Name: **KEY**
BSAD 210—Montgomery College

EXAM 1

KEY

- There are 110 possible points on this exam. The test is out of 100.
- You have one class period to complete this exam, but you should be able to complete it in less than that
- Please turn off all cell phones and other electronic equipment.
- Be sure to read all instructions and questions carefully.
- Remember to show all your work. You may print your formulas in Excel using the Show Formulas option in the Formulas tab. Printed versions of your work showing formulas ***and*** showing the results counts as showing your work. But you must include both with your test for "showing your work" to count this way. Write your name on both print outs.
- Try all questions! You get zero points for questions that are not attempted.
- Note the last sheet lists all the equations you will need for this exam.
- *Please print clearly and neatly.*

Part I: Matching. *Write the letter from the column on the right which best matches each word or phrase in the column on the left. You will not use all the options on the right and you cannot use the same option more than once.*

2 points each.

1. **C** Coefficient of variation
2. **F** Mean
3. **B** Median
4. **H** Mode
5. **A** Sample
6. **E** Standard deviation
7. **D** Population

A. Problem: will have error
B. Problem: when samples with unusual observations change
C. Problem: additional calculation needed
D. Problem: impractical to get data for
E. Problem: when used to compare sample with very different averages
F. Problem: samples with large outliers
G. Problem: will have accuracy issues
H. Problem: samples of continuous data
I. Problem: samples with a low standard deviation

1. *Simply relying on the standard deviation is what we do most of the time. The reason is because the CoV required an additional step: divide by the mean.*
2. *The mean has problems with outliers. It can make the central tendency seem higher or lower than it actually is.*
3. *"Unusual" typically means any value that's higher or lower than the central tendency. Observations whose value changes, but doesn't change where the median is, means the median won't change. While this answer also works for Mode, it would render this question without an answer. Remember: each option cannot be used more than once.*
4. *The mode is not good at describing data that is not categorical (i.e. continuous). The most common answer might be way off from the central tendency because all other observations differ slightly.*
5. *All samples have error; no sample is 100% precise. It's an unavoidable problem, but it's still a problem. Accuracy, however, is avoidable.*
6. *Why we use the coefficient of variation; larger averages have larger standard deviations.*
7. *Why we sample.*

Part II: Multiple Choice. *Choose the best answer to the following.*

4 points each.

Number of Exterior Doors	*Percent of Doors in the U.S.*

KEY

8. Consider this hypothetical sample data on how many exterior doors (in other words, all doors that lead outside) each house in the United States has. Based on this information available in the sample, what is the mean number of doors? (Data are also available in Practice Exam 1 Data Set).

1	10%
2	40%
3	25%
4	11%
5	3%
6	1%
Unknown	10%

   a. 2.00
   b. 2.30
   c. 2.56
   d. 3.50
   e. None of the above

The first step is to multiply each percent—in decimal form—by the number of doors. (1)(0.10) = 0.10; (2)(0.40) = 0.80; (3)(0.25) = 0.75; (4)(0.11) = 0.44; (5)(0.03) = 0.15; (6)(0.01) = 0.06. Added together this equals 2.3 (0.10 + 0.80 + 0.75 + 0.44 + 0.15 + 0.06). Normally, we would be done, but we don't have data on 100% of the sample. We only have data on 90% (10% + 40% + 25% + 11% + 3% + 1%) of the sample, or 0.90. So instead of dividing by 1, we divide by 0.90 to get 2.56.

9. Which of the following is an example of categorical data?
   a. A person's income at the age of 25.
   b. Which state a person was born in.
   c. If a person owns a car or not.
   d. B & C
   e. None of the above

Categorical data is best represents by discrete groups, states in the case of B and "yes, I own a car" or "no, I do not own a car" in the case of C. C, though, is an interesting case because you could meaningfully express C as quantitative data using a "1" for yes and "0" for no. The average would be the same as the percent of respondent who own a car. (This is called a "dummy variable," an idea we'll talk more about in the last unit.)

10. Imagine you're tallying the results of customer surveys. After averaging three surveys, the average customer rating is 7.8 out of ten. After including a fourth survey, the average increases to 8.1. Based on the law of large numbers, what would expect the average of all customer surveys to be?
    a. Above 8.1
    b. 8.1
    c. Below 8.1 but above 7.8
    d. 7.8

KEY

e. Below 7.8

The best guess is 8.1, as that is the average with the largest sample size. That the average increased from a sample of three to a sample of four doesn't mean it'll increase again when you get your fifth sample.

