is? Would you recommend this venture?





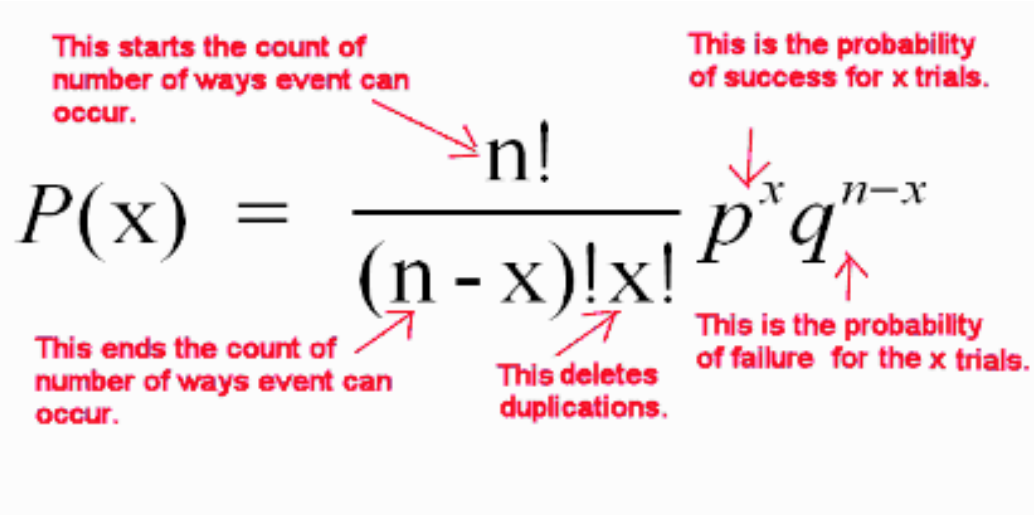
- Answer: The expected profit is $3330. Recommend, because the standard deviation is not so large

4.3 - Discrete Random Variables

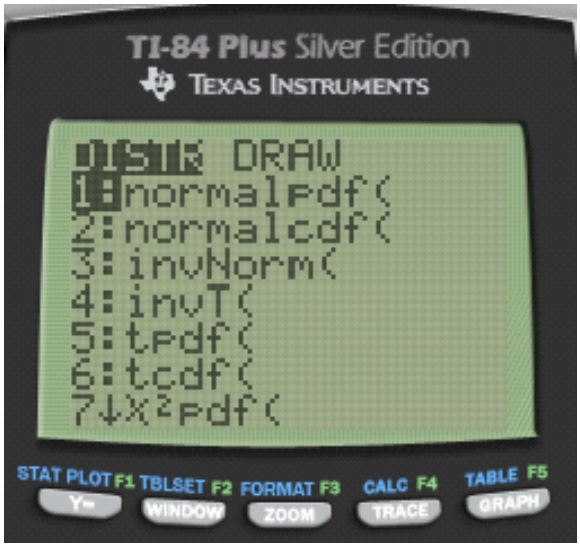
Monday, February 13, 2017 10:24 PM

Binomial Distribution

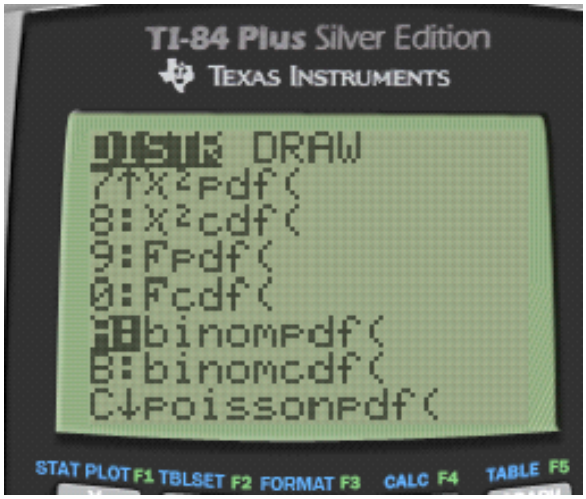
- BINP
  - B = Binary process = 2 process
  - I = Independent event
  - N = Number of trials
  - P = Probability of success
- Binomial Probability
  - binompdf(n, p, x)
  - n=trials
  - p=probability
  - x=value



- Calculator
  - 2ND + VARS (DISTR)



- A: binompdf / B: binomcdf



- binompdf vs binomcdf

$$P(X = c) = \text{binompdf}(n, p, c)$$

n -> number of trials

p -> probability of success

This finds the probability of exactly c successes, for some number c.

$$P(X \le c) = \text{binomcdf}(n, p, c)$$

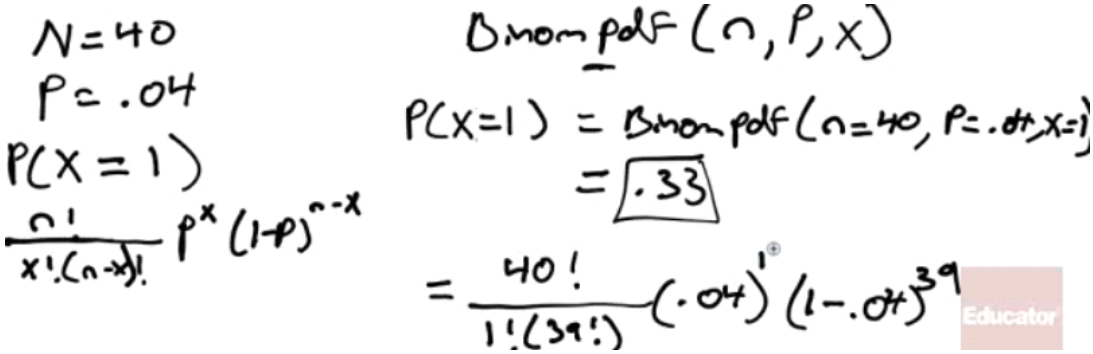
n -> number of trials

p -> probability of success

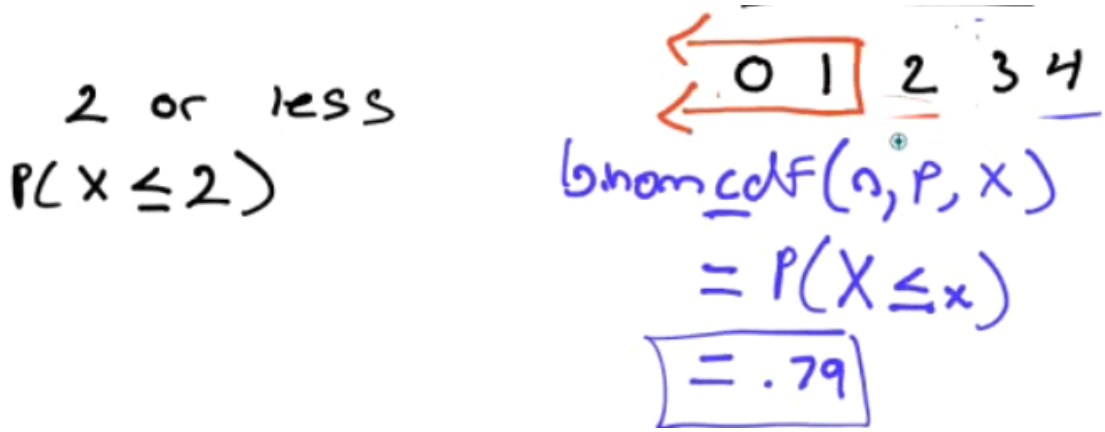
This finds the probability of
c or fewer successes.

Practice Questions for Binomial Distribution

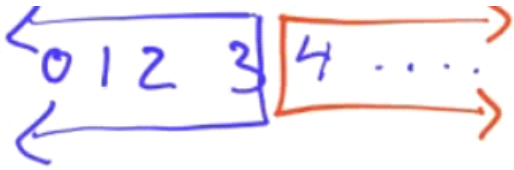
- A manufacturer produces a large number of toasters. From past experience, the manufacturer knows that approximately 4% are defective. In a quality control procedure, we randomly select 40 toasters for testing.
  - Determine the probability that exactly one of the toasters is defective



  - Find the probability that at most two of the toasters are defective



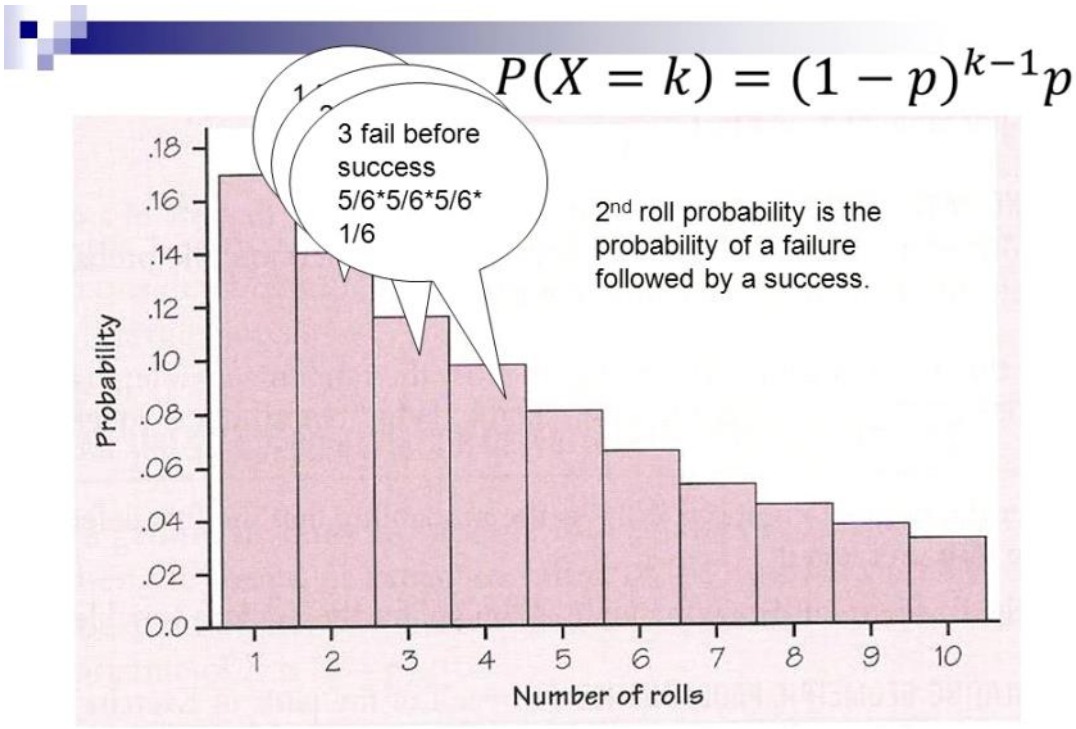
  - Find the probability that more than three toasters are defective



$1 - \text{Binomcdf}(40, .04, 3) = .07$

Geometric Distribution

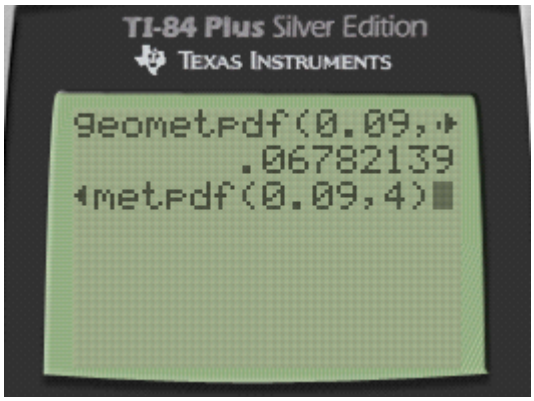
- BI~~N~~P
  - Not given the number of trials
- Question Format
  - How many trials until a success
- Geometric Probability
  - geometpdf(p,x)
  - geometcdf(p,x)
  - p=probability of success
  - x=number of trials until 1 success



AP Statistics, Section 8.2.2 6

Practice Questions for Geometric Distribution

- There is a probability of 0.09 that a vaccine will cause a certain side effect. Suppose that a number of patients are inoculated with the vaccine. We are interested in the number of patients vaccinated until the first side effect is observed
    - Find the probability that exactly 4 patients must be vaccinated in order to observe the first side effect.



$$P(X=4)=(.09)(1-.09)^{4-1}$$

    - What is the probability that the number of patients vaccinated until the first side effect is observed at most 5?

$$P(X \leq 5)$$

$$= \text{Geometcdf}(p, x) = .38$$

$p = .09$, $x = 5$

Mean and Standard Deviation

Binomial	Mean	$\mu = np$
	Variance	$\sigma^2 = npq$
	Standard Deviation	$\sigma = \sqrt{npq}$
Geometric	Mean	$\mu = \frac{1}{p}$
	Variance	$\sigma^2 = \frac{q}{p^2}$
	Standard Deviation	$\sigma = \sqrt{\frac{q}{p^2}}$

4.4 - Combining Independent Random Variables

Tuesday, February 14, 2017 2:11 PM

Mean and SD of Two Random Variables

- Mean

$$\mu_{x+y} = \mu_x + \mu_y$$

$$\mu_{x-y} = \mu_x - \mu_y$$

- Variance

$$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$$

$$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y$$

Transforming Random Variables (Linear Transformations)

- Mean

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

- Variance

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

Practice Questions

- You have a sample of male/female couples. The mean height of all the woman is 65 inches, with a standard deviation of 4 inches. The mean height of all the men is 70 inches, with a standard deviation of 6 inches.
    - What is the average combined height of the couples? What is the standard deviation of the combined height?

$$\mu_{W+M} = \mu_W + \mu_M$$
$$= 65 + 70$$
$$= 135$$

$$\sigma^2_{W+M} = \sigma^2_W + \sigma^2_M$$
$$= (4)^2 + (6)^2$$
$$= 16 + 36$$
$$= 52 \rightarrow \sqrt{52} = \boxed{7.2} = \sigma_{W+M}$$

- What is the average difference in the height of the couples? What is the standard deviation of the difference

$$\mu_{W-M} = \mu_W - \mu_M$$
$$65 - 70 = -5$$

$$\sigma^2_{W-M} = \sigma^2_W + \sigma^2_M$$
$$4^2 + 16^2 = 52$$
$$\sigma_{W-M} = \sqrt{52} = 7.2$$

- A report of the National Center for Health Statistics says that the height of 20-year-old men have mean 176.8 cm and standard deviation 7.2 cm. There are 2.54 cm in an inch. What are the mean and standard deviation in inches?

$$\mu_{inches} = \frac{\mu_{cm}}{2.54}$$

$$= \frac{176.8}{2.54} = 69.6$$

$$\sigma^2_{inches} = \left(\frac{1}{2.54}\right)^2 (7.2)^2$$

$$\sigma^2_{inches} = \left(\frac{1}{2.54}\right)^2 (7.2)^2$$

$$= 8.035$$

$$\sigma_{inches} = \sqrt{8.035} = 2.834$$

- The number of calories in a one-ounce serving of a breakfast cereal is a random variable with mean 110. The number of calories in a full cup of whole milk is a random variable with mean 140. For breakfast you eat one ounce of the cereal with 1/2 cup of whole milk. Let Z be the random variable that represents the total number of calories in this breakfast. What is the mean and SD of Z?

$$Z_{\text{total calories}} = \text{Cereal} + \tfrac{1}{2}\,\text{milk}$$

$$\mu_Z = \mu_{\text{cereal}} + \tfrac{1}{2}\mu_{\text{milk}}$$

$$= 110 + \frac{140}{2} = 180$$

$$\sigma^2_Z = \sigma^2_{\text{cereal}} + \frac{\sigma^2_{\text{milk}}}{(2)^2}$$

- Your school has the best men's swimming team in the region. The 400-meter freestyle relay team is undefeated this year. In the 400-meter relay, each swimmer swims 100 meters. The times, in seconds, for the four swimmers this season are approximately Normally distributed with means and standard deviations as shown

	Mean	SD
John	55.2	2.8

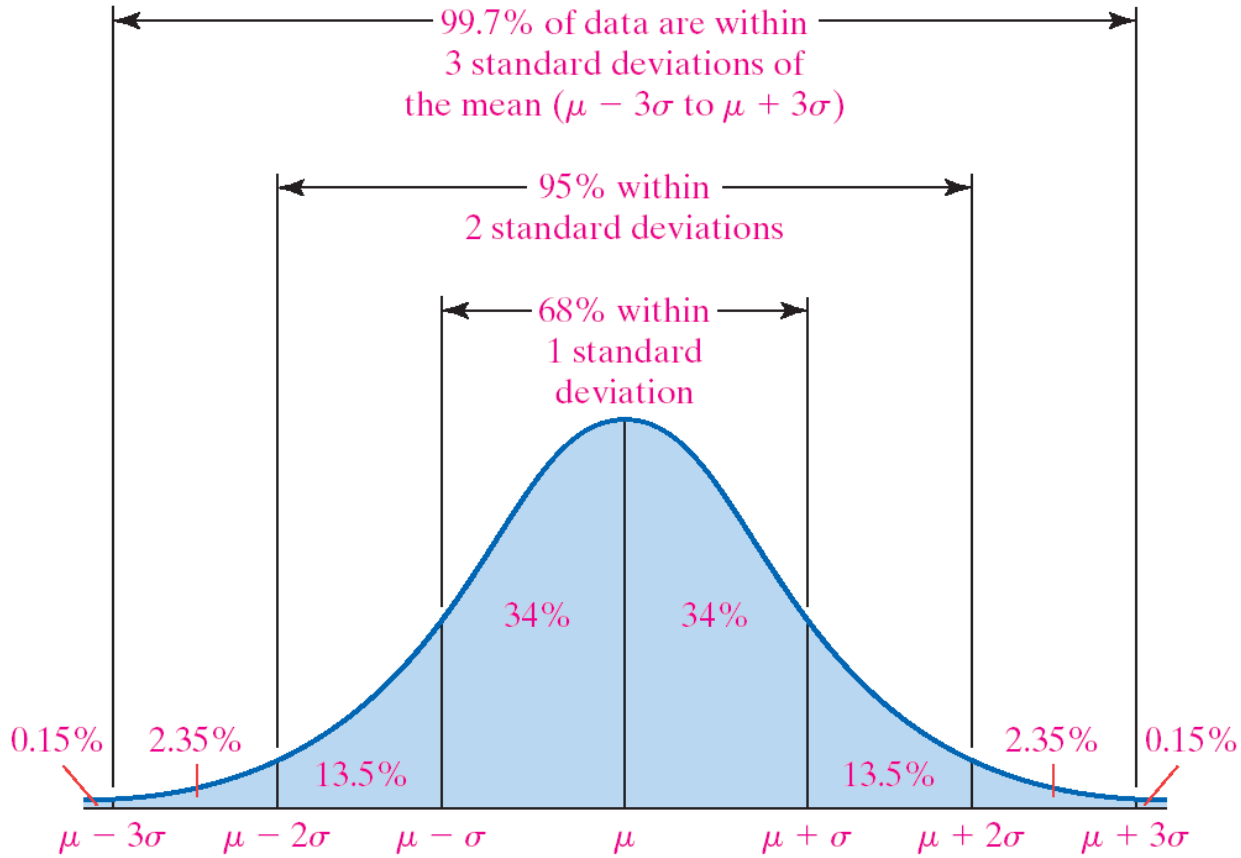
Jerry	58	3
Jim	56.3	2.6
Joe	54.7	2.7

- Find the mean and standard deviation for the total team time in the 400-meter freestyle relay.
  - Mean = 55.2+58+56.3+54.7 = 224.2
  - Variance = 2.8^2+3^2+2.6^2+2.7^2 = 30.89
  - SD = 5.6
- Find the mean and standard deviation for the average time of a single swimmer
  - Mean = 224.2/4 = 56.05
  - Variance = 30.89/4^2 = 1.9306
  - SD = 1.39

4.5 - Normal Random Variables

Tuesday, February 14, 2017 2:12 PM

The Empirical Rule



Z Scores (Standardized Score)

- How many standard deviations away from the mean your value x is

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

$\mu =$ Mean

$\sigma =$ Standard Deviation

Using the Normal Table

Table entry for z is the probability lying below z.

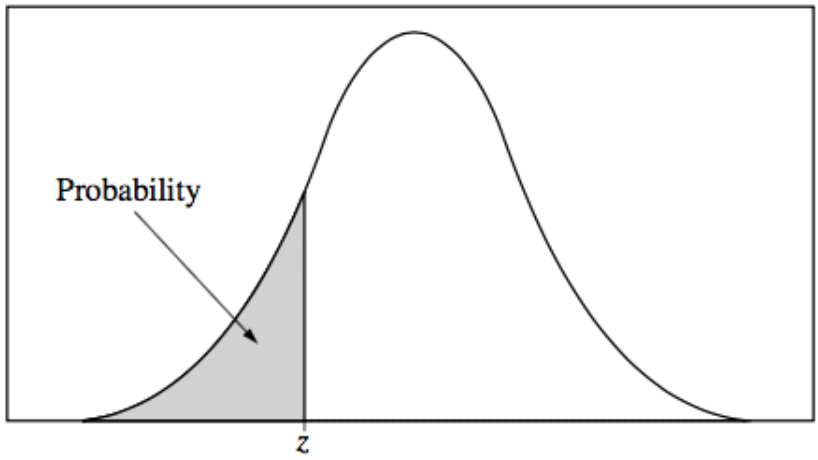


Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table entry for z is the probability lying below z.

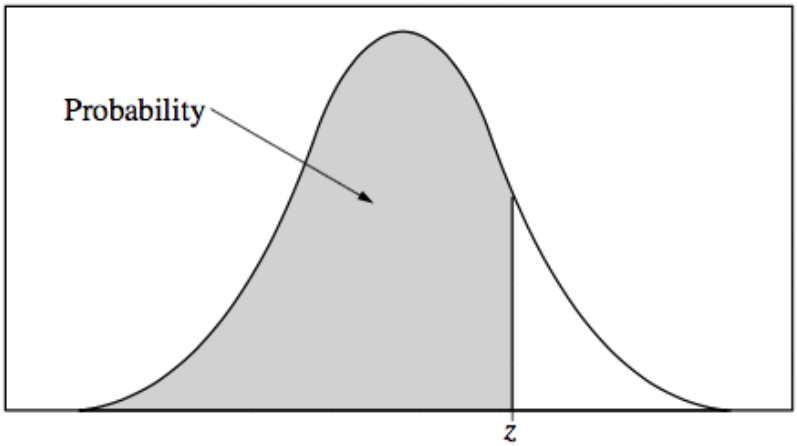


Table A ***(Continued)***

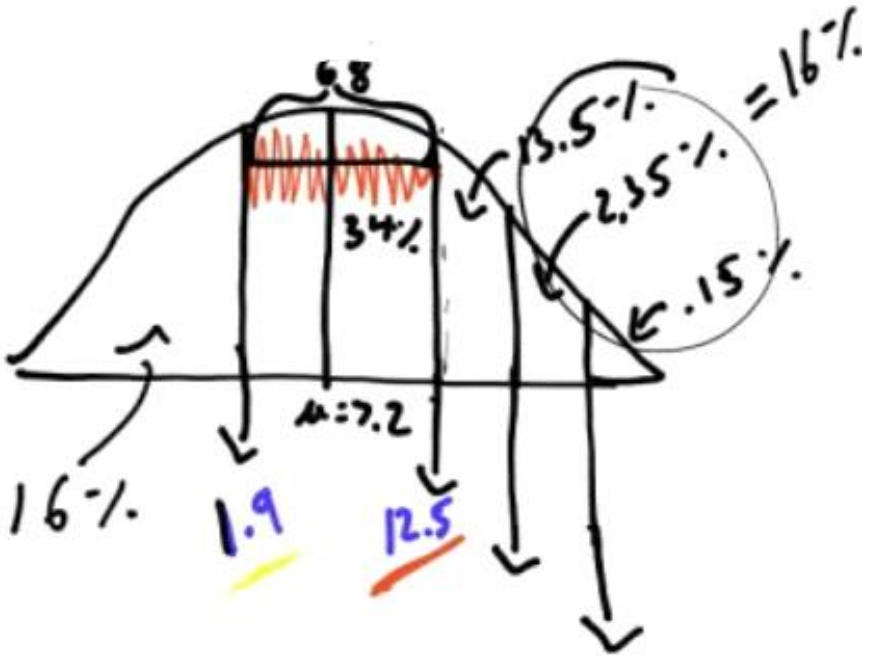
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Practice Questions

1. A study of college freshmen's study habits found that the time (in hours) that college freshmen use to study each week follows a normal distribution with a mean of 7.2 hours and a standard deviation of 5.3 hours

- How many hours do the students who study in the top 15% spend studying?
- The middle 68%?

- Top 15%: 12.5 hours
- Middle 68%: 1.9 hours to 12.5 hours



2. Suppose that the weight of navel oranges is normally distributed with mean of 8 ounces, and standard deviation of 1.3 ounces. And the weights of Valencia oranges is normally distributed with mean of 9 ounces, and standard deviation of 1.6 ounces

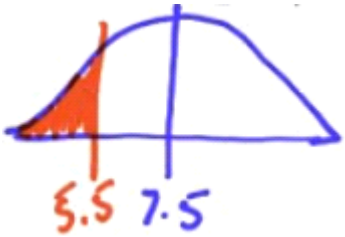
- You grow a navel orange that weighs 9.5 ounces and a Valencia orange that weight 10.5 ounces, which should you enter in the giant fruit contest?

$$\text{Navel} \sim N(8, 1.3)$$
$$\text{Valencia} \sim N(9, 1.6)$$

  - Z score for navel orange = (9.5-8)/1.3 = 1.1538
  - Z score for Valencia orange = (10.5-9)/1.6 = 0.9375

- The weights of newborn children in the United States vary according to the Normal Distribution with mean 7.5 pounds and standard deviation 1.25 pounds.
- What is the probability that a baby chosen at random weighs less than 5.5 pounds at birth?

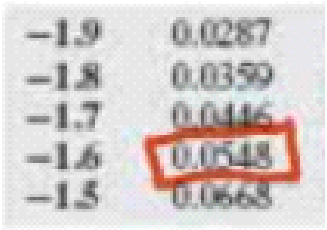
a. Draw a sketch



i. Calculate Z score

$$Z = \frac{X-\mu}{\sigma} = \frac{5.5-7.5}{1.25} = \frac{-2}{1.25} = -1.6$$

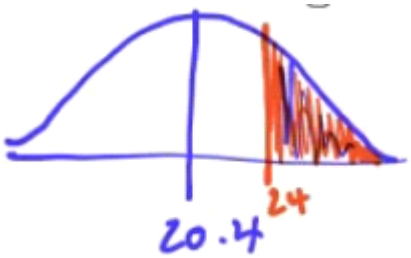
b. Look up probability on the normal table



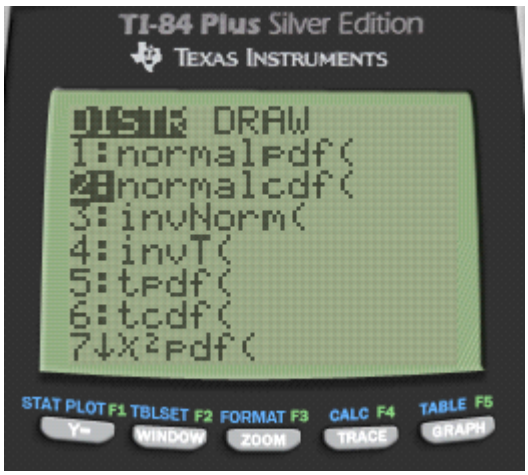
3. The composite score of students on the ACT college entrance examination in a recent year had a Normal distribution with mean of 20.4 and standard deviation of 5.8

- What is the probability that a randomly chosen students scored 24 or higher on the ACT?

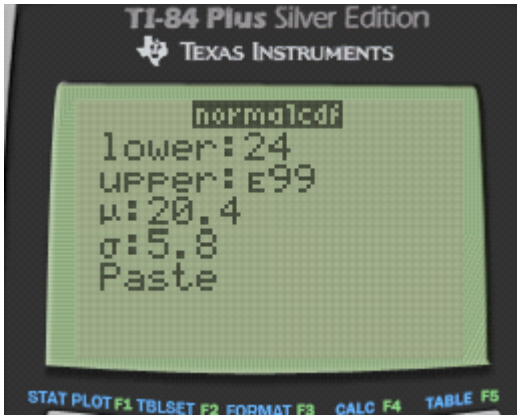
a. Sketch



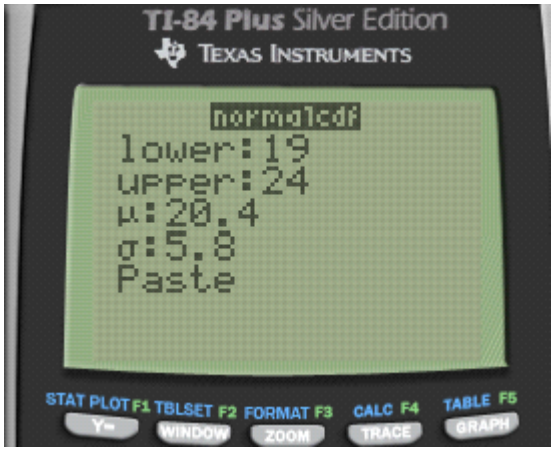
b. 2ND + VARS (DISTR) --> 2: normalcdf



c. Normalcdf(lower, upper, mean, standard deviation)

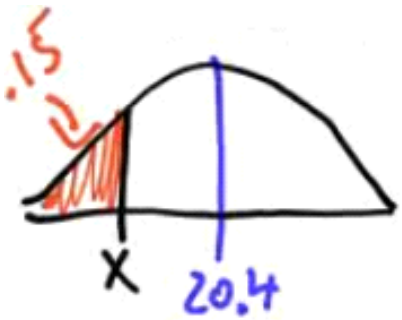


- What is the probability that a randomly chosen student scored between a 19 and a 24 on the ACT?



- What score would someone in the 15th percentile have scored?

a. Sketch



b. Find the z value on the normal table

Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.(
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.(
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.(
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.(
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.(
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.(
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.(
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.(
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.(
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.(
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.(
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.(
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.(
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.(
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.(
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.(
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.(
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.(
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.(
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.(
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.(
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.(
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1
[illegible]	.1357	.1335	.1314	[illegible]	[illegible]	.1251	.1
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1
[illegible]	.1841	.1814	.1788	[illegible]	[illegible]	.1711	.1
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2

Z≈-1.035

c. Solve for x

$$\frac{X-\mu}{\sigma} = -1.04 = \frac{X-20.4}{5.8}$$

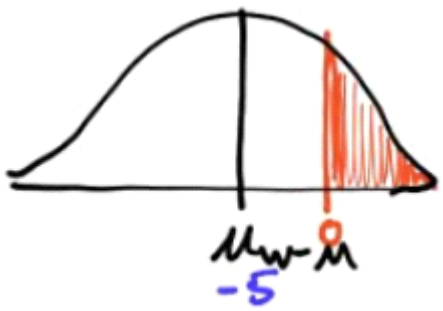
$X = 14.3$

d. Calculator: invNorm(area, mean, standard deviation)

4. Suppose that the mean height of men is 70 inches with a standard deviation of 3 inches. And suppose that the mean height for women is 65 inches with a standard deviation of 2.5 inches

- If the heights of men and women are Normally distributed, find the probability that a randomly selected woman is taller than a randomly selected man.

a. Sketch



b. Find the necessary information

	Mean	SD
Men	70	3
Women	65	2.5
W-M	65-70 = -5	Sqrt(3^2+2.5^2) = 3.9

c. Calculate Z score

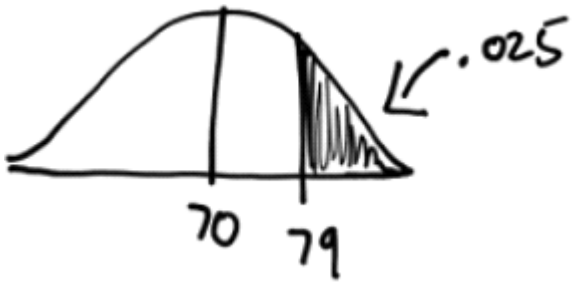
$$z = \frac{x-\mu}{\sigma} = \frac{0-(-5)}{3.9} = 1.28$$

d. Find the probability on the table and subtract that from 1

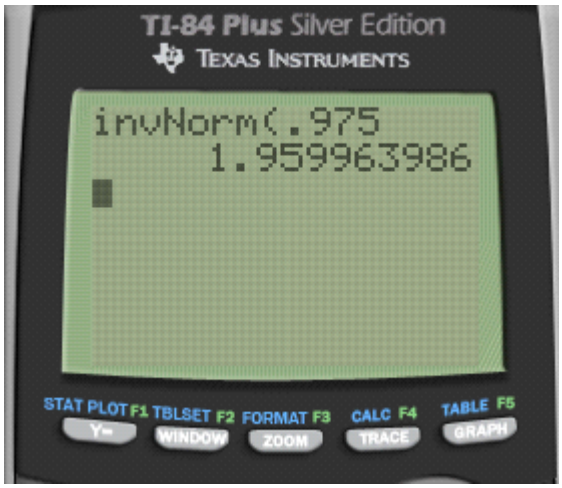
1-0.9015 = 0.0985 = 9.85%

- Suppose that the height (X) in inches, of adult men is a normal random variable with mean of 70 inches. If P (X>79) = 0.025
- What is the standard deviation of this random normal variable?

a. Sketch



b. Find the z score on the calculator: invNorm(area)



c. Solve for SD

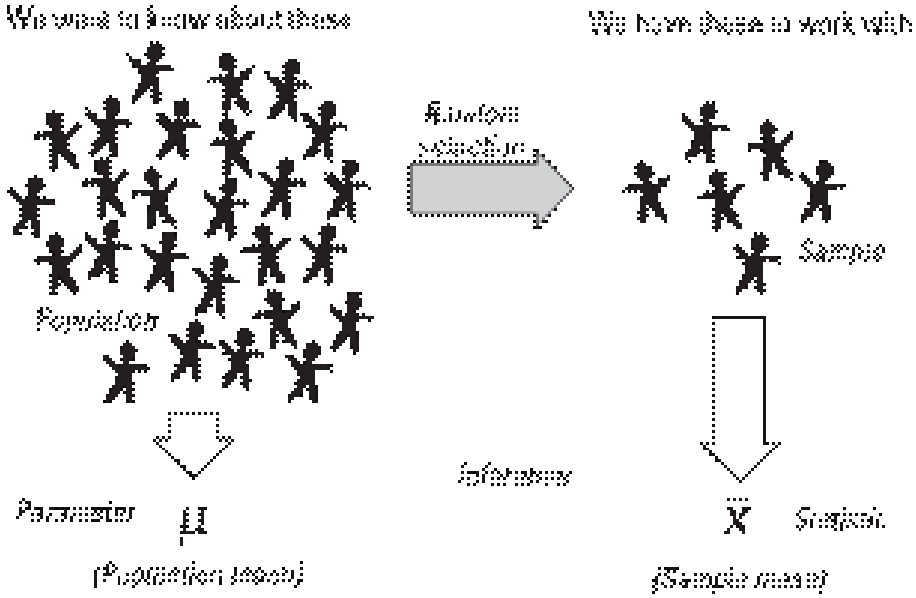
$$1.96 = \frac{79-70}{\sigma}$$

$$\sigma = 4.6$$

5.1 - Sampling Distributions

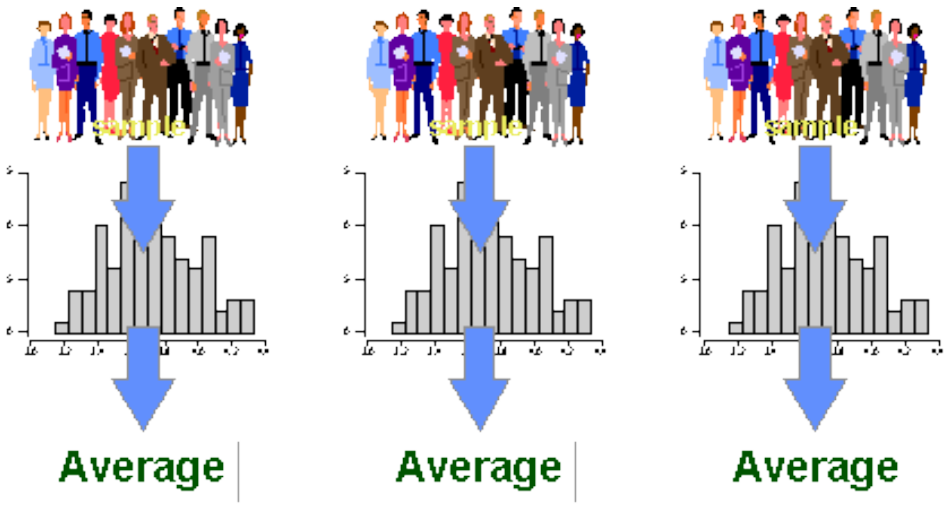
Tuesday, February 14, 2017 2:12 PM

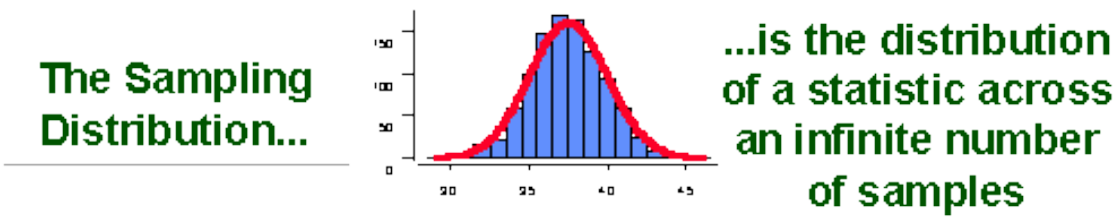
Parameter vs. Statistic



Sampling Distribution

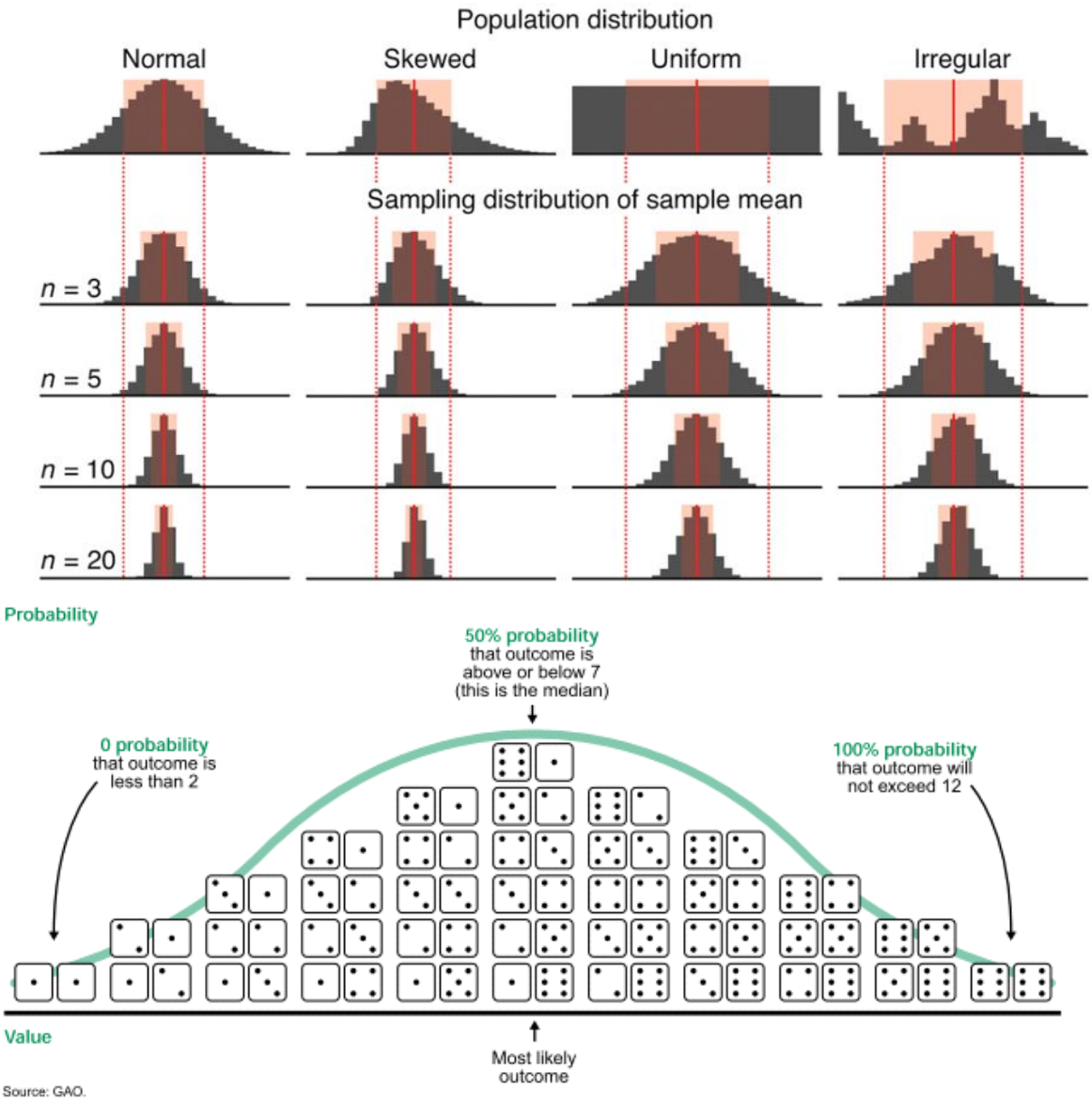
- The "sampling distribution" is the values taken by the statistic in all possible samples of the same size from that population
- The "sampling distribution" is always referring to the distribution of the sample





Central Limit Theorem

- The sampling distribution of the sample mean is normally distributed



Conditions (RIN)

- Random
  - How the sample is selected
- Independent
  - N≥10n
  - N: population size
  - n: sample size
- Normal

- For means
  - $n \geq 30$
  - If the population is normally distributed, n can < 30
- For proportions:
  - $np \geq 10$ AND $n(1\text{-}p) \geq 10$

Sampling Distribution of a Sample Mean

- $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Sampling Distribution of a Sample Proportion

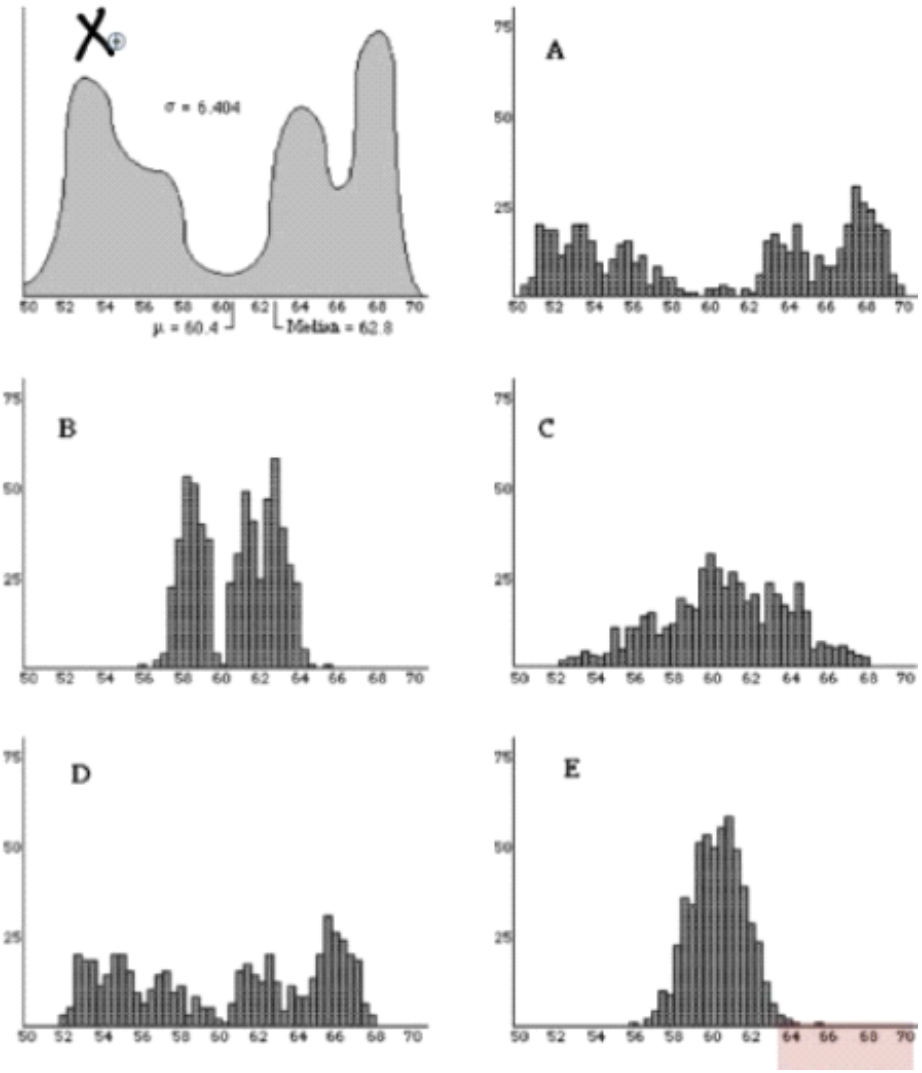
- $\hat{p} \sim N(p, \sqrt{\frac{pq}{n}})$

Review

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal IF np ≥ 10 and n(1 - p) ≥ 10
Quantitative (example: age)	μ = population mean, σ = population standard deviation	$\bar{x}$ = sample mean	μ	$\frac{\sigma}{\sqrt{n}}$	**When will the distribution of sample means be approximately normal ?**

Practice Questions

1. Assume graph X represents the actual distribution select which graph the sampling distribution of the sample mean look like, for a sample size of n = 50?



