Midterm 1 Review

Part I: Gathering Data

Basic Definitions

You'll want to be able to differentiate between the following four definitions: population, sample, parameter and statistic. If given an example of a statistical study, could you identify each part?

- **Population:** the entire group of individuals that we want information about. Any numeric value from the population is a **parameter.**
- **Sample:** subset of the population that we actually examine to gather information about the population. Any numeric value from the sample is a **statistic.**

1. Characteristics of a population are called _____, while those of a sample are termed _____.
   A. Parameters; statistics
   B. Statistics; variables
   C. Statistics; measures
   D. Statistics; parameters

2. Time magazine reported the results of a random sample of 800 adults contacted during the two days following the presidential election. According to poll results, 56% reported that they had voted in the election. In statistical terms, what do we call the 56% figure
   A. A simple random sample. B. A population C. A value of a statistic D. A value of a parameter

3. In March 2007, Consumer Reports published an evaluation of large screen, high definition television sets. The magazine purchased and tested 98 different models from a variety of manufacturers. The population in this study is:
   A. The 98 purchased large screen, high definition television sets.
   B. All large screen, high definition television sets.
   C. All television sets.
   D. All manufacturers of large screen, high definition television sets.

Types of Data

Be able to identify if a variable is **quantitative** (numerical, takes a quantity) or **categorical** (divides people into categories). Remember, not all numerical variables are quantitative – your Red ID, SSN or zip code are all numerical variables that divide you into categories.

This is mostly tested in regards to the methods we can use for each type of data:

Categorical	Quantitative
Single Variable: Bar Chart Pie Chart	Single Variable: Histogram Stem Plot Box Plot
Two Variables: Contigency Table	Two Variables: Scatterplot Correlation / Linear Regression

4. Which of the following would be an appropriate display for home zip codes of SDSU students?
A. Pie chart B. Histogram C. Box Plot D. Stem Plot E. B, C and D.

Use the following to answer questions 5-6:

An exercise physiologist is doing a research study on post-menopausal women and bone density. The researcher considers a variety of variables that could have a role in the bone density of this group of women. He looks at a group of 1529 post-menopausal women and asks them about whether or not they had taken oral contraceptives prior to menopause and how many minutes they exercise.

5. What would be the appropriate graphical display for these two variables, respectively?

   A. bar chart and histogram
   B. pie chart and bar chart
   C. histogram and boxplot
   D. boxplot and pie chart

6. What is the population in this study?
   A. Women
   B. Post-menopausal women
   C. 1539 post-menopausal women
   D. It can't be determined from the information given

7. The Gallup poll conducted a telephone survey of 1180 American voters during the first quarter of 2008. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person voted in the midterm congressional election. Which of the variables is quantitative?
   A. Voter's party affiliation.
   B. Whether or not the person voted in the midterm election.
   C. Voter's region.
   D. Voter's age.
   E. None of the variables is quantitative.

Sampling Designs

Make sure you have the big 4 memorized: SRS, Systematic, Cluster and Stratified. Remember, most students struggle with telling stratified from cluster samples:

Stratified	Cluster
Subjects within group are SIMILAR for some characteristic or set of characteristics	Subjects within group are DISSIMILAR
We choose a SAMPLE from EACH group	We choose ENTIRE group(s) at random

- **Simple Random Sampling (SRS):** We choose people at random (i.e. picking names out of a hat). Each member of the population has an equal chance of being included.
- **Systematic Sampling:** We choose every **nth** item.
- **Cluster Sampling:** The population is divided into groups that are **dissimilar** on characteristics. We choose **entire** clusters at random and combine 1 or more clusters to get our overall sample.
- **Stratified sampling:** The population is divided into groups that are **similar** on some characteristic or set of characteristics, we then choose people at random (by SRS) from each group (**we sample from every group rather than taking a few entire groups!**) and combine those samples into our overall sample.

- **Multistage Random Sampling:** Any combination of a variety of sampling methods
- V**oluntary Response Sampling:** Sample choose themselves. (Web survey, call-in polls)
- **Convenience Sampling:** We choose people that are easiest to reach.

8. For quality assurance, every tenth machine part is selected from an assembly line and measured for accuracy. This is an example of which sampling method:
   A. Cluster B. Systematic C. Simple Random Sample D. Stratified

9. In order to estimate the proportion of patients in a hospital who wake up regularly at night, a random sample of 10 patients is selected from each of the different wings of the hospital (OB-GYN, ICU, etc). This is an example of:

10. SDSU's Dining Services conducted a survey to find out how on-campus first-year students felt about the current required meal plans. They randomly sampled 10 floors from the residence halls and asked all students from the selected floors to answer a brief questionnaire. What type of sampling did they use?

11. A survey was conducted to assess student opinions about a proposed new student coffee shop on campus. The sample of 200 students was made up of a random sample of 50 students from each of the four student groups: freshmen, sophomores, juniors and seniors. The type of sampling method used is:

12. A large retail company randomly selected ten of its stores from a list of all the company's U.S. stores. All employees at these ten stores are asked to fill out an employee satisfaction survey. This is an example of:

13. An airline company randomly chooses two flights from a list of all international flights taking place on a particular day. All passengers on these two flights are asked to fill out a customer satisfaction survey. What sampling technique was used?

14. A large retail firm conducted a survey to find out how its employees felt about the available health benefits. The firm randomly selected 5 employees from each of its retail stores. This is an example of which of the following types of sampling:

Types of Bias

Any question about biases will ask for the most prevalent type of bias. Even if you can argue that a few of the types listed below are answers (they will be), be on the lookout for those keys to let you know what the problem being tested is!

- **Undercoverage:** Entire population targeted is not included in the design of the sample.
  - Be on the lookout for any mention of a certain group in the sample design (if they mention they only sampled females, or people of a certain age group, or people from a certain region) and check that the group mentioned is the same as the population they are interested in. If not, you have undercoverage!
- **Non-response:** An individual selected into the sample cannot be contacted or refuses to cooperate.
  - Any mention of choosing people into the sample and people not responding – whether not being home for interview or phone call, refusing to participate, etc…
- **Response Bias:** Interviewee's responses are influenced by the interviewer.
  - Any mention of inappropriate behavior on the interviewer or interviewee, if you can see any reason the interviewee may lie - the way the question was asked, who was asking the question, etc…
- **Voluntary Response Bias:** The type of bias associated with voluntary response samples.
  - Web surveys, mail in surveys, call in surveys…

15. In a survey to estimate potential voter support for a California ballot initiative, a random sample of registered voters in Los Angeles is selected. The most significant source of bias will be:
    A. Non-response bias
    B. Undercoverage
    C. Response Bias
    D. Voluntary Response Bias
    E. It will not be biased.