11. When making a histogram in Excel, it might be tempting to use the histogram button which makes a histogram with one click. This can be a helpful tool if you're trying to get an intuitive idea of what the data look like but it shouldn't be used for something others would see. Why?
    a. Because the bins would not be easy to read.
    b. Because the bins would be too large.
    c. Because the bins would be awkwardly defined.
    d. A & C
    e. None of the above

Excel will make reasonably sized bins but because those bins are the result of a mathematical formula, they will not be cleanly set. This will make the resulting histogram less intuitive and harder to read.

12. Suppose you had some data concerning daily oil production (in gallons) for 500 different wells in the United States. If you wanted to get an idea for the distribution of production (including if there are multiple modes), which data display would be most appropriate?
    a. Dot plot
    b. Histogram
    c. Stem-and-leaf
    d. Box plot
    e. None of the above

There's going to be a wide variety of production, including fractions since we're dealing with the daily production of a liquid, so a dot plot wouldn't be a good idea. We have too many observations for a stem-and-leaf plot and a box plot won't tell us if the distribution is bimodal. Histogram is the best option.

13. It's not unusual for lottery winners to win the lottery again. Lottery winners, unsurprisingly, tend to like playing the lottery and will often spending their winnings on even more tickets. Which of the following is this an example of?
    a. Gambler's fallacy
    b. Hot hand fallacy
    c. Central tendency
    d. Standard deviation
    e. None of the above

KEY

Someone who buys a lot of lottery tickets (such as a previous winner) is more likely to win the lottery than someone who buys fewer tickets. There's no fallacy there; that's just mathematics. If anything, it's an application of the law of large numbers.

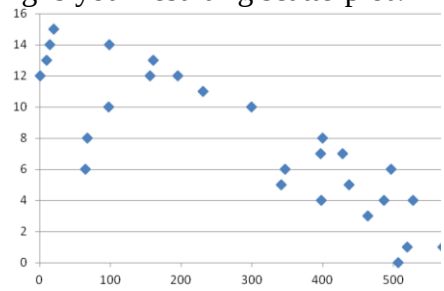
14. Use the Practice Exam 1 Data Set for this question. It includes hypothetical data on a hypothetical grocery store chain called The Happy Spud. What is the mean for amount (in thousands) that Happy Spud spent on advertising in the East region?
    a. 7.10
    b. 8.20
    c. 8.36
    d. 9.36
    e. None of the above

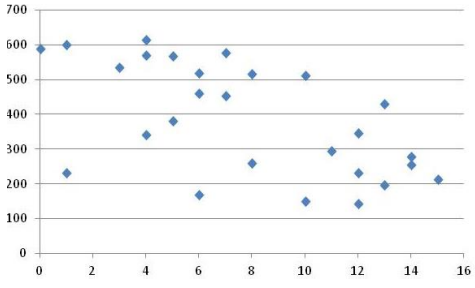
Keep in mind you're only interested in the East region so you'll only use E3 to E14. It's asking for the mean so you'll use the average function: "=AVERAGE(E3:E14)" and the answer is 7.1

15. Which of the following is a weakness of presenting data in a pie chart?
    a. It is intuitively difficult to tell what the chart represents
    b. People have trouble interpreting round objects
    c. **It's hard to tell which section is largest**
    d. A & C
    e. None of the above

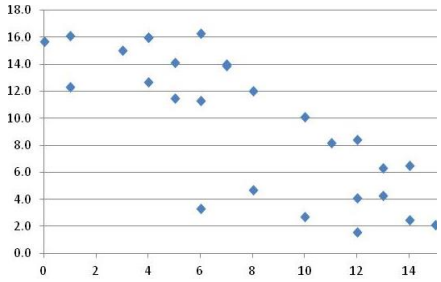
Pie charts are common because people intuitively know what they represent—each wedge is a percent of total something—but it's difficult to tell which section is largest. Option (B) is kind of correct—what matters for charts is the share of each section's circumference but the chart gives you an area—but because of how it is worded, it isn't correct.

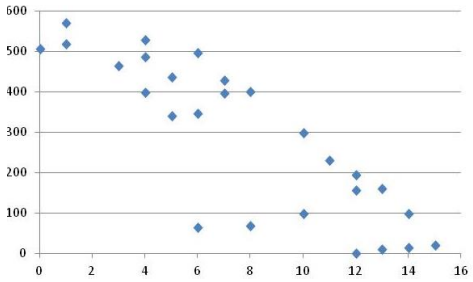
16. Use the Practice Exam 1 Data Set for this question with Happy Spud franchises. Create a scatterplot with Annual profit, in thousands and Number of competing stores in district with Number of competing stores in district on the horizontal axis. Which of the following is your resulting scatterplot?

a. 

b. 

KEY

c. 

d. 