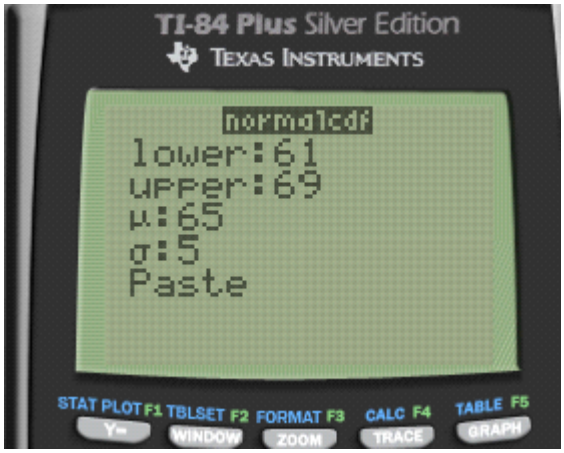
- Answer: E

2. The weight of the eggs produced by a certain species of chicken is Normally distributed with mean 65 g and standard deviation 5 g.

- If a farmer selects a random sample of 10 every morning to check the health of his laying hens, what is the mean and SD of the sampling distribution of the weight of the eggs?

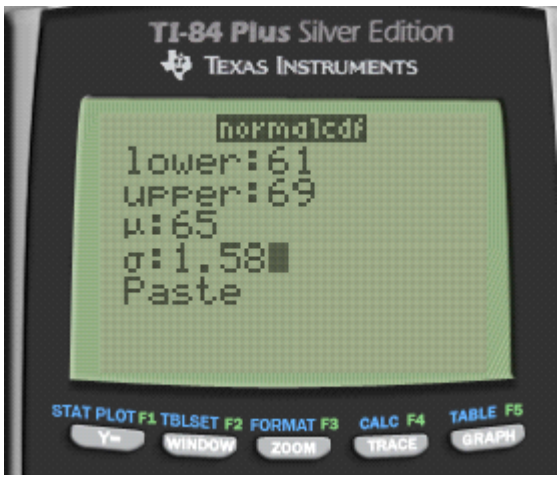
$$\bar{X} \sim N\left(\mu_{\bar{x}} = \mu = 65, S_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}}\right)$$

- Calculate the probability that a randomly selected egg weighs between 61g and 69g



- Calculate the probability that the mean weight of the farmers 10 eggs falls between 61g and 69g.

$$\bar{X} \sim N\left(\mu_{\bar{x}} = 65, \sigma_{\bar{x}} = \frac{5}{\sqrt{10}}\right)$$



3. A survey asks a random sample of 500 adults in California if they support an increase in the state sales tax of 1%. Suppose that 40% of all adults in California support the increase.

- If p hat is the proportion of the sample who are in favor of the increase, what is the mean of the sampling distribution of p hat? The SD?

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

$$\mu_{\hat{p}} = .40$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(.40)(.60)}{500}} = .022$$

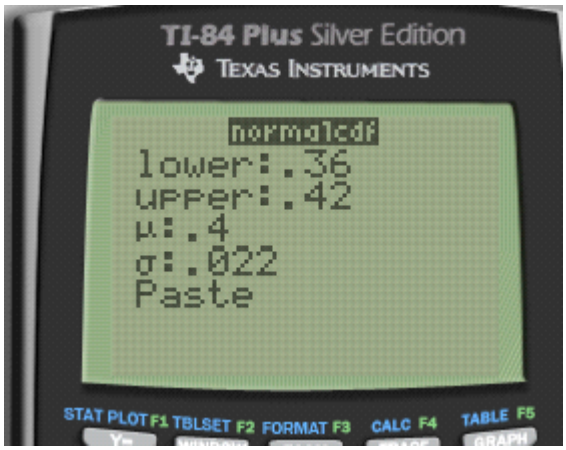
- How large a sample would be needed to guarantee that the standard deviation of it is no more than 0.02?

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

$$.02 \geq \sqrt{\frac{(.4)(.6)}{n}}$$

$$n = 600$$

- Find the probability that p hat is between 0.36 and 0.42

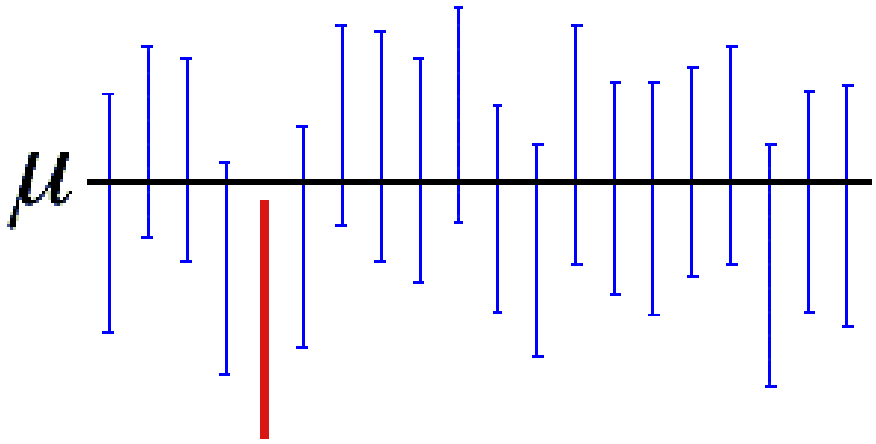


6.1 - Confidence Intervals

Tuesday, February 14, 2017 2:12 PM

What is a CI?

- Using a Statistic to estimate a Parameter
- It is NOT a probability
- It is an interval that will cover the true parameter X% of the time



A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

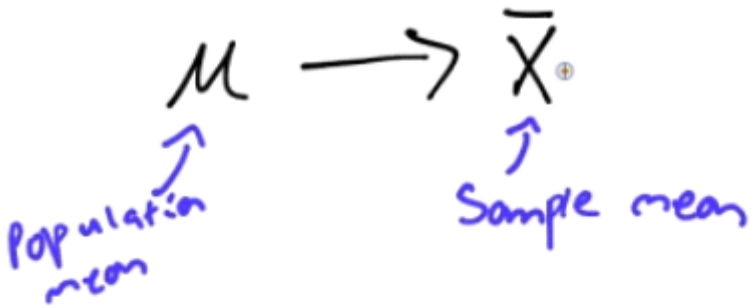
- So we can interpret a CI as
  - "We are X% confident that the true population parameter lies within A and B"

General Math Behind a CI

- Formula
  - Point Estimate ± Margin of Error
  - Point Estimate ± Critical Value * Standard Error

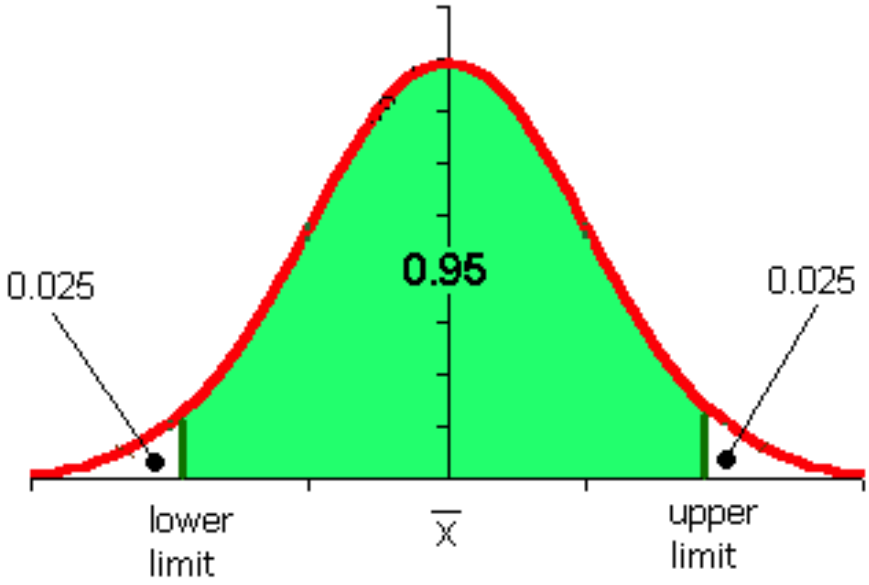
Population Parameter	***Sample Estimate***	***Conditions for Use***	***Formula***
mean μ	$\bar{x}$	Simple Random Sample	$\bar{x} \pm z_c \frac{s}{\sqrt{n}}$
proportion p	$\hat{p}$	Simple Random Sample $n\hat{p} \geq 5, n\left(1-\hat{p}\right) \geq 5$	$\hat{p} \pm z_c \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$
difference of means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	Independent Simple Random Samples	$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
difference of proportions $p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	Independent Simple Random Samples $n_1\hat{p}_1 \geq 5,\ n_1\left(1-\hat{p}_1\right) \geq 5,$ $n_2\hat{p}_2 \geq 5,\ n_2\left(1-\hat{p}_2\right) \geq 5$	$\left(\hat{p}_1 - \hat{p}_2\right) \pm z_c \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$

- Point Estimate



- Critical Value

Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291



- Standard Error

Mean: ← population SD ← Sample SD

$$\frac{\sigma}{\sqrt{n}} \quad \text{OK} \quad \frac{s}{\sqrt{n}}$$

Proportion:

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

- Confidence Interval

CI For	Sample Statistic	Margin of Error
Population mean (μ)	$\bar{x}$	$\pm z^* \frac{\sigma}{\sqrt{n}}$
Population mean (μ)	$\bar{x}$	$\pm t^*_{n-1} \frac{s}{\sqrt{n}}$
Population proportion (p)	$\hat{p}$	$\pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Steps to Calculating a CI

- Read the problem and outline the STASTICS
- Check your CONDITIONS
  - Random
  - Independent: N≥10n
  - Normal: n>30
- CALCULATE
  - Point Estimate ± Critical Value * Standard Error
- INTERPRET
- **S**ome **C**an **C**alculate **I**ntervals

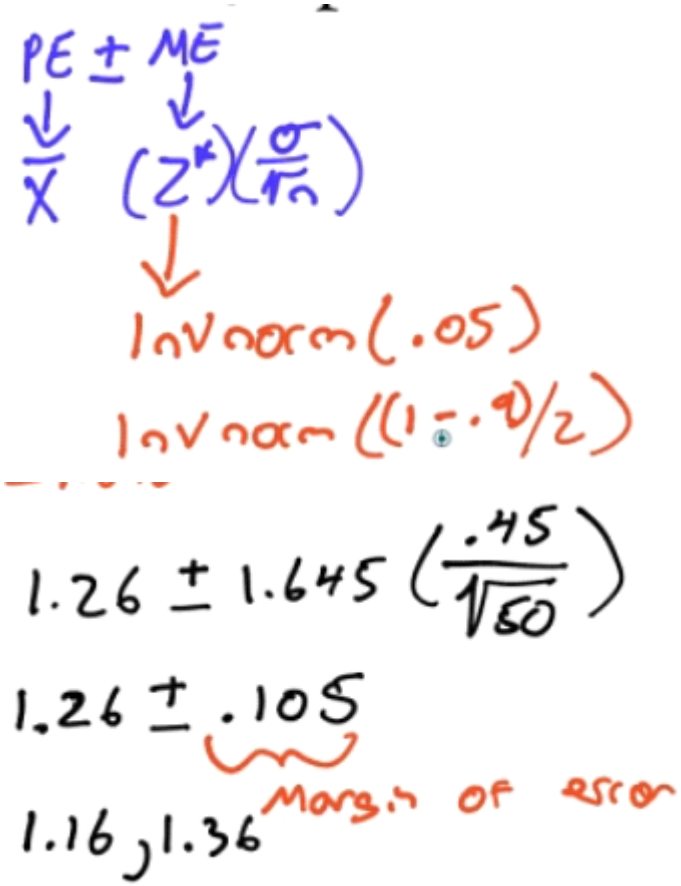
Practice Questions

1. The effect of drugs and alcohol on the nervous system have been the subject of considerable research. Suppose a neurologist is testing the effect of a drug on response time by injecting 50 rats with a dose, subjecting each to a stimulus, and recording the response time. The average response time for the 50 injected rats was 1.26s. Assuming the mean response time for a rat that has not been injected with the drug is 1.4s with standard deviation of 0.45, construct a 90% confidence interval to determine if the drug has an effect on response time.
   - Statistics
     - Mean = 1.26
     - Population SD = 0.45
     - n = 50

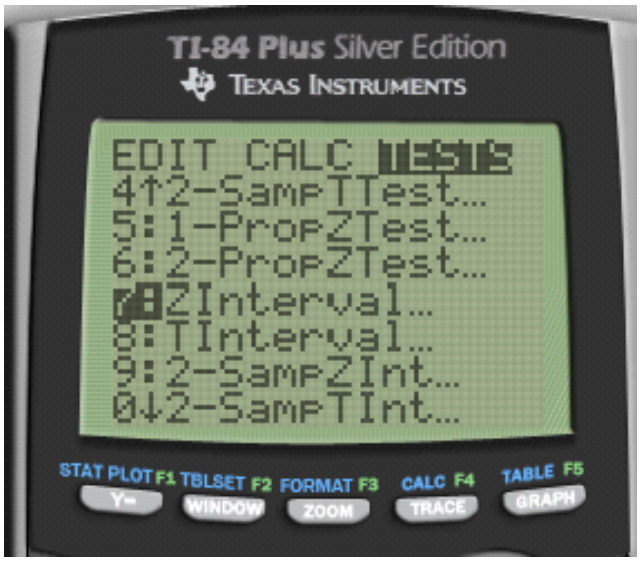
- CL = 90%

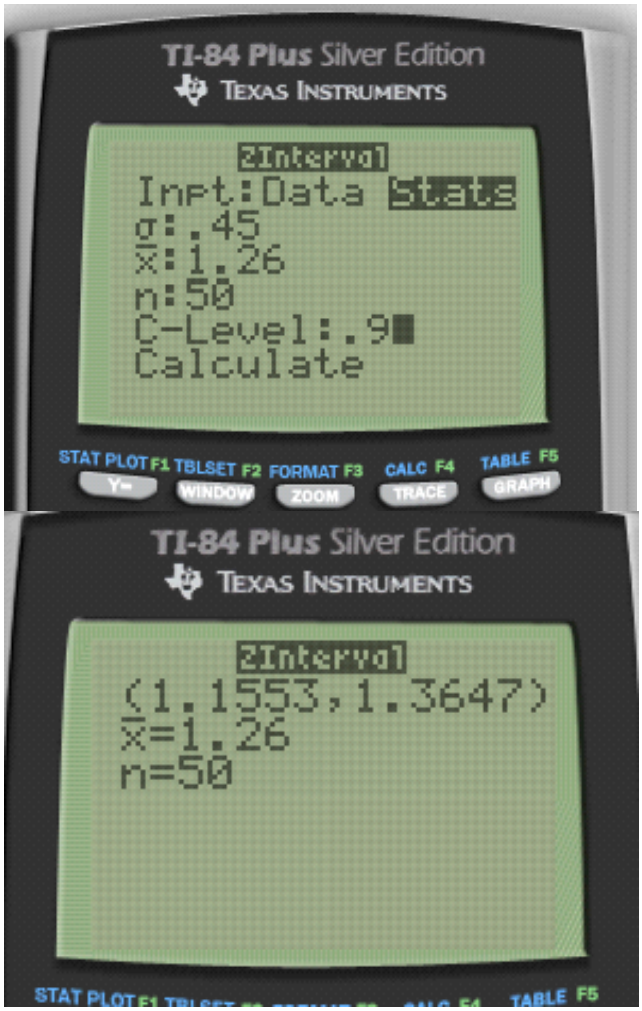
- Conditions
  - Random: Assume rats are a random sample
  - Independent: N > 10n
  - Normal: n > 50

- Calculate



- Calculate by calculator





- Calculated using Z test

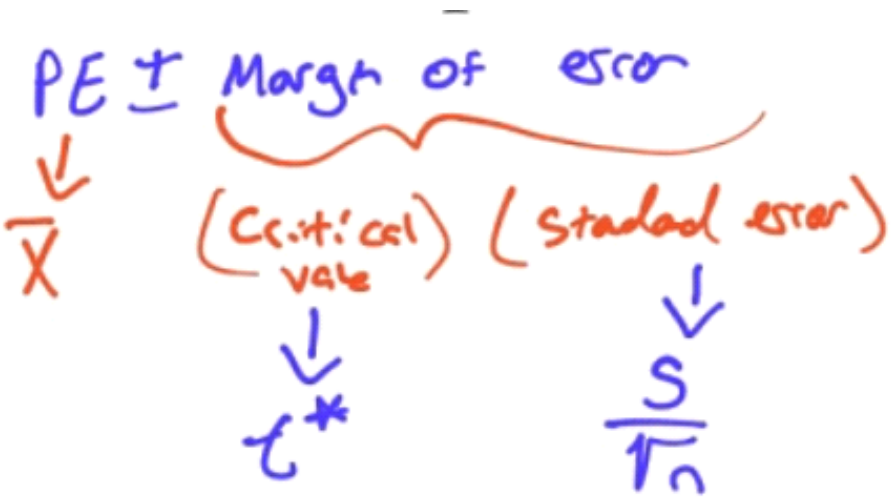
- Interpret
  - We are 90% confident that the true mean response time for rats given the new drug is between 1.16s and 1.36s.
  - 1.4s is not in the interval, so we have evidence the new drug make rats faster

2. There are two fire stations in a town, one in the northern half and one the southern half. The one in the northern part is known to respond to calls within 4 min. The council members in the town are worried that the southern fire station isn't as good so they hire a statistician. The statistician collects a random sample of 50 call/responses from the southern fire station. The mean response time is 5.3 min with a standard deviation of 3.1. Construct a 95% confidence interval to determine if the council members have cause to worry about the southern station
   - Statistics
     - Mean = 3.5
     - Sample SD = 3.1

- n = 50
- CL = 95%

- Conditions
  - Random: Yes
  - Independent: N > 10n = 500
  - Normal: n > 30

- Calculate



invt(area, df)

$df = n - 1$

$df = 49$

invt((1-.95)/2, 49)

$t^* = 2$

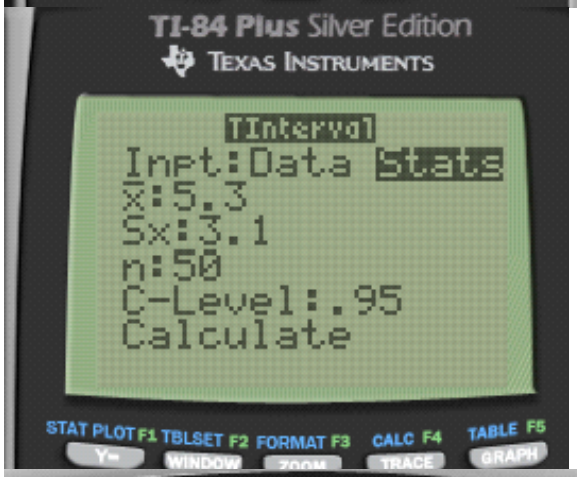
$5.3 \pm (2)\left(\frac{3.1}{\sqrt{50}}\right)$

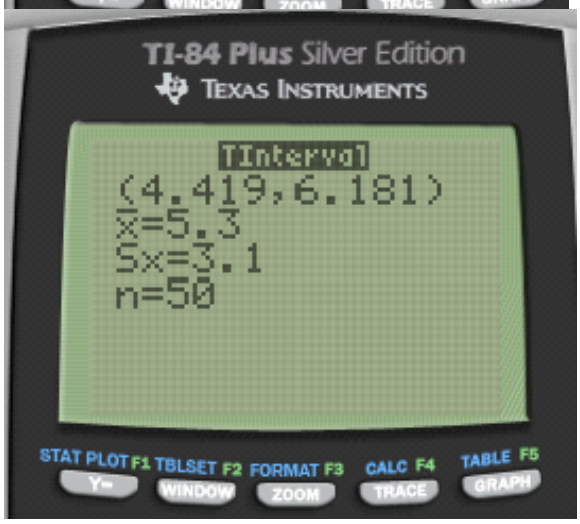
$5.3 \pm (2)(.438)$

$5.3 \pm .877$

4.42, 6.18

- Calculate by calculator





- Interoperate
  - We are 95% confident that the true population mean response time for the southern fire station is between 4.42 and 6.18 mins
  - 4 is not in the interval, so we do have reason to be concerned.

3. The US Department of Transportation reported that 75% of all fatally injured automobile drivers were intoxicated. A random sample of 32 records in Carson County, Colorado, showed that 16 involved a drunk driver. Use a 99% confidence interval to determine whether or not there is evidence that indicates the population proportion of driver

fatalities related to alcohol is different than 75%

- Statistics
  - P hat = x / n = 16/32 = 0.5
  - n = 32
  - CL = 99%
- Conditions
  - Random: Yes
  - Independent: N > 10n = 320
  - Normal: n * p hat > 10 and n * (1-p hat) >10
- Calculate

$$\text{InvNorm}((1-.99)/2)$$

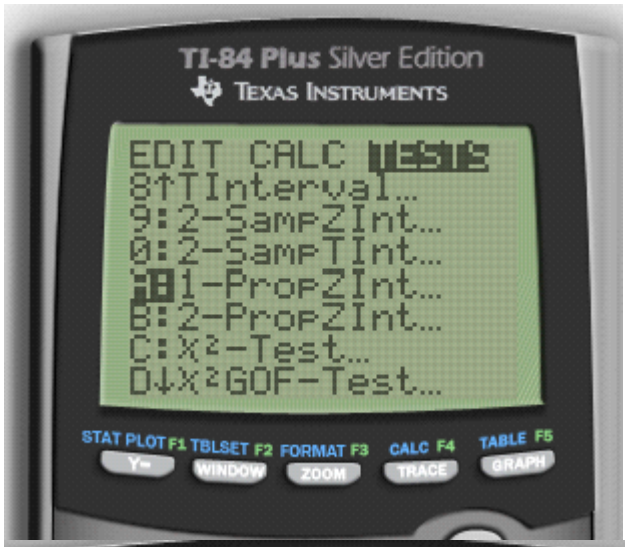
$$z^* = 2.58$$

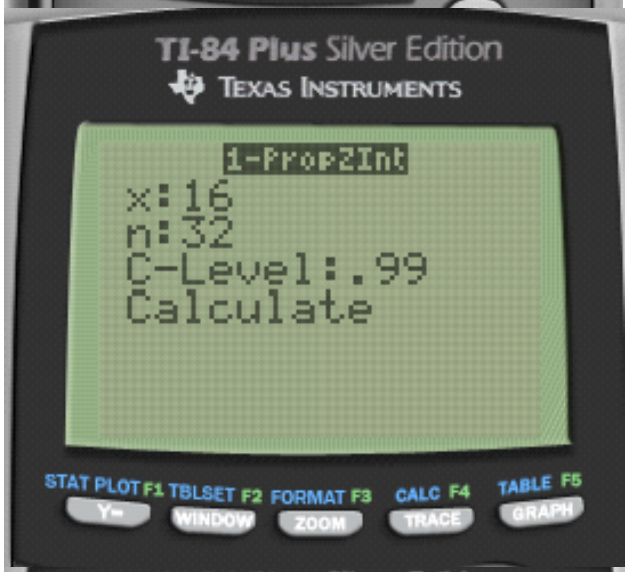
$$PE \pm \text{Margin of error}$$

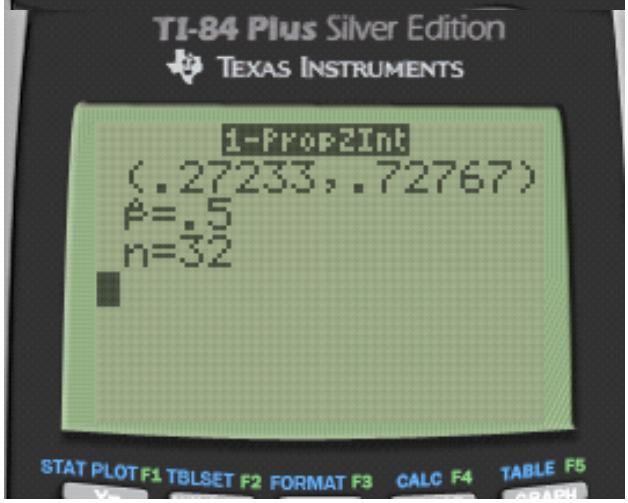
$$\hat{p} \pm (z^*)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.5 \pm (2.58)\sqrt{\frac{(.5)(1-.5)}{32}}$$

- Calculate by calculator







- Interperate
  - We are 99% confident that the true population proportion of driver fatalties in Carson County is between 27.2% and 72.8%
  - 75% is not in our interval, so it appears that Carson County is lower in the US

6.2 - Hypothesis Testing

Wednesday, February 15, 2017 1:42 PM

Hypothesis Test

- Using a Statistic to test a claim about a Parameter
- Steps (**W**hy **C**an't **C**at **P**lay **I**nstruments)
  - **W**rite the hypothesis
    - Null hypothesis (H_0): Parameter = ___
    - Alternative hypothesis (H_1/H_a): Parameter > or < or ≠ _____
  - **C**keck conditions (RIN)
    - Random Sample
    - Independent: N >10n
    - Normal:
      - μ: n≥30
      - p: $\frac{p}{np} > 10$ and $\frac{p}{n(1-p)} > 10$
  - **C**alculate the test statistic

$$\text{Test stat} = \frac{(\text{Point Estimate}) - (\text{Null Hypothesis})}{(\text{Standard Error})}$$

    - Mean

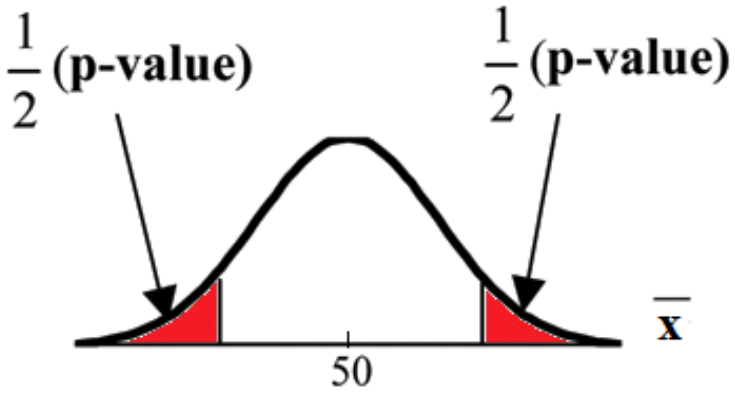
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

    - Proportion

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

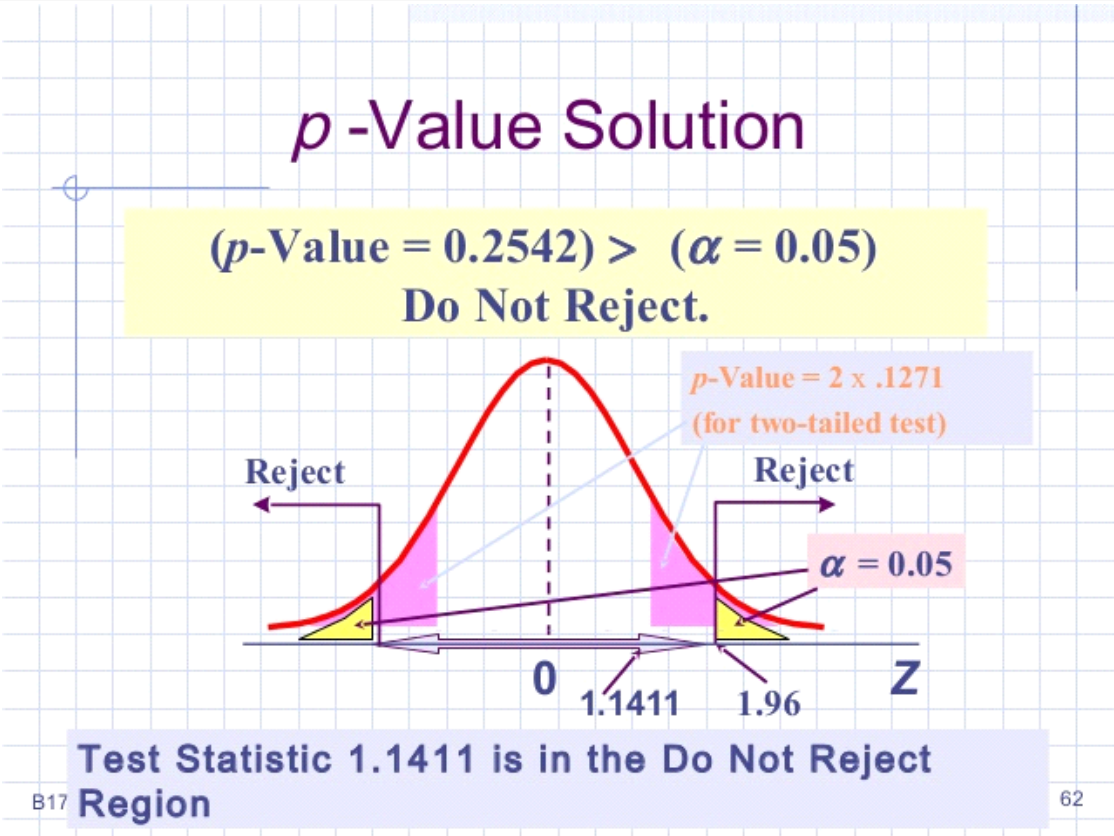
  - Look up the **P**-value (from Z table)

- Probability that the null hypothesis (H_0) is true, given the sample data you collected

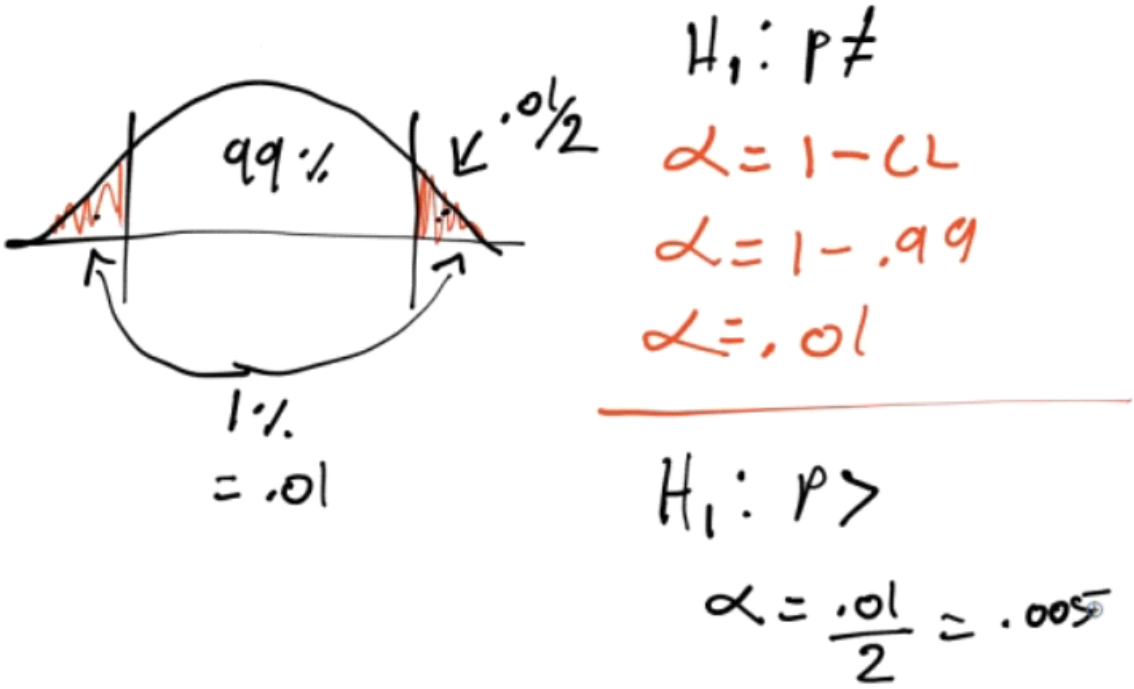


- Interpret

p < α	Reject the null hypothesis	do have evidence to support the claim
p > α	Fail to reject the null hypothesis	do not have evidence to support the claim



Confidence Interval vs α



Practice Question 1

DDT is an insectcide. It is believed that DDT causes birds' eggshells to be thinner and weaker than normal and makes the eggs more prone to breakage.

An experiment was conducted where 21 hawks were fed a food mixture 14 ppm DDT. The first egg laid by each bird was measured. The mean shell thickness was found to be 0.17 mm. The standard deviation of birdshell eggs is known to be 0.14 mm. A "normal" eggshell has a mean thickness of 0.2 mm.

Are the eggs with DDT the same as normal eggs?

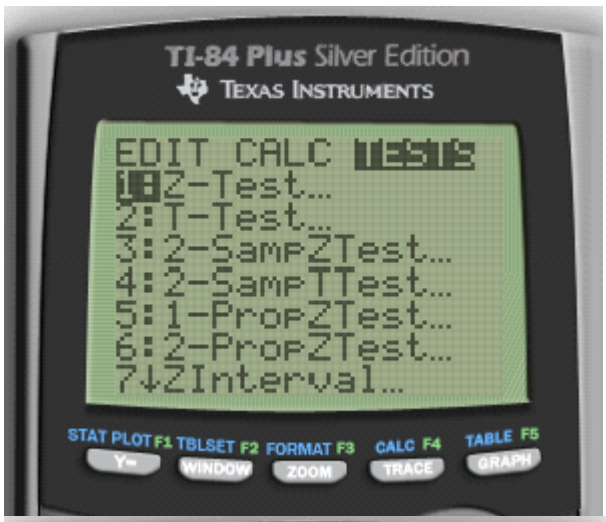
- Data
  - n = 21
  - Mean = 0.17
  - Population SD = 0.14
  - Normal μ= 0.2
- Write the hypothesis
  - Null hypothesis (H_0): μ = 0.2
  - Alternative hypothesis (H_1/H_a): μ ≠ 0.2
- Check conditions
  - Random: Assume random sampling
  - Independent: N > 10n = 210

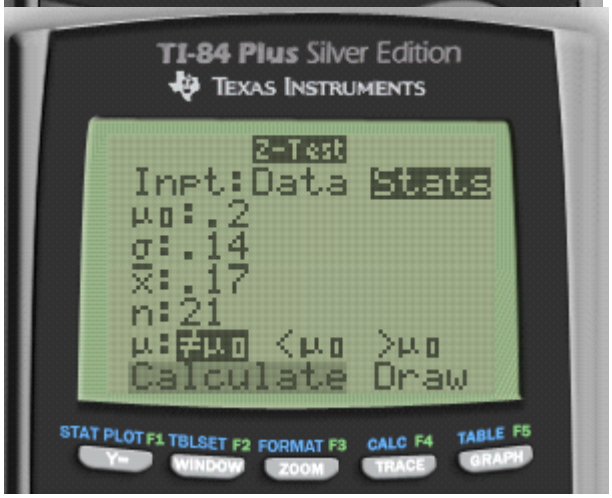
- Normal: sample size is not larger than 30. So, proceed with caution for interpreting results

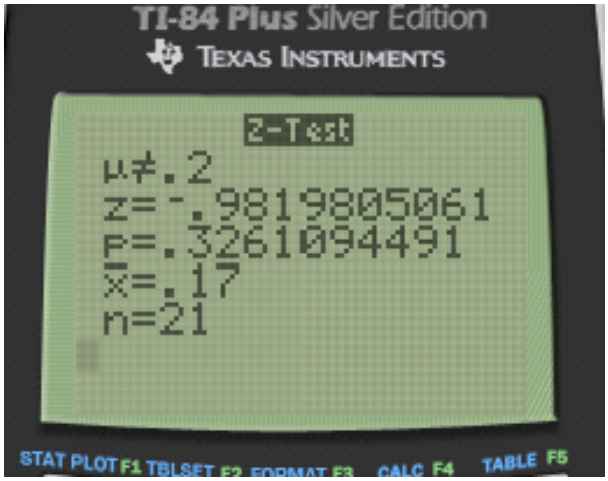
- Calculate the test statistic

$$Z = \frac{\hat{p} - H_0}{SE} \leftarrow \frac{\sigma}{\sqrt{n}}$$

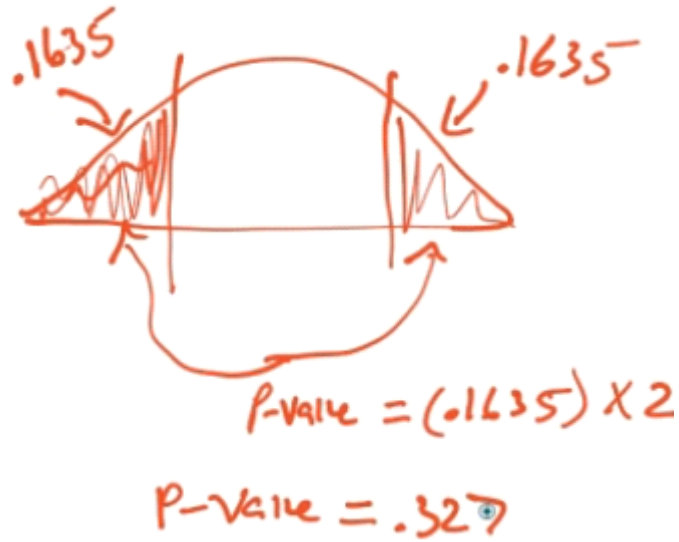
$$Z = \frac{.17 - .2}{.14/\sqrt{21}} = -.98$$



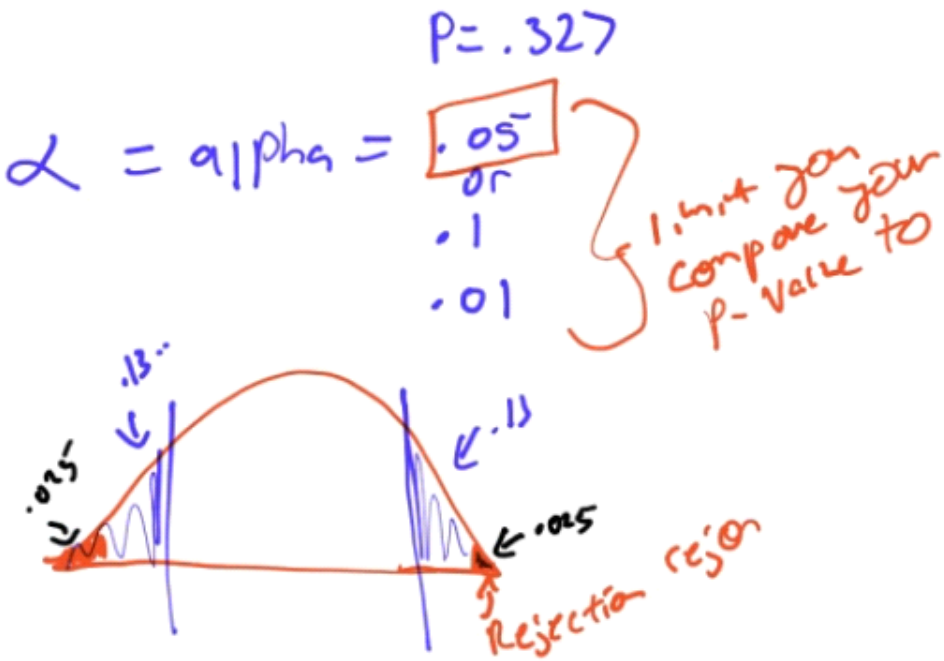




- Look up the P-value



- Interpret



  - Because our p-value = 0.327 is larger than α = 0.05, we fail to reject H0, which

means we do not have evidence to support the alternative/claim that the egg shells are different than normal

Practice Question 2

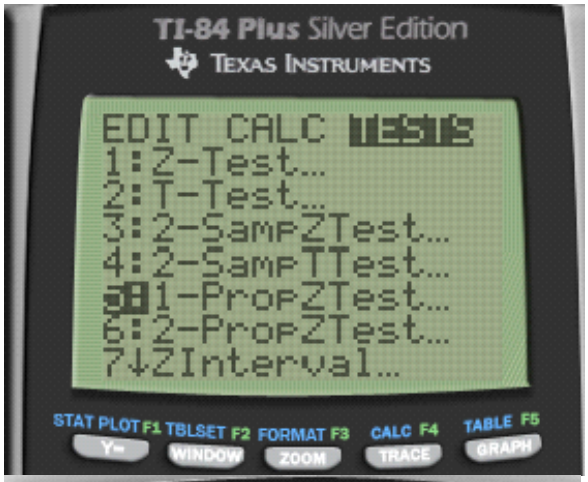
An opinion poll asks a simple random sample of 100 college seniors how they view their job prospects. In all, 56 say "good."