16. The IFOCE, the governing body for competitive eating, wants to know what competitive eaters think about Kobayashi's labor dispute. Kobayashi is currently unable to compete in IFOCE events because he won't compete exclusively in IFOCE events which is a stipulation in his contract. To see what other competitive eaters thought, the IFOCE asked each of the eaters competing at the Nathan's hot dog competition if they think Kobayashi should be banned. What is the most prominent source of bias in this survey?
    A. Voluntary Response Bias
    B. Nonresponse Bias
    C. Response Bias
    D. Undercoverage

17. A random sample of 200 drivers was asked the question, "Do you ever text while you are driving?" 7% of the sample responded "yes." However, when observers were placed at intersections, the percent of drivers who were observed texting was 24%. Which of the following is the most likely reason for the difference in results between the verbal and observational surveys?
    A. The sample size for the verbal survey was too small.
    B. The survey was voluntary, and only those who wanted to answer the survey did.
    C. The verbal survey question was sensitive, and respondents did not answer honestly.
    D. The difference can be attributed to random choice.

18. Which of the following are true statements?

    I. Voluntary response samples often under-represent people with strong opinions.
    II. Convenience samples often lead to undercoverage bias.
    III. Questionnaires with non-neutral wording are likely to have response bias.

    A. I and II B. I and III C. II and III D. I, II, and III E. None of the above

Experiments & Observational Studies

Be able to tell the difference between an observational study and an experiment. If we assign individuals to a treatment group (this can be as innocuous as us asking them to eat a serving of vegetables or watch a commercial), it's an experiment. If we just ask them about or observe their behavior, then it's an observational study.

19. In one study on the effect eating meat products has on weight level, an SRS of 500 subjects who admitted to eating meat at least once a day had their weights compared with those of an SRS of 500 people who claimed to be vegetarians. In a second study, an SRS of 500 subjects were served at least one meat meal per day for 6 months, while an independent SRS of 500 others were chosen to receive a strictly vegetarian diet for 6 months, with weights compared after 6 months. Which of the following is true:
    A. Both studies are controlled experiments.
    B. Both studies are observational studies.
    C. The first study is an observational study, while the second is a controlled experiment.
    D. The first study is a controlled experiment, while the second is an observational study.

Observational Studies

Observational studies require that we only observe! We cannot apply a treatment to the subjects.

- **Retrospective study:** looks **backward** in time, we ask subjects about what happened in the past to try to determine possible risk factors.
- **Prospective study:** looks **forward** in time, normally we follow subjects to see if an outcome occurs.

Principles of Experimental Design

You may be asked questions regarding the basic principles of a good experiment, so make sure you know these 3!

- **Control:** The experimental conditions for all treatment groups to assure that lurking variables do not bias the results, part of this is including a control (nontreatment) group.
- **Randomization:** The experimental units must be **RANDOMLY** assigned to treatments
- **Replication:** Replications of our experiment must be used to reduce chance variation in the results.

Experiment Terminology

Of all the experiment terminology, it's most likely you'll be asked about number of treatments or to list all the treatments. Remember that treatments include **all possible combinations** of the levels of each factor. If given a description of an experiment, you should be able to identify each of the following:

- **Experimental units/Subjects:** Individuals you are studying in the experiment.
- **Response variable:** Outcome or dependent variable. This is what we are ultimately measuring or interested in (our y variable in a regression context).
- **Factors:** The **explanatory (independent, x) variables** that are thought to influence the response variable studied. Combine specific values (**levels**) of each factor to form a treatment.
- **Treatment:** A specific condition applied to the subjects. (A combination consisting of a level of each factor.)

A good experiment will usually employ the following to help **CONTROL** for lurking variables. Know these definitions:

- **Placebo:** A dummy treatment. Used to control for the placebo effect.
- **Blinding:** A **double-blind** experiment is one where neither the participant nor the researcher taking the measurements knows who had which treatment. A **single-blind** experiment is one where the participants do not know which treatment they have been assigned. **Helps reduce bias!**

A last definition associated with experiments that you may be asked about:

- **Statistically Significant:** An observed effect so large that it would rarely occur by chance.

Use the following information to answer questions 6 and 7
Does type of potato and preparation method affect the taste of potato chips? An experiment was conducted in which tasters rated the flavor of potato chips. The researcher used three different types of potato in the experiment. Potatoes were either deep fried or oven-baked, and the flavor of the resulting chip was rated.

20. Which of the following is the response variable?
    A. Preparation method.
    B. Flavor of potato chip.
    C. Type of potato.
    D. Not enough information provided.

21. How many treatments could this experiment have?
    A. 6 B. 9 C. 2 D. 3 E. 5

22. In order to test the effects of new drug on treating acid reflux, 150 patients with chronic acid reflux were assigned to one of four different dosage levels of the drug (none, low, moderate and high) and one of two diets (restricted or normal) for one month. They then reported back the number of times in the month they experienced acid reflux. How many treatments were used in the above experiment?
    A. 4 B. 2 C. 8 D. 1

Experimental Designs "How did we determine who got what treatment?"

If you're asked a question about experimental design, cross out anything that isn't what you're being asked for! Some experimental designs have methods almost identical to our sampling designs (SRS v. completely randomized and stratified v. block design). These will distract us if we aren't careful.

- **Completely Randomized Design:** Similar to an SRS setup, each subject is equally likely to get assigned to any treatment.
- **Matched Pairs:** Subjects are paired according to variables that affect the response and then randomly assigned with one to the treatment and the other to the control group.
    - Keep an eye out for before and after studies. They are a matched pair design where each subject is their own control! If it seems like every got the treatment, check to see if you've been given a before and after study.
- **Block Design:** Blocks of similar subjects are formed and within each block, they are randomly assigned to treatment groups.
    - Similar to a stratified set up, if we first **divide** our subjects into groups that are **similar**, then randomly assign from **each** group into our treatment groups, we have a block design.

23. A high school track coach wants to determine if strength training will improve the times of the runners on the team. The team consists of 20 long distance runners and 20 sprinters. If the coach randomly assigns half of the long-distance runners and half of the sprinters to a strength training regime and the other half continue with only track practices. What type of experimental design has the track coach used?
    A. Stratified
    B. Completely Randomized
    C. Block
    D. Matched Pairs

24. A study was conducted to compare the effectiveness of 3 advertisements for the same product. Since men and women respond differently to advertising, the researcher randomly assigned the women to three groups, one to view each advertisement. Then the men were separately assigned to 3 groups to watch the advertisements. What time of experiment is this?
    A. Stratified sample
    B. Randomized block design experiment
    C. Completely randomized experiment
    D. Matched pairs experiment
    E. This is not an experiment

25. In a study, 40 subjects who claimed to follow a vegetarian lifestyle and 40 who said they followed a non-vegetarian lifestyle were followed for 6 months and then had their cholesterol levels measured. Which type of experimental design was used here?
    A. Completely randomized
    B. Block design
    C. Matched pairs design
    D. None of the above; this is an observational study.