Correct Answer

e. None of the above

Note that the range of competitors is in the teens while the profit's in the 100s; answer A is actually the inverse of the correct one. Profit is on the wrong axis. (You'd have to move the profit column to the right of competition column so it ends up on the correct axis.)

17. Ira wants to learn how people view video game violence and decides to collect a sample. Which of the following methods would result in an accurate sample?
    a. **While standing outside a major metro station on a weekend, offer people a chance to win a prize if they complete a short survey.**
    b. While standing outside a video game store on a weekend, ask people to take a detailed survey.
    c. While standing outside a busy bus stop Monday morning and evening, request people take a short survey.
    d. All of these will result in a very inaccurate sample.
    e. It is impossible to get an accurate sample; there is always inaccuracy.

The video game store option has an obvious bias: you're going to get a disproportionate number of video game fans. A "detailed" survey will also result in many people not bothering unless they have strong opinions. For similar reasons, the bus stop isn't great (but it's a big improvement). Because people will be going to work or home, you run the risk of getting a disproportionate number of strong opinions.

It's notable that the first option is also probably the most expensive thanks to the prize. But this prize incentivizes moderates to take the survey. (This is also why I give extra credit if you fill out an evaluation.) The metro also has a better chance of capturing a more accurate cross section of the population compared to a bus stop. The weekend also hinders the chance people will be too busy to bother. It's not perfect—it doesn't include people who drive a lot or stay in on the weekends—but it's not clear that would bias the results.

It is impossible to have a completely precise sample, but it is possible to have one that's accurate, or that has no systematic bias.

KEY

18. Which of the following is an example of cross-sectional data?
    a. Unemployment rates in the United States, from 1800 to 1850.
    b. Profitability of various German bakeries founded between 1976 and 2016.
    c. A company's monthly revenues for the past five years.
    d. A & C
    e. None of the above

Cross-sectional data has elements of different kinds of things—companies, states, countries, people—collected at the same time. While B is cross-sectional (it's data on various bakeries), it wasn't complied at the same time and this is a 40 year gap. It's one thing if the data's taken from a short span of time—there's little chance of data changing (the memo data sets are like this)—but forty years is too broad.

19. How does breast milk compare with feeding infants formula? Due to ethical constraints, studies addressing this question cannot randomly assign families one source of food or another. Instead, they compare outcomes (e.g. IQ) of children whose parents breast fed with those of children whose parents bottle fed. But parents who breast feed are very different from parents who bottle feed. What kind of sample bias is this?
    a. Self-selection
    b. Survivorship
    c. Undercoverage
    d. B *or* C; it is impossible to tell with the information provided
    e. It could be any; it is impossible to tell with the information provided

By their actions, parents are choosing to be in one sample or another. Since parents who breast feed are very different than parents who don't (breast feeding babies tend to have parents that are wealthier and better educated), those who opt into one sample are materially different than those who opt into a different sample.

Part III: Short Answer. *Answer the following.*
16 points each.

20. You've been put in charge of promotion and advertising for a new line of energy drinks, XTREME CAFFINE!, at the beverage company you work for. One of your first tasks is to create a website for XTREME CAFFINE!. The data below indicate the growth rate of unique visits each month after launch (the data are also available in the Practice Exam 1 Data Set):

Month	*Growth of Visits*	*Month*	*Growth of Visits*
May	48%	September	11%
June	30%	October	6%
July	25%	November	-2%
August	12%	December	4%

KEY

The website was launched in April, with 15,000 unique visitors. By the end of the year, how many unique visits are there? What is the average growth rate over this eight-month time span? (For the second question, round to the nearest two decimal places.)

Show your work; if you used Excel to answer this question, write what you put into Excel so I know how you got the answer you did.

Let's do the first question first. Recall to calculate the end-result number using a growth rate, you convert the percent value to decimal form. You then add one. So 48% becomes 1.48, 11% becomes 1.11, and -2% becomes 0.98.

Now you multiply:

(15,000)(1.48)(1.3)(1.25)(1.12)(1.11)(1.06)(0.98)(1.04) = 48,452.1, or 48,452. (Since this is number of visits, it makes sense to round this to the nearest whole number. You can't have 0.1 visits.)

To find the average growth rate, divide your result by 15,000 (leaving only the combined growth rates): about 3.2301.

Now take the eighth root. You can do this easily in Excel using the ^ symbol (=3.2301^(1/8)). You should get about 1.1579. Subtract one and convert back to percent: 15.79%.

Once you've converted by adding one, you could also use the GEOMEAN function to multiply the values and take the nth root, resulting in 1.1579. Again, you'd have to subtract one to undo the manipulation from earlier.