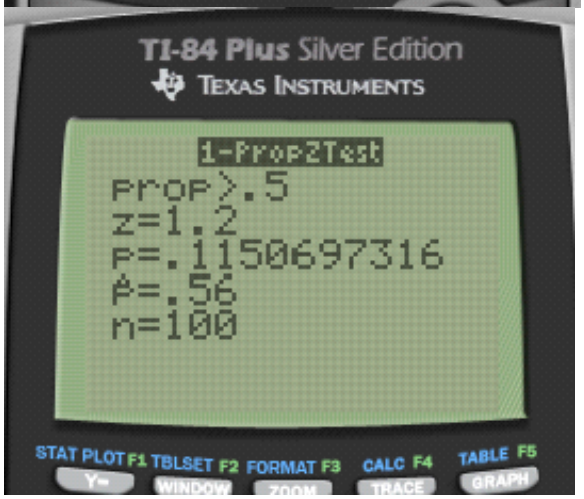
Does the pool give reason to conclude that more than half of all seniors think their job prospects are good.

- Hypothesis
  - H_0: p = 0.5
  - H_1: p > 0.5
- Conditions
  - Random: given, SRS
  - Independent: N > 10n = 1000
  - Normal:
    - N*p0 = 100 * 0.5 = 50 > 10
    - N*(1-p0) = 100 * 0.5 = 50 > 10
- Calculate

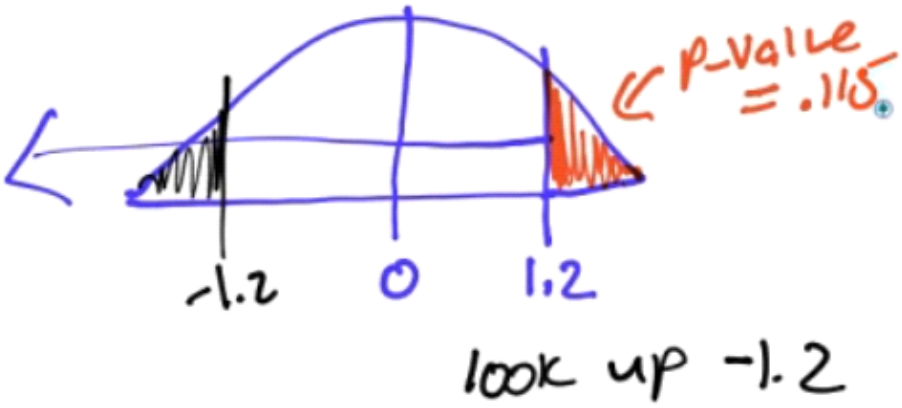
$$Z = \frac{pE - \text{Null hyp } p_0}{SE \leftarrow \sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$= \frac{.56 - .5}{\sqrt{\frac{(.5)(1-.5)}{100}}}$$

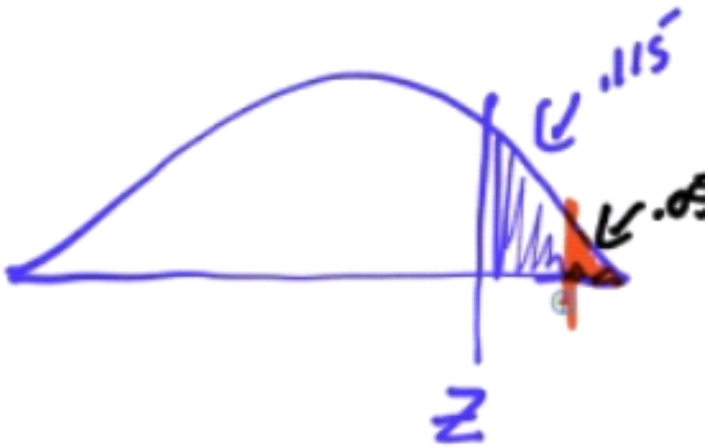




- Look up the P-value



- Interpret
  - Because p-value =0.1151 > $\alpha = 0.05$, we fail to reject H_0.
  - This means that we do not have evidence the claim that the true population proportion of college seniors with "good" prospect is above 0.5



Practice Question 3

When the manufacturing process is working properly, the light bulbs have lifetimes that follow a right-skewed distribution with $\mu = 10$ hours and SD = 1 hours.

A quality control statistician selects a simple random sample of n=100 light bulbs every hour and measures the lifetime of all light bulbs produced that hour is less than 10 hours at the 5% significance level, then all those light bulbs are discarded.

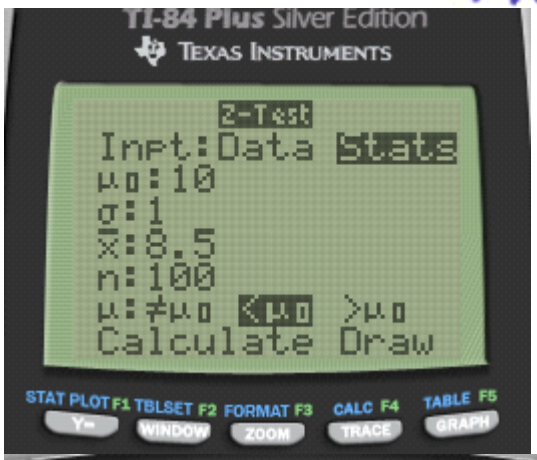
The current sample of 100 has a mean of 8.5, should this batch be discarded?

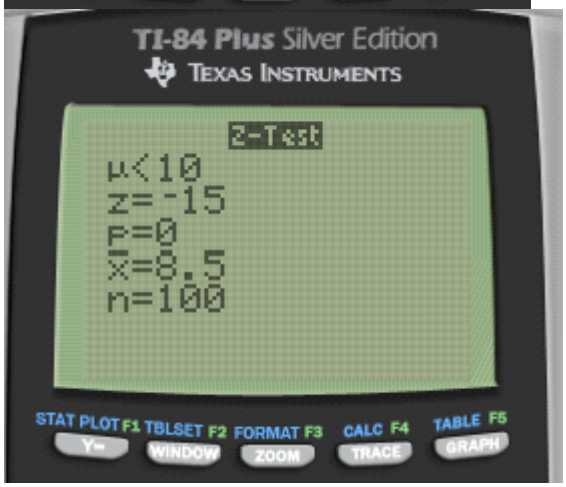
- Write the hypothesis
  - H_0: $\mu = 10$
  - H_1: $\mu < 10$
- Check conditions

- Random: given, random sample
- Independent: N > 10n = 1000
- Normal: n = 100 > 30

- Calculate

$$Z = \frac{PE - H_0}{SE} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{8.5 - 10}{1/\sqrt{100}}$$





- P-value
  - P < 0.0002
- Interpret
  - P < 0.0002 < α = 0.05
  - We reject the null hypothesis test.
  - We do have evidence to support the claim that the mean life of light bulbs is less than 10 hours, so we should discard this batch

Practice Question 4

The US Department of Transportation reported that 75% of all fatally injured

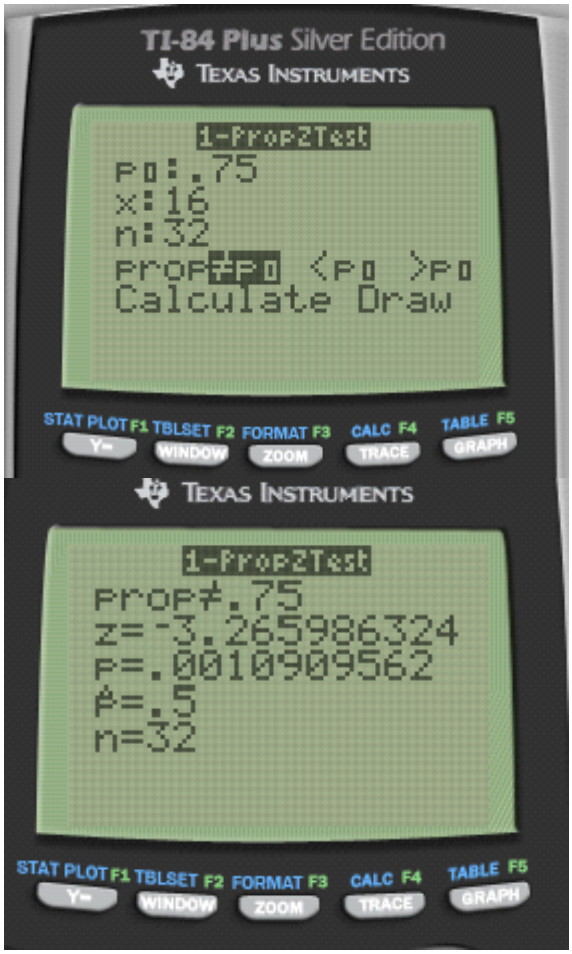
automobile drivers were intoxicated. A random sample of 32 records in Carson County, Colorado, showed that 16 involved a drunk driver. Use a 99% confidence interval to determine whether or not there is evidence that indicates the population proportion of driver fatalities related to alcohol is different than 75%.

- Write the hypothesis
  - H_0: p = 0.75
  - H_1: p ≠ 0.75
- Conditions
  - Random: given
  - Independent: N > 10n = 320
  - Normal:
    - n*p = 32 * 0.75 = 24 > 10
    - n*(1-p) = 32 * 0.25 = 8 < 10
    - May not be normal, proceed with interpret with caution
- Calculate

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

$$= \frac{.5 - .75}{\sqrt{\frac{(.75)(.25)}{32}}}$$

- P-value



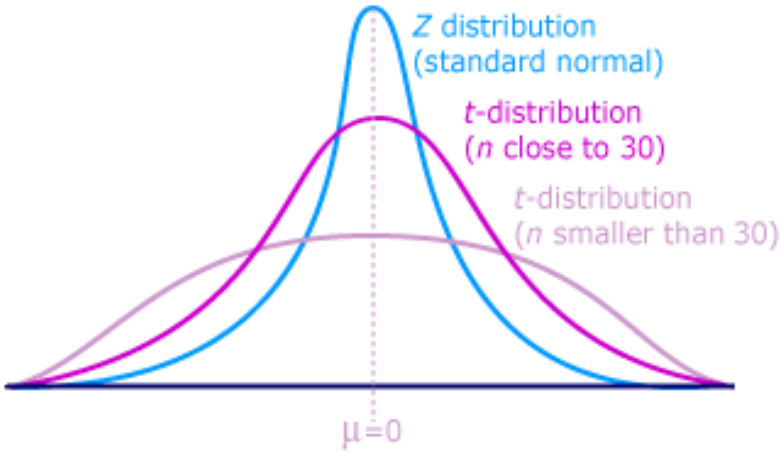
- Interpret
    - Because p = 0.001 < α, we reject H_0 and do have evidence to support the claim that the population proportion of driver fatalities related to alcohol is different than 75%

6.3 - The T Distribution

Wednesday, February 15, 2017 1:43 PM

T Distribution

- When do we use the T distribution
  - Know the mean
  - Do not know the population SD



- Formula

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

- Degrees of freedom (df)
  - df = n-1

Confidence Interval Example

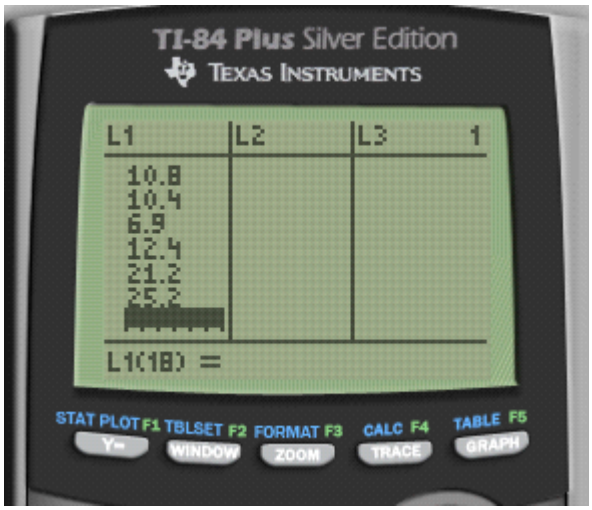
The super Corporation manufactures a device they claim "may increase gas mileage by 23%"

Here are the percent changes in gas mileage for 17 identical vehicles, as presented in one of the company's advertisements

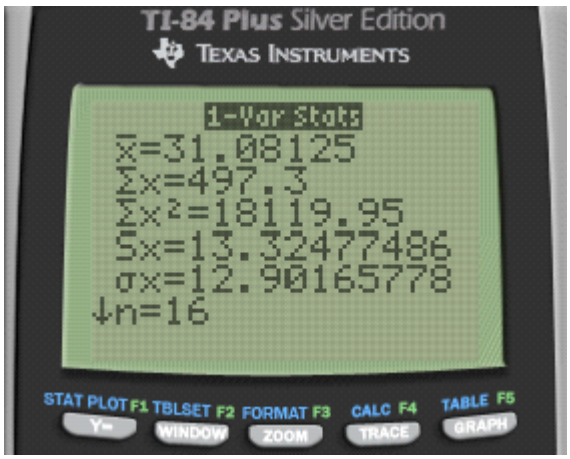
48.3	46.9	46.8	44.6	40.2	38.5
34.6	33.7	28.7	28.7	24.8	10.8
6.9	12.4	21.2	25.2		

Construct and interpret a 90% confidence interval to estimate the mean fuel savings in the population of all such vehicles.

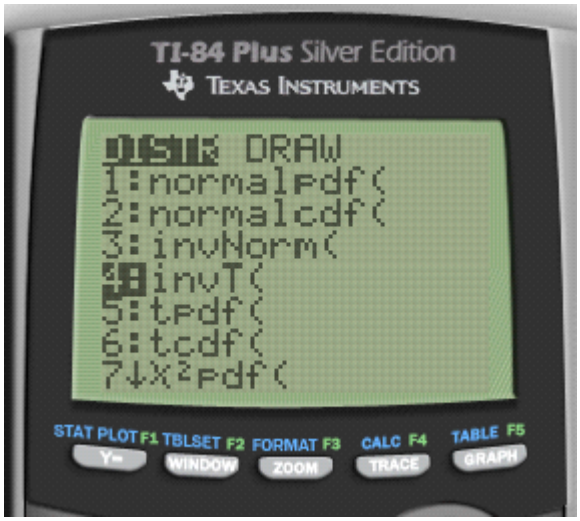
- Type in the data

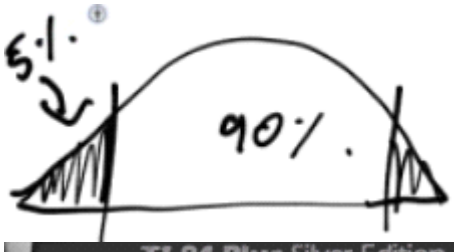


- 1-Var Stats for the sample

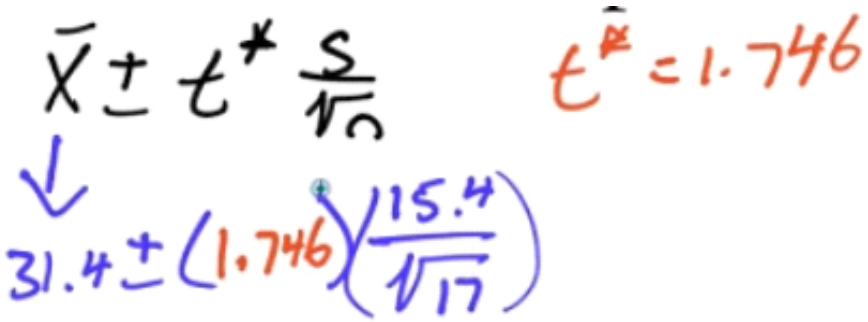


- Looking for t





- Calculate



- Calculate by calculator

- Interpret
    - We are 90% confident that the mean percent change in gas mileage is between 24.68% and 38.16%
    - Since 23% is in the interval, it appears that the machine does even better

than 23% savings

Hypothesis Test Example

When the manufacturing process is working properly, the batteries have lifetimes that follow a right-skewed distribution with μ = 10 hours. A quality control statistician selects a simple random sample of n = 20 batteries every hour and measures the lifetime of each.

If she is convinced that the mean lifetime of all batteries produced that hour is less than 10 hours at the 5% significance level, then all those batteries are discarded.

The current sample of 20 has a mean of 8.5 and a SD of 3, should this batch be discarded?

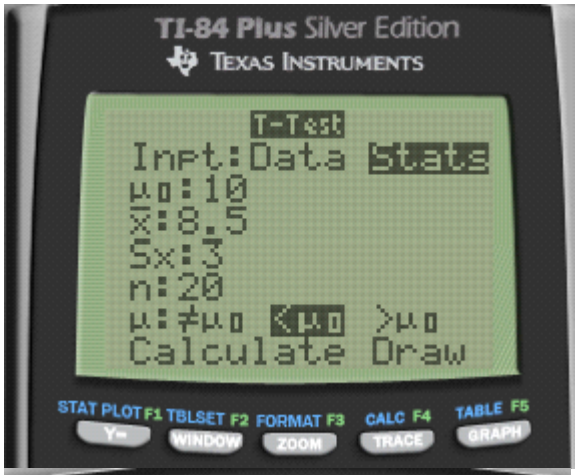
- Hypothesis
  - H0: μ = 10
  - H1: μ < 10
- Conditions
  - Random: given
  - Independent: N > 10n = 200
  - Normal: not larger than 30, but still ok since we are using a T distribution
- Calculate

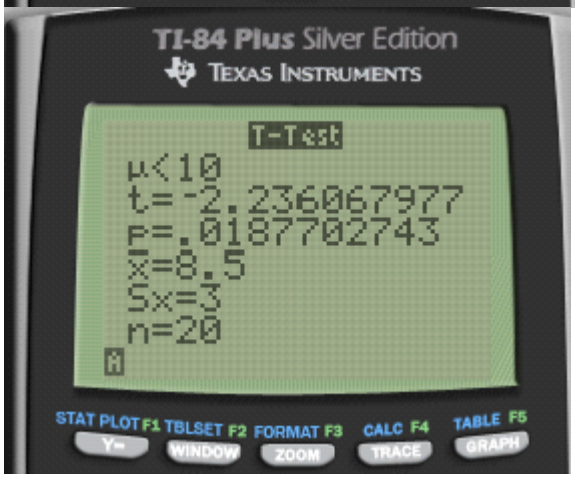
$$\text{Test stat} = \frac{(\text{Point Estimate}) - (\text{Null Hypothesis})}{(\text{Standard Error})}$$

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

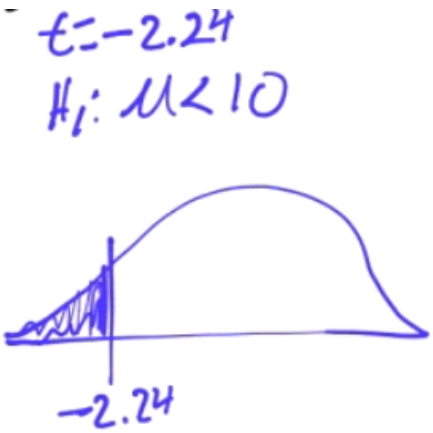
$$= \frac{8.5 - 10}{3/\sqrt{20}} = -2.24$$

- Calculate by calculator





- P-value



  - df = n - 1 = 19
- Interpret
  - Because p < α, we reject the null hypothesis. Thus, we do have evidence to support the claim that the mean battery life in this whole batch is less than 10 hours and so should be discarded

Matched Pairs T-Test

The average weekly loss of study hours due to consuming too much alcohol on the weekend is studied on 10 students before and after a certain alcohol aware ness

program is put into operation. Do the data provide evidence that the program was effective?

Paired T-Test and CI: Before, After

```
Paired T for Before - After

               N      Mean     StDev   SE Mean
Before        10   5.38000   3.20583   1.01377

After         10   4.86000   3.10312   0.98129
Difference    10  0.520000  0.407704  0.128927


95% lower bound for mean difference: 0.283662
T-Test of mean difference = 0 (vs > 0): T-Value = 4.03  P-Value = 0.001
```

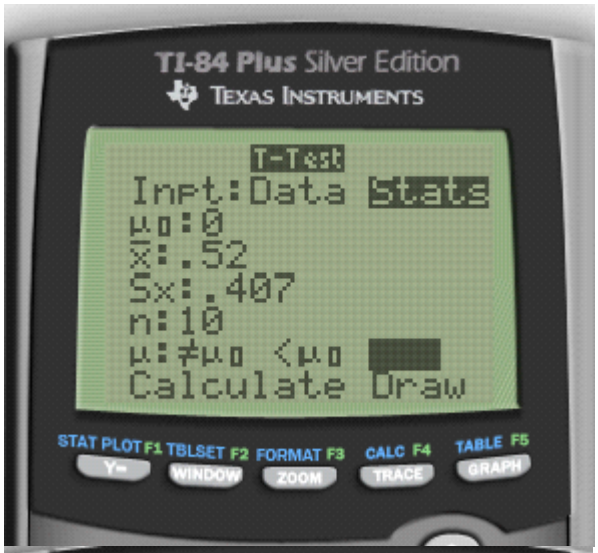
- Formula

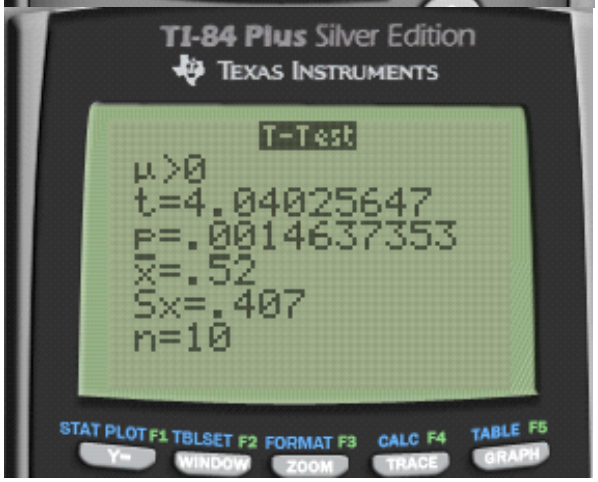
$$t = \frac{\mu_d - 0}{s_d / \sqrt{n}}$$

- Data
  - N=10
  - Difference of mean = 0.52
  - Difference of SD = 0.407
- Hypothesis (d = before - after)
  - H_0: $\mu_d = 0$
  - H_1: $\mu_d > 0$
- Conditions
  - Random: assume random selection
  - Independence: N > 10n = 100
  - Normal: t-distribution
- Calculate

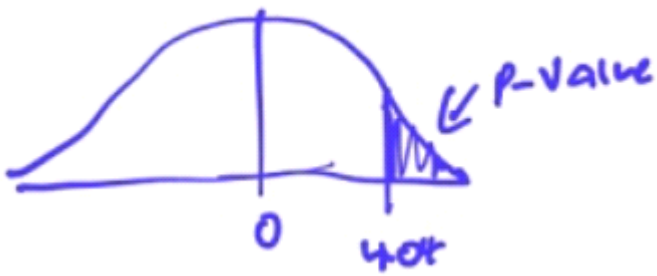
$$t = \frac{PE - H_0}{SE} = \frac{\bar{X}_d - \mu_{od}}{\frac{S_0}{\sqrt{n}}} = \frac{.52 - 0}{.407/\sqrt{10}}$$

- Calculate by calculator





- P-value
  - df = 10-1 = 9



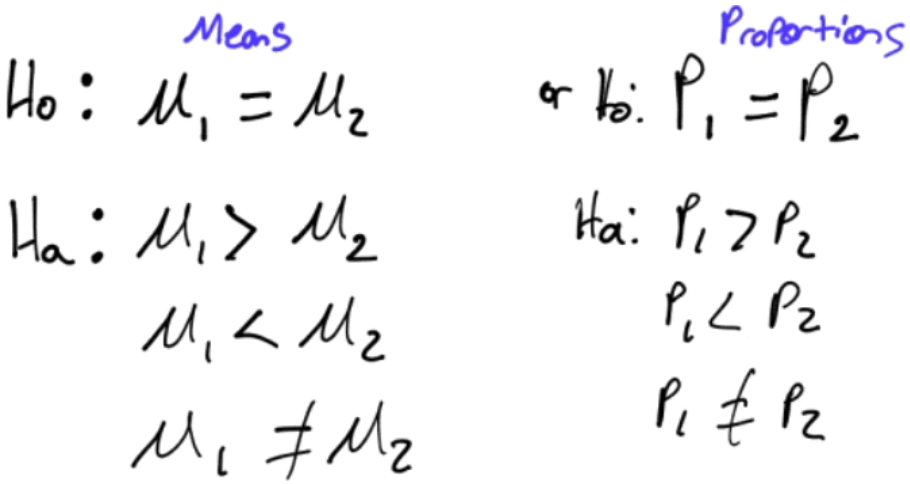
- Interpret
  - P < α, Reject, do have evidence

6.4 - Two Samples

Wednesday, February 15, 2017 1:43 PM

Two Samples

- Hypothesis



Practice Question 1

In a test of the reliability of products produced by two machines, machine A produced 15 defective parts in a run of 280, while machine B produced 10 defective parts in a run of 200.

Do results imply a difference in the reliability of these two machines?

- Data
  - P_1 = 15/280 = 0.0536
  - P_2 = 10/200 = 0.05
  - 1: Machine A
  - 2: Machine B
- Hypothesis
  - H_0: $P_1 = P_2$
  - H_1: $P_1 \neq P_2$
- Conditions
  - Random: assumed
  - Independent: Each part produced is independent

- Normal: n1*p1 > 10, n1*(1-p1) > 10, n2*p2 > 10, n2*(1-p2) > 10

- Calculate

$$\text{Test stat} = \frac{(\text{Point Estimate}) - (\text{Null Hypothesis})}{(\text{Standard Error})}$$

$H_0: p_1 = p_2$

$(\hat{p}_1 - \hat{p}_2)$

$\emptyset$

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

↖ standard deviation of the statistic

- Calculate by calculator

- Interpret
  - P > α, so we fail to reject the null hypothesis. We do not have evidence to support the claim that two machines have different reliabilities

Practice Question 2

The use of helmets among recreational alpine skiers and snowboarders are generally low. A study wanted to examine if helmet use reduces the risk of head injury.

In the study, they compared the ski/board related injury costs for those who wore helmets and those who did not wear helmets. The helmet wearers had a mean injury cost of $10,200 per person (SD = $25,000) and the non-helmet wearers had a mean injury cost of $45,500 (SD = $10,000).

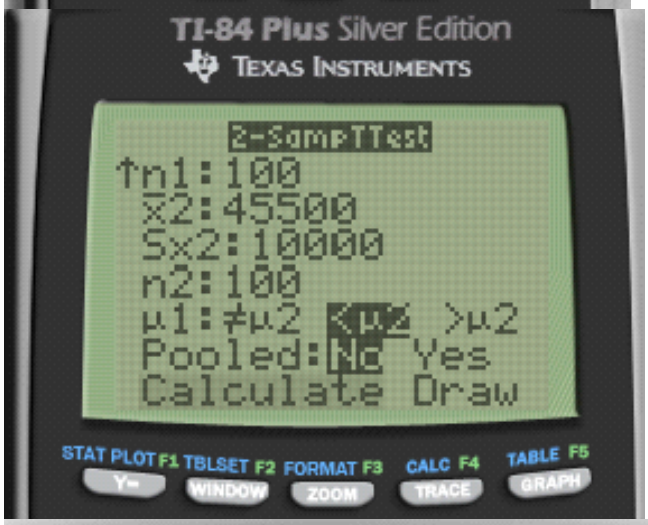
Are ski/board injuries less severe among those who wear helmets?

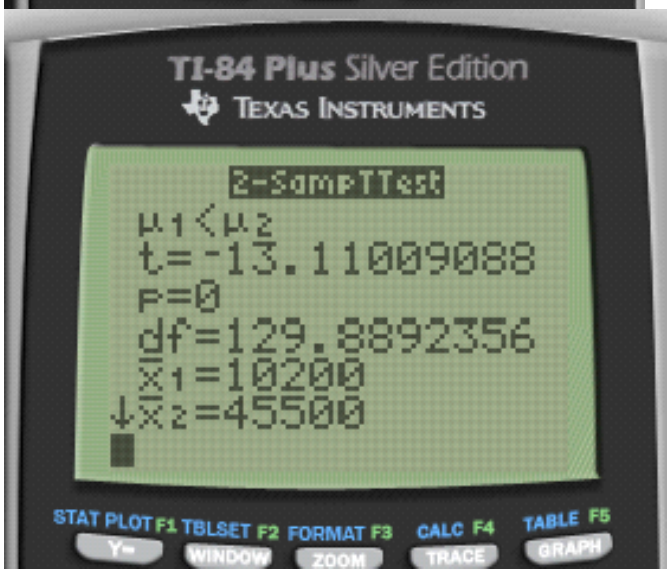
- Hypothesis
  - H_0: $\mu_1 = \mu_2$
  - H_1: $\mu_1 < \mu_2$
- Conditions
  - Random: assume SRS
  - Independent: assume each person independent
  - Normal: $n_1 > 30$, $n_2 > 30$
- Calculate

$$\text{Step 3: Test stat} = \frac{(\text{Point Estimate}) - (\text{Null Hypothesis})}{(\text{Standard Error})}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{10{,}200 - 45{,}50}{\sqrt{\frac{(25{,}000)^2}{100} + \frac{}{}}}$$

  - df = smaller n - 1
- Calculate by calculator





- Interpret
  - P < α
  - So we reject the null hypothesis and have evidence to support the claim that the injury cost of group 1 (helmet wearers) is less than group 2 (non-helmet wearers)

Practice Question 3

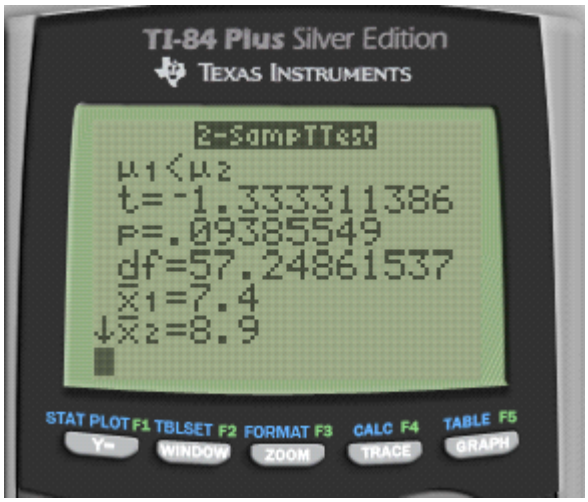
A pharmaceutical company claims its new drug will relieve headaches faster than any other drug on the market. To determine whether this claim is valid, the new drug is given to each of the 30 randomly selected persons and a standard drug is given to another 30 randomly selected persons.

The number of minutes required for each to recover from the headache is recorded. The sample results are:

$$\text{New drug: } \overline{X}_1 = 7.4 \quad S_1 = 4.1 \quad n_1 = 30$$
$$\text{Old drug: } \overline{X}_2 = 8.9 \quad S_2 = 4.6 \quad n_2 = 30$$

- Hypothesis
  - H_0: $\mu_1 = \mu_2$
  - H_1: $\mu_1 < \mu_2$
- Conditions
  - Random: given
  - Independent: assumed
  - Normal: n = 30
- Calculate

$$t = \frac{PE - H_0}{SE} = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



- Interpret

- P > α
- We fail to reject H0, so we do not have evidence to support the claim that the new drug works better

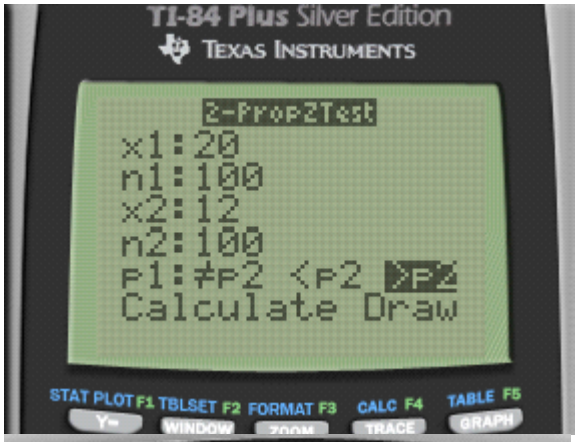
Practice Question 4

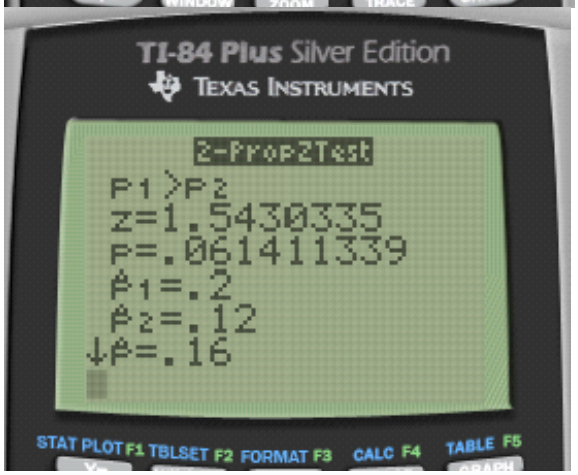
In a large university in the year 2000 an SRS of 100 entering freshmen found that 20 finished in the bottom third of their high school class. Admission standards at the university have become more stringent and in 2015 and a new SRS of 100 entering freshmen found that 12 finished in the bottom third of their high school class.

Does it appear that t he admission standards have actually become more stringent?

- Data
  - P1 = x1 / n1 = 20/100 = 0.2
  - P2 = x2 / n2 = 12/100 = 0.12
- Hypothesis
  - H0: P1 = P2
  - H1: P1 > P2
- Conditions
  - Random: given, SRS
  - Independent: assumed
  - Normal: n1*p1 > 10, n2*p2 > 10, n1*(1-p1) > 10, n2*(1-p2) > 10
- Calculate

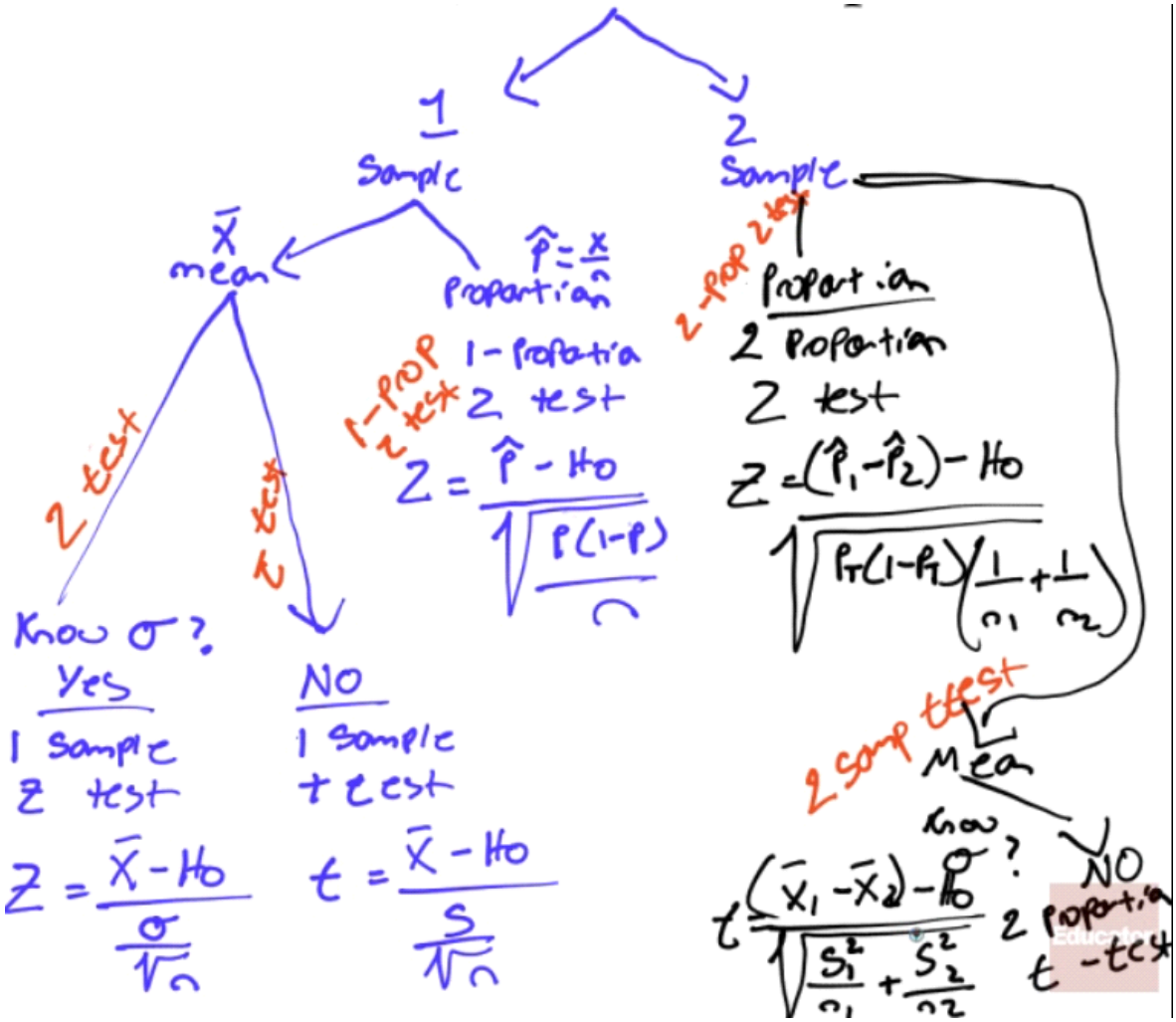
$$Z = \frac{PE - H_0}{SE} = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$





- Interpret
  - P > α
  - So we fail to reject the null hypothesis and do not have evidence to support the claim that acceptance is more stringent

"Pick Your Test" Map


1 Sample
2 Sample
x̄ mean
p̂ = x/n
Proportion
2 test
1-PROP z test
1-Proportion 2 test
Z = (p̂ - Ho) / √(p(1-p)/n)
Know σ?
Yes
1 Sample Z test
Z = (x̄ - Ho) / (σ/√n)
NO
1 Sample t test
t = (x̄ - Ho) / (s/√n)
2-PROP Z test
Proportion
2 Proportion
2 test
Z = ((p̂1 - p̂2) - Ho) / √(p̄(1-p̄)(1/n1 + 1/n2))
2 samp t test
Mean
Know σ?
NO
2 Proportion t-test
t = ((x̄1 - x̄2) - Ho) / √(s1²/n1 + s2²/n2)

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of $n_1 - 1$ and $n_2 - 1$	$n_1, n_2 < 30$; and/or σ_1, σ_2 unknown
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d} - \mu_d)}{s_d/\sqrt{n}}$	t_{n-1}	$n < 30$ pairs of data and/or σ_d unknown
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$ for each group

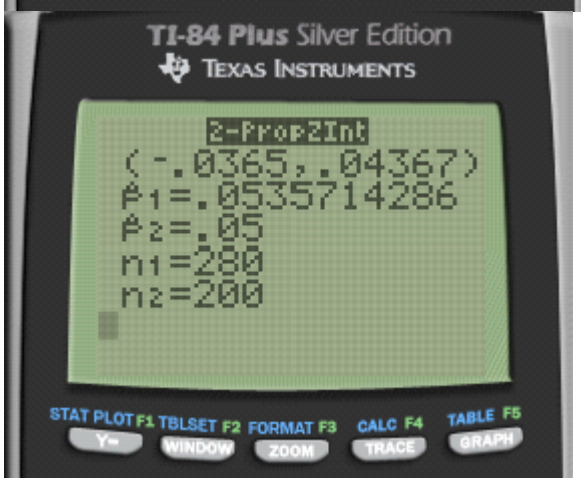
Confidence Interval Practice Question

In a test of the reliability of products produced by two machines, machine A produced 15 defective parts in a run of 280, while machine B produced 10 defective parts in a run of 200.

Do results imply a difference in the reliability of these two machines?

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

$$\left(\frac{15}{280} - \frac{10}{200}\right) \pm 1.96 \sqrt{\frac{\frac{15}{280}\left(1-\frac{15}{280}\right)}{280} + \cdots}$$



- 0 is in the interval so there does not appear to be difference

6.5 - Hypothesis Testing of Least-Squares Regression Line

Wednesday, February 15, 2017 1:43 PM

Tests for the Regression Line

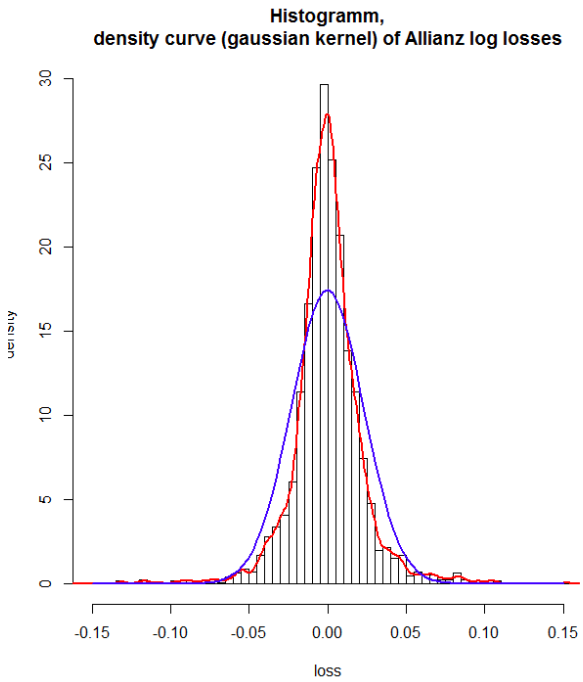
- Is there a correlation?