26. An advertising firm is interested in seeing how an advertisement affects people's opinion of a certain brand. They ask 100 randomly selected people to fill out a brief questionnaire indicating how they feel about the brand. They then view a 45 second advertisement and fill out the questionnaire again. What type of experiment did the advertising group use?
    A. Stratified B. Completely Randomized C. Matched Pairs D. SRS

Confounding v. Lurking Variables

Remember the drawings from class. Confounding variables add another relationship with the response variable (in addition to the one with the existing explanatory variable) while lurking variables have a relationship with both existing explanatory and response variables (which is the entire cause of the relationship we saw between the explanatory and response variable).

- Two variables are **confounded** when you can't tell which of them (or whether it's the combination) had an affect on the response variable.
    - Example: You might want to test if a fertilizer helps your garden produce more tomatoes. Suppose you spread it on half the plants and record the number of tomatoes they produce. But you spread it on the sunny half, leaving the shady half unfertilized! Now, even if you have more tomatoes on your fertilized plants, you don't know whether that's because of fertilizer or sunshine (or the two together, which is actually the most likely case).
- A **lurking variable** is sometimes referred to as common response. It's a variable that drives two other variables, creating the impression of an association between them while in reality the two variables are BOTH response variables.
    - Example: Suppose a researcher finds a strong association between the number of computers per capita and life expectancy - countries with fewer computers have lower life expectancy. Do we think that the computers affect life expectancy in some way? NO! The general socioeconomic status (which could be measured in something like gross domestic product (GDP)) is likely causing both the number of computers and the life expectancy to rise.

27. Foods rich in omega-3 fatty acids are rumored to help with allergies. Suppose if to study this, a matched pairs experiment is run where 100 allergy sufferers record their allergy symptoms and then change their diet to increase the amount of food containing omega-3 fatty acids. After a week, they record their symptoms again and find that there's a statistically significant decrease in symptoms. However, someone also took pollen readings and found that the pollen levels dropped from the beginning of the week to the end of the week. Pollen levels would be an example of a:
    A. Lurking variable B. Confounding variable C. All of the above. D. None of the above.

Part II: Exploring and Understanding Data

Contingency Tables

Most of these questions can also be answered using what we learned in probability, make sure to read carefully for what is being asked (see more in the Part IV review notes).

- Know how to make a **conditional contingency table**, whatever is **conditioned on** or **conditioned by** will become the denominator for every cell in that row or column

Use the following to answer problems 1-4:

The table below shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

	Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

1. What percent of adults surveyed are male and don't plan to watch the Super Bowl?

2. What percent of adults are looking forward to the commercials?

3. Of those looking forward to watching the game, what percent are females?

4. Create a conditional table of gender, conditioning on viewing preferences.

	Male	Female
Game		
Commercials		
Won't watch		

Histograms

Remember, first step should be filling any missing information into your histogram – putting any missing x values down and writing how many observations are in each bar at the top.

- Know how to find median, Q1, and Q3 intervals for a histogram
- Know how to find probabilities using a given histogram.

Calculator shortcuts for finding Q1, median and Q3 intervals:

Method 1:

1. Enter column numbers in L1 and frequencies in L2.
2. 1-VAR-STATS L1, L2 - 5 number summary gives out column location of each part

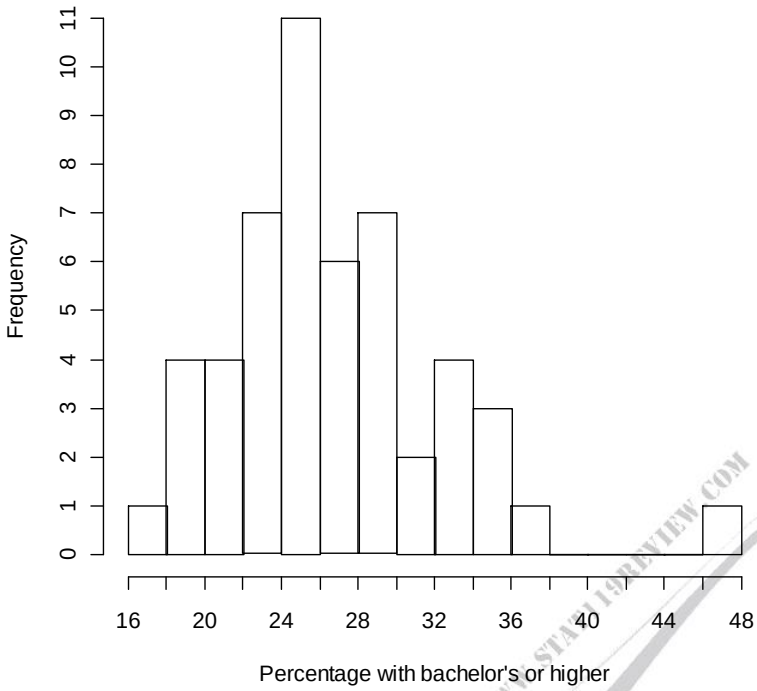
Note: Check to be sure the n= is correct!

Method 2:

1. Enter the values 1 through n into L1.
2. 1-VAR-STATS - 5 number summary gives out location of item
3. Count over to determine column.

Use the following to answer problems 5-6:

The Chronicle of Higher Education published the accompanying data on the percentage of the population with a bachelor's degree or graduate degree in 2007 for each of the 50 U.S. states and the District of Columbia.

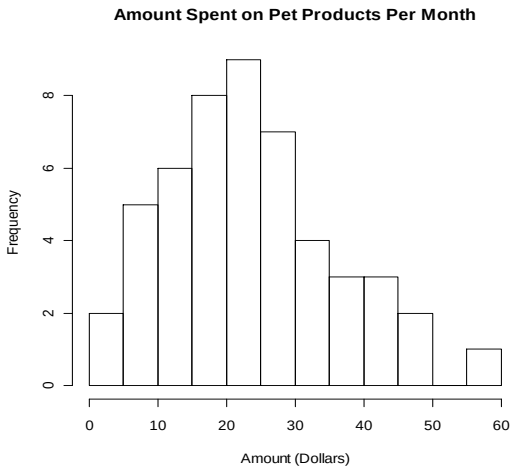


5. What percent of states have over 30% of people with a bachelor's degree or higher?
   A. 21.6% B. 35.3% C. 17.6% D. 22.0% E. 36.7%

6. Which interval is the third quartile located in?
   A. [22,24) B. [24,26) C. [26,28) D. [28,30) E. [30,32)

Use the following information to answer questions 7-8

A marketing consultant randomly surveyed 50 shoppers at a particular pet store to see how much they spent each month on pet products. A histogram of the survey data was constructed.