21. Using the information below (the data are also available in the Practice Exam 1 Data Set), determine which stock has a more volatile price.[1] (When doing your calculations, round to the nearest cent.) Show your work.

Month (2014)	*Microsoft (Start of the Month)*	*Apple (Start of the Month)*
January	$35.99	$555.68
February	$37.74	$502.61
March	$37.92	$523.42
April	$41.15	$537.76

[1] Data from Yahoo! Finance; Apple issued a 7:1 stock split in June 2014 which is why its stock price is much lower now.

KEY

Show your work; if you used Excel to answer this question, write what you put into Excel so I know how you got the answer you did.

The first step is to determine the average stock price of each stock.

$$Microsoft = \frac{\$35.99 + \$37.74 + \$37.92 + \$41.15}{4} = \$38.20$$

$$Apple = \frac{\$555.68 + \$502.61 + \$523.42 + \$537.76}{4} = \$529.87$$

Or use the AVERAGE function.

Now, determine the standard deviation.

$$Microsoft = \sqrt{\frac{(\$35.99 - \$38.20)^2 + (\$37.74 - \$38.20)^2 + (\$37.92 - \$38.20)^2 + (\$41.15 - \$38.20)^2}{4 - 1}} = \$2.15$$

$$Apple = \sqrt{\frac{(\$555.68 - \$529.87)^2 + (\$502.61 - \$529.87)^2 + (\$523.42 - \$529.87)^2 + (\$537.76 - \$529.87)^2}{4 - 1}} = \$22.46$$

Or use the STDEV.S function.

It might appear that Microsoft is clearly more consistent but we can't tell for until we adjust for the average using the coefficient of variation.

$$Microsoft = \frac{\$2.15}{\$38.20} \times 100 = 5.63\%$$

$$Apple = \frac{\$22.46}{\$529.87} \times 100 = 4.24\%$$

It turns out Apple's stock price is more consistent.

KEY

22. As China grows wealthier, alcohol consumption is on the rise. (*Economist* "The Spirit Level")[2] China's alcohol consumption per person rose from 2.5 liters in 1978 to 6.7 liters in 2010. For the *Economist* article:

> But the countrywide statistics hide a grimmer picture. More than half the Chinese population [does not drink alcohol at all]. Those who do drink often do so to great excess. Male Chinese drinkers down far more than Japanese ones, and almost as much as notoriously sozzled British, Australian or Irish boozers.

Based on the information provided, what is the median alcohol consumption per person in China? What's advantageous about using the median here? What's disadvantageous about using it?

The median is zero—if more than half the population doesn't drink, then the median consumption will be zero liters per person. If you chose 6.7 liters per person, you reported the mean (total number of liters consumed divided by population), not the median.

The good part about using the median here is that it describes the typical person. Most people consume zero alcohol and the number would reflect that fact.

The deceptive part about using the median is that it would mask any *alcohol consumption. Drinkers drink a lot and you wouldn't get any inkling of that information if you focused just on the median.*

This is why reporting both the median and the mean are useful. If there's a big gap between the two, it tells you a lot about the distribution of the data.

KEY

[2] http://www.economist.com/news/china/21611118-chinese-are-drinking-more-spirit-level

Exam 1 Equation and Information Reference

Function	*Output*
ABS	The absolute value of an input
AVERAGE	Arithmetic mean of a dataset
CTRL + `	Show formulas
CTRL + F	Find
CTRL + P	Print
CTRL + X	Cut highlighted area
CTRL + C	Copy highlighted area
CTRL + V	Paste highlighted area
CTRL + Z	Undo
F4	Makes cell reference absolute
GEOMEAN	Geometric mean of a dataset (adjustments must be added manually)
LARGE	Larger values of a dataset (k=1 is largest, k=2 is second largest, k=3 is third largest…)
MAX	Maximum value of a dataset
MEDIAN	Median of a dataset
MIN	Minimum value of a dataset
MODE	Mode of a dataset
QUARTILE	The 0^{th} to 4^{th} quartile of a dataset
SQRT	Finds the square root of the value in question.
SMALL	Smaller values of a dataset (k=1 is smallest, k=2 is second smallest, k=3 is third smallest…)
STDEV.S	Standard deviation of a sample

Geometric Mean

$$Geometric\ Mean = \sqrt[n]{\prod_{i=1}^{n}(1 + x_i)} - 1$$

Weighted Average

$$Weighted\ Average = \frac{\sum_i^n (w_i x_i)}{\sum_i^n w_i}$$

Coefficient of Variation

$$CV_{sample} = \frac{s}{\bar{x}}(100)$$

KEY