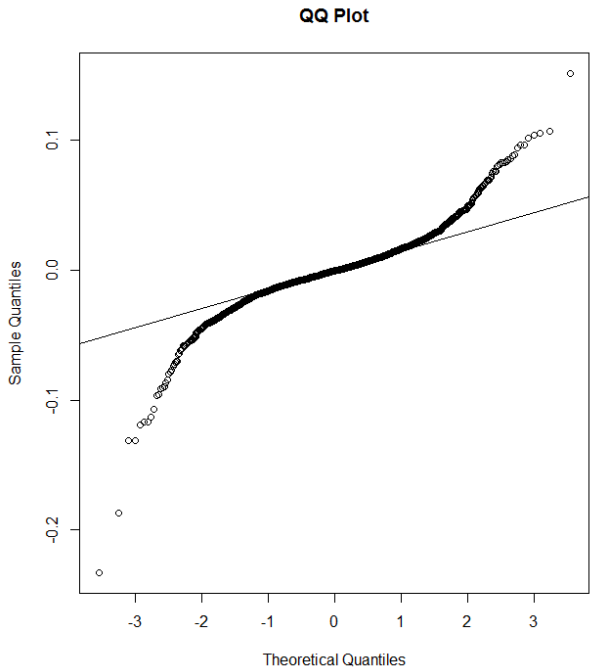
H0	r=0	b=0
H1	r≠0	b≠0

- Is the y intercept = 0
  - H0: a = 0
  - H1: a ≠ 0

Conditions for Hypothesis Testing

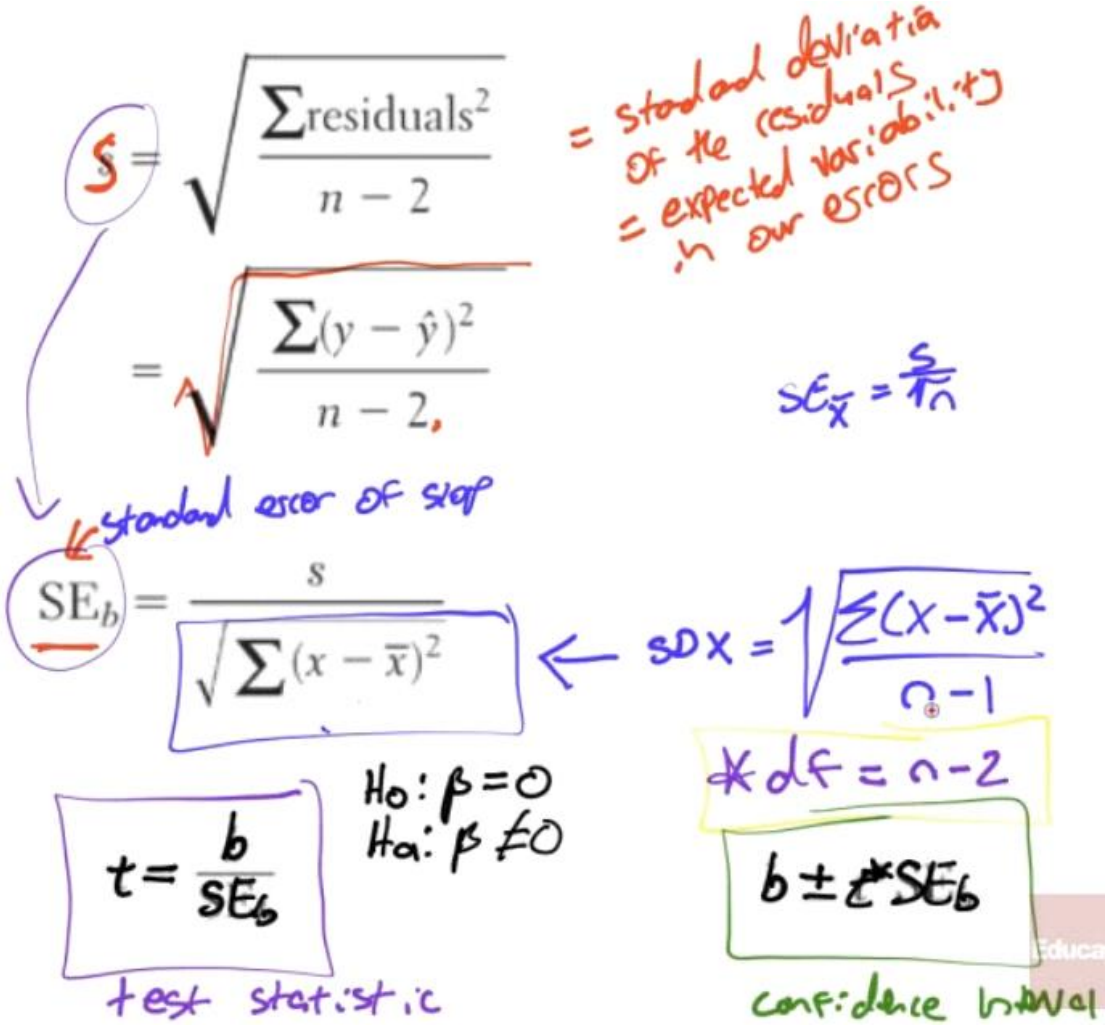
- Linearity
  - Linear relationship between x and y
- Constant Variability (homoscedasticity)
- Normality
  - The residuals should be normally distributed (from Histogram and QQ plot)





- Independence
  - All the Y are independent

Hypothesis Testing



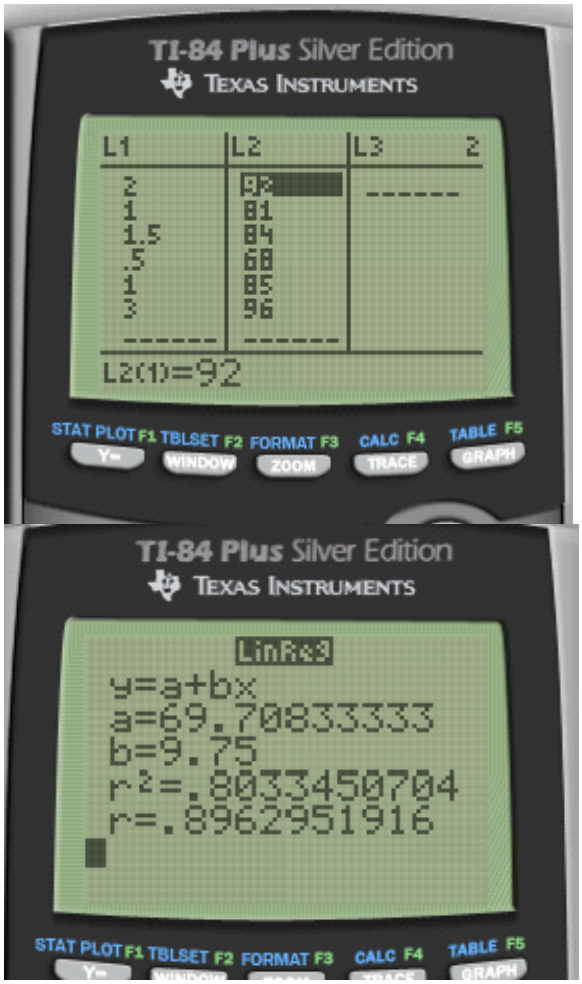
Practice Question 1

A teacher asked her students to record the total amount Of time they spent studying for a particular test.

The amounts of study time x (in hours) and the resulting test grades y are given below

x	2	1	1.5	0.5	1	3
y	92	81	84	68	85	96

1. Obtain the equation of the least-squares regression line and the correlation.

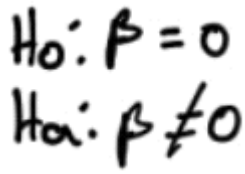


- y = 69.7 + 9.75x
- r = 0.896 (strong correlation)
- r^2 = 0.803 (80.3% of the change in grade can be explained by the study time)

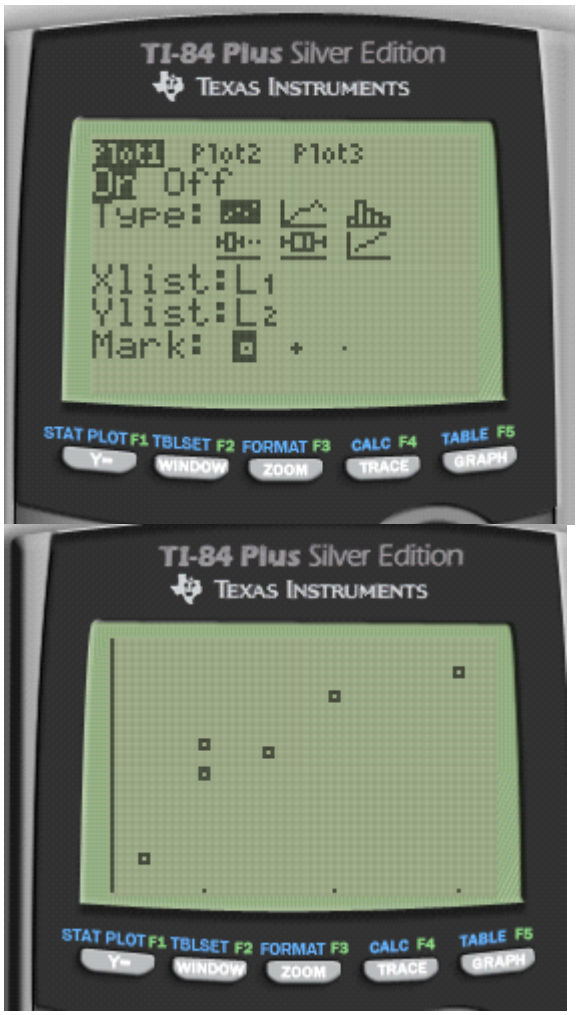
2. Explain in words what the slope b of the least-squares line says about hours studied a nd grade awarded.
    - For every 1 hour increase in study time, the grade is expected to go up by 9.75 points
3. Test the hypothesis that the amount of study time is correlated to the test grade
    - Data

L1	L2	L3	L4
x	y	y hat	Residual

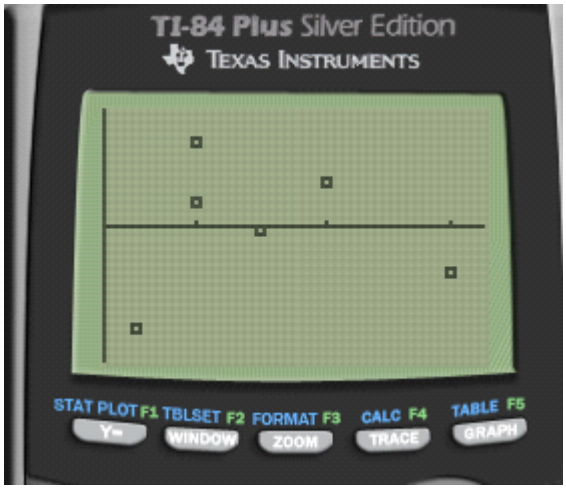
    - Hypothesis



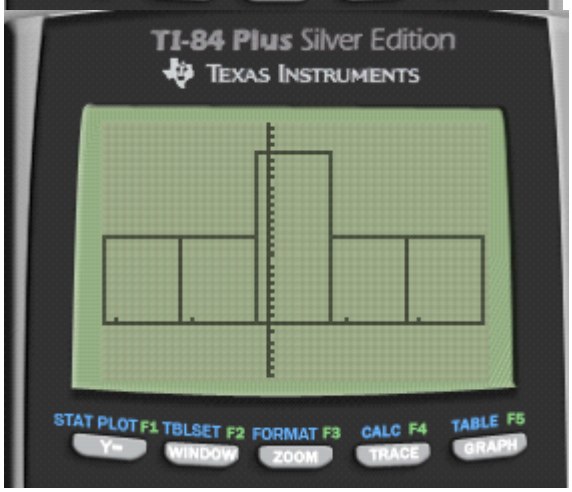
- Conditions
  - Linearity



  - Constant Variance



- Normal Residuals



- Independence: each observation is independent

○ Calculate

$$t = \frac{b}{SE_b} = 4.04$$

$$df = \begin{matrix} 4 \\ 6-2=4 \end{matrix}$$

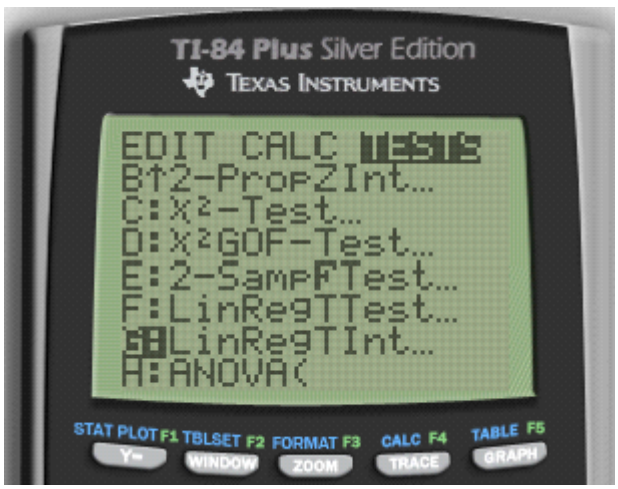
$$P\text{-value} = .016 < \alpha = .05$$

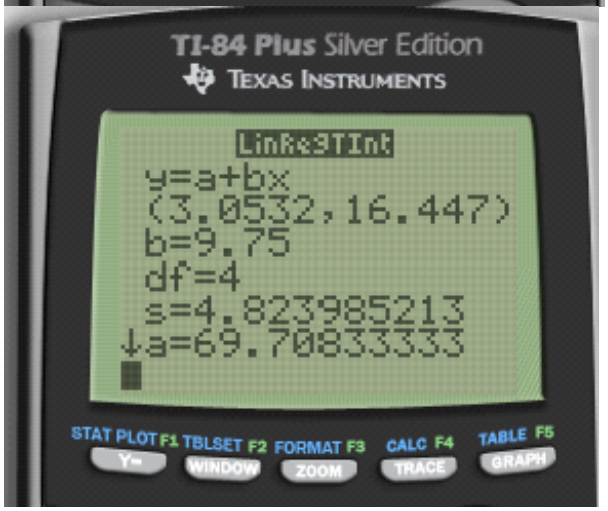
- Interpret
  - So we reject the null hypothesis and have evidence to support the claim that the slope is not equal to zero. There is a correlation between study time and test grades

4. What is the 95% confidence interval of the slope?
   - Equation

$$b \pm t^*(SE_b)$$

   - Calculate





- Interpret
  - We are 95% confident that on average, for every 1 hour increase in study time, the final grade will go up between 3.05 and 16.45 points

Interpreting Computer Output

Regression Analysis: Final versus Quiz Average

The regression equation is
Final = 12.1 + 0.751 Quiz Average

Predictor	Coef	SE Coef	T	P
Constant	12.12	11.94	1.01	0.315
Quiz Average	0.7513	0.1414	5.31	0.000

S = 9.71152 R-Sq = 37.0% R-Sq(adj) = 35.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2663.7	2663.7	28.24	0.000
Residual Error	48	4527.1	94.3		
Total	49	7190.7			

$\hat{y} = 12.12 + .7513(x)$

Y = final grade
X = quiz grade average

Practice Question 2

An economics professor wishes to analyze whether a person's income can predict the cost of their car

Predictor	Coef	SE Coef	T	P
Constant	438.525	3.341	131.25	0.000
Income	0.51145	0.02325	22.00	0.000

S = 12.2225 R-Sq = 91.0% R-Sq(adj) = 90.8%

- What's the least-squares regression equation
  - y hat = 438.535 + 0.511 * x
  - y = cost of car
  - x = income

- What is the standard error about the line (aka the standard deviation of the regression model)? Interpret this value in context
    - On average, we expect our prediction of cost is off by 12.22.
- Interpret the slope of the least-squares regression line in the context of this problem
    - For every $1 increase in income, car cost increases, on average, $0.51
- What are the null and alternative hypotheses to test if there is an association between income and car cost?
    - $H_0: \beta = 0$ ← Is not an association
      $H_a: \beta \neq 0$ ← IS an association
- What is the value of the test statistic for testing the hypotheses
    - $H_0: \beta = 0$
      testing b
      $t = \frac{b}{SE_b} = \frac{.511}{.0232} = 22$
- What is the P-value for the test
    - $P < 0.001$
- Is income useful for predicting the cost of a person's car? Use a significance level of 0.01. Explain briefly

P-Value < 0.001 < $\alpha = .01$

Reject H_0, we have evid. to support the claim that there is

- an ass. between income and car cost

So income does appear to be useful in predicting car cost

Practice Question 3

Test if the number of beers is associated with the BAC

```
The regression equation is
BAC = - 0.0127 + 0.0180 Beers
```

n = 24

Predictor	Coef	StDev
Constant	-0.01270	0.01264
Beers	0.017964	0.002402

S = 0.02044 R-Sq = 80.0%

$H_0: \beta = 0$
$H_a: \beta \neq 0$

$\rightarrow t = \frac{b}{SE_b} = \frac{.018}{.0024} = 7.48$

$df = 24 - 2 = 22$

P-Value < .001 < $\alpha = .$

Assume all conditions are satisfied:
linearity
constant variance
normality of the residuals
independence

Reject the null and do have evidence to support that # of beers is associated with the BAC

6.6 - Hypothesis Tests for Categorical Data (Chi-Squared Tests)

Wednesday, February 15, 2017 1:43 PM

Chi-Squared Goodness of Fit Test

- One categorical variable with counts (or proportion) in each category
- We have seen: products are produced by two machines, machine A produced 15 defective parts in a run of 280, while machine B produced 10 defective parts in a run of 200. Is there a difference in the reliability of these two machines?
- New question type: products are produced by three machines, machine A produced 15 defective parts in a run of 280, while machine B produced 10 defective parts in a run of 200. Is there a difference in the reliability of these machines?

Practice Question 1

In the past, for a large introductory statistics course, the proportions of students that received grades of A, B, C, D, or F have been, respectively, 0.15, 0.35, 0.30, 0.10, and 0.10

This year, there were 200 students in the class, and following grades were given:

Grade	A	B	C	D	F
Number	51	79	61	8	1

Test to see whether the distribution of grades this year was different from the distribution in the past?

- Hypothesis
  - H0: PA = 0.15, PB = 0.35, PC = 0.30, PD = 0.10, PF = 0.10
  - H1: at least one p does not fit the distribution
- Calculate expected values

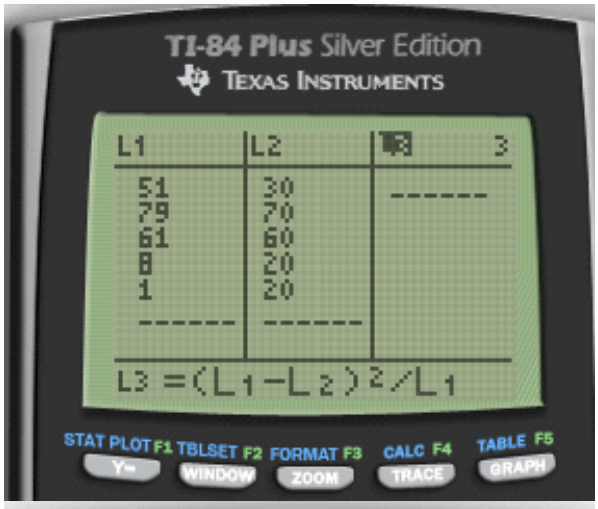
Grade	A	B	C	D	F
Observed	51	79	61	8	1
Expected	30	70	60	20	20

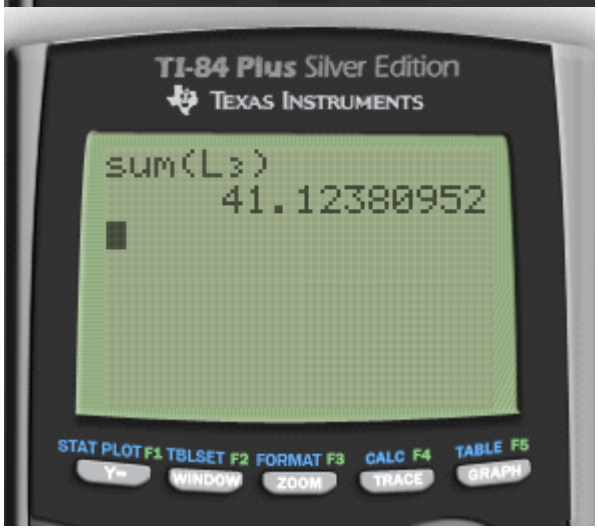
- Conditions
  - Random
  - Independent

- Count: At least 80% of the expected counts are greater than 5 and none are less than 1

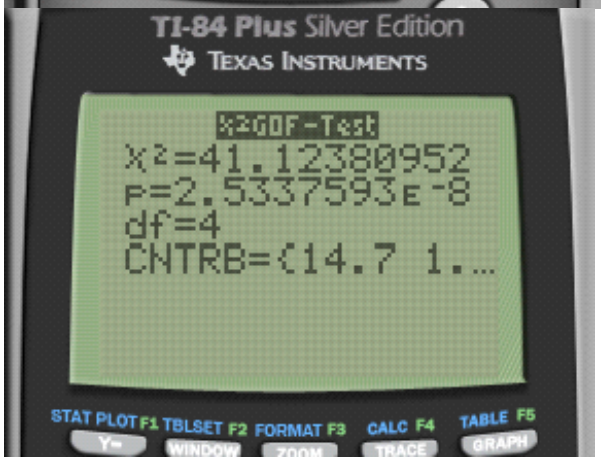
- Calculate

$$X^2 = \sum \frac{(o-e)^2}{e}$$



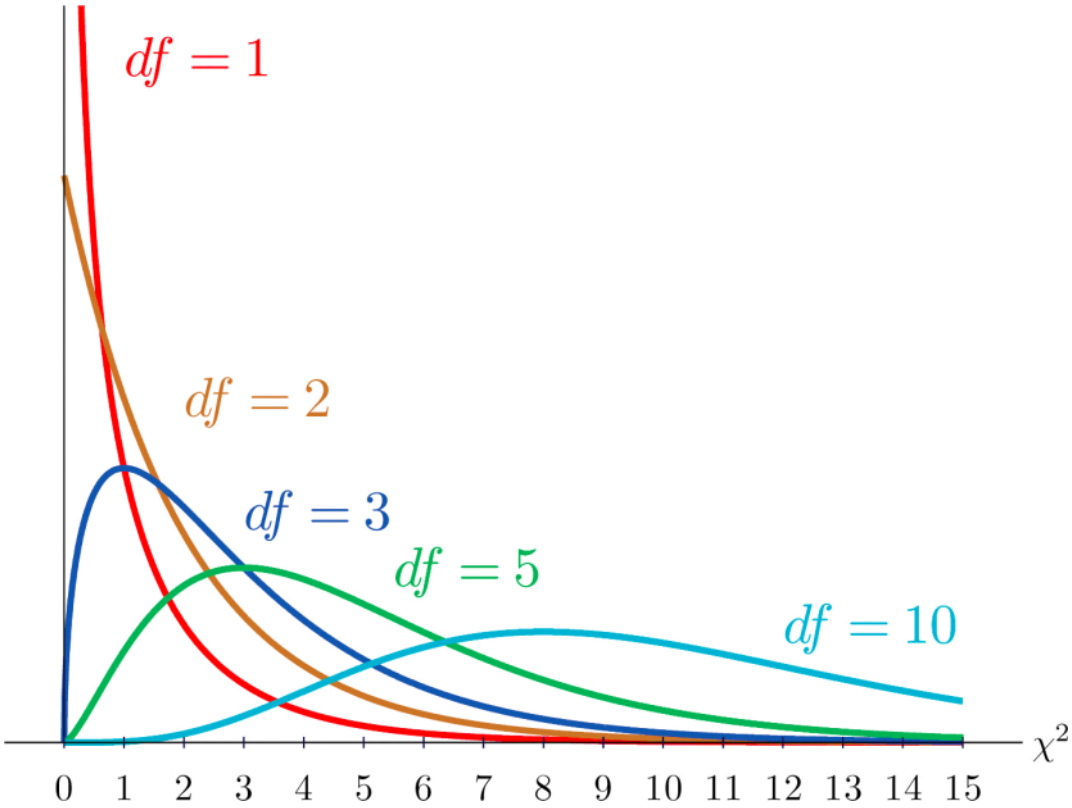


- Calculate by calculator



- P-value

  df = k - 1 (k: number of categories)



- Interpret
    - $P < \alpha$
    - So we reject the null hypothesis and have evidence to support the claim that at least one grade proportion does not fit the expected distribution

Chi-Squared Test of Homogeneity or Independence/Association

- Two categorical variables
- Homogeneity
    - Do two or more sub-groups of a population share the same distribution of a categorical variable (each group has its own sample)
    - Do people of different races have the same proportion of smokers to non-smokers.
    - Do different education levels have different proportions of Democrats, Republicans, and Independent
- Independence/Association
    - Determining whether two categorical variables are associated (variables from a single SRS)
    - Is there an association between race and smoking status
    - Is there an association between education and voting preference

Practice Question 2

Girls and boys at an elementary school were sampled and asked about their favorite subject

1. Does favorite subject differ by gender?

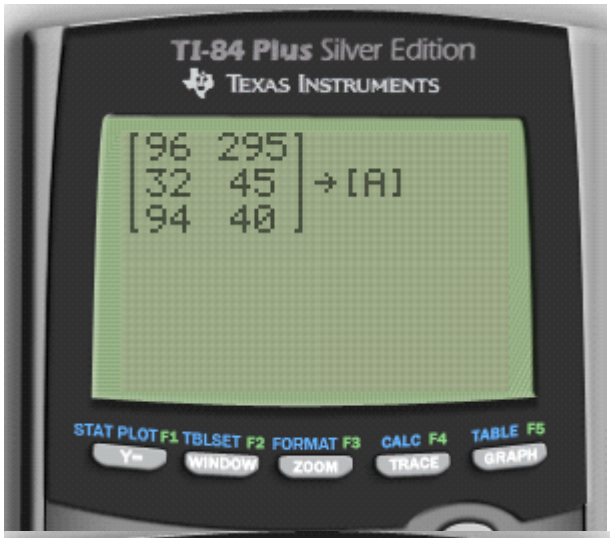
Favorite subject	Boys	Girls	Total
Math	96	295	391
English	32	45	77
Social Studies	94	40	134
Total	222	380	602

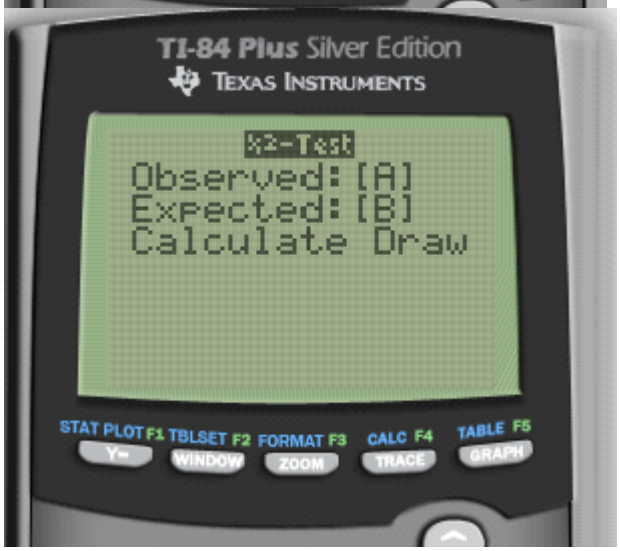
- Hypothesis
    - H0: favorite subject does not differ by gender
    - H1: favorite subject does differ by gender

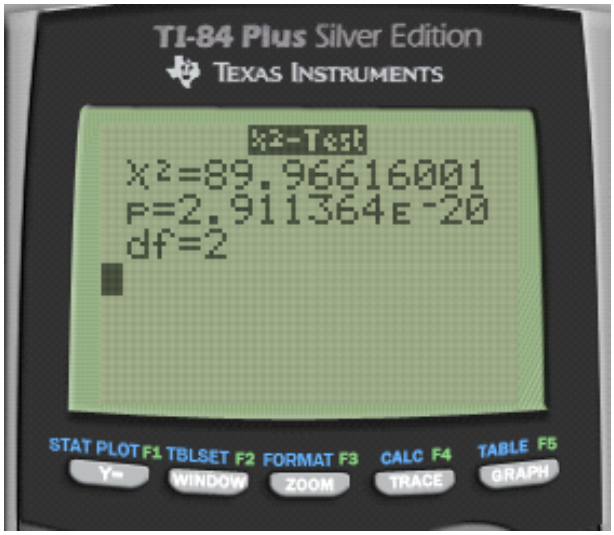
- Expected
  - Row Total * Colum Total / Total
- Conditions
  - For each sub group, the sample is a SRS
  - NO expected cell counts are < 5
- Calculate

$$X^2 = \sum \frac{(o-e)^2}{e}$$

- Calculate by calculator







- P-value
  - df = (r-1)*(c-1)
- Interpret
  - P < α
  - So we reject the null hypothesis and have evidence to support the claim that favourite subject is different between boys and girls

2. Is favourite subject associated with gender?

- H0: There is no association between favorite subject and gender
- H1: There is an association between favorite subject and gender

Practice Question 3

You are playing a dice game with a friend. They brought a 6 sided die that you think may not be fair. You conduct an experiment to determine if it is fair. You roll the die 100 times and get following:

Side	1	2	3	4	5	6
Frequency	17	24	15	22	12	10

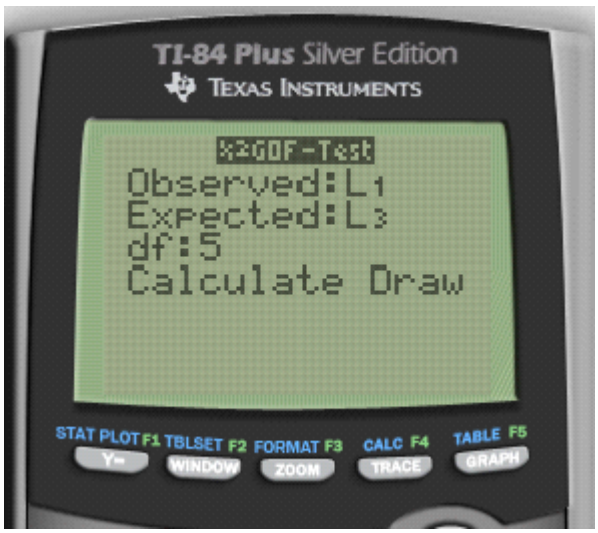
- Hypothesis
  - H0: P1 = P2 = P3 = P4 = P5 = P6
  - H1: at least one is not equal
- Expected
  - 1/6 = 0.1667
- Conditions

- Random
- Independent
- Expected counts are greater than 5

- Calculate

$$X^2 = \sum \frac{(o-e)^2}{e}$$

- Calculate by calculator



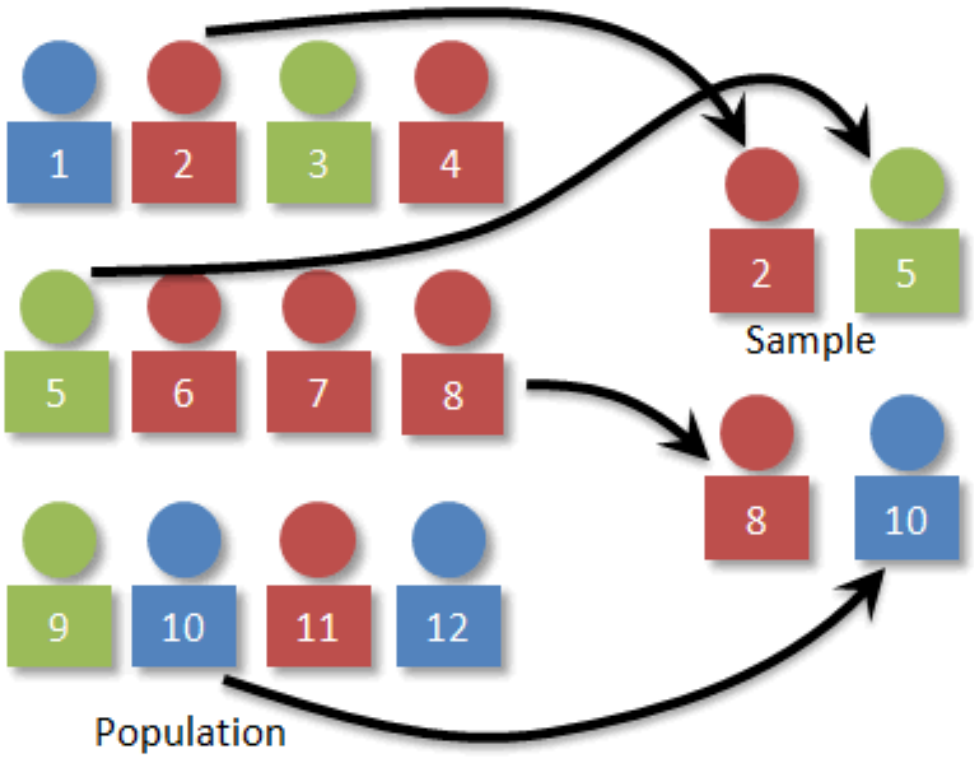
- Interpret
  - P > α
  - So we fail to reject the null hypothesis and do not have evidence to support the claim that the die is unfair

Sample Questions

Wednesday, April 5, 2017 11:28 AM

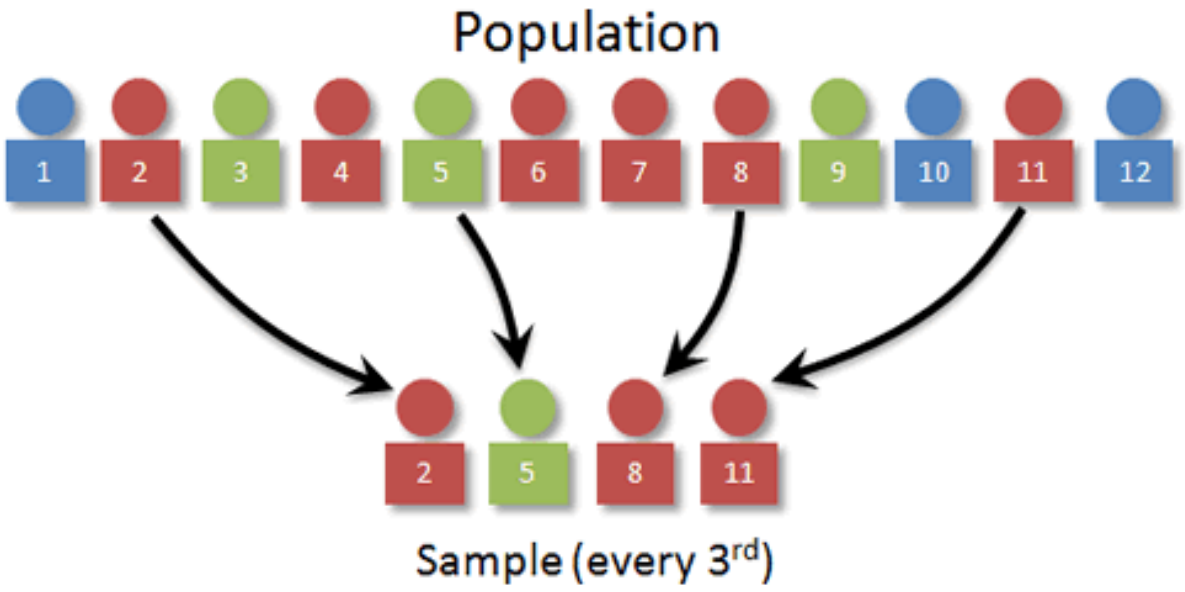
Question 2

- Simple Random Sampling

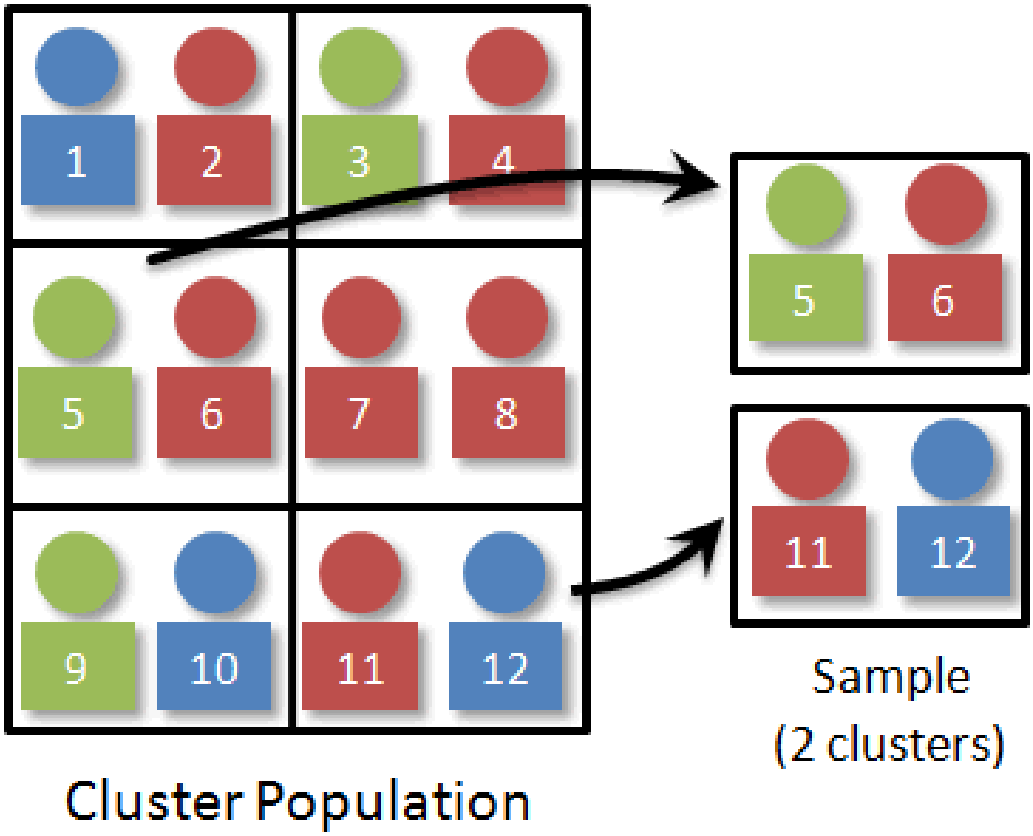


- Stratified Sampling

- Systematic Sampling



- Cluster Sampling


1
2
3
4
5
6
7
8
9
10
11
12
5
6
11
12
Sample
(2 clusters)
Cluster Population

Question 6

P-Value Approach

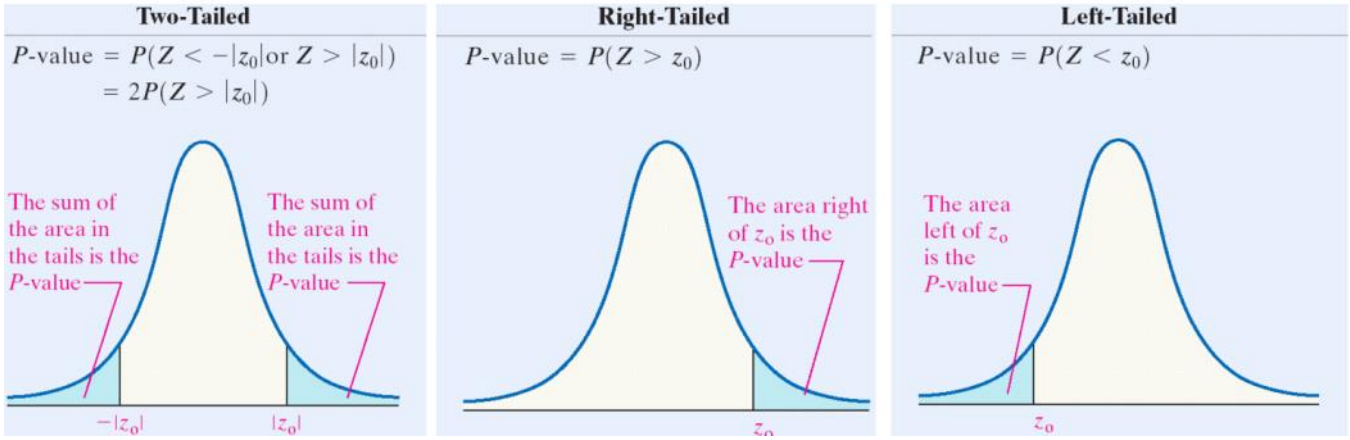
Assume that the null hypothesis is true.

The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.

How to compute the P-Value for each type of test:

Step 1: Compute the test statistic $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Two-tail | Right Tail | Left Tail



Question 14

- -1 < r < 1

Formula for the Least Squares Regression Line (LSRL)		
LSRL Formula	**Slope and Intercept**	**Variables in Slope and Intercept Formulas**
$\hat{y} = a + bx$	Slope: $b = r\dfrac{s_y}{s_x}$ Intercept: $a = \bar{y} - b\bar{x}$	r = correlation coefficient $\bar{x}$ = mean of independent variable $\bar{y}$ = mean of dependent variable s_x = standard deviation of independent variable s_y = standard deviation of dependent variable

Question 15

- The power of a test is affected by sample size (bigger sample, more power) and alpha level (larger alpha, i.e. .05 compared to .01, more power)

SIGNIFICANCE LEVEL

There is a trade-off between the *significance level* and power: the more stringent (lower) the significance level, the lower the power. Figure 3 shows that power is lower for the 0.01 level than it is for the 0.05 level. Naturally, the stronger the evidence needed to reject the null hypothesis, the lower the chance that the null hypothesis will be rejected.

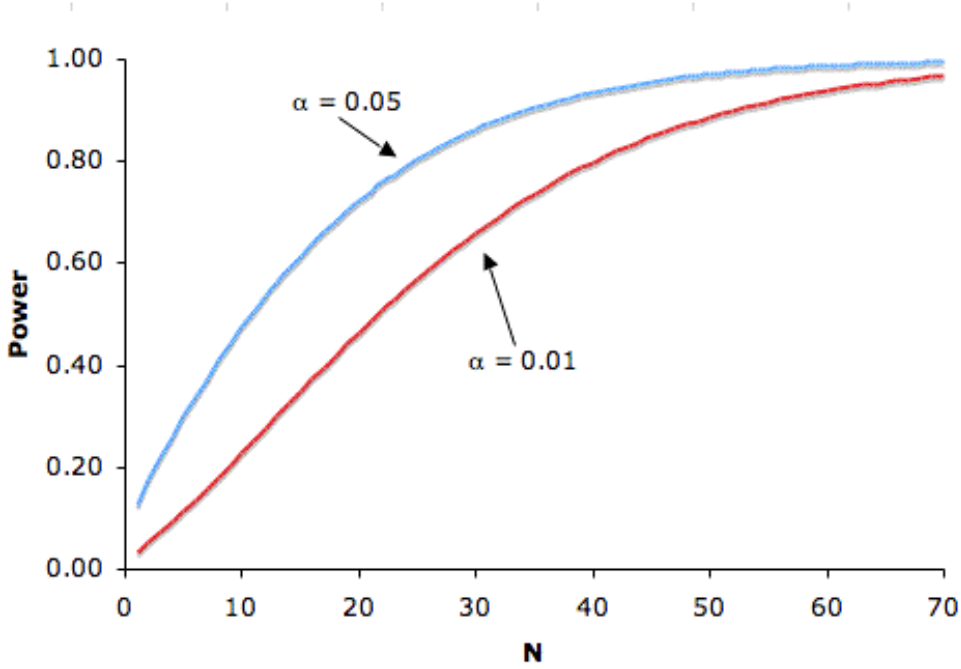


Figure 3. The relationship between significance level and power with one-tailed tests: μ = 75, real μ = 80, and σ = 10.