7. What percent of individuals spend between $30 and $50 per month?
   A. 12% B. 24% C. 19% D. 38%

8. What interval is Q3 located in?
   A. [10,15)
   B. [15,20)
   C. [20,25)
   D. [25,30)
   E. [30,25)

Stem plots

Usually just a method for us to give you data, not many questions about creating them.

- On back-to-back stemplots, always make sure you read away from the stem and that you're answering for the group the question is about!

9. A stemplot for the weights of a group of college women is given below

09 | 9	
10 | 2 2 5 6 7	
11 | 2 4 4 8 9 9	
12 | 1 3 5 5	
13 | 0 7	
14 | 9	
15 | 2	

What is the value of the median for this data set?

A. 114 B. 116 C. 118 D. 118.5 E. 119

Which measure of spread would be best for this data set?

A. Standard deviation because it is always the best measure of spread
B. Standard deviation because the data is skewed
C. IQR because it is always the best measure of spread
D. IQR because the data is skewed

Boxplots

Know how to make a boxplot by hand!

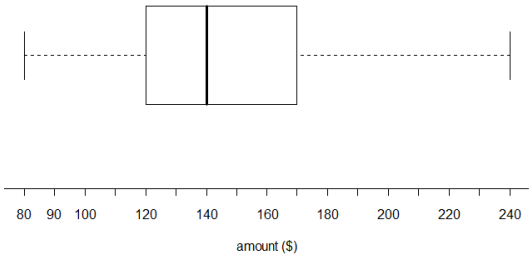
- Memorize your fence formulae!
  - Mild lower fence: Q1 – 1.5(IQR)
  - Mild upper fence: Q3 + 1.5(IQR)
  - Extreme lower fence: Q1 – 3(IQR)
  - Extreme upper fence: Q3 + 3(IQR)
- Know how to use your fences to determine if a value is an outlier
- Each portion of the boxplot represents 25% of the data, the box itself represents the interquartile range (middle 50%) of the data.

Calculator shortcut for creating boxplot:

1. Enter all data values in L1.
2. Press 2nd, then Y = to access the STAT PLOT
3. Turn on Stat Plot 1, make sure it's set for a modified boxplot (only necessary once).
4. Go to ZOOM, select 9. ZOOMSTAT.
5. Press TRACE to view the different portions of the boxplot.

Note: The calculator only checks mild fences, so you still need to memorize the fence formula to determine if an outlier is mild or extreme.

10. A chain of sports shops in Lake Tahoe wants to study how much a beginning skier spends on his or her initial purchase of ski equipment. Data was collected from 48 sales and is displayed below.



75% of sales receipt totals were at least what amount? _____

75% of sales receipt totals were less than what amount? _____

Approximately what is the IQR for the sales data? _____

Use the following to answer problems 11-13:

Big Mac prices in U.S. dollars for 44 different countries are given below:

1.84	1.86	1.90	1.95	2.17	2.19	2.19	2.28	2.33	2.34	2.45
2.46	2.50	2.51	2.60	2.62	2.67	2.71	2.80	2.82	2.99	3.09
3.33	3.34	3.43	3.48	3.54	3.56	3.59	3.67	3.73	3.74	3.84
3.84	3.86	3.89	4.00	4.33	4.39	4.90	4.91	6.19	6.56	7.20

11. Calculate and label the 5-number summary for the Big Mac pricing data.

12. Are there any outliers in the Big Mac pricing data? **Be sure to show all necessary calculations and state your conclusion.**

13. Construct and label a modified boxplot for the Big Mac pricing data.

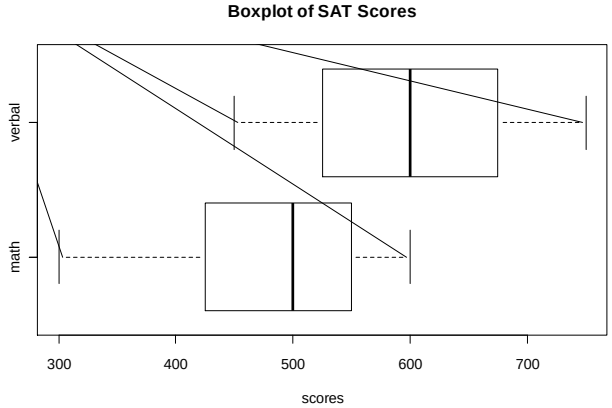
14. The five-number summary for the size (acres) of a sample of farms is:

Minimum	quartile #1	Median	quartile #3	Maximum
6.0	19	35.5	62	150

Is the largest farm an outlier?

A. No, it is within 1.5(IQR) of the third quartile.
B. No, it is within 1.5(IQR) of the first quartile
C. Yes, it is more than 1.5(IQR) above the median.
D. Yes, it is more than 1.5(IQR) above the third quartile.
E. Yes, it is more than 1.5(IQR) above the first quartile.

15. A box plot below summarizes the distribution of SAT verbal and math scores for students at an Arizona high school



Which of the following statements is FALSE?

A. The interquartile range for the math scores is smaller than the interquartile range for the verbal scores.
B. The range of the math scores equals the range of the verbal scores.
C. The highest math score equals the median verbal score.
D. The verbal scores are roughly symmetric while the math scores are skewed to the right.

Note: We also learned two other types of graphs for quantitative data: Timeplots which can be appropriate for a single quantitative variable if we have measure of time. And scatterplots which require TWO quantitative variables.

Graph Features
The quantitative graphs covered above can help you see the following aspects of a distribution:

- Modes (unimodal, bimodal, multimodal)
- Symmetry (symmetric, skewed to left, skewed to right)
- Usual Features (outliers, gaps)

Describing Distributions Numerically

All our measures in this section can be found in the 1-VAR STATS output for your TI-83 or TI-84. Make sure you know how to enter a list and use this feature – even if only to check your work you did by hand!

Know how to find all numeric measures.

	Shape of Distribution	
	Symmetric distribution	Skewed distribution or one with outliers
Measure of Center	mean	median
Measure of Spread	standard deviation	IQR

Expect to be asked which measure of center or spread to use for a described (or graphically presented) data set. Use the above chart to determine and pay close attention to if they're asking for center or spread.

16. The salaries of professional athletes are right-skewed with a number of high outliers. Which measure would be used to describe the spread of the salaries?
    A. Mean B. Median C. Standard Deviation D. IQR E. Range

Measures of Center

All of these give us a measure of where the center of our data is.

- **Mean:** average of all data values
- **Median:** middle value of all data values
    - Remember to use the median location formula = (n+1)/2
    - Robust to outliers or skew since it doesn't matter at all what the values to either side of the median are.

Remember, **your mean will move in the direction of a tail (skew) or outliers**. Assume you will have at least one question giving you a mean and median and asking about the shape of the distribution.

- A mean larger than the median means the distribution is skewed right
- A mean smaller than the median means the distribution is skewed left
- A mean close to the median means the distribution is approximately symmetric

17. The salaries of professional athletes are right-skewed with a number of high outliers. Which of the following is true?
    A. The mean will be larger than the median.
    B. The median will be larger than the mean.
    C. The mean and median will be close to one another.
    D. It cannot be determined from the information given.

18. Below are the summary statistics for average daily wind speed for a region of Massachusetts:

Min	Q1	Median	Q3	Max	Mean	Standard Deviation
0.20	1.15	1.90	2.93	8.67	2.13	1.33

Which of the following is a correct statement?

A. The sample of average wind speeds is left skewed
B. The sample of average wind speeds is right skewed
C. The sample of average wind speeds is symmetric
D. We cannot tell the shape of the distribution from summary statistics

19. According to a recent article, the mean hourly wage for general managers at a large firm was $92.33 and the median was $69. We can conclude

A. More than half the managers earn less than $92.33/hour
B. More than half the managers earn more than $69/hour
C. The distribution of hourly wages is symmetric
D. The distribution of hourly wages is left-skewed

20. The census found that in 2003, households headed by persons between the ages of 45 and 54 had a median household income of $61,111 and a mean household income of $77,634. What does this tell us about household incomes for households headed by persons between the ages of 45 and 54?

A. Over half of the households earn less than $61,111
B. Over half of the households earn more than $61,111
C. Over half of the households earn less than $77,634
D. Over half of the households earn more than $77,634

Measures of Spread / Variability

All of these give us a measure of how spread out our data is (or how much our values vary from one another).

- **Standard Deviation (and Variance):** measures the spread or variability of the data.
  - $s \geq 0$
  - The closer s is to 0, the less spread out our data is
  - For s to be exactly 0, all data values must be identical
  - Will get very large with a skewed distribution or with outliers
- **Interquartile Range (IQR) = Q3 - Q1**
  - Memorize the formula, it won't be given to you
  - Robust to outliers or skew since it cuts off the 25% of data on either end
- **Range = max – min**
  - Unlikely you'll be asked for it and it's never the right choice when asked for the appropriate measure of spread, but know the formula in case

- **First Quartile (Q1):** Middle value of the smallest half of the data, the 25^{th} percentile.
- **Third Quartile(Q3):** Middle value of the largest half of the data, the 75^{th} percentile.

Five Number Summary

Consists of minimum, Q1, median, Q3, maximum. Remember to label them!

You can use the five number summary to create a boxplot.

21. A weather station recorded wind speeds every day for 50 days. When the researcher looked at the data using a histogram, there was one low outlier - a day where the wind speed was 0 mph. It was later determined that the instrument that measured wind speed malfunctioned on that day.

    This incorrect data point was removed from the data and a new histogram was made and all new statistics were calculated. Which of the following statements is true?

    When the incorrect data point is removed:
    A. The mean and standard deviation will both be larger
    B. The mean and standard deviation will both be smaller
    C. The mean will be larger, but the standard deviation will be smaller
    D. The mean will be smaller, but the standard deviation will be larger
    E. None of the above

22. If 75% of values in our distribution are at least 200 and the IQR is 50, which of the following is true?
    A. 50% of the values are between 200 and 250.
    B. The median of our distribution is 225.
    C. 50% of the values are between 150 and 200.
    D. Both A and B are true.
    E. Both B and C are true.

Part III: Exploring Relationships Between Variables

For everything in Part III (scatterplots, correlation, linear regression), always be on the lookout for a categorical variable to be slipped into a question. If the question has a categorical variable, none of the methods in Part III are applicable!

- Anything with the word CAUSE in it will be FALSE. As statisticians, we never get to say one thing causes another, we can only talk about associations and statistical significance.

Scatterplots

Displays the relationship between two quantitative variables, including the four aspects below

- **Form:** linear, curved, and clustered.
- **Direction:** positive association and negative association.
- **Strength:** how close points lie to a simple form (such as a line).
- **Outliers:** extreme observations.

Correlation

Be on the lookout for correlation coefficients from categorical variables, assuming causation from correlation, and forgetting that correlation only describes the **linear** relationship. All this common mistakes are regularly tested.

- **Correlation Coefficient (r):** measures the strength and direction of a **linear** relationship between two variables.

	Positive	Negative
Weak	$0 < r < 0.3$	$-0.3 < r < 0$
Moderate	$0.3 < r < 0.7$	$-0.7 < r < -0.3$
Strong	$0.7 < r < 1$	$-1 < r < -0.7$

  - **Be able to interpret r.**
    - The interpretation for a correlation coefficient is: "There is a **strong negative** linear relationship between **x** and **y**."
    - You'll want to use the above table to determine the first bit, the strength and direction, then insert what the x and y variables actually are

- $-1 \leq r \leq 1$
  - The closer r is to -1 or 1, the closer the points on the scatterplot fall on a straight line
- **Be able to calculate r from R^2**
  - $r = \pm\sqrt{R^2}$
  - The sign on r will be the same as the sign of the slope for our regression line, don't forget to write R^2 as a decimal before taking the square root!

*Be on the lookout for correlation coefficients from categorical variables, assuming causation from correlation, and forgetting that correlation only describes the **linear** relationship. All these common mistakes are regularly tested.*

1. The National Opinion Research Center at the University of Chicago conducted a survey where they obtained data involving the number of hours of watching television in a typical day and the age of the survey participants. The resulting regression equation is $\hat{y} = 2.19 + 0.0173x$

   The percentage of the variation in hours watching television that can be explained by the relationship between hours watching television and age is 37.49%

   The correlation coefficient describing the relationship between watching TV and age is:
   A. 0.3749 B. 0.6123 C. 0.1406 D. 0.0173

2. A researcher finds that the correlation coefficient between people's IQ and their favorite type of music is 0.67. What should he conclude?
   A. There is a strong non-linear relationship.
   B. There is a moderate positive linear relationship.
   C. Those who listen to classical music have a higher IQ.
   D. A correlation coefficient is not appropriate for this data.