Question 18

the width of a confidence interval is dependent on the z (or t)* and the standard deviation of the statistic. Assuming the z* is unchanged, the question is which is the smaller standard deviation
sqrt(.7*.3/50) is approximately 0.0648 while sqrt([37/60*23/60]/60) is approximately 0.0628
so the confidence interval based on 37 out of 60 will be slightly narrower than the confidence interval based on 35 out of 50

2002 Multiple Choice

Tuesday, February 21, 2017 3:19 PM

Question 10

The lengths of individual shellfish in a population of 10,000 shellfish are approximately normally distributed with mean 10 centimeters and standard deviation 0.2 centimeter. Which of the following is the shortest interval that contains approximately 4,000 shellfish lengths?

4,000 = 40%

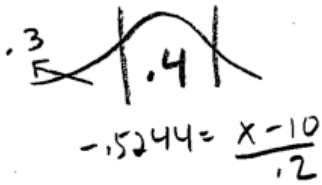
(A) 0 cm to 9.949 cm

(B) 9.744 cm to 10 cm

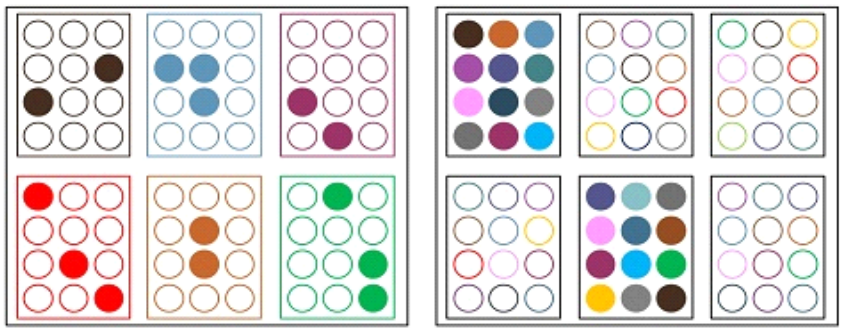
(C) 9.744 cm to 10.256 cm

(D) 9.895 cm to 10.105 cm

(E) 9.9280 cm to 10.080 cm



Question 15



Stratified Sampling Vs Cluster Sampling

Question 16

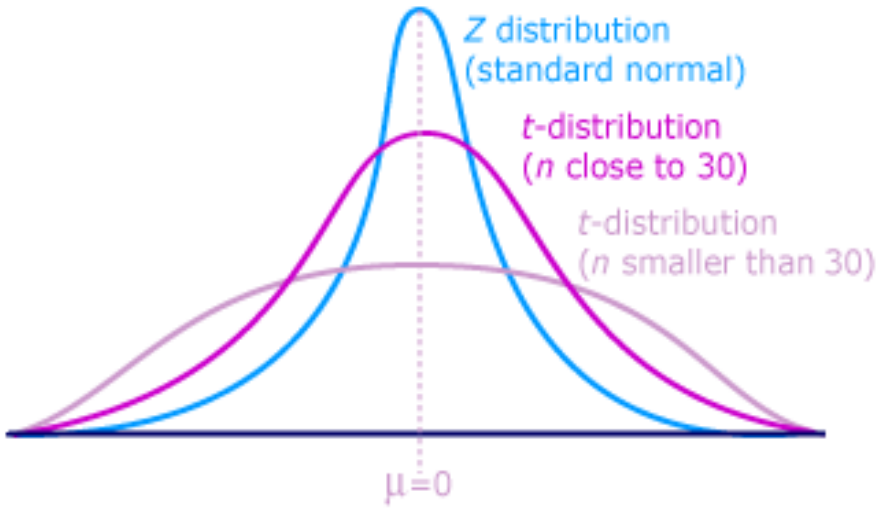
Jason wants to determine how age and gender are related to political party preference in his town. Voter registration lists are stratified by gender and age-group. Jason selects a simple random sample of 50 men from the 20 to 29 age-group and records their age, gender, and party registration (Democratic, Republican, neither). He also selects an independent simple random sample of 60 women from the 40 to 49 age-group and records the same information. Of the following, which is the most important observation about Jason's plan?

(A) The plan is well conceived and should serve the intended purpose.
(B) His samples are too small.
(C) He should have used equal sample sizes.
(D) He should have randomly selected the two age groups instead of choosing them nonrandomly.
(E) He will be unable to tell whether a difference in party affiliation is related to differences in age or to the difference in gender. He's considering 2 different variables

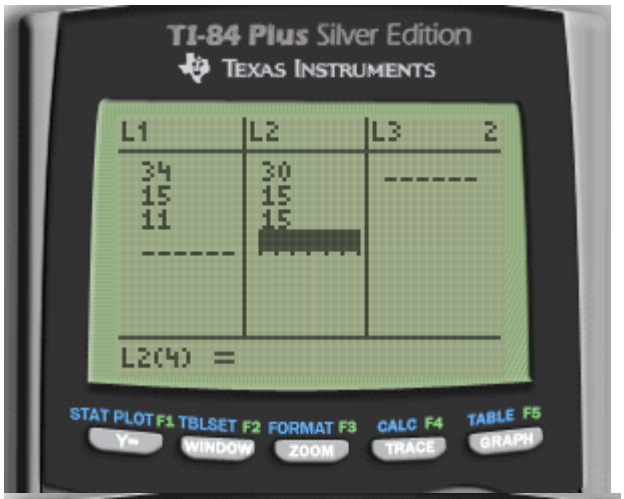
Question 17

- Residuals = observed/actual y - predicted y
- See Chi-square test statistic formula

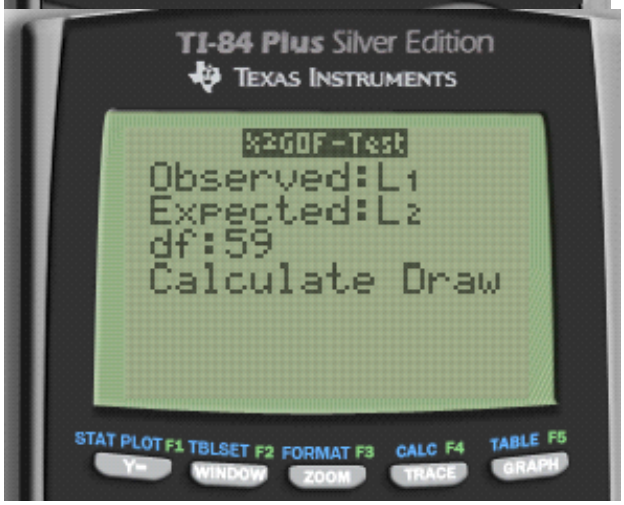
Question 18

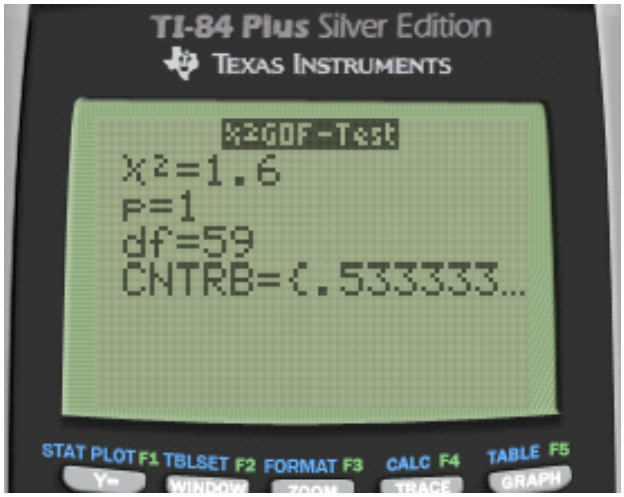

Z distribution
(standard normal)
t-distribution
(n close to 30)
t-distribution
(n smaller than 30)
μ=0

Question 19


TI-84 Plus Silver Edition
TEXAS INSTRUMENTS
L1 L2 L3 2
34 30 ------
15 15
11 15

L2(4) =
STAT PLOT F1 TBLSET F2 FORMAT F3 CALC F4 TABLE F5
Y= WINDOW ZOOM TRACE GRAPH


TI-84 Plus Silver Edition
TEXAS INSTRUMENTS
χ2GOF-Test
Observed:L1
Expected:L2
df:59
Calculate Draw
STAT PLOT F1 TBLSET F2 FORMAT F3 CALC F4 TABLE F5
Y= WINDOW ZOOM TRACE GRAPH



Question 23

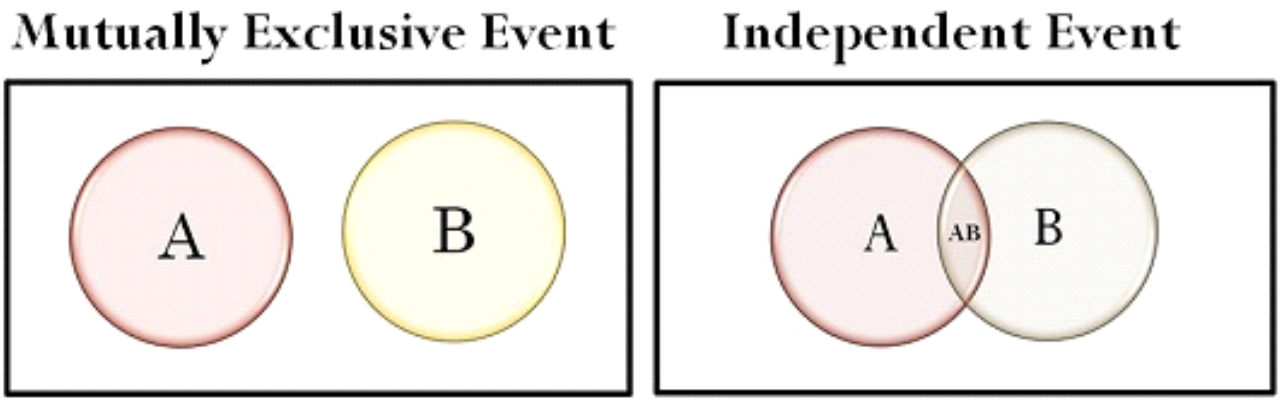
Events are mutually exclusive if the occurrence of one event excludes the occurrence of the other(s). Mutually exclusive events cannot happen at the same time. For example: when tossing a coin, the result can either be `heads` or `tails` but cannot be both.

$$\left.\begin{aligned} P(A \cap B) &= 0 \\ P(A \cup B) &= P(A) + P(B) \\ P(A \mid B) &= 0 \\ P(A \mid \neg B) &= \frac{P(A)}{1 - P(B)} \end{aligned}\right\} \text{mutually exclusive } A, B$$

Events are independent if the occurrence of one event does not influence (and is not influenced by) the occurrence of the other(s). For example: when tossing two coins, the result of one flip does not affect the result of the other.

$$\left.\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(A \cup B) &= P(A) + P(B) - P(A)P(B) \\ P(A \mid B) &= P(A) \\ P(A \mid \neg B) &= P(A) \end{aligned}\right\} \text{independent } A, B$$

This of course means mutually exclusive events are not independent, and independent events cannot be mutually exclusive. (Events of measure zero excepted.)



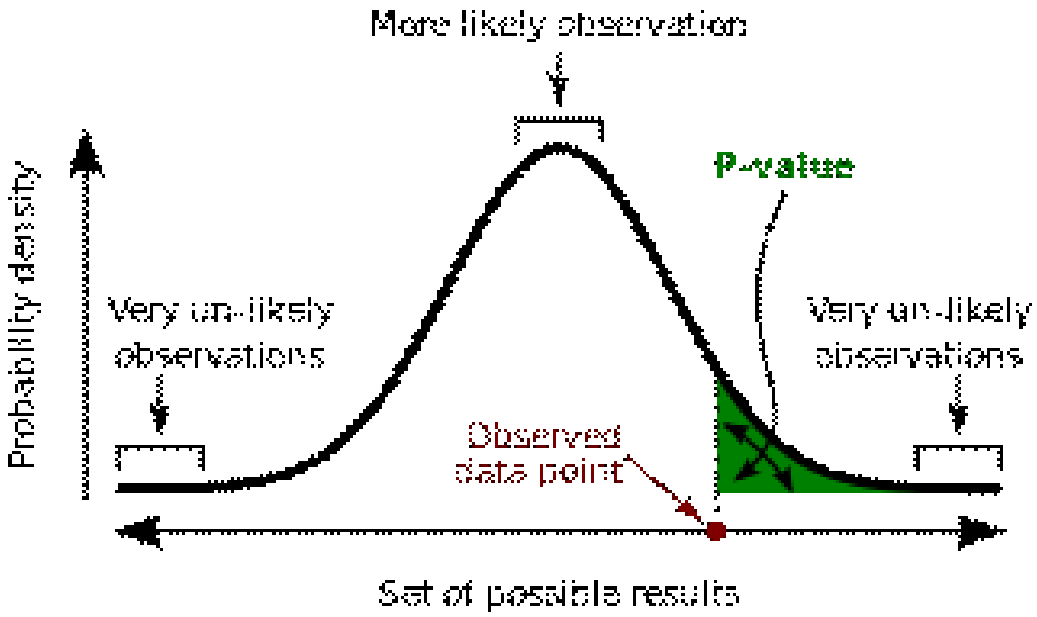
Question 24

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Question 34

Coefficient of Determiation

- The **coefficient of determination** is the ratio of the explained variation to the total variation.
- The symbol for the coefficient of determination is r^2.
- $$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$
- Another way to arrive at the value for r^2 is to square the correlation coefficient.

Bluman, Chapter 10 49

Question 35

. In a test of the hypothesis $H_0: \mu = 100$ *versus* $H_a: \mu > 100$, the power of the test when $\mu = 101$ would be greatest for which of the following choices of sample size n and significance level α?

(A) $n = 10, \alpha = 0.05$

(B) $n = 10, \alpha = 0.01$

(C) $n = 20, \alpha = 0.05$

(D) $n = 20, \alpha = 0.01$

(E) It cannot be determined from the information given.

as $n \uparrow$, less variability

as $\alpha \uparrow$, reject more

Question 37

A simple random sample produces a sample mean, $\bar{x}$, of 15. A 95 percent confidence interval for the corresponding population mean is 15 ± 3. Which of the following statements must be true?

(A) Ninety-five percent of the population measurements fall between 12 and 18. The pop is the pop → doesn't change.

(B) Ninety-five percent of the sample measurements fall between 12 and 18. you are estimating μ, not individ measurements

(C) If 100 samples were taken, 95 of the sample means would fall between 12 and 18.

(D) $P(12 \leq \bar{x} \leq 18) = 0.95$ This is only based off 1 sample.

(E) If $\mu = 19$, this $\bar{x}$ of 15 would be unlikely to occur.

If $\mu = 19$, it should be in the C.I.,
but it's not, so $\bar{x} = 15$ is unlikely

Question 40

A student working on a history project decided to find a 95 percent confidence interval for the difference in mean age at the time of election to office for former American Presidents versus former British Prime Ministers. The student found the ages at the time of election to office for the members of both groups, which included all of the American Presidents and all of the British Prime Ministers, and used a calculator to find the 95 percent confidence interval based on the t-distribution. This procedure is not appropriate in this context because

(A) the sample sizes for the two groups are not equal

(B) the entire population was measured in both cases, so the actual difference in means can be computed and a confidence interval should not be used You know the actual difference ⇒ No need to estimate

(C) elections to office take place at different intervals in the two countries, so the distribution of ages cannot be the same

(D) ages at the time of election to office are likely to be skewed rather than bell-shaped, so the assumptions for using this confidence interval formula are not valid

(E) ages at the time of election to office are likely to have a few large outliers, so the assumptions for using this confidence interval formula are not valid

2011 Free Response

2017年4月27日 星期四 下午9:26

Question 1 (a)

No, it is not reasonable to believe that the distribution of 40-yard running times is approximately normal, because the minimum time is only 1.33 standard deviations below the mean $\left(z = \frac{4.4 - 4.6}{0.15} \approx -1.33\right)$. In a normal distribution, approximately 9.2 percent of the z-scores are below -1.33. However, there are no running times less than 4.4 seconds, which indicates that there are no running times with a z-score less than -1.33. Therefore, the distribution of 40-yard running times is not approximately normal.

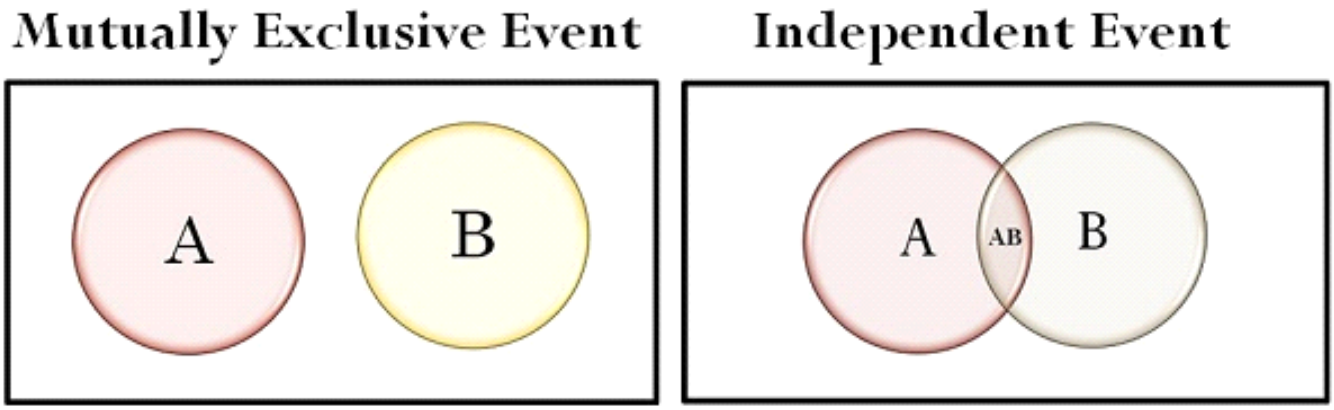
Question 1 (b)

- How to interpret the z-score

The z-score for a player who can lift a weight of 370 pounds is $z = \frac{370 - 310}{25} = 2.4$. The z-score indicates that the amount of weight the player can lift is 2.4 standard deviations above the mean for all previous players in this position.

Question 2 (b)

- Independence vs. Dependence
  - A conditional probability is the probability of some event occurring, given that some other event has already occurred. The conditional probability of event X occurring, given that some other event Y has already occurred, is written as P(X|Y).
  - For example, P(M|N) would be the probability of the occurrence of event M given that event N has already occurred. It would be read as "the probability of M, given N."
  - As stated earlier, two events are considered independent if the occurrence of one of the events does not change the probability of the other event from what it would have been had the first event not occurred. Thus, two events, X and Y, are independent if P(X|Y) = P(X) or P(Y|X) = P(Y)
  - Actually, these two conditional relationships are related. If one is true, the other must be true. If one is false, the other must be false.
  - If **P(X|Y) = P(X)**, then **P(Y|X) = P(Y)**, and the events are **independent**.
  - If **P(X|Y) ≠ P(X),** then **P(Y|X) ≠ P(Y)**, and the events are **dependent**.

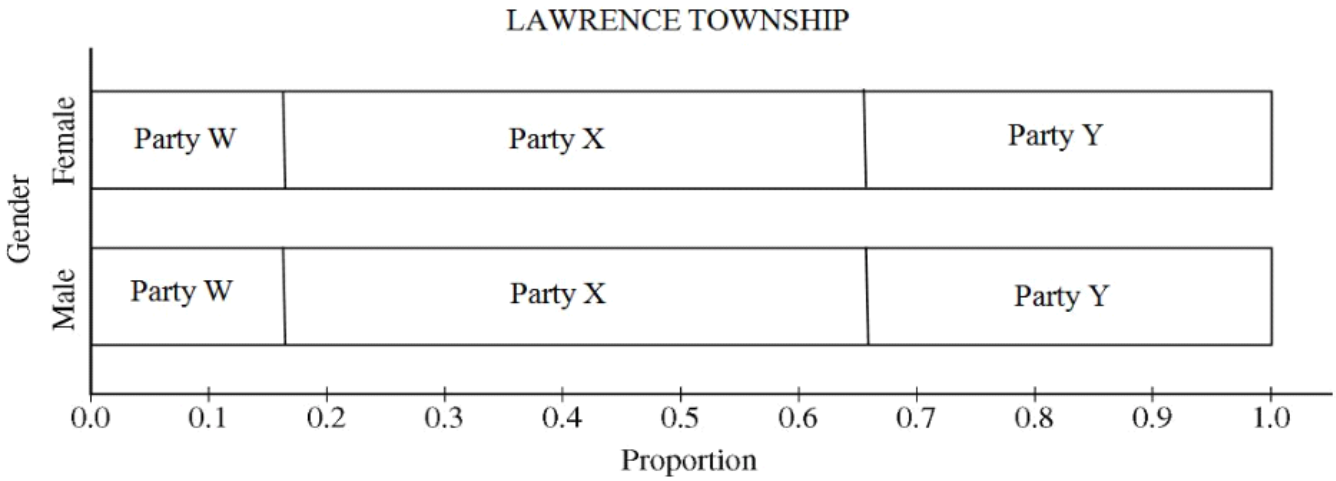


Question 2 (c)

The marginal proportions of voters registered for each of the three political parties (without regard to gender) are given below.

$$\text{Party W: } \frac{88}{500} = 0.176$$
$$\text{Party X: } \frac{244}{500} = 0.488$$
$$\text{Party Y: } \frac{168}{500} = 0.336$$

Because party registration is independent of gender in Lawrence Township, the proportions of males and females registered for each party must be identical to each other and also identical to the marginal proportion of voters registered for that party. Using the order Party W, Party X, and Party Y, the graph for Lawrence Township is displayed below.



Question 3 (a)

- Process for randomly selecting 2 numbers from 1 to 9

Step 1: Generate a random integer between 1 and 9, inclusive, using a calculator, a computer program, or a table of random digits. Select all four apartments on the floor corresponding to the selected integer.

Step 2: Generate another random integer between 1 and 9, inclusive. If the generated integer is the same as the integer generated in step 1, continue generating random integers between 1 and 9 until a different integer appears. Again select all four apartments on the floor corresponding to the second selected integer.

Question 3 (b)

Because the amount of wear on the carpets in apartments with children could be different from the wear on the carpets in apartments without children, it would be advantageous to have apartments with children represented in the sample. The cluster sampling procedure in part (a) could produce a sample with no children in the selected apartments; for example, a cluster sample of the apartments on the third and sixth floors would consist entirely of apartments with no children. Stratified random sampling, where the two strata are apartments with children and apartments without children, guarantees a sample that includes apartments with and without children, which, in turn, would yield sample data that are representative of both types of apartments.

Question 4

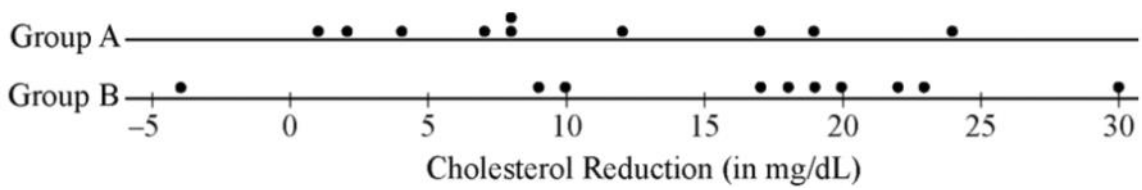
- Degrees of freedom for two-sample t-test.

$$t\text{-}statistic = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

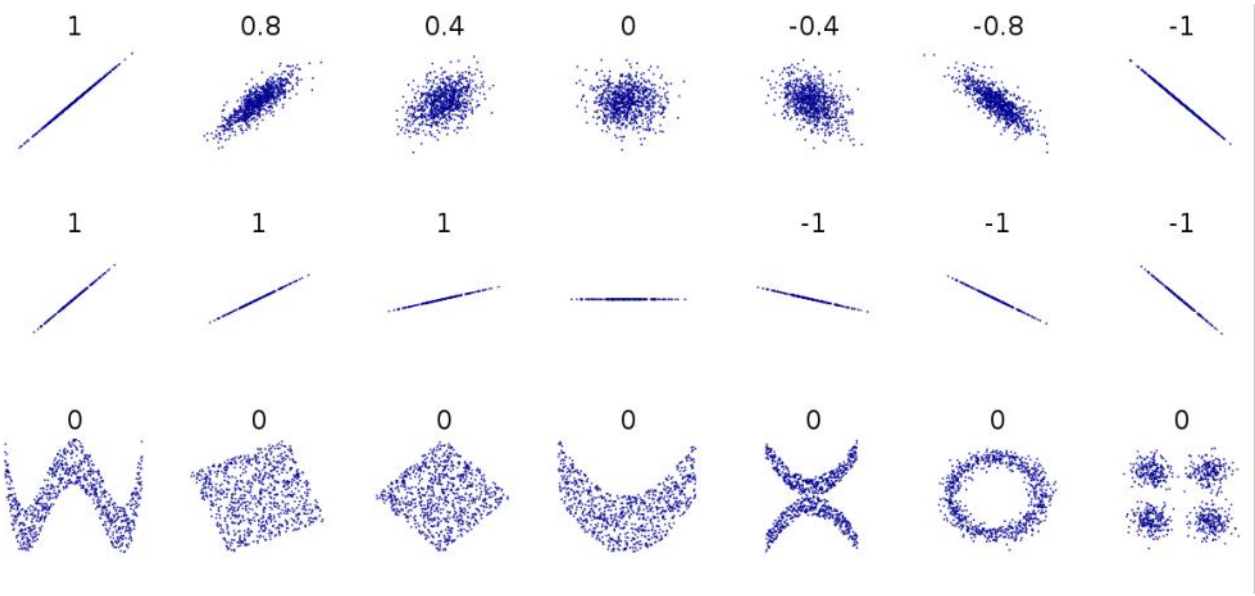
- Graphs of both distributions must be produced and described to check the normality condition.

The second condition is that the two populations are approximately normally distributed or the sample sizes are sufficiently large. Because of the small sample sizes (10 in each treatment group), we need to check whether it is reasonable to assume that the samples came from populations that are normally distributed. The following dotplots reveal slight skewness and a possible outlier for group B, but it appears reasonable to proceed with the two-sample *t*-test.



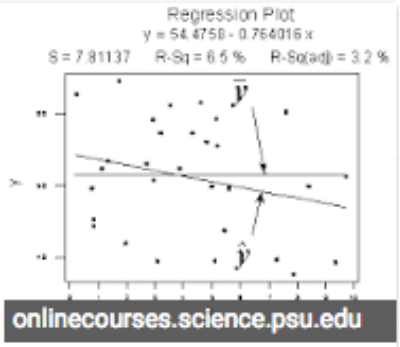
Question 5 (c)

- Correlation coefficient



- Coefficient of determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. ... An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable.



Coefficient of Determination: Definition - Stat Trek
stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination

Question 5 (d)

- Explaining relationship

Yes, there is very strong statistical evidence that the population slope differs from zero, so electricity production is linearly related to wind speed. For testing the hypotheses $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$, where β represents the population slope, the output reveals that the test statistic is $t = 12.63$ and the p-value (to three decimal places) is 0.000. Because the p-value is so small (much less than both 0.05 and 0.01), the sample data provide very strong statistical evidence that electricity production is linearly related to wind speed.

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}}$$

= standard deviation of the residuals = expected variability in our errors

$$= \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$$

$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

standard error of slope

$$SE_b = \frac{s}{\sqrt{\sum (x-\bar{x})^2}}$$

$SD_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

* $df = n-2$

$H_0: \beta = 0$

$H_a: \beta \neq 0$

$$t = \frac{b}{SE_b}$$

test statistic

$$b \pm t^* SE_b$$

confidence interval

Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses

H_0: $\beta_1 = 0$	(no linear relationship)
H_1: $\beta_1 \neq 0$	(linear relationship does exist)

- Test statistic

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

$df = n - 2$

Question 6 (a)

- Checking condition for one-sample z-interval
  - Random
  - Normal: Large sample size
    - To satisfy the third component, the response:
      - Must check both the number of successes and the number of failures.
      - Must use a reasonable criterion (for example, ≥ 5 or ≥ 10).
      - Must provide numerical evidence (for example, $2{,}688 \geq 10$ and $6{,}912 \geq 10$, or $9{,}600 \times 0.28 \geq 10$ and $9{,}600 \times 0.72 \geq 10$).

2011 Free Response (Form B)

2017年4月27日 星期四 下午9:26

Question 1 (a)

- Estimation for median

The median is the value with half of the P-T ratios at or below it and half of the values at or above it. For n observations in a group, use $\frac{n+1}{2}$ to find the position of the median in the ordered list of observations.

For states west of the Mississippi $(n=24)$ the median falls between the 12th and 13th value in the ordered list, and both the 12th and 13th values fall in the interval 15–16. For states east of the Mississippi $(n=26)$ the median falls between the 13th and 14th value in the ordered list, and both of these values also fall in the interval 15–16.

From the histogram, cumulative frequencies for the two groups are shown in the table below.

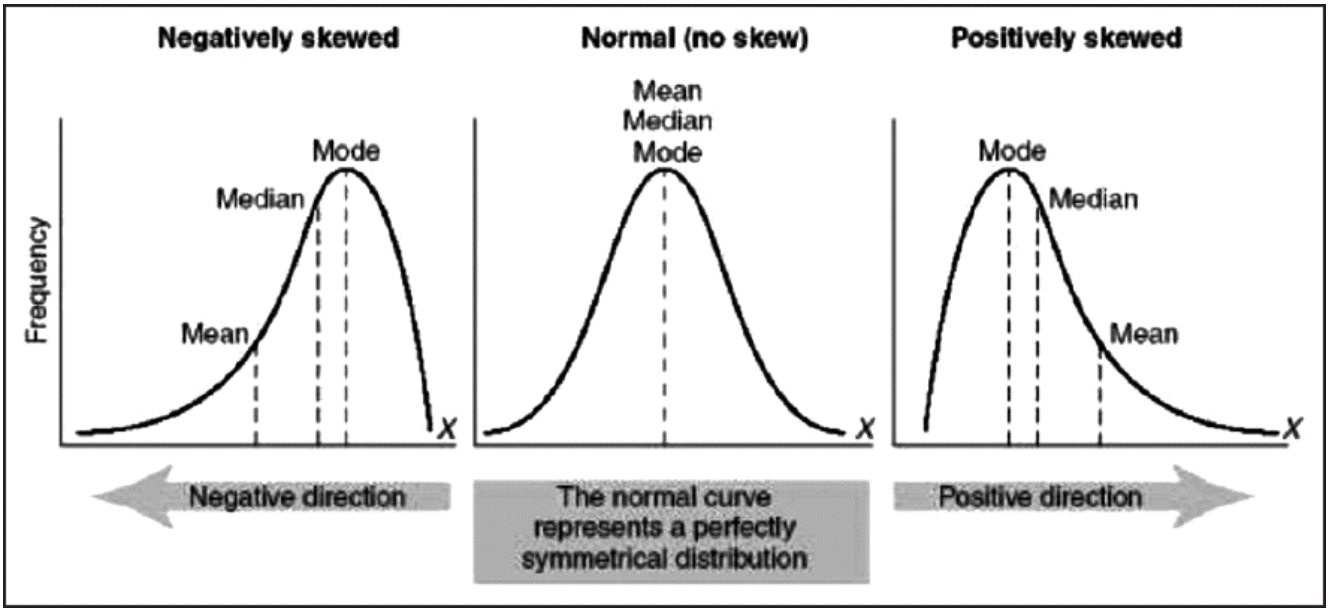
Interval	West	East
12–13	1	2
13–14	$1+4=5$	$2+4=6$
14–15	$1+4+6=11$	$2+4+4=10$
15–16	$1+4+6+3=14$	$2+4+4+11=21$

Thus, the median P-T ratio for both groups is at least 15 students per teacher and at most 16 students per teacher.

Question 1 (c)

The medians of the two distributions are about the same, as determined in part (a). The distribution of P-T ratios for states that are west of the Mississippi River is skewed to the right, indicating that the mean will probably be higher than the median. The rough symmetry for the east group indicates that the mean will be close to the median. Thus, the mean for the west group will probably be greater than the mean for the east group.

- Mean, Median, and Skew



Question 1 (a)

The study was an experiment because treatments (D-cycloserine or placebo) were imposed by the researchers on the people with acrophobia.

Question 3 (a)

- Pay attention to the notation

Let Y denote the number of flights Sam must make until he receives his first upgrade. The random variable Y follows a geometric distribution with $p = 0.1$. The probability that Sam's upgrade will occur after his third flight is calculated below.

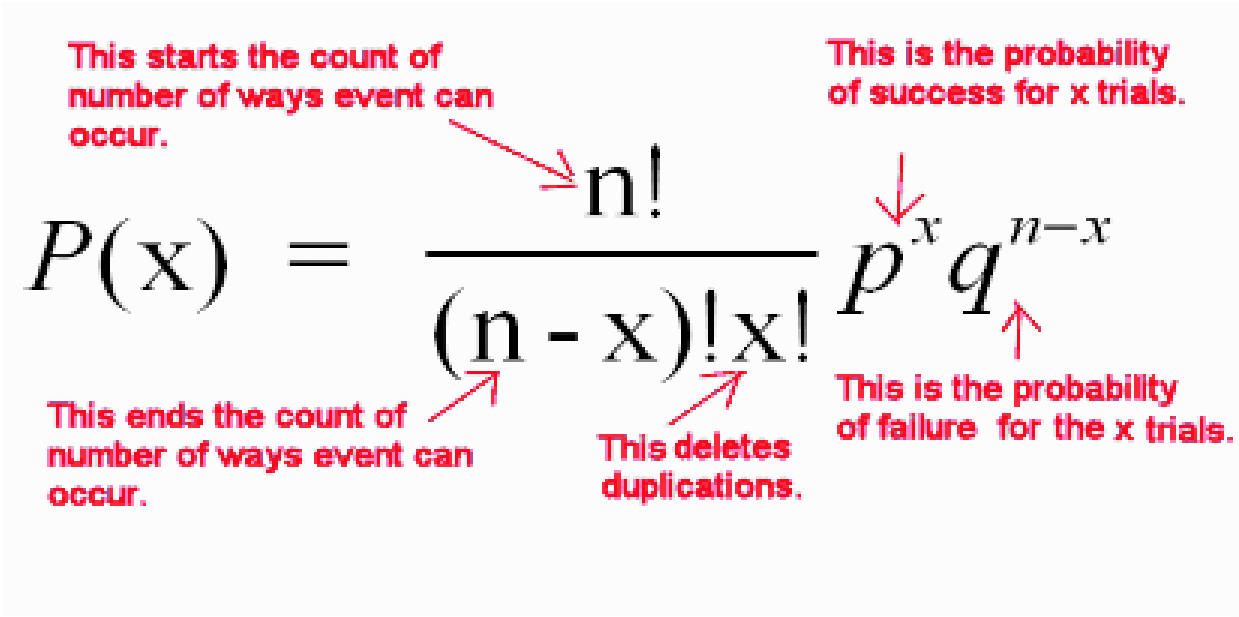
$$\begin{aligned} P(Y \geq 4) &= 1 - P(Y \leq 3) \\ &= 1 - \left[P(Y=1) + P(Y=2) + P(Y=3)\right] \\ &= 1 - \left[0.1 + 0.9 \times 0.1 + (0.9)^2 \times 0.1\right] \\ &= 1 - \left[0.1 + 0.09 + 0.081\right] \\ &= 0.729 \end{aligned}$$

Question 3 (b)

Let p denote the probability that Sam will be upgraded to first class on a particular flight. Let X denote the number of upgrades Sam will receive in 20 flights. The random variable X follows a binomial distribution with $n = 20$ independent trials and $p = 0.1$. The probability that Sam will be upgraded exactly 2 times in his next 20 flights is calculated as follows.

$$\begin{aligned} P(X=2) &= \binom{20}{2}(0.1^2)(0.9^{18}) \\ &\approx 0.2852 \end{aligned}$$

- Binomial distribution



$$P(X = c) = \text{binompdf}(n, p, c)$$

n -> number of trials

p -> probability of success

This finds the probability of exactly c successes, for some number c.

$$P(X \leq c) = \text{binomcdf}(n, p, c)$$

n -> number of trials

p -> probability of success

This finds the probability of c or fewer successes.

Question 3 (c)

Let X denote the number of upgrades Sam will receive in 104 flights. The random variable X follows a binomial distribution with $n = 104$ independent trials and $p = 0.1$. Thus,

$$\begin{aligned} P(X > 20) &= 1 - P(X \leq 20) \\ &\approx 1 - 0.9986 \\ &\approx 0.0014. \end{aligned}$$

Because this probability is so small, it is very unlikely that Sam would receive more than 20 upgrades in 104 flights if the airline's claim is correct. This would be expected to happen less than 1 percent of the time, indicating that one should be surprised if Sam receives more than 20 upgrades during the next year.

Question 4 (b)

- Conditions for a chi-square inference procedure

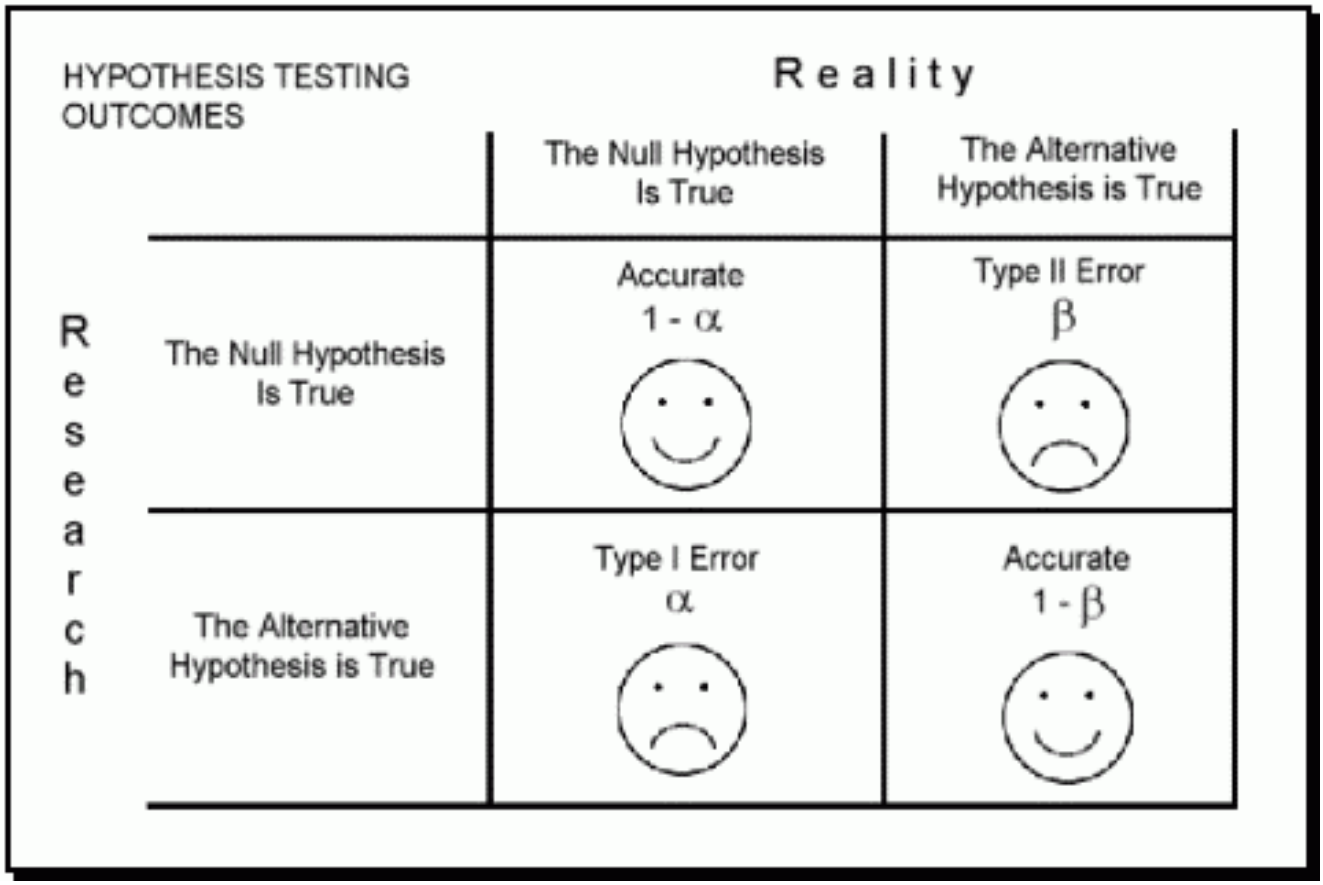
The following conditions for inference are met:

1. The students were randomly selected.
2. The expected cell counts should be at least 5. The computer output indicates that all expected counts are greater than 5. The smallest expected cell count is 6.825.

Question 4 (d)

Because the null hypothesis was rejected, a Type I error may have been made. A Type I error is concluding that there is an association between the perceived effect of part-time work on academic achievement and the average time spent on part-time jobs when, in reality, there is no association between the two variables.

- Type I error and Type II error



Question 5 (a)

- Conditions for one-proportion Z interval

1. Random sample
2. Large sample ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$)

- Calculation for one-proportion Z interval

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

$$\left(\frac{978}{2,350}\right) \pm 2.57583 \sqrt{\frac{0.41617(1-0.41617)}{2,350}}$$

$$0.41617 \pm 2.57583 \times 0.01017$$

$$0.41617 \pm 0.02619$$

$$(0.38998, 0.44236)$$

- Interpretation for one-proportion Z interval

Based on the sample, we are 99 percent confident that the proportion of the vaccine-eligible people in the United States who actually got vaccinated is between 0.39 and 0.44. Because 0.45 is not in the 99 percent confidence interval, it is not a plausible value for the population proportion of vaccine-eligible people who received the vaccine. In other words, the confidence interval is inconsistent with the belief that 45 percent of those eligible got vaccinated.

Question 5 (b)

The sample-size calculation uses 0.5 as the value of the proportion in order to provide the minimum required sample size to guarantee that the resulting interval will have a margin of error no larger than 0.02.

$$n \geq \frac{(2.576)^2(0.5)(0.5)}{(0.02)^2} = \left(\frac{2.576}{2(0.02)}\right)^2 = 4{,}147.36$$

Thus, a sample of at least 4,148 vaccine-eligible people should be taken in Canada.

- Partially correct (P) if supporting work is shown, *BUT* the response includes one or both of the following errors:
  1. 0.41617 (the sample proportion) or 0.45 is used instead of 0.5.
  2. An incorrect critical z-value is used — unless the same incorrect value was used in part (a).