Linear Regression

Know each of following for a free response AND/OR multiple choice question:

- Be able to interpret slope
  - For each 1 unit increase in x, y increases/decreases by the slope.
- Make predictions
  - Plug given x into the regression equation to find prediction ($\hat{y}$)
- Calculate a residual
  - **Memorize formula:** $e = y - \hat{y}$ (error is true value minus predicted value)

The first step for all three of those questions, which you will do each of on the exam, is figuring out which variable is x (e**X**planatory, independent) and which is y (response, dependent). Here are some methods to do that:

1. Use the regession line.
2. We **predict y.** So if you're asked to predict number of wins, that's your y variable.
3. Worst case, you won't have either of these, then you need to remember that **x influences/affects y.**

3. Highway planners investigated the relationship between traffic density (x) and the average speed of traffic (y) on a moderately large city thoroughfare. The researchers found the regression equation to be:

   $\hat{y} = 50.55 - 0.352x$

   When the traffic density was 25, the average speed of traffic was 40 mph. calculate the residual.
   A. 41.75 B. 1.75 C. 15 D. -1.75 E. None of the above

4. A regression line is computed relating a car's fuel efficiency (mpg) and the cost of the car (in dollars): Fuel efficiency = 35.06 – 0.0002(cost). What is the correct interpretation for the slope?
   A. For every dollar increase in car cost, fuel efficiency increases by 0.0002 mpg.
   B. For each mpg increase in fuel efficiency, care cost decreases by 0.0002 dollars.
   C. For every dollar increase in car cost, fuel efficiency decreases by 0.0002 mpg.
   D. For each mpg increase in fuel efficiency, care cost increases by 35.06 dollars.

Measures of Predictive Power

- **R^2 / Coefficient of Determination / Percent of Variability…**
  - All three of the above are ways **R^2** can be referred to, don't let yourself be confused!
  - Can find from correlation coefficient r, by squaring and multiplying by 100%
    - **$R^2 = (r)^2 \times 100\%$**
  - "The percentage of variability in Y that is explained by the regression line"
  - As R^2 gets closer to 100%, the predictive power for our model gets BETTER

- **Residuals**
  - **e** = observed y (truth) – predicted y (predicted) = $y - \hat{y}$
  - Negative residual: prediction is too high (overpredict)
  - Positive residual: prediction is too low (underpredict)

- **Residual Plots**
  - **e** from above is calculated and plotted for each observation as (x,e)
  - Four problems to watch out for – all indicate that our model is a poor fit
    - Pattern / curvature (possible nonlinear relationship)
    - Fanning (possible nonlinear relationship)
    - Large values
    - Outliers

5. A homebuilder's association lobbying for various home subsidy programs argued that, during periods of high interest rates, the number of building permits issued decreased drastically, which in turn reduced the availability of new housing. Data relating housing loan interest rates (%) and the number of building approvals in thousands were collected. The data was analyzed to produce the results below:
   The regression equation is: building approvals (1000s) = 239.5 – 8.057 (interest rate %)
   Correlation = -0.843

   A correct interpretation of the slope is:
   A. For each 1000 building approvals, the interest rate decreases by 239.5.
   B. For each 1% increase in interest rate, the number of building approvals (1000s) decreases by 239.5.
   C. For each 1000 building approvals, the interest rate decreases by 8.057%
   D. For each 1% increase in interest rate, the number of building approvals (1000s) decreases by 8.057.

   What percent of variability in building approvals can be explained by the regression model?
   A. 8.057% B. 84.3% C. 71.06% D. 91.82%

Other Definitions

- **Extrapolation:** Don't make predictions outside of the range for which you have data.

You may have to determine which of the below characteristics a certain point has on a scatterplot of residual plot:

- **Outlier:** Far away from other points, on either x or y.
- **Leverage:** Far away from other points, only in regards to the x value.
- **Influential Point:** Removing the point would result in a very different regression line.

Part IV: Probability
If you can answer the question intuitively without notation, do it that way! If you're not sure where to start, translate everything into probability notation and see if a contingency table or the formula sheet can help out.

Types of Probability Questions

Conditional Probability
A conditional probability just means that we already know some outcome has occurred. There are number of ways to realize you're dealing with a conditional probability question:

- Both variables have been mentioned, but there is no "and" word (remember "but" is another word that means "and")
- The key words to let you know which variable we already know are **of**, **if**, and **given**.

The "probability of A **given** B" or "**Of** those in B, what is the probability A would occur" are two ways to denote a conditional probability $P(A \mid B) = P(A \cap B)/P(B)$

- As soon as you realize they've told you something is given, write the line in your notation and write whatever is given **BEHIND** the line.
- The formula for conditional probability is given on the formula sheet, just be careful changing the As and Bs to whatever your letters are. Note: The given portion is **ALWAYS** the denominator!

1. Employment data at a large company show that 74% of workers are married, 42% are college graduates and 20% are college graduates and married. Given that an employee is a college graduate, what is the probability that the employee is married?
   A. 0.4762 B. 0.2703 C. 0.74 D. 0.96

2. Of the participants at a conference, 50% attended breakfast, 70% attended dinner and 40% attended both breakfast and dinner. If a participant attended breakfast what is the probability she also attended dinner?

   A. 0.20 B. 0.71 C.0.57 D. 0.80 E.0.70

Contingency Tables
There should be at least one contingency problem. It will either by one which can be solved by aid of a contingency table or one with a given table.

- For the created table, you can then be asked for one of the other "and" probabilities – just pick it off the table!
- For both, the most commonly asked for things or the "or" probabilities and the conditional probabilities.

3. A survey of students showed that 30% are in favor of a ban of alcohol on campus (event A), 62% are in favor of a ban on tobacco products on campus (event B) ; 25% are opposed to both measures.

	Ban tobacco?		
Ban alcohol?	yes (B)	no (B^c)	**Total**
yes (A)			
no (A^c)			
Total			1

What is the probability a student selected at random is on favor of banning both alcohol and tobacco products?
A. 0.186 B. 0.13 C. 0.17 D. 0.75 E. None of these

Given that a student is opposed to a tobacco ban, what is the probability he is opposed to an alcohol ban?
A. 0.6579 B. 0.3571 C. 0.25 D. 0.2742 E. None of these

Use the following to answer questions 3-5:
Real estate ads suggest that 57% of homes for sale in San Diego County have central air conditioning (event A), 23% have swimming pools (event B), and 29% of houses for sale had neither air conditioning nor a swimming pool.

4. What is the probability that a randomly selected home for sale in San Diego County has both central air conditioning and a swimming pool?

A. 0.1311 B. 0.51 C. 0.8 D. 0.09 E. None of these

5. If a randomly selected home for sale has a swimming pool, what is the probability it does not have air conditioning?

A. 0.3913 B. 0.2456 C. 0.14 D. 0.6087 E. None of these

6. Are events A and B independent
   A. No, because A and B overlap.
   B. No, because $P(A \cap B) \neq P(A)P(B)$
   C. Yes, because A and B overlap.
   D. Yes, because $P(A \cap B) = P(A)P(B)$
   E. There is not enough information provided to answer this question.

7. A survey of the entering MBA students at a university in the United States classified the students by gender and type of MBA program.

	Two-year MBA	Evening MBA	Total
Female	116	66	182
Male	48	38	86
Total	164	104	268

What percent of the evening MBAs are women?

What is the probability a student selected at random is male or a two-year MBA student?

Disjoint v Independent

It is very important to understand the difference between disjoint and independent events. Expect at least one question that involves you understanding the difference.

- **Disjoint:** Mutually exclusive. Two events cannot happen at same time.
  - $P(A \cap B) = 0$
- **Independent:** The outcome of one event does not influence the outcome of the other.
  - $P(A|B) = P(A)$
  - $P(A \cap B) = P(A) \bullet P(B)$
  - Any time you see independence given in a problem, write down $P(A \cap B) = P(A) \bullet P(B)$, it will probably be used in the problem.

8. The probability of event A is 0.3. The probability of event B is 0.6. If events A and B are disjoint, then:
   A. P(A or B) = 0.9 B. P(A and B) = 0.18
   C. P(A or B) = 0.72 D. P(A or B) = 0

9. If P(A) = 0.30 and P(B) = 0.25, what is P(A or B) if A and B are independent?
   A. 0.075 B. 0.55 C. 0.475 D. 0 E. None of these

10. Two events, X and Y, are independent, such that P(X) = 0.41 and P(Y) = 0.52. What is the value of $P(X \cup Y)$?
    A. 0.2132 B. 0.9300 C. 0.1100 D. 0.7168

Given Formulae

- $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Should be used for any "or" question.
  - Can also be used to solve for "and" with a little algebra:;
    $P(A \cup B) = P(A) + P(B) - \mathbf{P(A \cap B)}$

- $P(A \text{ and } B) = P(A \cap B) = P(A|B)*P(B)$
  - Only use if you have a conditional! Otherwise, try using the formula above.
  - If you have P(B|A), don't fret: $P(A \cap B) = P(B \cap A)$,
    so $P(A \cap B) = P(B|A)*P(A)$

11. 39% of new model cars have a built in DVD player, 21% have a GPS navigation system, and 47% have at least one of those features. What is the probability that a new model car has both these features?
    A. 0.13 B. 0.08 C. 0.26 D. 0.0819 E. 0.6

12. Advertisements for statistics texts suggest that 79% come with a CD with statistical tools, 35% have online help, and 18% have both aids. Find the probability that a randomly selected statistics text has statistical tools or online help.
    A. 1.14 B. 0.18 C. 0.96 D. 0.28 E. 0.61

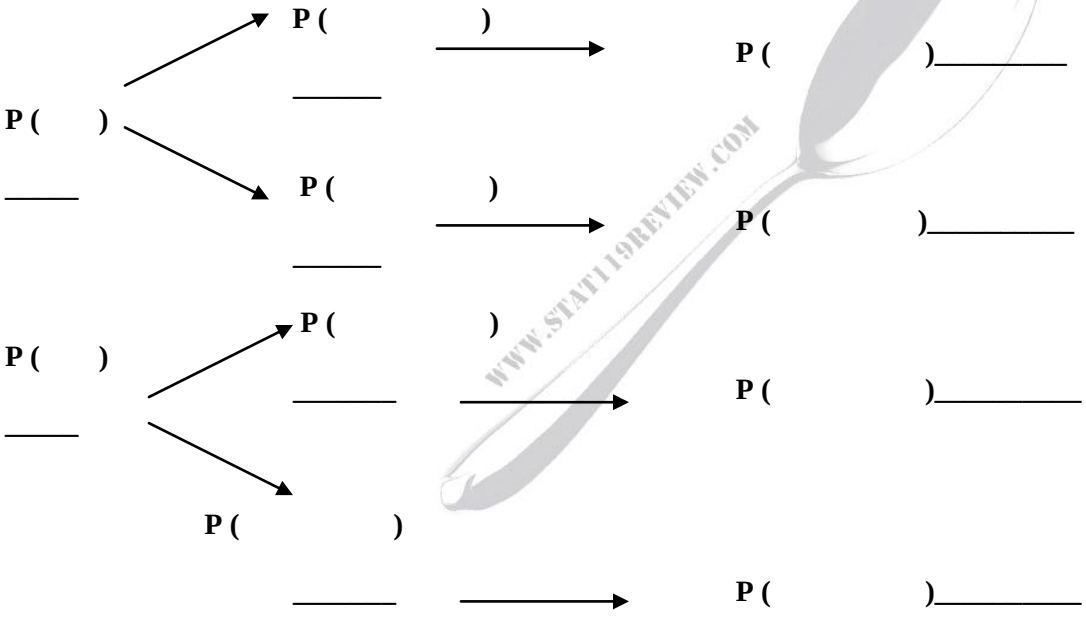
Tree Diagrams

If you're given two conditional probabilities and one marginal probability, it's likely a tree diagram will help you solve the question! If it is stated on the test that a tree diagram might help, DRAW A TREE DIAGRAM!

- **Memorize** the probability notation and construction of the tree. The items in blue below will be given (or can be calculated by subtracting a given probability from 1).
- You can use the probabilities in green to fill in a contingency table and then answer ANY question!

Hints and Tips:
- The most common questions asked in tree diagrams are for P(B) and P(A|B) or a similar type of probability.
- For P(B), just add the two joints in the green section that have a B in them.
- For P(A|B), the given item is almost always the thing you calculated first, so you can reuse your work!

Use the following information for questions 12-13
Suppose a campus computer store sells laptop and desktop computers. The probability a computer sold is a laptop computer is 0.56 (event A). In the first year, 15% of laptop computers require service, while 5% of desktops require service. Let B be the event that a computer requires service.
Complete a tree diagram for this scenario using the correct notation and values:



13. What is the probability a computer required service and was a laptop?

    A. 0.028 B. 0.084 C. 0.28 D. 0.2 E. None of these

14. What is the probability that a computer will require service?

15. Given that a computer required service, what is the probability that it was a laptop?

    A. 0.792 B. 0.084 C. 0.298 D. 0.42 E. None of these

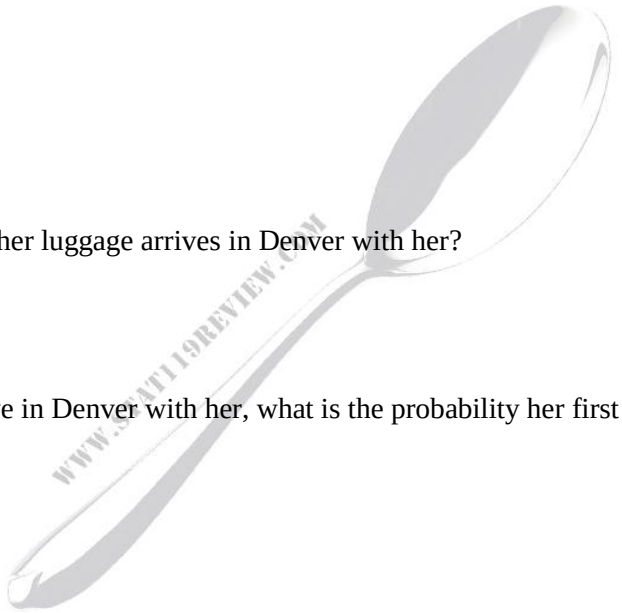
Use the following to answer questions 12-14

Leah is flying from Boston to Denver with a connection in Chicago. The probability her first flight leaves on time is 0.15. If the flight is on time, the probability that her luggage will make the connecting flight in Chicago is 0.95, but if the first flight is delayed, the probability that the luggage will make it is only 0.65. Let A be the event that the first flight leaves on time, let B be the event that the luggage will make the connecting flight.