- Calculating Required Sample Size to Estimate Population Mean

If prior estimate of population proportion exists:

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2 \qquad n = 0.76(1-0.76)\left(\frac{1.96}{0.030}\right)^2 = 778.564$$

If prior estimate of population proportion does not exist:

$$n = 0.25\left(\frac{z_{\alpha/2}}{E}\right)^2 \qquad n = 0.25\left(\frac{1.96}{0.030}\right)^2 = 1067.111$$

Sample Size Formulas

$$Margin\ of\ Error\ (ME) = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$ME = 1.96\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$n = \frac{p(1-p)z^2}{ME^2}$$

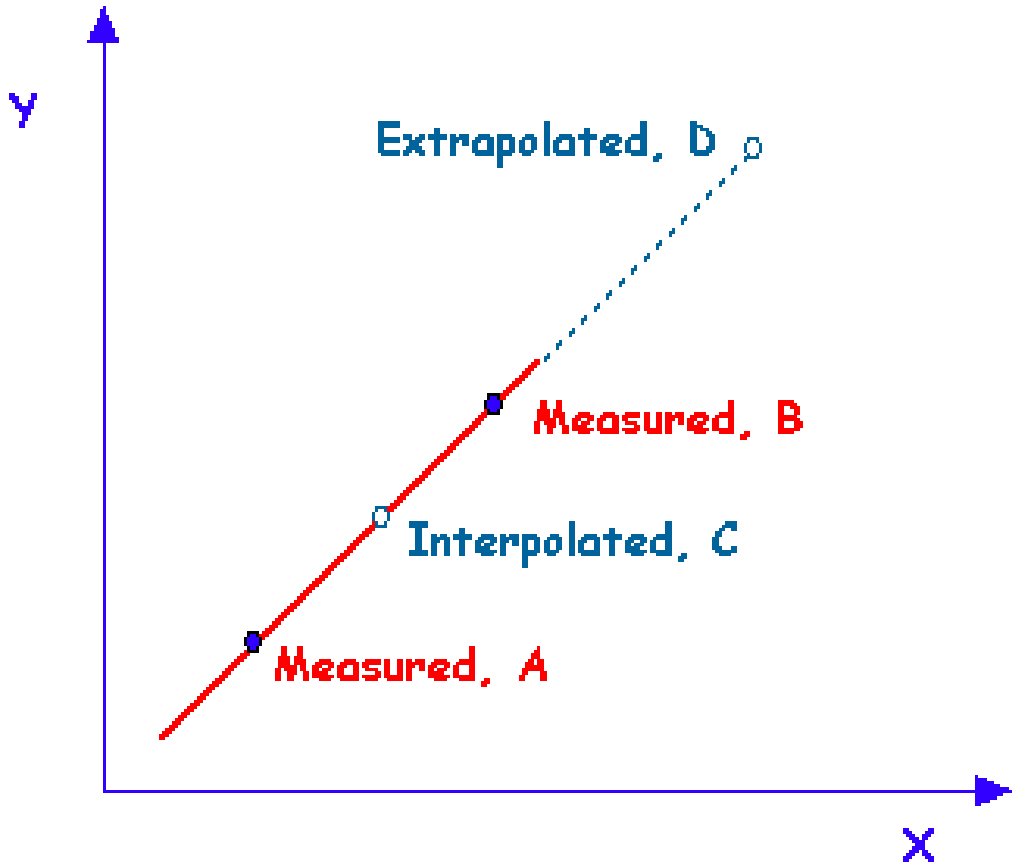
Formula based on $t-score$

$$ME = t\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$n = \left(\frac{\sigma t}{ME}\right)^2$$

Question 6 (b)

No. This is extrapolation beyond the range of data from the experiment. Buffer strips narrower than 5 feet or wider than 15 feet were not investigated.



Question 6 (c)

- Describe distribution: Mean & Standard deviation

Because the distribution of nitrogen removed for any particular buffer strip width is normally distributed with a standard deviation of 5 parts per hundred, the sampling distribution of the mean of four observations when the buffer strips are 6 feet wide will be normal with mean $33.8+3.6\times6=55.4$ parts per hundred and a standard deviation of $\frac{\sigma}{\sqrt{n}}=\frac{5}{\sqrt{4}}=2.5$ parts per hundred.

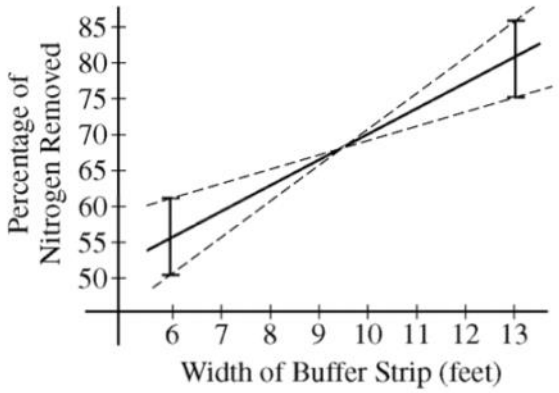
Question 6 (d)

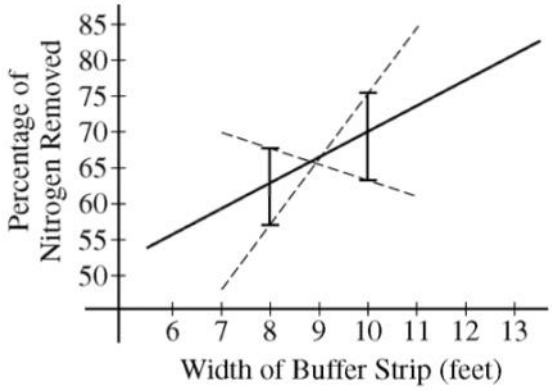
The distribution of the sample mean is normal, so the interval that has probability 0.95 of containing the mean nitrogen content removed from four buffer strips of width 6 feet extends from $55.4-1.96\times2.5=50.5$ parts per hundred to $55.4+1.96\times2.5=60.3$ parts per hundred.

- Confidence interval

	Distribution	Minitab Path	Formula
Mean (σ known)	Z-distribution	Stat > Basic Stat > 1-sample Z > Options	$\mu = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Mean (σ unknown)	T-distribution	Stat > Basic Statistics > 1-Sample t	$\mu = \bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$
Standard Deviation	Chi-squared (X^2) distribution	Stat > Basic Statistics > Display Descriptive Statistics	$s\sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}} \leq \sigma \leq s\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}$
Proportion (exact)	F-distribution	Stat > Basic Statistics > 1-Proportion	$P_{lower} = \frac{\nu_1 F_{\alpha/2,(\nu_1,\nu_2)}}{\nu_2 + \nu_1 F_{\alpha/2,(\nu_1,\nu_2)}}$ $P_{upper} = \frac{\nu_1 F_{1-\alpha/2,(\nu_1,\nu_2)}}{\nu_2 + \nu_1 F_{1-\alpha/2,(\nu_1,\nu_2)}}$
Proportion (estimate)	Z-distribution		$p = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Question 6 (e)





If we think that the sample mean nitrogen removed at a particular buffer width might reasonably be any value in the intervals shown, a sample regression line will result from connecting any point in the interval above 6 to any point in the interval above 13. With this in mind, the dashed lines in the plots above represent extreme cases for possible sample regression lines. From these plots, we can see that there is a wider range of possible slopes in the second plot (on the right) than in the first plot (on the left). Because of this, the variability in the sampling distribution of b, the estimator for the slope of the regression line, will be smaller for the first study plan (with four observations at 6 feet and four observations at 13 feet) than it would be for the second study plan (with four observations at 8 feet and four observations at 10 feet). Therefore, the first study plan (on the left) would provide a better estimator of the slope of the regression line than the second study plan (on the right).

Question 6 (f)

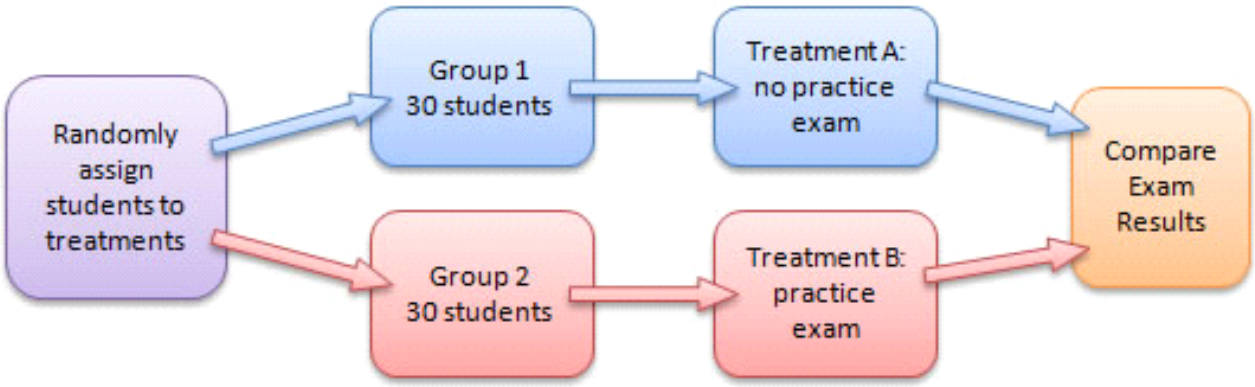
To assess the linear relationship between width of the buffer strip and the amount of nitrogen removed from runoff water, more widths should be used. To detect a nonlinear relationship it would be best to use buffer widths that were spaced out over the entire range of interest. For example, if the range of interest is 6 to 13 feet, eight buffers with widths 6, 7, 8, 9, 10, 11, 12 and 13 feet could be used.

2012 Multiple Choice

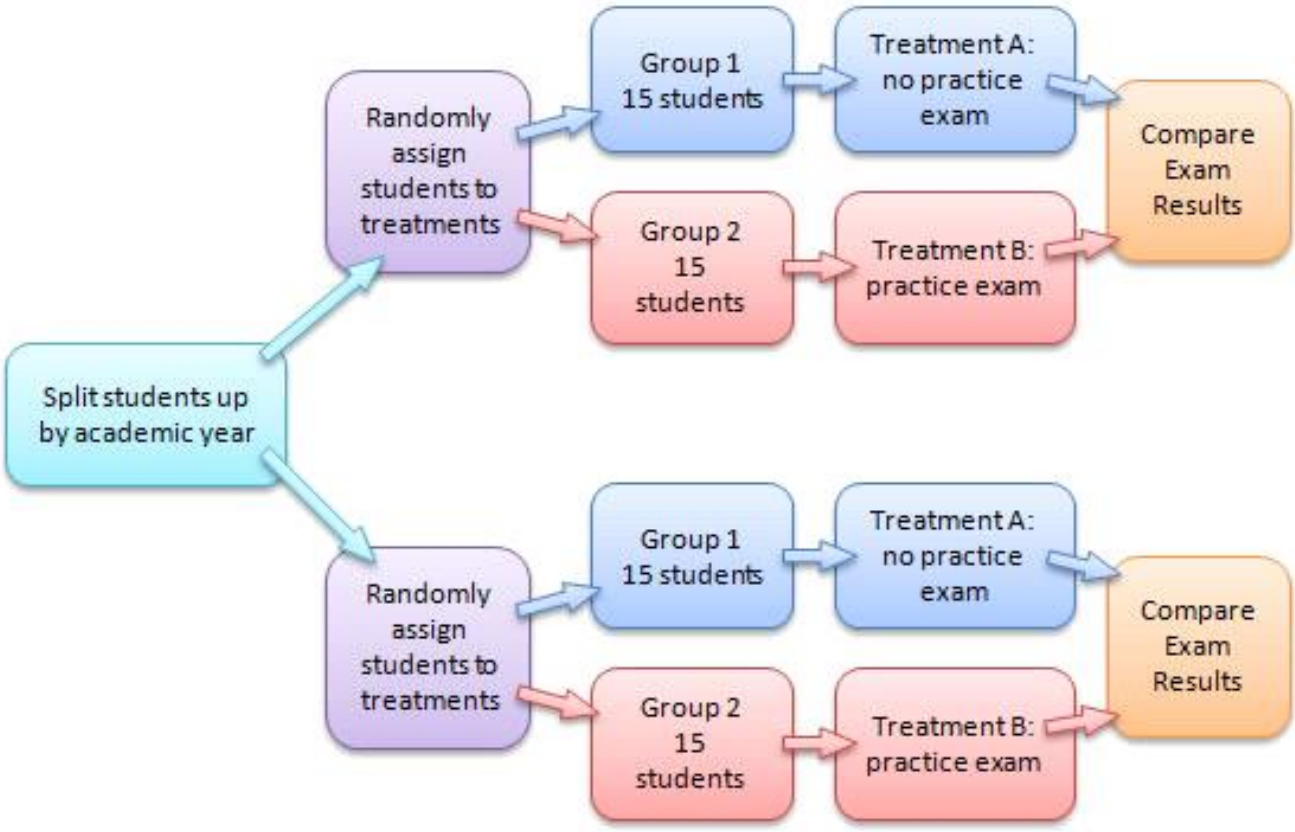
Monday, April 17, 2017 11:26 AM

Question 11

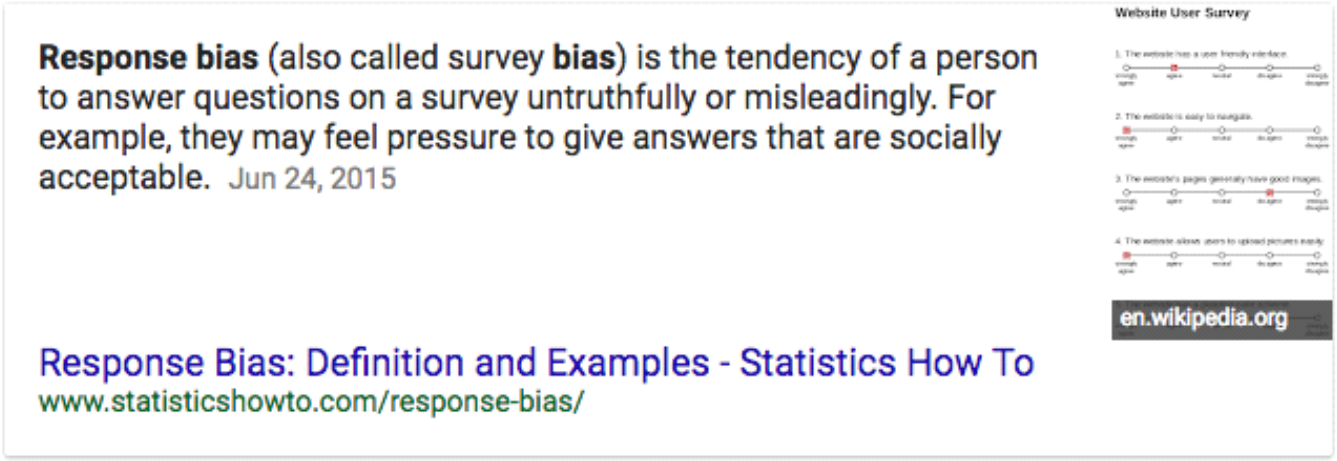
- Completely Randomized Design

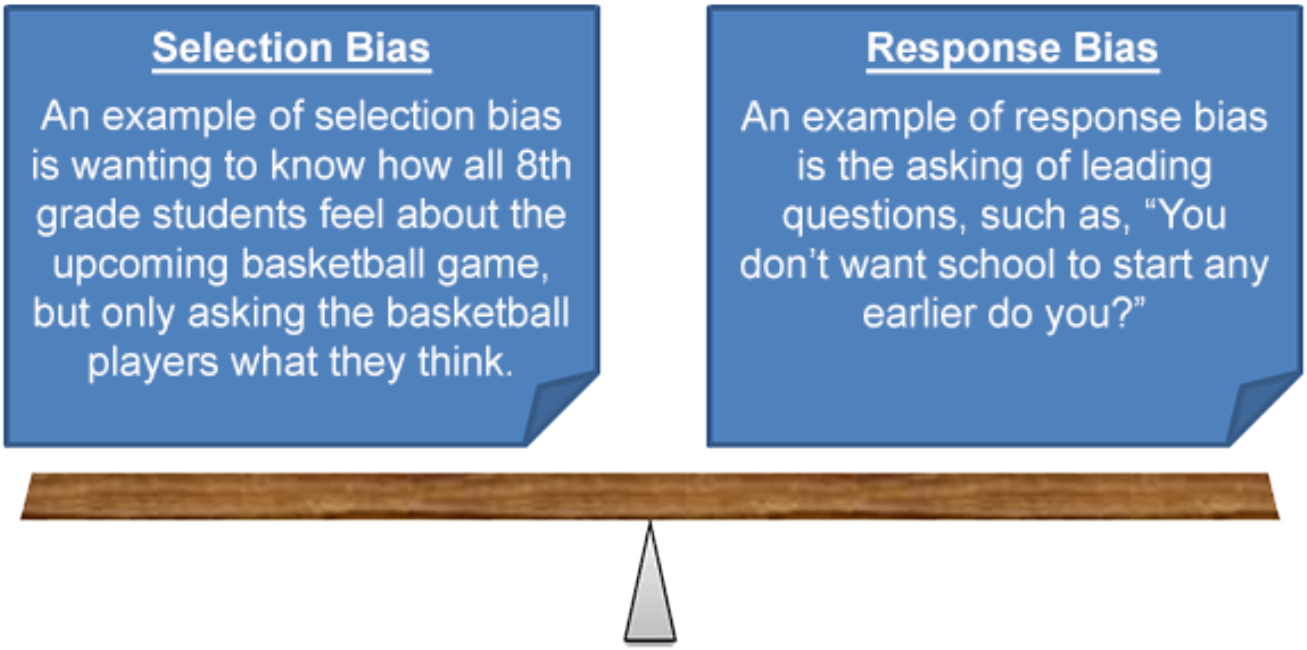


- Randomized Block Design



Question 12





- How to minimize response bias
  - Use Clear Language
  - Choose Words and Phrases With Care
  - Know How To Frame Your Questions
  - Provide Just the Right Amount of Options
  - Plan a Neutral Survey Structure
  - Keep Styling At a Minimum
  - Be Honest

Question 13

13. For a sample of 42 rabbits, the mean weight is 5 pounds and the standard deviation of weights is 3 pounds. Which of the following is most likely true about the weights for the rabbits in this sample?

(A) The distribution of weights is approximately normal because the sample size is 42, and therefore the central limit theorem applies.

(B) The distribution of weights is approximately normal because the standard deviation is less than the mean.

(C) The distribution of weights is skewed to the right because the least possible weight is within 2 standard deviations of the mean.

(D) The distribution of weights is skewed to the left because the least possible weight is within 2 standard deviations of the mean.

(E) The distribution of weights has a median that is greater than the mean.

Question 18

- Population size should be at least 10 times the sample size so that the degree of dependence among observations is negligible.

Question 22

- Confidence Interval Interpretation

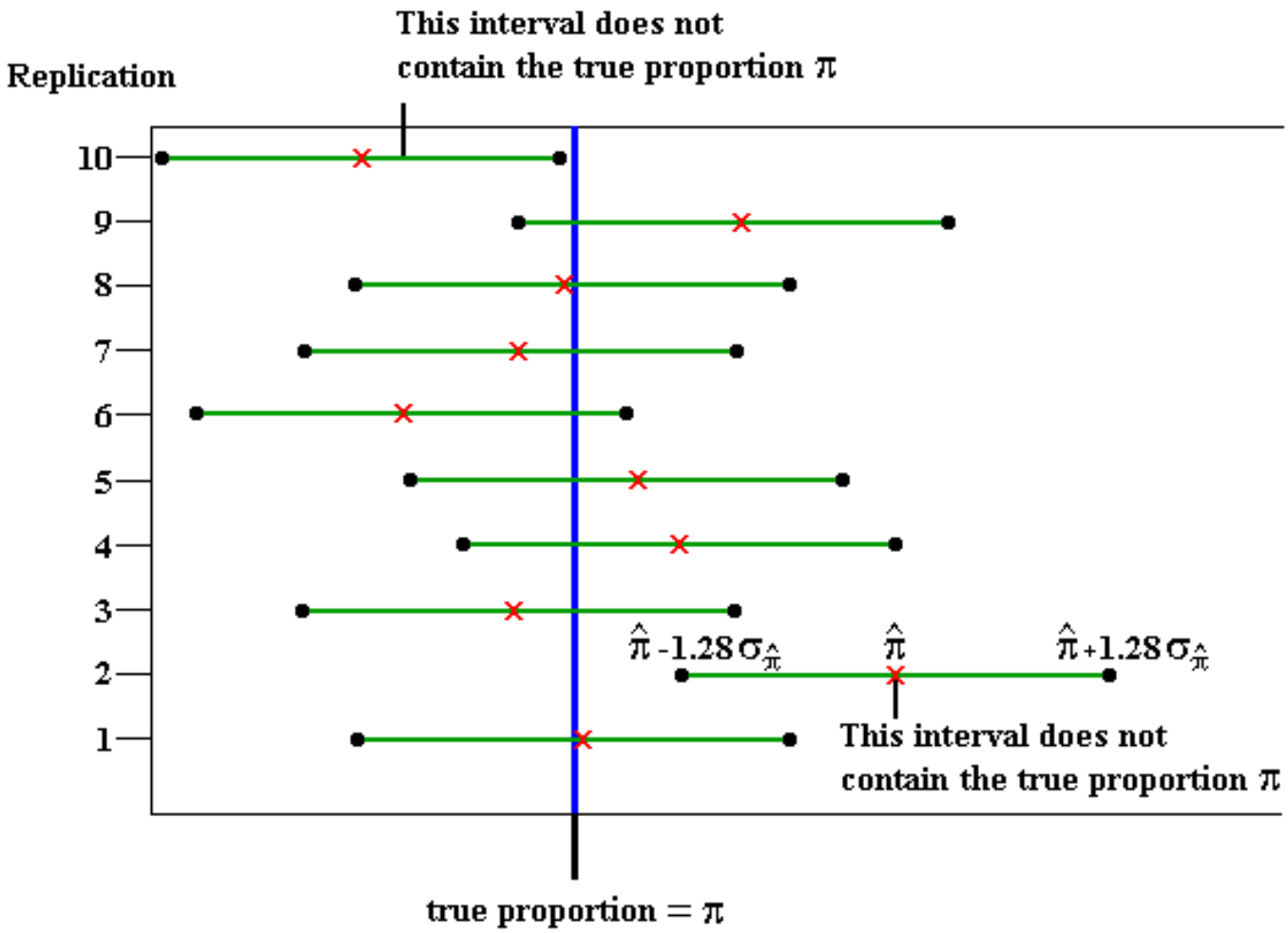
> If repeated samples were taken and the 95% **confidence interval** was computed for each sample, 95% of the **intervals** would contain the population mean. A 95% **confidence interval** has a 0.95 probability of containing the population mean. 95% of the population distribution is contained in the **confidence interval**.
>
> Confidence Intervals Introduction - Online Stats Book
> onlinestatbook.com/2/estimation/confidence.html

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure (presented in the next section) that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.



Question 28

28. An experimenter conducted a two-tailed hypothesis test on a set of data and obtained a *p*-value of 0.44. If the experimenter had conducted a one-tailed test on the same set of data, which of the following is true about the possible *p*-value(s) that the experimenter could have obtained?

(A) The only possible *p*-value is 0.22.

(B) The only possible *p*-value is 0.44.

(C) The only possible *p*-value is 0.88.

(D) The possible *p*-values are 0.22 and 0.78.

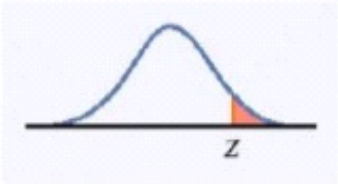
(E) The possible *p*-values are 0.22 and 0.88.

P-value in one-sided and two-sided tests

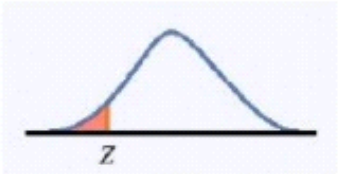
13

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

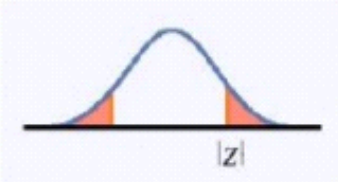


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.

- $p_{one-tailed} = \frac{1}{2} p_{two-tailed}$
- $p_{one-tailed} = 1 - \frac{1}{2} p_{two-tailed}$

Question 32

- SE Coef = Standard Deviation of **Statistic** (not population)

2012 Free Response

Friday, April 7, 2017 4:07 PM

Question 1 (a)

The data show a weak but positive association between price and quality rating for these sewing machines. The form of the association does not appear to be linear. Among machines that cost less than $500, there appears to be very little association between price and quality rating. But the machines that cost more than $500 do generally have better quality ratings than those that cost less than $500, which causes the overall association to be positive.

Question 1 (b)

The sewing machine that most affects the appropriateness of using a linear regression model is the one that costs about $2,200 and has a quality rating of about 65. Although the other four sewing machines costing more than $500 generally have higher quality ratings than those costing under $500, their prices and quality ratings follow a trend that suggests that quality ratings may not continue to increase with higher prices, but instead may approach a maximum possible quality rating. The $2,200 sewing machine is the most expensive of all but has a relatively low quality rating, which is consistent with a nonlinear model that approaches a maximum possible quality rating and then perhaps decreases. If a linear model were fit to all of the data, this one machine would substantially pull the regression line toward it, resulting in a poor overall fit of the line to the data.

Question 3 (a)

Household size tended to be larger in 1950 than in 2000. The histograms reveal a much larger proportion of small (1-, 2-, and 3-person) households in 2000 than in 1950. Similarly, the histograms reveal a much smaller proportion of large (5-person and larger) households in 2000 than in 1950. Also, the median household sizes can be calculated to be 5 people per household in 1950 compared with 3 or 4 people per household in 2000. The year 1950 displayed slightly more variability in household sizes than the year 2000. Although the interquartile ranges for both years are the same (3 people), the standard deviation (1950: about 2.6 people; 2000: about 2.1 people) and the range (1950: 13 people; 2000: 11 people) are larger for 1950 than for 2000. Both distributions of household size are skewed to the right. In both years, there are a few households with very large families, as large as 14 people in 1950 and 12 people in 2000.

- Summarizing Distribution (SOCS)
  - Shape
    - Skewed left/right
  - Outlier
    - Q1 - 1.5 * IQR
    - Q3 + 1.5 * IQR
  - Center
    - Mean or Median
  - Spread

        - SD or IQR

Question 3 (b)

The conditions for applying a two-sample *t*-procedure are:

1. The data come from independent random samples or from random assignment to two groups;
2. The populations are normally distributed, or both sample sizes are large;
3. The population sizes are at least 10 (or 20) times the sample sizes.

The first condition is satisfied because independent random samples were selected for the years 1950 and 2000. The second condition is satisfied because the sample sizes (500 in each group) are quite large, despite the right skewness of the distributions of household sizes in the sample data. The third condition is satisfied because the number of households in the large metropolitan area in both 1950 and 2000 would easily exceed $10 \times 500 = 5{,}000$.

- Conditions for Sampling Distribution (RIN)
    - Random
        - How the sample is selected
    - Independent
        - N≥10n
        - N: population size
        - n: sample size
    - Normal
        - For means
            - $n \geq 30$
            - If the population is normally distributed, n can < 30
        - For proportions:
            - $np \geq 10 \text{ AND } n(1\text{-}p) \geq 10$

Question 4

- Hypothesis Test
    - Using a Statistic to test a claim about a Parameter
    - Steps (**W**hy **C**an't **C**at **P**lay **I**nstruments)
        - **W**rite the hypothesis
            - Null hypothesis (H_0): Parameter = ___
            - Alternative hypothesis (H_1/H_a): Parameter > or < or ≠ _____

- **C**kеck conditions (RIN)
  - Random Sample
  - Independent: N >10n
  - Normal:
    - μ: n≥30
    - p: $\frac{p}{np} > 10$ and $\frac{p}{n(1-p)} > 10$
- **C**alculate the test statistic

$$\text{Test stat} = \frac{(\text{Point Estimate}) - (\text{Null Hypothesis})}{(\text{Standard Error})}$$

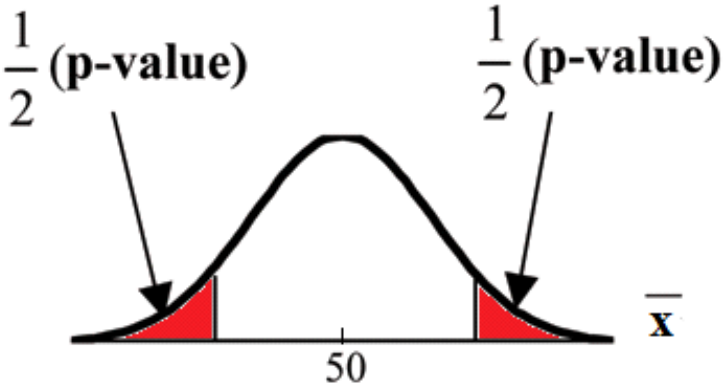
  - Mean

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

  - Proportion

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

- Look up the **P**-value (from Z table)
  - Probability that the null hypothesis (H_0) is true, given the sample data you collected



- **I**nterpret

$p < \alpha$	Reject the null hypothesis	do have evidence to support the claim
$p > \alpha$	Fail to reject the null hypothesis	do not have evidence to support the claim

- Step 1

States a correct pair of hypotheses.

Let p_{07} represent the population proportion of adults in the United States who would have answered "yes" about the effectiveness of television commercials in December 2007. Let p_{08} represent the analogous population proportion in December 2008.

The hypotheses to be tested are $H_0: p_{07} = p_{08}$ versus $H_a: p_{07} \neq p_{08}$

- Step 2

Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a two-sample z-test for comparing proportions.

Because these are sample surveys, the first condition is that the data were gathered from independent random samples from the two populations. This condition is met because we are told that the subjects were randomly selected in the two different years. Although we are not told whether the samples were selected independently, this is a reasonable assumption given that they are samples of different sizes selected in different years.

The second condition is that the sample sizes are large, relative to the proportions involved. This condition is satisfied because all sample counts (622 "yes" in 2007; $1{,}020 - 622 = 398$ "no" in 2007; 676 "yes" in 2008; $1{,}009 - 676 = 333$ "no" in 2008) are all at least 10 (or, are all at least 5).

An additional condition may be checked: The population sizes (more than 200 million adults in the United States) are much larger than 10 (or, 20) times the sample sizes.

- Step 3

Correct mechanics, including the value of the test statistic and p-value (or rejection region).

The sample proportions who answered "yes" are:

$$\hat{p}_{07} = \frac{622}{1{,}020} \approx 0.6098 \text{ and } \hat{p}_{08} = \frac{676}{1{,}009} \approx 0.6700.$$

The combined proportion, $\hat{p}_c$, who answered "yes" in these two years is:

$$\hat{p}_c = \frac{622 + 676}{1{,}020 + 1{,}009} = \frac{1{,}298}{2{,}029} \approx 0.6397.$$

The test statistic is:

$$z = \frac{\hat{p}_{07} - \hat{p}_{08}}{\sqrt{\hat{p}_c\left(1 - \hat{p}_c\right)\left(\frac{1}{n_{07}} + \frac{1}{n_{08}}\right)}} = \frac{0.6098 - 0.6700}{\sqrt{(0.6397)(1 - 0.6397)\left(\frac{1}{1{,}020} + \frac{1}{1{,}009}\right)}} \approx -2.82.$$

The p-value is $2P(Z \leq -2.82) \approx 0.0048$.

- Step 4

State a correct conclusion in the context of the study, using the result of the statistical test.

Because this p-value is smaller than any common significance level such as $\alpha = 0.05$ or $\alpha = 0.01$ (or, because this p-value is so small), we reject H_0 and conclude that the data provide convincing (or, statistically significant) evidence that the proportion of all adults in the United States who would answer "yes" to the question about the effectiveness of television commercials changed from December 2007 to December 2008.

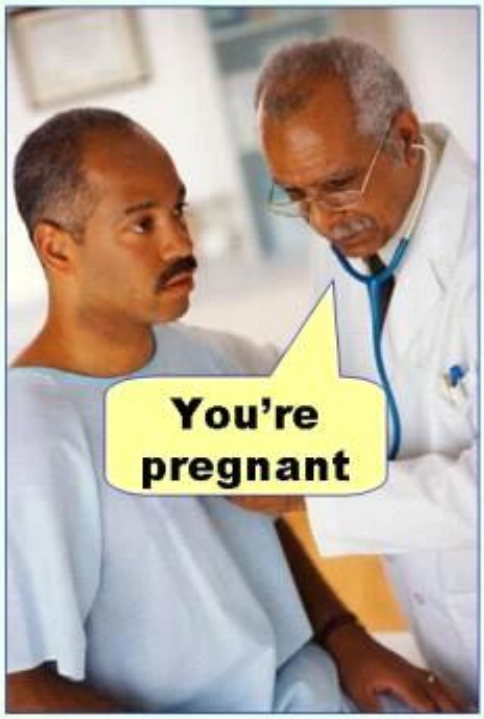
Question 5 (a)

In the context of the study, a Type II error means failing to reject the null hypothesis that 35 percent of adult residents in the city are able to pass the test when, in reality, less than 35 percent are able to pass the test. The consequence of this error is that the council would not fund the program, and the city would continue to have a smaller proportion of physically fit residents than the council would like.

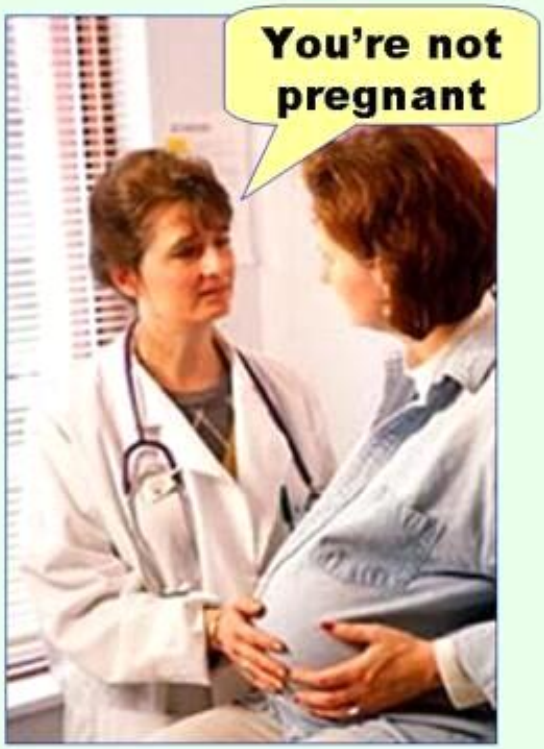
- Type I: falsely think alternative hypothesis is true (1 false), DO reject the null hypothesis (1 word)
- Type II: falsely think alternative hypothesis is false (2 falses), DO NOT reject the null hypothesis (2 word)

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
	The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

Type I error
(false positive)



Type II error
(false negative)



Question 5 (b)

Because the p-value of 0.97 is larger than $\alpha = 0.05$, we fail to reject the null hypothesis. There is not convincing evidence that the proportion of adult residents in the city who are able to pass the physical fitness test is less than 0.35. After all, the sample proportion of $\hat{p} = 0.416$ is actually higher than 0.35, which is in the opposite direction of the alternative hypothesis.

Essentially correct (E) if the response correctly completes the following three components:

1. Links the p-value to the conclusion by stating that the p-value is greater than $\alpha = 0.05$,
   OR
   by stating that the p-value is large.
   OR
   by correctly interpreting the p-value.
2. Uses context by referring to the proportion of adult residents who are able to pass the test,
   OR
   by referring to the funding of the program.
3. Makes a correct conclusion that describes the lack of evidence for the alternative hypothesis ($H_a : p < 0.35$).

Question 5 (c)

This is not a randomly selected sample because the sample was selected by recruiting volunteers. It seems reasonable to think that volunteers would be more physically fit than the population of city adults as a whole. Therefore, the sample proportion will likely overestimate the population proportion of adult residents in the city who are able to pass the physical fitness test.

Question 6 (a)

Peter can number the students from 1 to 2,000 and then use a calculator or computer to generate 100 unique random numbers between 1 and 2,000 without replacement. If non-unique numbers are generated, the repeated numbers are ignored until 100 unique numbers are obtained. The students whose numbers correspond to the randomly generated numbers are then selected for the sample.

Question 6 (c)

The variance of Rania's estimator is $(0.6)^2 \text{Var}(\bar{X}_f) + (0.4)^2 \text{Var}(\bar{X}_m)$, where $\text{Var}(\bar{X}_f) = \frac{\sigma_f^2}{n_f}$ represents the variance of the point estimator for females and $\text{Var}(\bar{X}_m) = \frac{\sigma_m^2}{n_m}$ represents the variance of the point estimator for males.

The estimated standard deviation is the square root of the variance. Using the respective sample standard deviations s_f and s_m for the population parameters, Rania's estimate is calculated as:

$$\sqrt{(0.6)^2 \frac{s_f^2}{n_f} + (0.4)^2 \frac{s_m^2}{n_m}} = \sqrt{(0.6)^2 \frac{(1.80)^2}{60} + (0.4)^2 \frac{(2.22)^2}{40}} \approx \sqrt{0.01944 + 0.01972} \approx 0.198.$$

Question 6 (d)

The comparative dotplots from Rania's data reveal that the distribution of the number of soft drinks for females appears to be quite different from that of males. In particular, the centers of the distributions appear to be significantly different. Additionally, the variability of values around the center *within* gender in each of Rania's dotplots appears to be considerably less than the variability displayed in the dotplot of Peter's data. Rania's estimator takes advantage of the decreased variability within gender because her data were obtained by sampling the two genders separately. Peter's estimator has more variability because his data were obtained from a simple random sample of all the high school students.

2013 Free Response

Tuesday, April 11, 2017 10:02 PM

Question 1 (b)

Step 1: Identifies the appropriate confidence interval (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a one-sample t-interval for a population mean.

Conditions: 1. The sample is randomly selected from the population.
2. The population has a normal distribution, or the sample size is large.

The first condition is met because we were told that the crows were randomly selected. The sample size of 23 is not considered large, so we need to examine the sample data to assess whether it is reasonable to assume that the population distribution of lead levels for all crows in this region is normal. The stem-and-leaf plot shows no strong skewness or outliers, so we will consider the second condition to be met.

Step 2: Correct mechanics

A 95% confidence interval for the population mean μ is given by: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$. The critical value for 95% confidence, based on $23 - 1 =$ 22 degrees of freedom, is $t^* = 2.074$. The 95% confidence interval for μ is therefore

$$4.90 \pm 2.074 \times \frac{1.12}{\sqrt{23}} \approx 4.90 \pm 0.484,$$

which is the interval $(4.416, 5.384)$ ppm.

Using the raw data rather than the given summary statistics, the 95% confidence interval for μ is (4.411, 5.3803).

Step 3: Interpretation

We can be 95% confident that the population mean lead level among all crows in this region is between 4.416 and 5.384 parts per million.

Question 2 (c)

Stratifying by campus would be more advantageous than stratifying by gender provided that opinions about appearance of university buildings and grounds between the two campuses differ more than the opinions about appearance of university buildings and grounds between the two genders.

Question 4

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate test is a chi-square test of independence.

The conditions for this test were satisfied because:
1. The question states that the sample was randomly selected.
2. The expected counts for all six cells of the table were all at least 5, as seen in the following table that lists expected counts in parentheses beside the observed counts:

	Five or more servings of fruit and vegetables	Four or fewer servings of fruit and vegetables	Total
18–34 years	231 (240.2)	741 (731.8)	972
35–54 years	669 (719.4)	2242 (2191.6)	2911
55+ years	1291 (1231.4)	3692 (3751.6)	4983
Total	2191	6675	8866

Statistic/Parameter	**Condition/Assumption**	**How do we check?**
More two Sample Proportions (Test for Homogeneity)	1. Count Condition: The data are **counts**. 2. Independent Condition: Data in groups are **independent**. 3. **Large sample**	1. Verify this. 2. SRS and 10n < N 3. Count > 5

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the p-value is very small (for instance, much smaller than $\alpha = 0.05$), we would reject the null hypothesis at the 0.05 level and conclude that the sample data provide strong evidence that there is an association between age group and consumption of fruits and vegetables for adults in the United States. In particular, older (55+ years of age) people were more likely to eat five or more servings of fruits and vegetables, and middle-aged people (35–54 years of age) were less likely to eat five or more servings of fruits and vegetables.

Question 5 (a)

No, it would not be reasonable to conclude that meditation causes a reduction in blood pressure for men in the retirement community. Because this is an observational study and not an experiment, no cause-and-effect relationship between meditation and lower blood pressure can be inferred. It is quite possible that men who choose to meditate could differ from men who do not choose to meditate in other ways that were also associated with blood pressure.

Question 5 (b)

The sample sizes were too small, relative to the overall sample proportion of successes, to justify using a normal approximation. One way to check this is to note that the combined sample proportion of successes is $\hat{p} = \frac{0+8}{11+17} = \frac{8}{28} \approx 0.286$, so neither $n_m\hat{p} = 11 \times \frac{8}{28} \approx 3.143$ nor $n_c\hat{p} = 17 \times \frac{8}{28} \approx 4.857$ is at least 10.

Statistic/Parameter	Condition/Assumption	How do we check?
Two Sample Proportions (Independent)	1. **Randomization** Condition: Samples in each group are random samples (SRS) or representatives of their populations or in experiments the treatments are randomly assigned. 2. Normality Condition: **n_1p_1 and $n_2p_2 \geq 10$ and n_1q_1 and $n_2q_2 \geq 10$**. 3. Independent Condition: The selection of each subject is independent of each other (**10n < N**) for each sample. In some experiments this is not necessary. 4. Independence of Groups Condition: The groups are **independent** of each other.	1. Based on the information provided. 2. Show that the inequalities are true. 3. Show that the inequality is true. 4. Based on the information provided.

Question 5 (c)

The observed value of the sample statistic $\hat{p}_m - \hat{p}_c$ is $\frac{0}{11} - \frac{8}{17} \approx -0.47$. The graph of simulation results reveals that a difference of -0.47 or more extreme was very rare. In fact, the value -0.47 was the smallest possible outcome and occurred in only 76 of the 10,000 repetitions in the simulation. Thus, assuming that all men in the retirement community were equally likely to have high blood pressure whether they meditate or not, there is an approximate probability of 0.0076 of getting a difference of -0.47 or smaller by chance alone. Because this approximate *p*-value is very small, there is convincing evidence that men in this retirement community who meditate were less likely to have high blood pressure than men in this retirement community who do not meditate. However, because this is an observational study, even though we can conclude that meditation is *associated* with a lower chance of having high blood pressure, we cannot conclude that meditation *causes* a reduction in the likeliness of having high blood pressure.

Question 6 (a)

The Western Pacific Ocean had more typhoons than the Eastern Pacific Ocean in all but one of these years. The average seems to have been about 31 typhoons per year in the Western Pacific Ocean, which is higher than the average of about 19 typhoons per year in the Eastern Pacific Ocean. The Western Pacific Ocean also saw more variability (in number of typhoons per year) than the Eastern Pacific Ocean; for example, the range of the frequencies for the Western Pacific is about 21 typhoons and only 10 typhoons for the Eastern Pacific.