12. Draw a tree diagram below: use correct notation to label events and fill in all appropriate probabilities (including intersections)

13. What is the probability that her luggage arrives in Denver with her?

14. If her luggage does not arrive in Denver with her, what is the probability her first flight did not leave on time?



Sampling without Replacement

Recognizing Sampling without Replacement

- The simplest question will say, "without replacement."
- A more difficult question will leave these keywords out. Instead, you'll need to notice a few key things.
  - The number of total items will be given, and likely fairly small < 30.
  - We will want to come away with two items, not just recording a characteristic of those two items. Selecting a committee and purchasing items are good examples of this.

Hints and Tips for Sampling without Replacement:

- Most common thing asked for is the probability of getting EXACTLY 1 of a type of item, do not forget that there are two ways to pick this if we choose 2 items!!
- Remember, in sampling without replacement, you need to determine how the first pick effects the probability in the second pick.

16. In a group of eleven adults, there are three who are allergic to peanuts. If we select two adults at random from the group (without replacement), what is the probability we select exactly one who is allergic to peanuts?
    A. 0.3967 B. 0.2182 C. 0.2727 D. 0.4364 E. None of these

17. A box contains 16 soccer jerseys: 4 medium, 7 large and 5 extra large. Two jerseys are drawn from the box, without replacement. What is the probability exactly 1 medium jersey is selected?
    A. 0.375 B. 0.20 C. 0.05 D. 0.40

Use the following to answer 18 and 19:
Suppose a fun-size bag of skittles has 5 red, 7 green, 6 purple and 8 yellow (a total of 26 skittles). You select two skittles from the bag, one at a time, without replacement.

18. What is the probability that both of the skittles are green?
    A. 0.0646 B. 0.0725 C. 0.2692 D. 0.5092

19. The probability that exactly one of the two skittles is red is:
    A. 0.037 B. 0.162 C. 0.323 D. 0.311 E. 0.155

20. Suppose you win a contest and as a prize are allowed to draw, while blindfolded, two bills from a container. Inside the container are seven $1 bills, four $10 bills, and eight $20 bills. What is the probability of drawing at least 1 $20 dollar bill?
    A. 0.515 B. 0.257 C. 0.421 D. 0.678

Sampling with Replacement

- The simplest question will say "independence" somewhere in the question.
- A more difficult question will say something about picking a small number of people using a simple random sample (or the word normal), but no total number of people is mentioned.

Hints and Tips for Sampling with Replacement:

- One of the most common things asked for is P(at least one), remember, you should calculated this using P(at least one) = 1 – P(none)
- Remember also, P(none) is P(not whatever) raised to the power of however many things we selected. A lot of students mistakenly use the given probability and forget to subtract it from 1 first!

17. Eighty-four percent of households in the United States own a computer. A random sample of six households is selected. What is the probability that exactly four of the households own a computer?

A. 0.4979 B. 0.0076 C. 0.1912 D. 0.2472 E. None of these

18. Assume we know that the probability that a college student has a Twitter account is 0.18. Three students are selected at random (therefore we can assume independence). What is the probability that at least one of these has a Twitter account?

A. 0.5514 B. 0.4486 C. 0.0058 D. 0.9942

Probability Formulas:

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

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

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

Z-Scores

A Z-score gives us the number of standard deviations away from the mean a value is, as well as if that value is below or above the mean. You can find the Z-score for a value with the given formula:

- A Z-score of 2 means the value is 2 standard deviations above the mean. A Z-score of -1 means the value is 1 standard deviation below the mean.
- Two types of questions:
  - Unusual: Scores further away from zero are more unusual, regardless of sign.
  - Better: Be careful about direction if asking who performed "better" – for a test, a higher Z-score would be better since scoring higher on the test is better; however, for a race, a lower Z-score is better since a smaller time is actually a better performance.

21. An incoming freshman took her college's placement exams in French and mathematics. In French she scored 82 and in math 86. The overall results on the French exam had a mean of 72 and a standard deviation of 8. On the math exam the mean score was 68 with a standard deviation of 12. On which exam did she do better compared with the other incoming freshmen?
    A. French because that has the higher z-score.
    B. Mathematics because that has the higher z-score
    C. French because that has the lower z-score
    D. Mathematics because that has the lower z-score

22. John is enrolled in a calculus course and has received his score on the midterm exam. His score on the midterm was 114, and the professor has indicated that his z-score was 0.78. What is the best interpretation of the meaning of this z-score?
    A. His score was 0.78 IQR above the medial of all scores.
    B. He got 78% of the problems on the test correct.
    C. Seventy-eight percent of all students taking the exam had scores below his.
    D. Seventy-eight percent of all students taking the exam has scores above his.
    E. His score was 0.78 standard deviations above the mean of all scores.

23. The distribution of math SAT scores is normally distributed with a mean score of 500 and standard deviation of 100. The distribution of math ACT scores is normally distributed with the mean score of 18 and standard deviation of 6. If Jake scored a 680 on the math section of the SAT and Liz scored a 12 on the math section of the ACT, whose score is more unusual?
    A. Jake, because the Z-score is larger.
    B. Liz, because the Z-score is negative.
    C. Jake, because the Z-score is further from 0.
    D. Liz, because the Z-score is smaller.

True / False Review

1. A z-score of -3 represents an observation 3 standard deviations below the mean.

2. If 10 is added to each value in a data set, the IQR will remain unchanged.

3. Outliers should always be removed from the data set before doing any calculations.

4. Standard deviation is the best measure of spread for a skewed distribution.

5. A point is influential if removing it changes the regression model.

6. If a calculated residual is negative, then the predicted value must be too low.

7. In a residual plot, random fluctuation of points around zero indicates a linear model is not appropriate.

8. A negative correlation indicates that the two variables are unrelated.

9. A scatterplot can be used to show the relationship between two categorical variables.

10. When making predictions, it is appropriate to use extrapolation.

11. A standard deviation of zero indicates all the data values must be the same.

12. If two data sets have the same mean and median, then they must be identical.

13. A point has high leverage if including it results in a different regression model.

14. In an experiment, a statistically significant result occurs when an observed difference is too large to occur by chance.

15. A correlation of zero always means there is no relationship between x and y.

16. If A and B are independent, $\mathbf{P(A \mid B) = P(B)}$

17. An IQR of 0 means all the data values must be the same.