2014 Multiple Choice

Wednesday, February 22, 2017 2:09 PM

Question 16

- Notice the condition

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	Z	$n\hat{p},\ n(1-\hat{p}) \geq 10$
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of $n_1 - 1$ and $n_2 - 1$	$n_1, n_2 < 30$; and/or σ_1, σ_2 unknown
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d} - \mu_d)}{s_d/\sqrt{n}}$	t_{n-1}	$n < 30$ pairs of data and/or σ_d unknown
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Z	$n\hat{p},\ n(1-\hat{p}) \geq 10$ for each group

Question 18

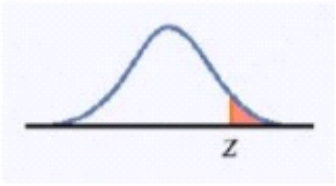
- P-value for two-sided t-test = P-value for one-sided t-test * 2

P-value in one-sided and two-sided tests

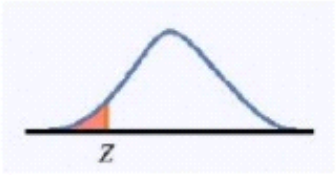
13

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

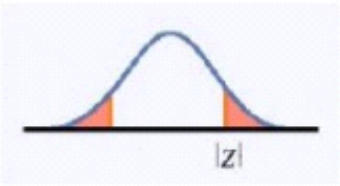


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the P-value for a two-sided test, use the symmetry of the normal curve. Find the P-value for a one-sided test and double it.

Question 21

A good sample is **representative**. This means that each sample point represents the attributes of a known number of population elements.

Bias often occurs when the survey sample does not accurately represent the population. The bias that results from an unrepresentative sample is called **selection bias**. Some common examples of selection bias are described below.

- **Undercoverage**. Undercoverage occurs when some members of the population are inadequately represented in the sample. A classic example of undercoverage is the *Literary Digest* voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from undercoverage of low-income voters, who tended to be Democrats.

  How did this happen? The survey relied on a convenience sample, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent. Undercoverage is often a problem with convenience samples.

- **Nonresponse bias**. Sometimes, individuals chosen for the sample are unwilling or unable to participate in the survey. Nonresponse bias is the bias that results when respondents differ in meaningful ways from nonrespondents. The *Literary Digest* survey illustrates this problem. Respondents tended to be Landon supporters; and nonrespondents, Roosevelt supporters. Since only 25% of the sampled voters actually completed the mail-in survey, survey results overestimated voter support for Alfred Landon.

  The *Literary Digest* experience illustrates a common problem with mail surveys. Response rate is often low, making mail surveys vulnerable to nonresponse bias.

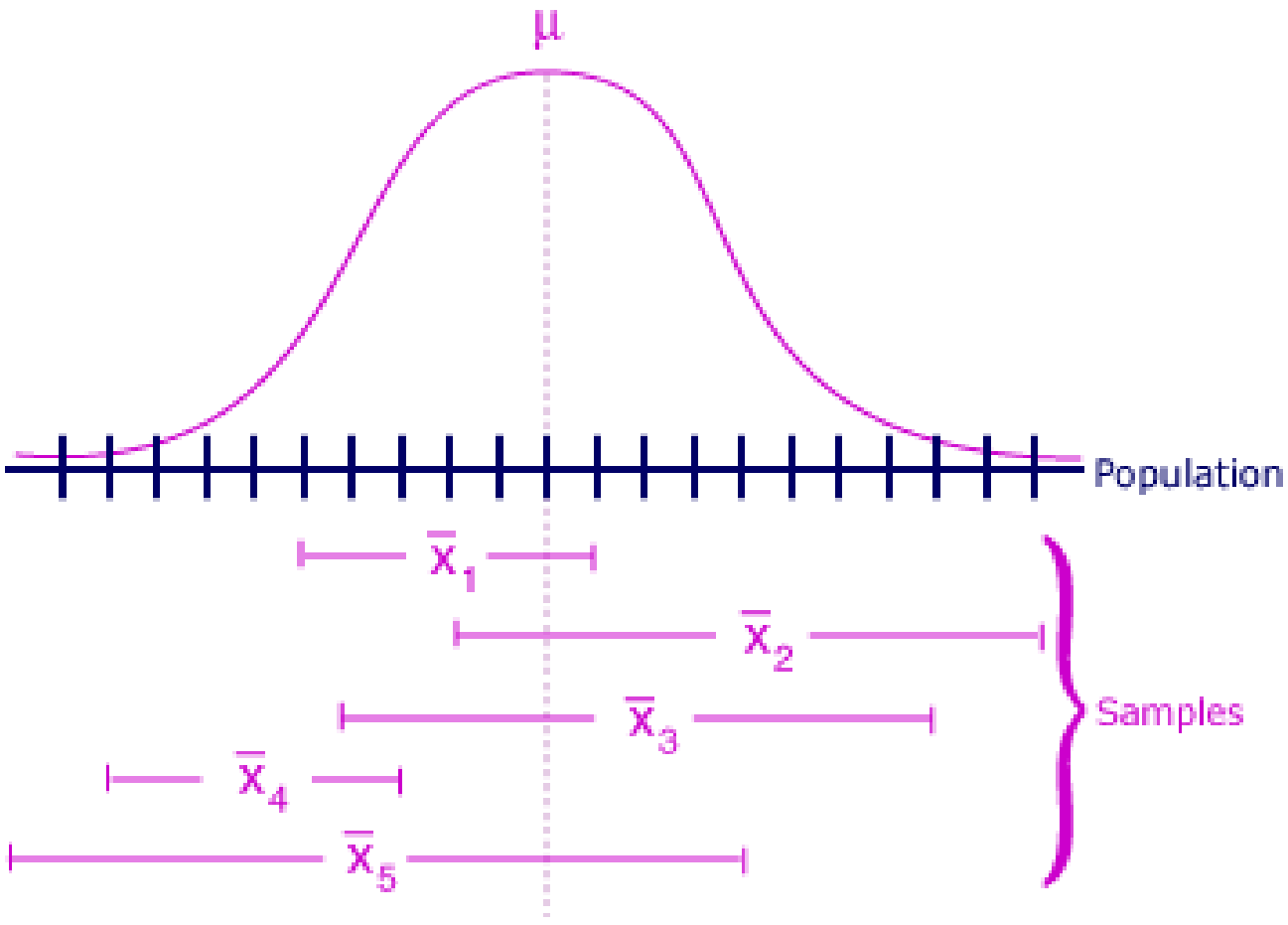
- **Voluntary response bias**. Voluntary response bias occurs when sample members are self-selected volunteers, as in voluntary samples. An example would be call-in radio shows that solicit audience participation in surveys on controversial topics (abortion, affirmative action, gun control, etc.). The resulting sample tends to overrepresent individuals who have strong opinions.

Response bias refers to the bias that results from problems in the measurement process. Some examples of response bias are given below.

- **Leading questions**. The wording of the question may be loaded in some way to unduly favor one response over another. For example, a satisfaction survey may ask the respondent to indicate where she is satisfied, dissatisfied, or very dissatified. By giving the respondent one response option to express satisfaction and two response options to express dissatisfaction, this survey question is biased toward getting a dissatisfied response.

- **Social desirability**. Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes or illegal activities in a survey, particularly if survey results are not confidential. Instead, their responses may be biased toward what they believe is socially desirable.

Question 25



Question 31

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

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

Question 32

- H0: There is no association between A and B / Variables are independent
- H1: There is an association between A and B / Variables are dependent
- P > α
    - We fail to reject the null hypothesis and do not have evidence to support the claim that there is an association between A and B
- P < α
    - We reject the null hypothesis and do have evidence to support the claim that there is an association between A and B

Question 37

The probability of making a type I error is called the **significance level** and is denoted by the Greek letter, α (alpha). The significance level should be chosen before data is collected. The probability of making a type II error is denoted by the Greek letter, β (beta). The quantity 1–β is called the **power of the test** and represents the probability of rejecting the null hypothesis when it is actually false, in other words, making the correct decision by rejecting the null hypothesis. The power of the test is the probability on not making a type II error.

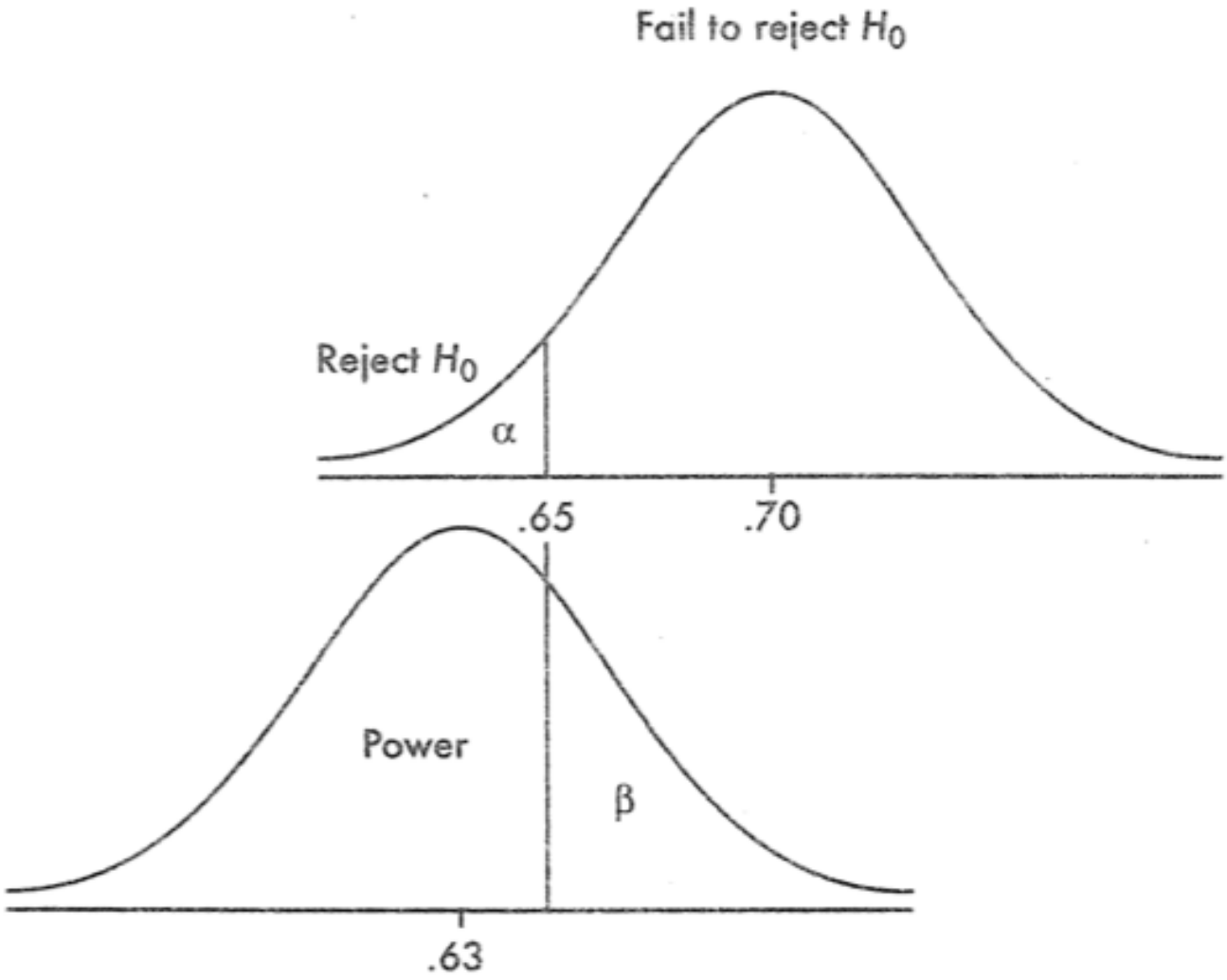
Any value of α may be chosen, although statisticians usually choose small values. Common values for α include

α = 0.01 – willingness to make a type I error 1% of the time
α = 0.05 – willingness to make a type I error 5% of the time
α = 0.10 – willingness to make a type I error 10% of the time

The person performing the test chooses the significance level α. How much power you get from a test depends on the significance level. The value of β is very difficult or impossible to calculate. The calculation of the value of β is beyond the scope of the AP Statistics Test, but you need to know what it means.

The power of a test (the probability of rejecting the null hypothesis when it is false) increases as the significance level also increases. A test performed at α = 0.10 has more power than a test performed at α = 0.05. Increasing the significance level, α, to increase the power can be counterproductive, since it increases the risk of making a type I error. Decreasing α also decreases the probability of rejecting the null hypothesis. This reduces the power of the test, which increases the probability of making a type II error, β. Choosing α properly requires finding a balance appropriate for the situation.

Decision	***H_0 Is Actually True***	***H_0 Is Actually False***
Fail to reject H_0	Correct decision	Type II error, β
Reject H_0	Type I error, α	Correct decision, 1 – β



The following points should be emphasized:

- Power gives the probability of avoiding a Type II error.
- Power has a different value for different possible correct values of the population parameter; thus it is actually a *function* where the independent variable ranges over specific alternative hypotheses.
- Choosing a smaller α (that is, a tougher standard to reject H_0) results in a higher risk of Type II error and a lower power—observe in the above graphs how making α smaller (in this case moving the critical cutoff value to the left) makes the power less and β more!
- The greater the difference between the null hypothesis p_0 and the true value p, the smaller the risk of a Type II error and the greater the power—observe in the above picture how moving the lower graph to the left makes the power greater and β less. (The difference between p_0 and p is sometimes called the *effect*—thus the greater the effect, the greater is the power to pick it up.)
- A larger sample size n will reduce the standard deviations, making both graphs

the effect, the greater is the power to pick it up.)

- A larger sample size n will reduce the standard deviations, making both graphs narrower resulting in smaller α, smaller β, and larger power!

2014 Free Response

Wednesday, April 12, 2017 9:58 AM

Question 2 (a)

The probability that all 3 people selected are women can be calculated using the multiplication rule, as follows:

$P(\text{all three selected are women})$

$= P(\text{first is a woman}) \times P(\text{second is a woman} | \text{first is a woman}) \times P(\text{third is a woman} | \text{first two are women})$

$= \frac{3}{9} \times \frac{2}{8} \times \frac{1}{7} \approx 0.012$

Question 2 (c)

No, the process does not correctly simulate the random selection of three women from a group of nine people of whom six are men and three are women. The random selection of three people among nine is done *without* replacement. However, in the simulation with the dice, the three dice rolls in any given trial are independent of one another, indicating a selection process that is done *with* replacement.

Question 3 (c)

For any one typical school week, the probability is $\frac{2}{5} = 0.4$ that the day selected is not Tuesday, not Wednesday, or not Thursday. Therefore, because the days are selected independently across the three weeks, the probability that none of the three days selected would be a Tuesday or Wednesday or Thursday is $(0.4)^3 = 0.064$.

Question 4 (a)

The median is less affected by skewness and outliers than the mean. With a variable such as income, a small number of very large incomes could dramatically increase the mean but not the median. Therefore, the median would provide a better estimate of a typical income value.

Question 5

Step 1: States a correct pair of hypotheses.

Let μ_{diff} represent the population mean difference in purchase price (woman – man) for identically equipped cars of the same model, sold to both men and women by the same dealer, in the county.

The hypotheses to be tested are $H_0 : \mu_{diff} = 0$ versus $H_a : \mu_{diff} > 0$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a paired *t*-test.

The conditions for the paired *t*-test are:

1. The sample is randomly selected from the population.
2. The population of price differences (woman – man) is normally distributed, or the sample size is large.

The first condition is met because the car models and the individuals were randomly selected. The sample size ($n = 8$) is not large, so we need to investigate whether it is reasonable to assume that the population of price differences is normally distributed. The dotplot of sample price differences reveals a fairly symmetric distribution, so we will consider the second condition to be met.

Step 3: Correct mechanics, including the value of the test statistic and *p*-value (or rejection region).

The test statistic is $t = \dfrac{585 - 0}{\dfrac{530.71}{\sqrt{8}}} \approx 3.12$.

The *p*-value, based on a *t*-distribution with $8 - 1 = 7$ degrees of freedom, is 0.008.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the *p*-value is very small (for instance, smaller than $\alpha = 0.05$), we reject the null hypothesis. The data provide convincing evidence that, on average, women pay more than men in the county for the same car model.

- Assumption in 2 sample independence T-test
  - **Normality**: Assumes that the population distributions are normal. The t-test is quite robust over moderate violations of this assumption. It is especially robust if a two tailed test is used and if the sample sizes are not especially small. Check for normality by creating a histogram.
  - **Independent Observations**: The observations within each treatment condition must be independent.
  - **Equal Variances**: Assume that the population distributions have the same variance. This assumption is quite important (If it is violated, it makes the test' s averaging of the 2 variances meaningless).
  - If it is violated, then use a modification of the t-test procedures as needed.
- Paired Sample T test
  - The matched-pair t-test (or paired t-test or paired samples t-test or dependent t-test) is used when the data from the two groups can be **presented in pairs**, For example where the same people are being measured in **before-and-after** comparison or when the group is given **two different tests at different times** (e.g. pleasantness of two different types of chocolate).

- Assumptions in paired sample t-test
  - The first assumption in the paired sample t-test is that only the **matched pair** can be used to perform the paired sample t-test.
  - In the paired sample t-test, **normal distributions** are assumed.
  - Variance in paired sample t-test: in a paired sample t- test, it is assumed that the **variance** of two sample is **same**.
  - The data is measurement data-interval/ratio
  - **Independence** of observation in paired sample t-test: in a paired sample t-test, observations must be independent of each other.
- Paired t-test vs two-sample t-test

A Paired t-test Is Just A 1-Sample t-Test

Many people are confused about when to use a paired t-test and how it works. I'll let you in on a little secret. The paired t-test and the 1-sample t-test are actually the same test in disguise! As we saw above, a 1-sample t-test compares one sample mean to a null hypothesis value. A paired t-test simply calculates the difference between paired observations (e.g., before and after) and then performs a 1-sample t-test on the differences.

How Two-Sample T-tests Calculate T-Values

The 2-sample t-test takes your sample data from two groups and boils it down to the t-value. The process is very similar to the 1-sample t-test, and you can still use the analogy of the signal-to-noise ratio. Unlike the paired t-test, the 2-sample t-test requires independent groups for each sample.

- Paired t- test

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

Question 6 (b)

(ii) Point B corresponds to a car with an actual FCR that is very close to the FCR that would be predicted for a car with its length by the regression model which predicts FCR using the explanatory variable length.

Question 6 (c)

Graph II reveals a moderate association that is positive and linear. In contrast, there is a weak association that is positive and linear in graph III. The association between engine size and residual (from predicting FCR based on length) is stronger than the association between wheel base and residual (from predicting FCR based on length).

Question 6 (d)

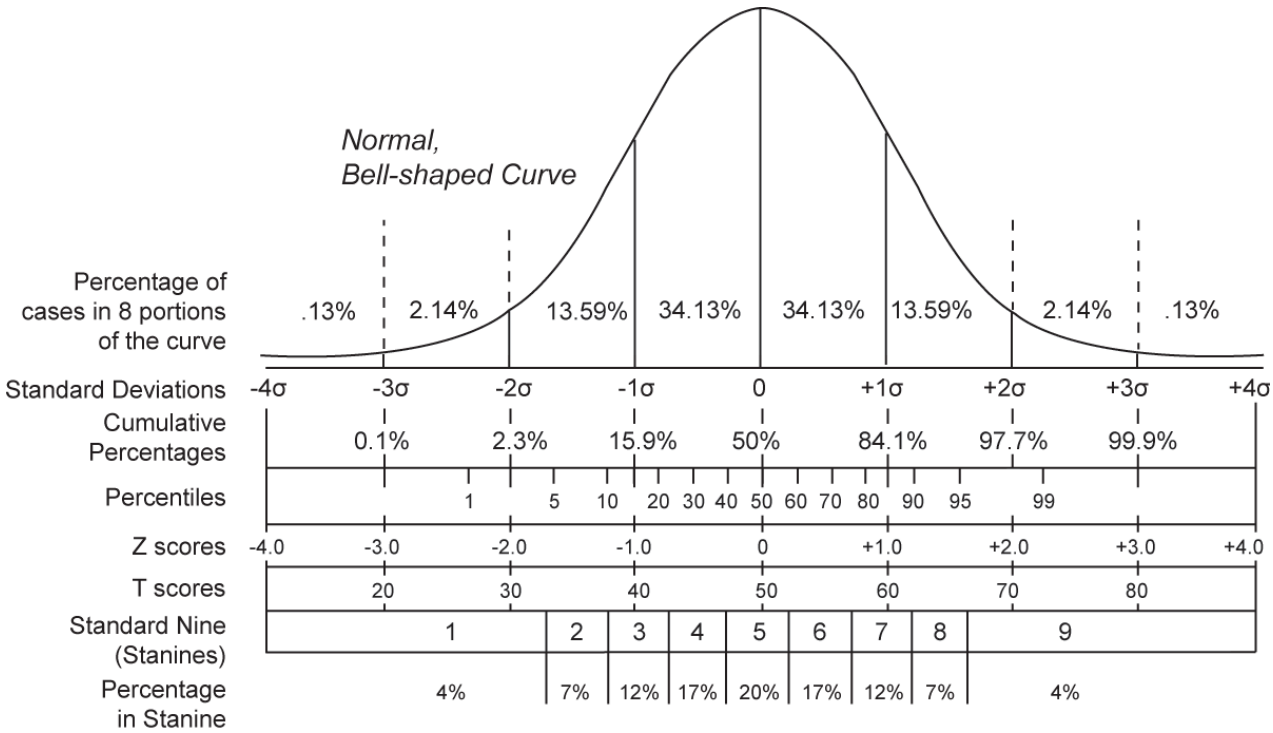
Engine size is a better choice than wheel base for including with length in a regression model for predicting FCR. The stronger association between engine size and residual (from predicting FCR based on length) indicates that engine size is more useful than wheel base for reducing the variability in FCR values that remains unexplained (as indicated by residuals) after predicting FCR based on length.

2015 Multiple Choice

Thursday, February 16, 2017 11:07 AM

Question 3

- Z-score will not be changed after conversion



Question 13

- Standard deviation of the sample distribution

MEAN AND STANDARD DEVIATION OF $\bar{x}$ cont.

Standard Deviation of the Sampling Distribution of $\bar{x}$

The ***standard deviation of the sampling distribution of*** $\bar{x}$ is

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

where σ is the standard deviation of the population and n is the sample size. This formula is used when $n/N \leq .05$, where N is the population size.

23

- Standard deviation of the sample proportion

Mean and Standard Deviation of $\hat{p}$ cont.

Standard Deviation of the Sample Proportion

The ***standard deviation of the sample proportion***, $\hat{p}$, is denoted by $\sigma_{\hat{p}}$ and is given by the formula

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Where p is the population proportion, q = 1 –p , and n is the sample size. This formula is used when $n/N \leq .05$, where N is the population size.

91

Question 20

22

Z-Test for testing difference between means

Test Condition

- Populations are normal
- Samples happen to be large,
- Population variances are known
- Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_{p1}^2}{n_1} + \frac{\sigma_{p2}^2}{n_2}}}$$

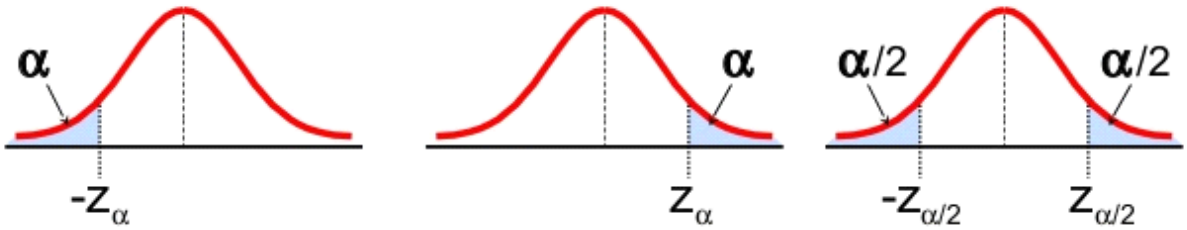
Question 20

Hypothesis Tests for Two Population Proportions

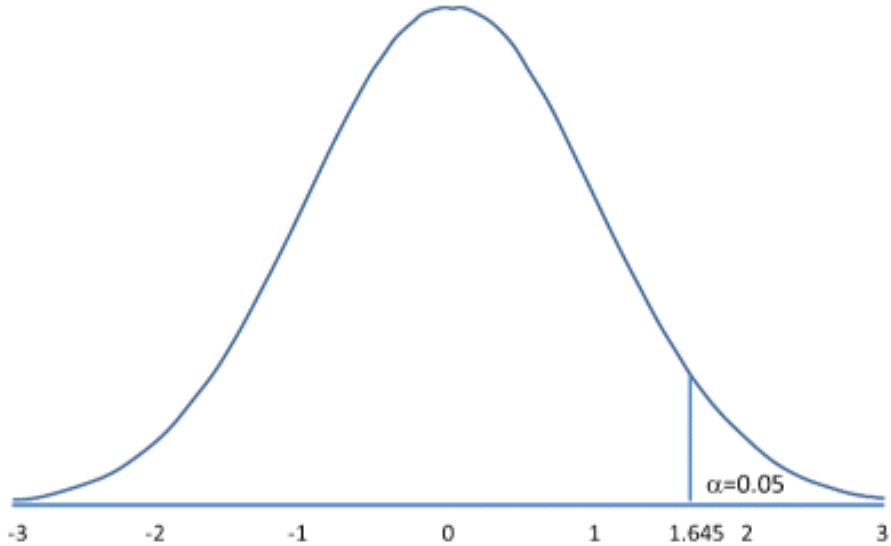
(continued)

Population proportions

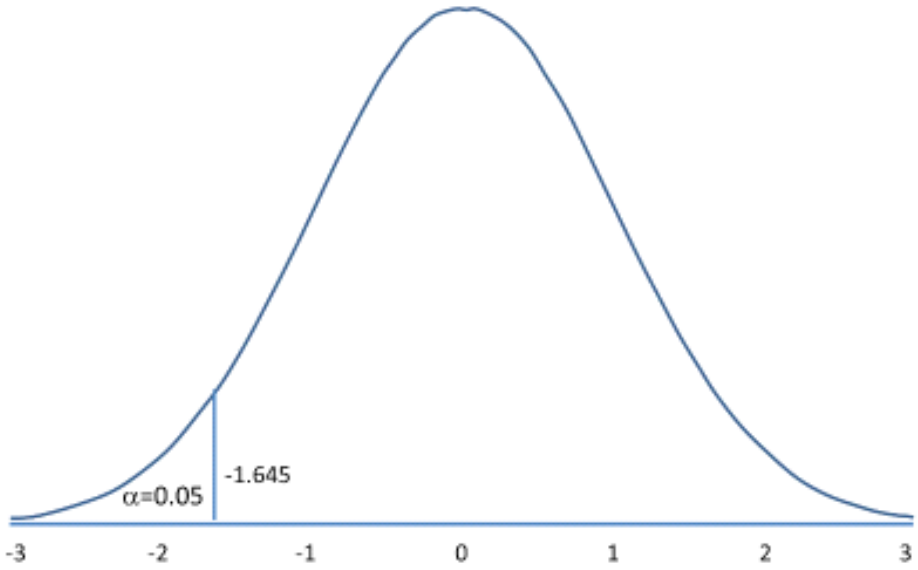
Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: p_1 - p_2 \geq 0$	$H_0: p_1 - p_2 \leq 0$	$H_0: p_1 - p_2 = 0$
$H_1: p_1 - p_2 < 0$	$H_1: p_1 - p_2 > 0$	$H_1: p_1 - p_2 \neq 0$



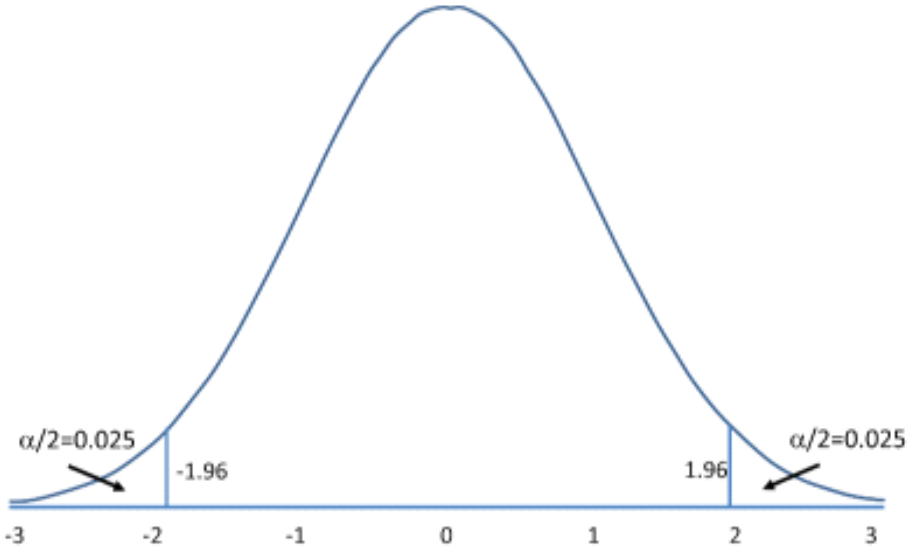
Reject H_0 if $Z < -Z_\alpha$	Reject H_0 if $Z > Z_\alpha$	Reject H_0 if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$



Rejection Region for Upper-Tailed Z Test ($H_1: \mu > \mu_0$) with α=0.05



Rejection Region for Lower-Tailed Z Test (H_1: $\mu < \mu_0$) with α =0.05



Rejection Region for Two-Tailed Z Test (H_1: $\mu \neq \mu_0$) with α =0.05

Question 23

In statistical hypothesis testing, a **type** I **error** is the incorrect rejection of a true null hypothesis (a "false positive"), while a **type II error** is incorrectly retaining a false null hypothesis (a "false negative").

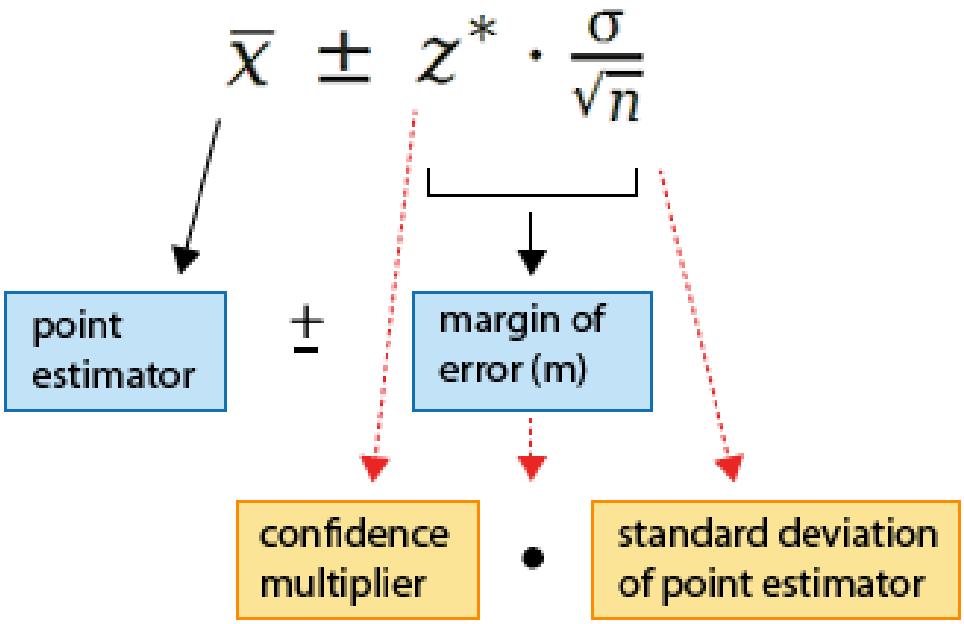
Type I and type II errors - Wikipedia
https://en.wikipedia.org/wiki/**Type**_I_and_**type**_II_**error**s

Question 24

- Standard error of the difference

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2+S_2^2}{N}}$$

Question 28



Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

Question 32

- $\sigma^2 = \sigma_1^2 \times n_1 + \sigma_2^2 \times n_2 + \sigma_3^2 \times n_3 + \square$

Question 37

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

2015 Free Response

2017年4月24日 星期一 下午2:32

Question 1 (a)

- Summarizing Distributions
  - Center: Median / Mean
  - Shape: Skew left / right
  - Spread: SD / IQR
  - Outlier

Question 1 (b)

(i) Five years after starting, at least 3 out of 30 (10%) of the salaries at Corporation A are greater than the maximum salary at Corporation B. If I accept the offer from Corporation A, I might be able to make a higher salary at Corporation A than at Corporation B.

(ii) Five years after starting, the minimum salary at Corporation B is greater than at Corporation A. In fact, at Corporation A it looks like some people are still making the starting salary of $36,000 and never received a raise in the five years since they were hired. So if I work at Corporation A, I might never receive a raise in salary.

Question 4

Step 1: States a correct pair of hypotheses.

Let p_{asp} represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a low-dose aspirin each day. Similarly, let p_{plac} represent the population proportion of adults similar to those in the study who would have developed colon cancer within the six years of the study if they had taken a placebo each day.

The hypotheses to be tested are $H_0 : p_{asp} = p_{plac}$ versus $H_a : p_{asp} < p_{plac}$ or equivalently, $H_0 : p_{asp} - p_{plac} = 0$ versus $H_a : p_{asp} - p_{plac} < 0$.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate procedure is a two-sample z-test for comparing proportions.

Because this is a randomized experiment, the first condition is that the volunteers were randomly assigned to one treatment group or the other. The condition is satisfied because we are told that the volunteers were randomly assigned to take a low-dose aspirin or a placebo.

The second condition is that the sample sizes are large, relative to the proportions involved. The condition is satisfied because all sample counts are large enough; that is, 15 with colon cancer in aspirin group, 26 with colon cancer in placebo group, $500 - 15 = 485$ cancer-free in aspirin group, and $500 - 26 = 474$ cancer-free in placebo group.

Step 3: Calculates the appropriate test statistic and p-value.

The sample proportions who developed colon cancer are $\hat{p}_{asp} = \frac{15}{500} = 0.030$ and $\hat{p}_{plac} = \frac{26}{500} = 0.052$.

The combined sample proportion who developed colon cancer is $\hat{p}_{combined} = \frac{15 + 26}{500 + 500} = 0.041$.

The test statistic is $z = \frac{0.030 - 0.052}{\sqrt{0.041(1 - 0.041)\left(\frac{1}{500} + \frac{1}{500}\right)}} \approx -1.75$ (-1.7542 from calculator).

The p-value is $P(Z \leq -1.75) = 0.0401$ (0.0397 from calculator), where Z has a standard normal distribution.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

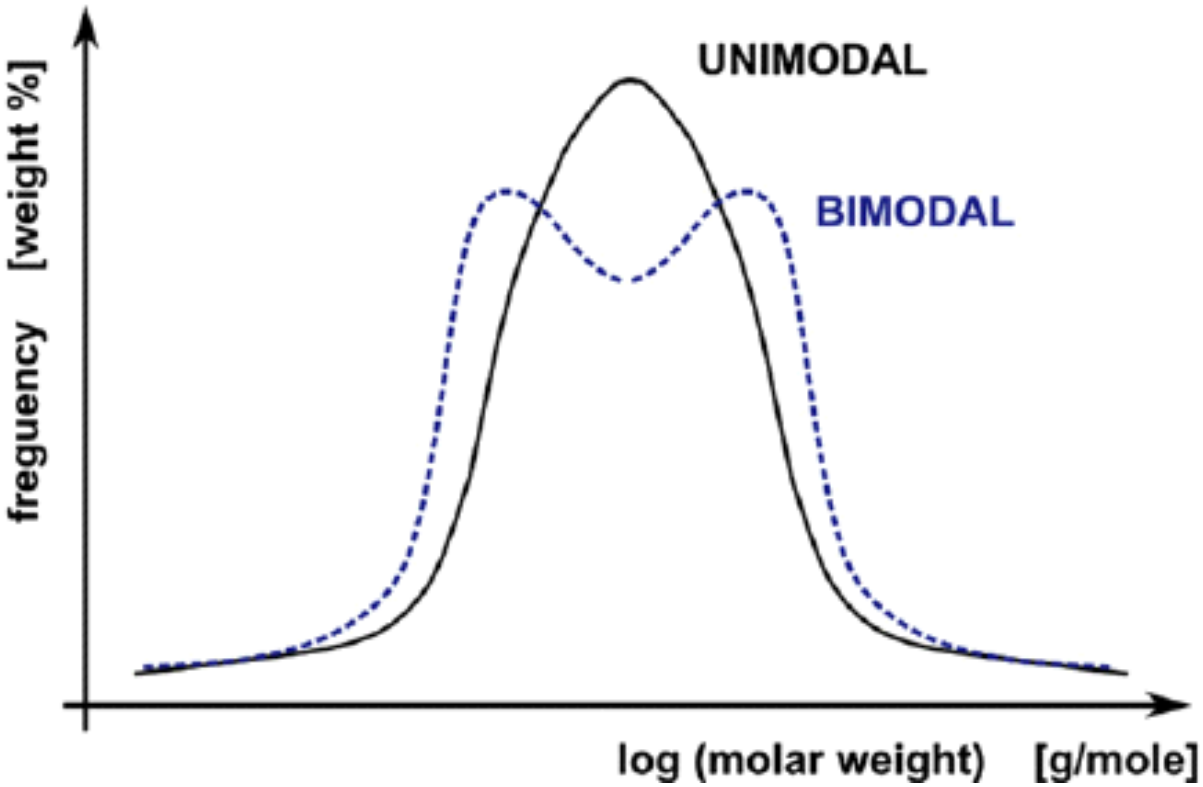
Because the p-value is less than the given significance level of $\alpha = 0.05$, we reject the null hypothesis. The data provide convincing statistical evidence that the proportion of all adults similar to the volunteers who would develop colon cancer if they had taken a low-dose aspirin every day is less than the proportion of all adults similar to the volunteers who would develop colon cancer if they had not taken a low-dose aspirin every day.

Question 5 (a)

- Response should be given in context

There is a moderately strong, positive, linear relationship between height and arm span so that taller students tend to have longer arm spans.

Question 6 (b)



Question 6 (c)

Method 2 would result in less variability in the sample of 200 tortillas on a given day because the sample comes from only one production line. Because the distributions of diameters are not the same for the two production lines, selecting tortillas from both lines as in Method 1 would result in more variable sample data.

2016 Free Response

2017年4月27日 星期四 上午12:32

Question 1 (a)

The distribution of Robin's tip amounts is skewed to the right. There is a gap between the largest tip amount (in the \$20 to \$22.50 interval) and the second largest tip amount (in the \$12.50 to \$15 interval), and the largest tip amount appears to be an outlier. The median tip amount is between \$2.50 and \$5.00. Robin's tip amounts vary from a minimum of between \$0 and \$2.50 to a maximum of between \$20.00 and \$22.50. About 78 percent of the tip amounts are between \$0 and \$5.

Essentially correct (E) if the response includes reasonable comments on the following five components:

1. Shape (skewed right)
2. Outlier (at least one) *OR* gap (one tip amount greater than \$20, next highest at most \$15)
3. Center between \$2.50 and \$5.00 (median) or between \$2.62 and \$5.13 (mean)
4. Variability, by noting that the tip amounts vary from about \$0 to at most \$22.50, or that a majority of tip amounts are between \$0 and a value greater than or equal to \$5, or by providing a correct numerical approximation of a measure of variability
5. Context (tip amounts)

Question 2 (a)

Step 1: States a correct pair of hypotheses.

H_0 : The proportion of children who would choose each snack is the same regardless of which type of ad is viewed.

H_a : The proportion of children who would choose each snack differs based on which type of ad is viewed.

Step 2: Identifies a correct test procedure (by name or formula) and checks appropriate conditions.

The appropriate procedure is a chi-square test of homogeneity.

The conditions for this test are satisfied because (1) the question states that the children were randomly assigned to groups, and (2) expected counts for the six cells of the table are all at least 5, as seen in the following table that lists expected counts beside observed counts.

Group	Choco-Zuties	Apple-Zuties	Total
A	21 (18.67)	4 (6.33)	25
B	13 (18.67)	12 (6.33)	25
C	22 (18.67)	3 (6.33)	25
Total	56	19	75

Step 3: Calculates the appropriate test statistic and p-value.

The test statistic is calculated as $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is

$$\chi^2 \approx 0.292 + 0.860 + 1.720 + 5.070 + 0.595 + 1.754 \approx 10.291.$$

The p-value is $P(\chi^2_{df2} \geq 10.291) \approx 0.006$.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the p-value is very small (for instance, much smaller than $\alpha = 0.05$), we reject the null hypothesis at the 0.05 level (and at the 0.01 level). The data provide convincing statistical evidence that the proportions who would choose each snack differ based on which ad is viewed.

Question 3 (a)

The explanatory variable is the person's degree of cigarette smoking. The response variable is whether the person develops Alzheimer's disease during the course of the study.

Essentially correct (E) if both variables are described correctly. A correct description includes some degree of status of the variables, such as smoking versus not smoking and developing Alzeimer's versus not developing Alzheimer's.

Question 3 (b)

This is an observational study because the people in the study were not assigned to a certain degree of cigarette smoking. Rather, the degree of cigarette smoking for each person was passively observed and recorded, not manipulated by the researchers.

Question 5 (a)

- Denote the procedure used

The appropriate procedure is a one-sample z-interval for a population proportion. The problem stated the conditions for inference have been met, so they do not need to be checked. A 95 percent confidence interval for the population proportion is given as $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is

$$0.37 \pm 1.96\sqrt{\frac{(0.37)(0.63)}{1{,}048}} \approx 0.37 \pm 0.03 = (0.34, 0.40).$$

We are 95 percent confident that the population proportion of all adults in the U.S. who would have chosen the economy statement is between 0.34 and 0.40.

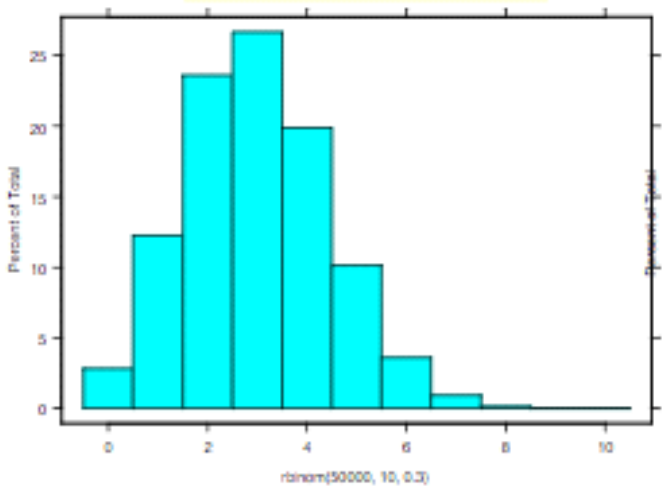
Question 5 (b)

One of the conditions for inference that was met is that the number who chose the economy statement and the number who did not choose the economy statement are both greater than 10. Explain why it is necessary to satisfy that condition.

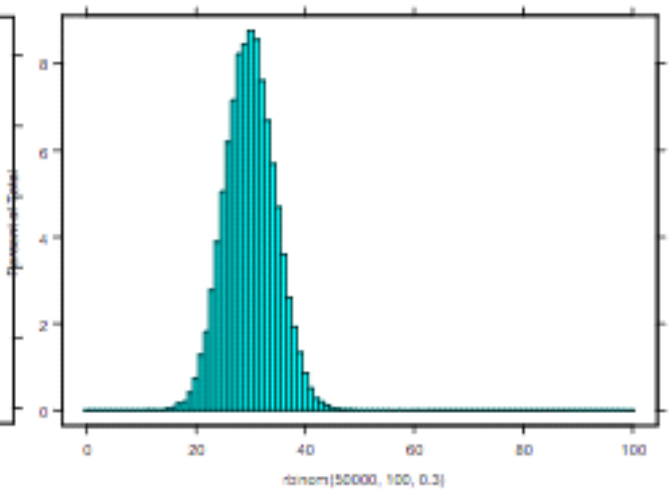
The condition is necessary because the formula for the confidence interval relies on the fact that the binomial distribution can be approximated by a normal distribution which then results in the sampling distribution of $\hat{p}$ being approximately normal. The approximation does not work well unless both $n\hat{p}$ and $n(1-\hat{p})$ are at least 10.

Normal approximation of binomial distribution

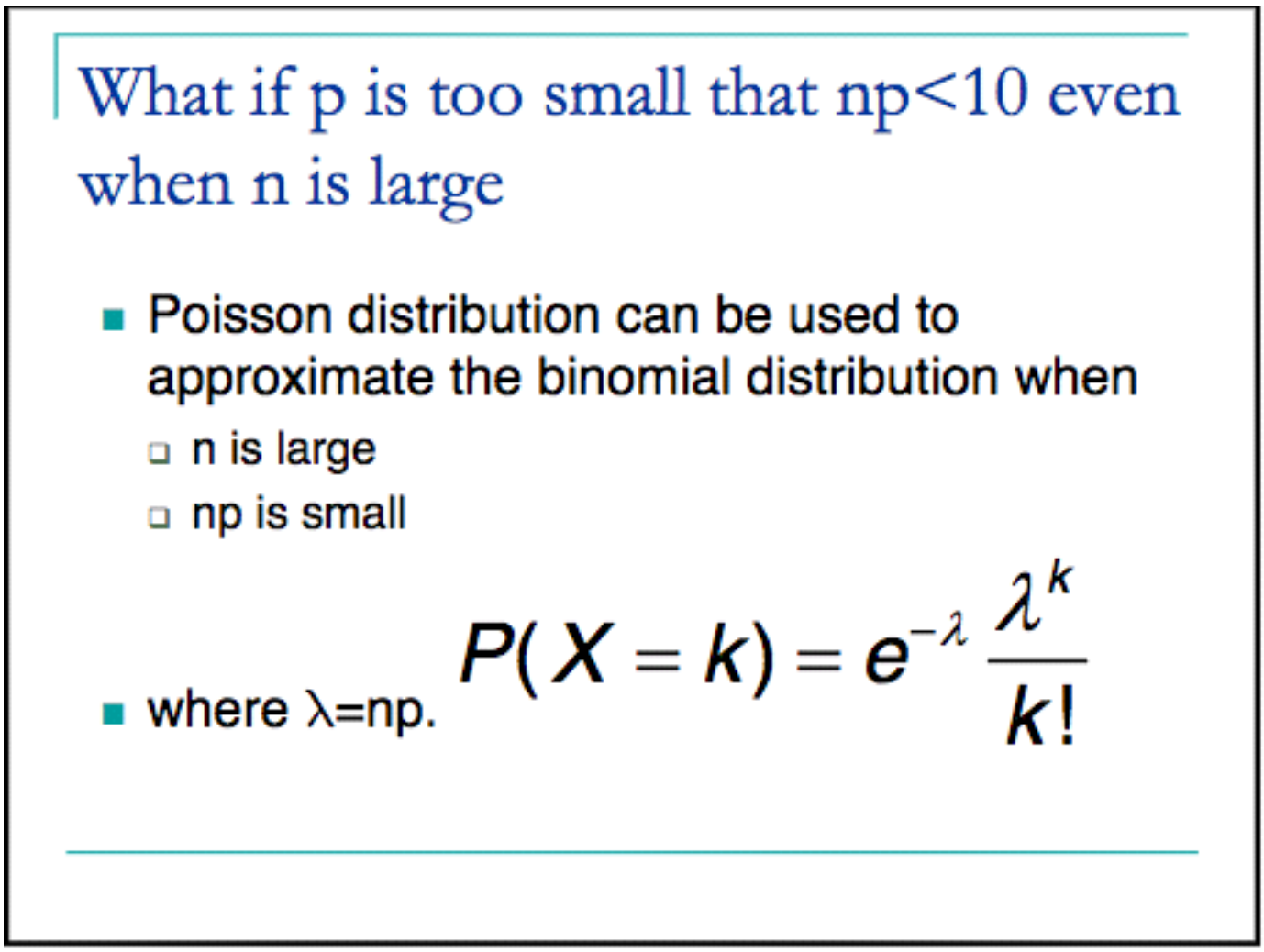
Binom(10, 0.3)



Binom(100, 0.3)



When n is large, so that (np>10) and (nq>10), the binomial distribution Binom(n, p) can be approximated by $N(np, \sqrt{np(1-p)})$



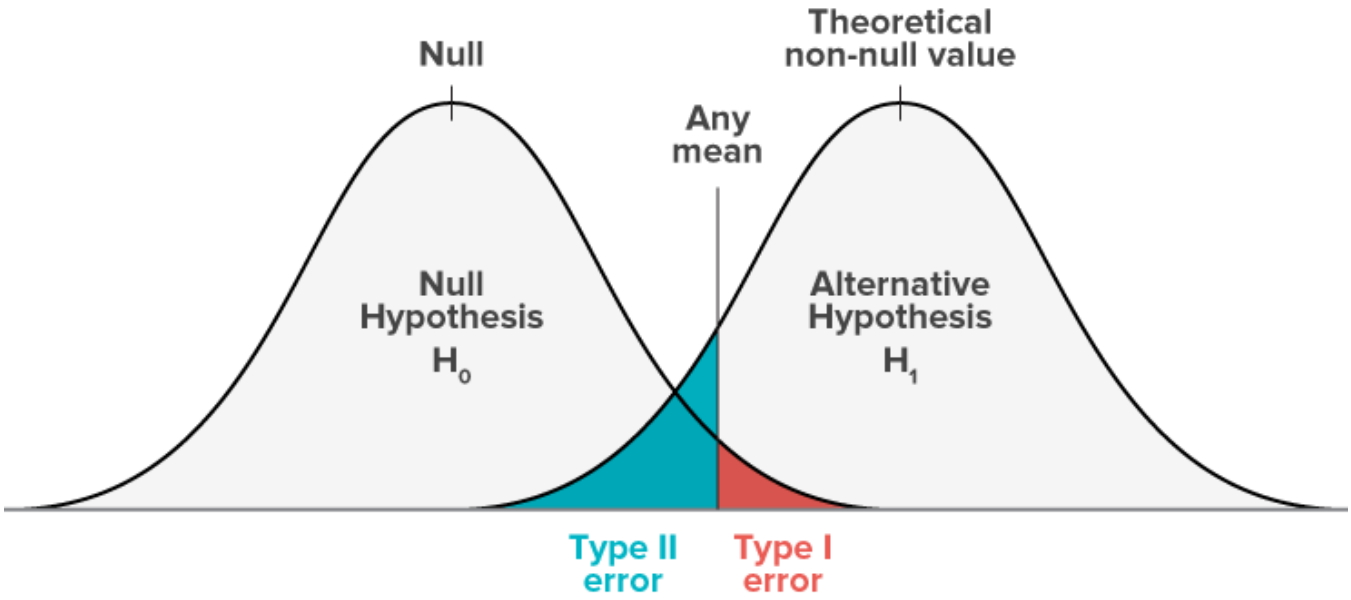
Question 5 (c)

The suggested procedure is not appropriate because one of the requirements for using a two-sample z-interval for a difference between proportions is that the two proportions are based on two independent samples. In the situation described the two proportions come from a single sample and thus are not independent.

Practice Exam Multiple Choice

Monday, February 20, 2017 10:50 PM

Question 12



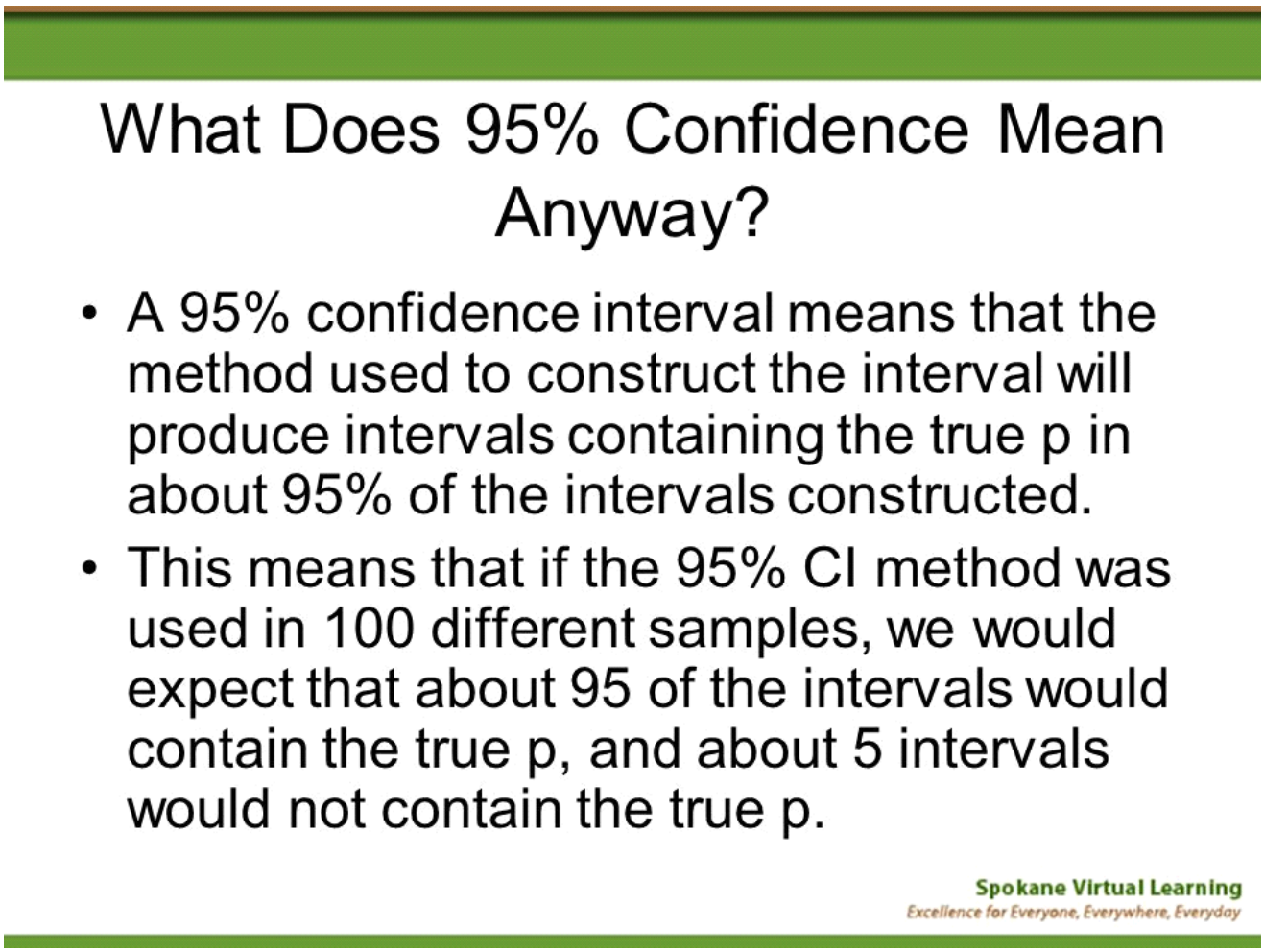
HYPOTHESIS TESTING OUTCOMES

Research \ Reality	The Null Hypothesis Is True	The Alternative Hypothesis is True
The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

Type I error, also known as a **"false positive"**: the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. Plainly speaking, it occurs when we are observing a difference when in truth there is none (or more specifically - no statistically significant difference). So the probability of making a type I error in a test with rejection region R is $P(R \mid H_0 \text{ is true})$.

Type II error, also known as a "**false negative**": the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. Plainly speaking, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is $1 - P(R \mid H_a \text{ is true})$. The power of the test can be $P(R \mid H_a \text{ is true})$.

Question 15

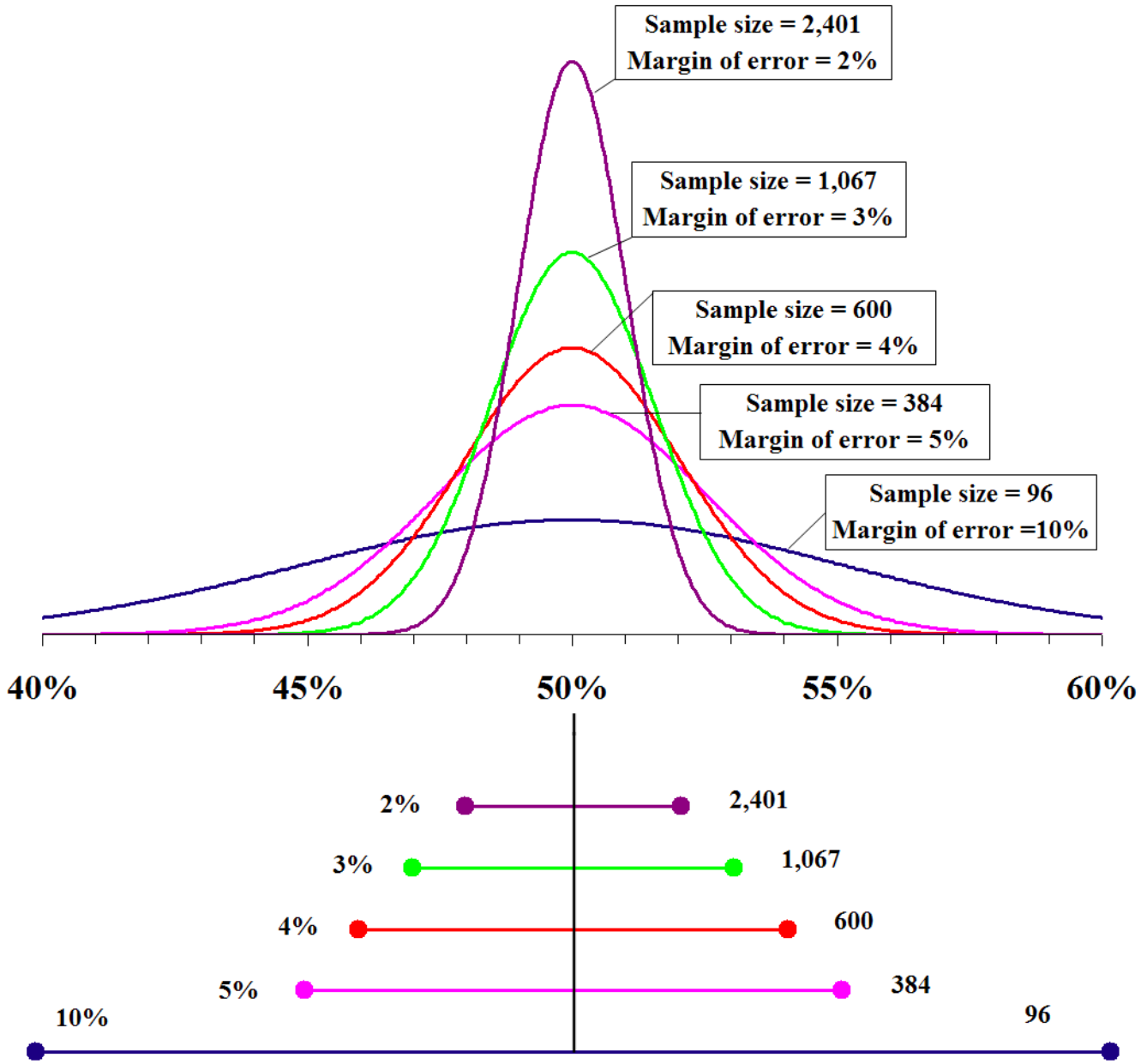


Question 22

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal IF np ≥ 10 and n(1 - p) ≥ 10
Quantitative (example: age)	μ = population mean, σ = population standard deviation	$\bar{x}$ = sample mean	μ	$\frac{\sigma}{\sqrt{n}}$	**When will the distribution of sample means be approximately normal ?**

Question 23

- 95% confidence interval of a sampling



- Margin of error vs. sample size at different confidence level

Sample	Confidence Level		
	80%	90%	95%
	% Margin of Error + / -	% Margin of Error + / -	% Margin of Error + / -
100	6.4	8.3	9.8
150	5.3	6.7	8
200	4.5	5.8	6.9
250	4.1	5.2	6.2
300	3.7	4.8	5.7
350	3.4	4.4	5.2
400	3.2	4.1	4.9
450	3.0	3.9	4.6
500	2.9	3.7	4.4
550	2.7	3.5	4.2
600	2.6	3.4	4.0
650	2.5	3.2	3.8
700	2.4	3.1	3.7
750	2.3	3.0	3.6
800	2.3	2.9	3.5
850	2.2	2.8	3.4
900	2.1	2.7	3.3
950	2.1	2.7	3.2
1000	2.0	2.6	3.1

- Equation
  - $\frac{invNorm(5\%, 0, 0.5)}{\sqrt{n}} = margin\ of\ error\ \%$
  - $n = \left(\frac{invNorm(5\%, 0, 0.5)}{margin\ of\ error\ \%}\right)^2$

Question 25

Standard error of difference

- To compare two groups, we have to calculate the standard difference error between the groups

Distribution	Parameter	Population value	Sample estimate	Standard error
Normal	mean	μ_1-μ_2	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}$
Binomial	Proportion	π_1-π_2	P_1-p_2	$\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$
Poisson	Rate	λ_1-λ_2	r_1-r_2	$\sqrt{\frac{r_1}{n_1} + \frac{r_2}{n_2}}$

Question 26

- $Sample\ distribution\ of\ the\ sum = Sample\ distribution\ of\ the\ mean \times sample\ size$

Question 27

$$E = \frac{(\text{row total})\,(\text{column total})}{(\text{grand total})}$$

One rainy Saturday morning, Adam woke up to hear his mom complaining about the house being dirty. "Mom is always grouchy when it rains," Adam's brother said to him.

So Adam decided to figure out if this statement was actually true. For the next year, he charted every time it rained and every time his mom was grouchy. What he found was very interesting - rainy days and his mom being grouchy were entirely *independent events*. Some of his data are shown in the table below.

Fill in the missing values from the frequency table.

	Raining	Not raining	Row total
Grouchy	7	66	73
Not grouchy	28	264	292
Column total	35	330	365

Question 30

In engineering, science, and **statistics**, **replication** is the repetition of an experimental condition so that the variability associated with the phenomenon can be estimated. ASTM, in standard E1847, defines **replication** as "the repetition of the set of all the treatment combinations to be compared in an experiment.

Replication (statistics) - Wikipedia
https://en.wikipedia.org/wiki/**Replication_(statistics)**

Question 33

Inferences about the Slope: Confidence Interval Example

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

Excel Printout for Produce Stores

	Lower 95%	*Upper 95%*
Intercept	475.810926	2797.01853
X Variable	1.06249037	1.91077694

At 95% level of confidence the confidence interval for the slope is (1.062, 1.911). Does not include 0.

Conclusion: There is a significant linear dependency of annual sales on the size of the store.

- $Confidence\ Intervel = Coef \pm invT\left(\frac{1 - Confidence\ Level}{2}, n - 2\right) \times (SE\ Coef)$

Question 38

"Blocking" vs "stratification"

"***Blocking***"

- word used in describing an experimental design

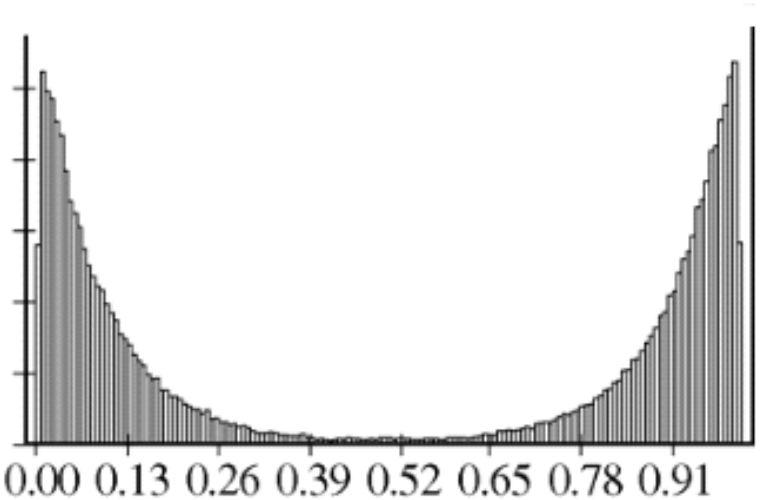
"***Stratification***"

- used in describing a survey or observational study
- Both refer to idea of only making comparisons within relatively similar groups of subjects

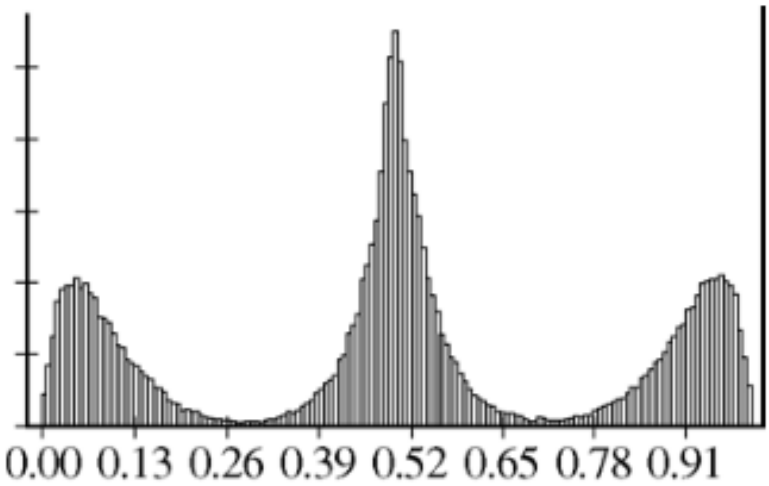
Chapter 6 8

Question 40

- Original data



- Sampling distribution of the sample mean with sample size = 2

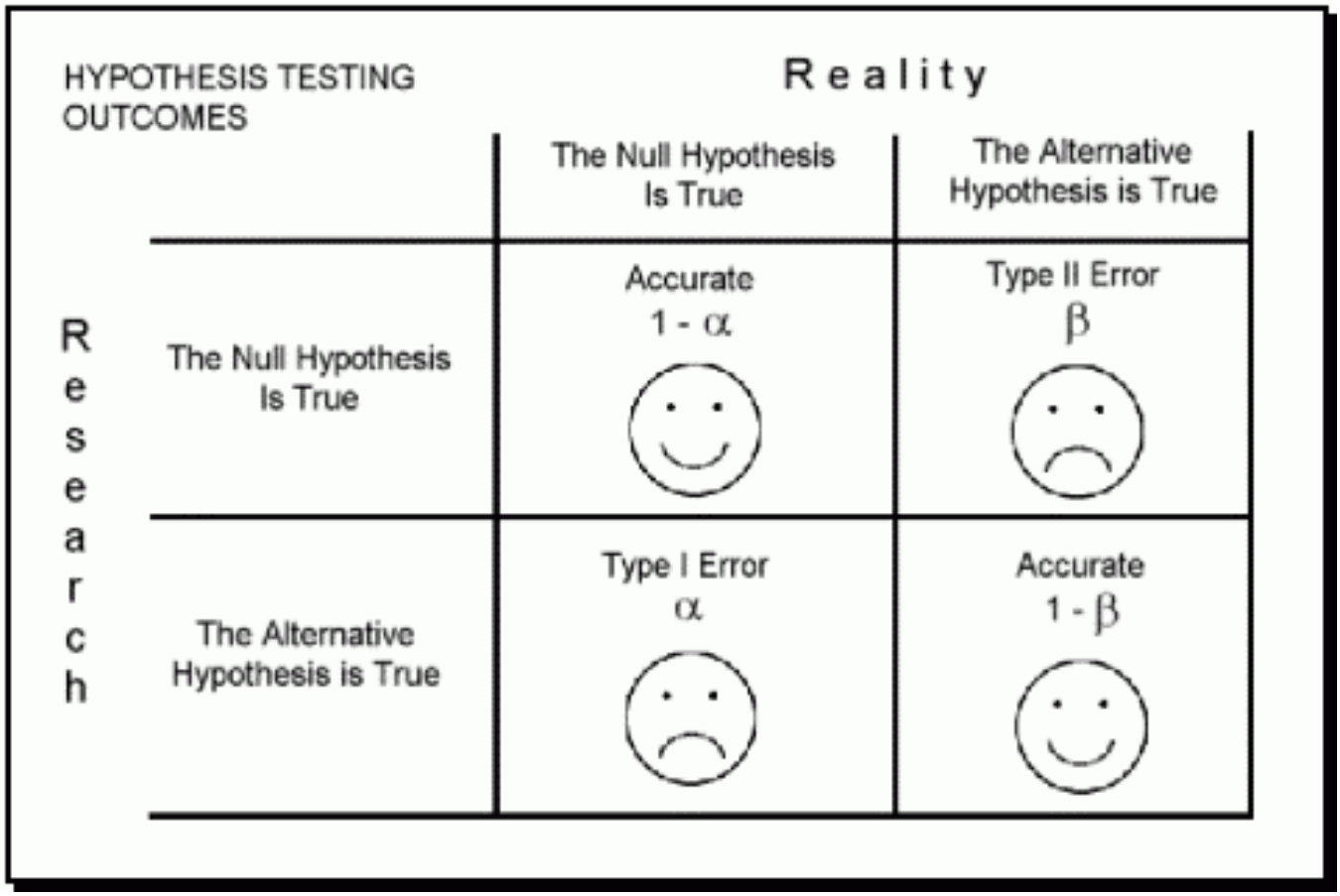

0.00 0.13 0.26 0.39 0.52 0.65 0.78 0.91

Princeton 1 Multiple Choice

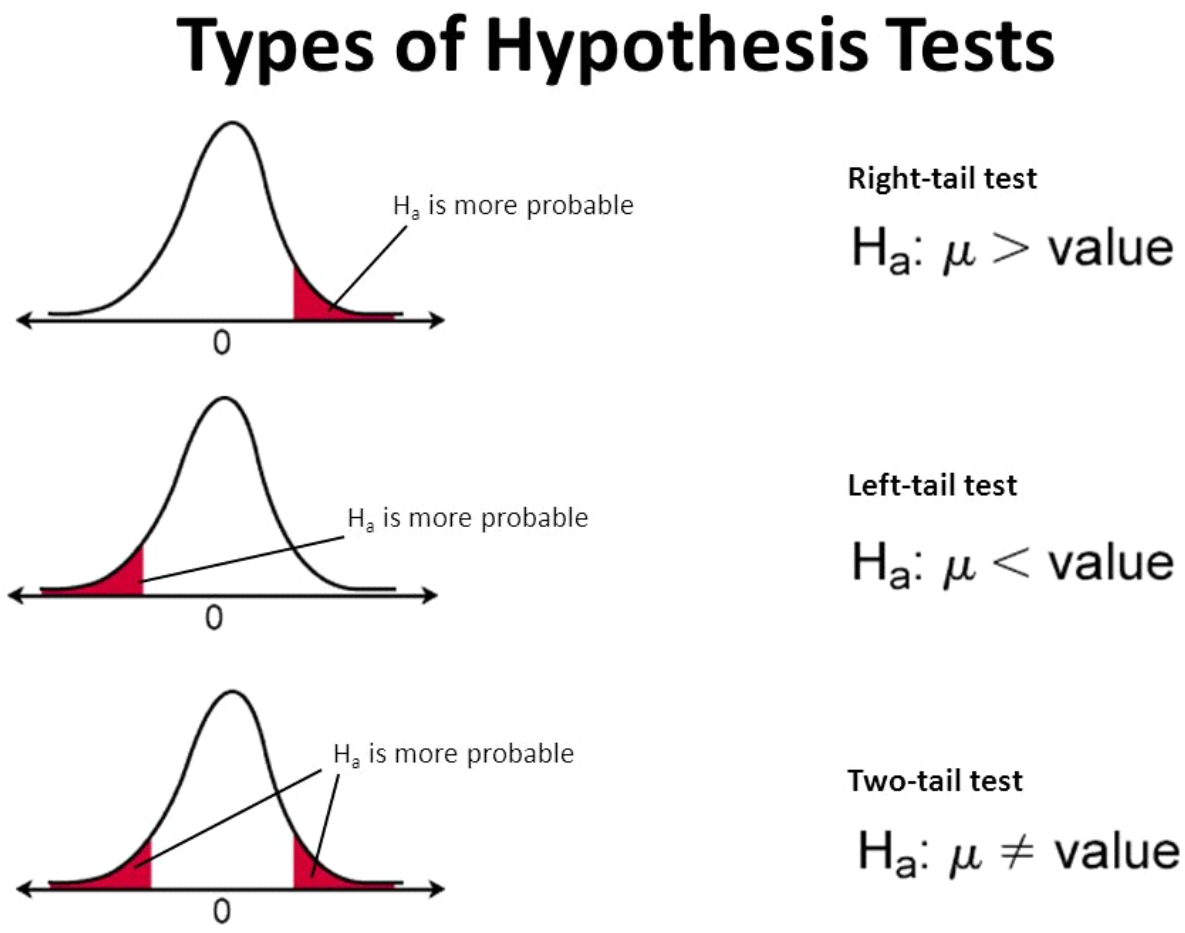
Thursday, February 23, 2017 10:39 AM

Question 10

- The type I error is the error of rejecting the null hypothesis when the null hypothesis is true.



Question 13



- The right-tailed test is used. So there is a 5 percent area in the rejection region in the right tail of the sampling distribution. If we construct a 90 percent confidence interval, then the upper confidence limit will match the critical value. If the test of hypothesis is rejected at a 5 percent level of significance, then the test statistic fell in the rejection region. In other words, the hypothesized value of mean did not belong to the 90 percent confidence interval.

Question 18

- A binomial model counts the number of successes out a fixed number of attempts at a task when each attempt has a constant probability of success

$$P(X = c) = \text{binompdf}(n, p, c)$$

n -> number of trials

p -> probability of success

This finds the probability of exactly c successes, for some number c.

$$P(X \leq c) = \text{binomcdf}(n, p, c)$$

n -> number of trials

p -> probability of success

This finds the probability of c or fewer successes.

Question 20

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D)P(D)}{P(T)}$$

$$= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)}$$

Question 25

- The population of interest is the population you are trying to draw an inference about from the collected data sets.

Question 31

- Response variable is about each sample, not the whole samples

Some examples of responding variables in different experiments—things to be observed or measured are:

The amount of water absorbed by two different brands of paper towels.

How far a ball rolls from different ramp angles.

The amount of feed eaten at a bird feeder in response to the type of seed in the feeder.

Question 33

- The margin of error $= \frac{\text{width}}{2}$
- Notice the difference between sample mean and sample proportion

Statistic	Standard Deviation of Statistic
Sample Mean	$\frac{\sigma}{\sqrt{n}}$
Sample Proportion	$\sqrt{\frac{p(1-p)}{n}}$

Question 37

- Equality of standard deviations is not necessary for a t-test to be valid. One of the conditions of a t-test is that the underlying populations must be normally distributed.

Princeton 2 Multiple Choice

Thursday, February 23, 2017 4:31 PM

Question 2

In the jury pool available for this week, 30 percent of potential jurors are women. A particular trial requires that, out of a jury of 12, at least three are women. If a jury of 12 is to be selected at random from the pool, what is the probability it meets the requirements of this trial?

(A) 0.168

(B) 0.843

(C) 0.915

(D) 0.949

(E) The answer cannot be determined without knowing the size of the jury pool.

- We cannot use a binomial model unless we know that the probability of drawing a woman for the pool is nearly constant.
- However, since we are drawing 12 jurors without replacement, this is not necessarily true unless the jury pool is very large (at least 120)

Requirements to be Binomial- B.I.N.S

- A binomial setting arises when we preform several independent trials of the same chance process and record the number of times that a particular outcome occurs.

The four conditions for a binomial setting are:

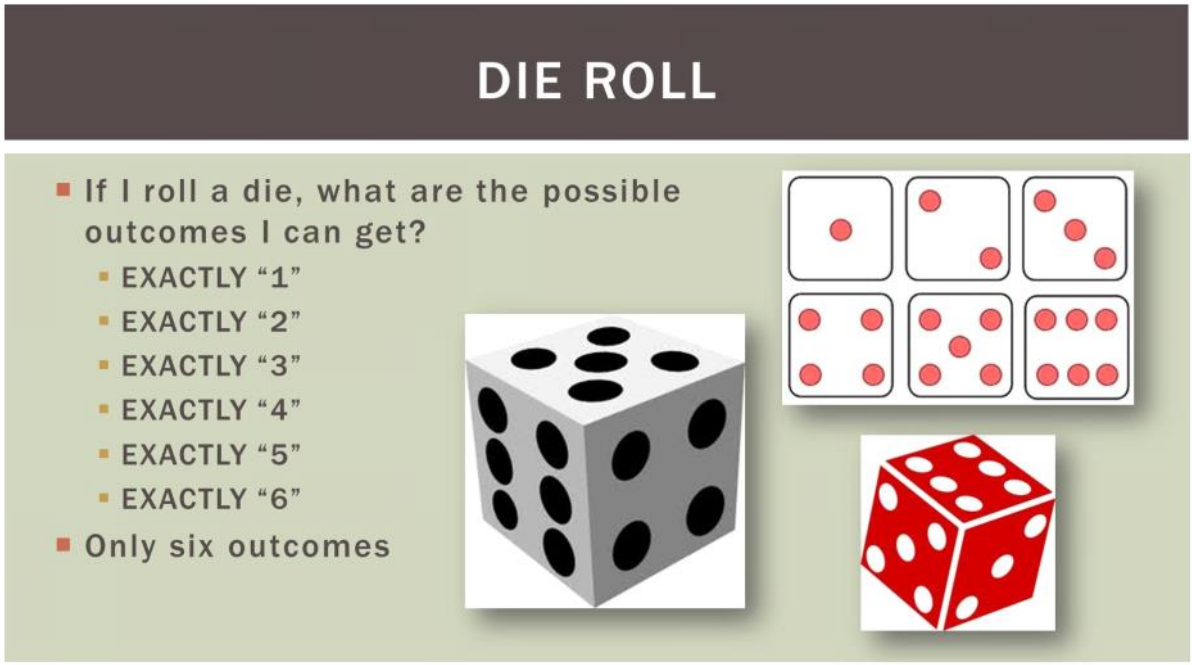
- **B**inary? The possible outcomes of each trial can be classified as "success" or "failure"
- **I**ndependent? Trials must be independent; that is, knowing the result of one trial must not have any effect on the result of any other trial
- **N**umber? The number of trials ***n*** of the chance process must be fixed in advance.
- **S**uccess? On each trial, the probability ***p*** of success must be the same.

Question 11

- A discrete variable takes only a countable number of values. The number of test questions a student guesses the answers to is a random variable with possible values 0, 1, 2,…n, where n is the number of questions on the test.

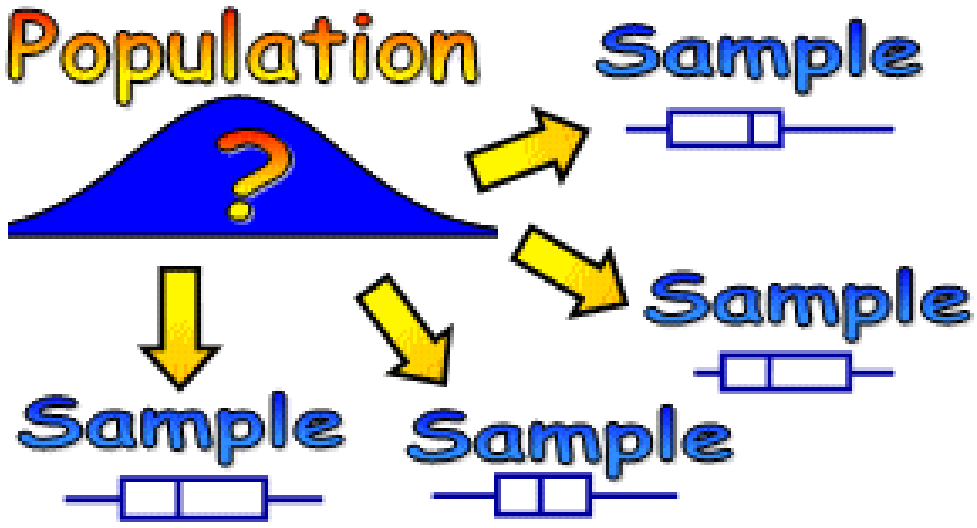
A **discrete variable** is a **variable** which can only take a countable number of values. In this example, the number of heads can only take 4 values (0, 1, 2, 3) and so the **variable** is **discrete**. The **variable** is said to be **random** if the sum of the probabilities is one. Probability Density Function.

Discrete Random Variables – Mathematics A-Level Revision
https://revisionmaths.com/advanced-level-maths-revision/.../discrete-random-variables



Question 29

- If we take different samples from the same population, the estimates from the different samples will be different. The difference in percentages may be entirely due to sampling variation.



Question 30

In a clinical trial, 30 sickle cell anemia patients are randomly assigned to two groups. One group receives the currently marketed medicine, and the other group receives an experimental medicine. Each week, patients report to the clinic where blood tests are conducted. The lab technician is unaware of the kind of medicine the patient is taking. This design can be described as

(A) a completely randomized design, with the currently marketed medicine and the experimental medicine as two treatments

(B) a matched-pairs design, with the currently marketed medicine and the experimental medicine forming a pair

(C) a randomized block design, with the currently marketed medicine and the experimental medicine as two blocks

(D) a randomized block design, with the currently marketed medicine and the experimental medicine as two treatments

(E) a stratified design with two strata, patients with sickle cell disease forming one stratum and those without sickle cell disease forming the other stratum

- This experiment consists of two treatments, the currently marketed medicine and the experimental medicine. Patients are not matched, and no blocks are formed. Only patients with sickle cell disease are involved in the experiment.

Question 31

- Confidence intervals are constructed as Statistic ± Margin of Error.
- Therefore, the statistic is always right in the center of the confidence interval.

Conditions/Assumption

2017年4月11日 星期二　　下午 10:30

Numerical Data

Statistic/Parameter	Condition/Assumption	How do we check?
One mean	1. Randomization Condition: Sample is a random sample (SRS) or representative of the population 2. Normality Condition: n ≥ 30 or the population distribution is fairly normal. 3. Independence Condition: The selection of each subject is independent of each other (10n < N)	1. Based on the information provided. 2. If n is small show that the sample is fairly normal. 3. Show that the inequality is true.
Two Sample Means (Independent)	1. Randomization Condition: Sample is a random sample (SRS) or representative of the population or in experiments the treatments are randomly assigned. 2. Normality Condition: n ≥ 30 for each sample or the population distributions are normal. 3. Independence of Groups Condition: The groups are independent of each other.	1. Based on the information provided. 2. If n is small show that the sample is fairly normal with a graph. 3. Based on the information provided.
Two Sample Means (Paired)	1. Paired Condition: The samples are dependent. 2. Randomization Condition: Sample is a random sample (SRS) or representative of the population or in experiments the treatments are randomly assigned to the subjects. 3. Normality Condition: n ≥ 30 or the population distribution is normal.	1. Based on the information provided. 2. Based on the information provided. 3. If n is small show that the distribution of the sample differences is fairly normal.
Slope of Regression Line	1. Linearity Assumption: The data seem to be a linear relationship. 2. Independent Assumption: The errors (Residuals) are independent. 3. Randomization Condition: The sample is a random sample (SRS) or representative of the population. This is necessary for generalizing. 4. Equal Variance Condition: The spread of the errors (residuals) for x value is about the	1. The scatterplot and the correlation coefficient. 2. No Pattern in the residual plot. 3. Based on the information provided. 4. Residual plot consistent. No outliers

	same. 5. Normal Condition: The distribution of the errors is normal.	or influential data. 5. Boxplot or histogram of the residuals. Check for extreme skewedn ess and outliers

Categorical Data

Statistic/Parameter	Condition/Assumption	How do we check?
One Proportion	1. Randomization Condition: Sample is a random sample (SRS) or representative of the population 2. Normality Condition: np≥ 10 and nq ≥ 10. 3. Independent Condition: The selection of each subject is “independent of each other (10n < N)	1.Based on the information provided. 2. Show that the inequalities are true. 3. Show that the inequality is true.
Two Sample Proportions (Independent)	1. Randomization Condition: Samples in each group are random samples (SRS) or representatives of their populations or in experiments the treatments are randomly assigned. 2. Normality Condition: n_1p_1 and $n_2p_2 \geq 10$ and n_1q_1 and $n_2q_2 \geq 10$. 3. Independent Condition: The selection of each subject is independent of each other (10n < N) for each sample. In some experiments this is not necessary. 4. Independence of Groups Condition: The groups are independent of each other.	1. Based on the information provided. 2. Show that the inequalities are true. 3. Show that the inequality is true. 4. Based on the information provided.
More than two Category Proportions (Goodness of Fit)	1. Count Condition: The data are counts. 2. Independent Condition: Data are sampled independently 3. Large sample	1. Verify this. 2. SRS and 10n < N 3. Count > 5
More two Sample Proportions (Test for Homogeneity)	1. Count Condition: The data are counts. 2. Independent Condition: Data in groups are independent. 3. Large sample	1. Verify this. 2. SRS and 10n < N 3. Count > 5
Relationship between Proportions of two Variables	1. Count Condition: The data are counts. 2. Independent Condition: Data are sampled independently	1. Verify this. 2. SRS and 10n < N 3. Count > 5

	3. Large sample	