

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

EXAM 1

- 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.
- Recall basic logic. “Water is wet” is a true statement. “Water is wet and leopards have stripes” is a false statement.
- *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 | A. Problem: will have error |
| 2. F Mean | B. Problem: when samples with unusual observations change |
| 3. B Median | C. Problem: additional calculation needed |
| 4. H Mode | D. Problem: impractical to get data for |
| 5. A Sample | E. Problem: when used to compare sample with very different averages |
| 6. E Standard deviation | F. Problem: samples with large outliers |
| 7. D Population | 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.

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?

<i>Number of Exterior Doors</i>	<i>Percent of Doors in the U.S.</i>
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. 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.

11. Use the Practice Exam 1 Dataset for this question. Which pair of variables is the most highly correlated?
- a. Number of competing stores in district & Annual profit, in thousands
 - b. Number of families in the sales area, in thousands & Annual profit, in thousands**
 - c. Number of competing stores in district & Advertising spent, in thousands
 - d. Square feet, in thousands & Annual profit, in thousands
 - e. None of the above

If you make a correlation table with Excel, you'll find this pair has a correlation coefficient of 0.95. Note option D is the second highest (0.89) and option third highest (-0.85). The higher the absolute value of the coefficient, the higher the correlation.

12. Use the Practice Exam 1 Dataset for this question. 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

13. 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.

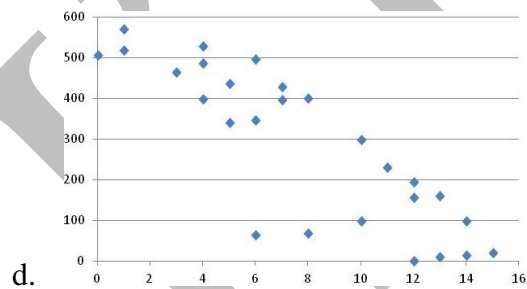
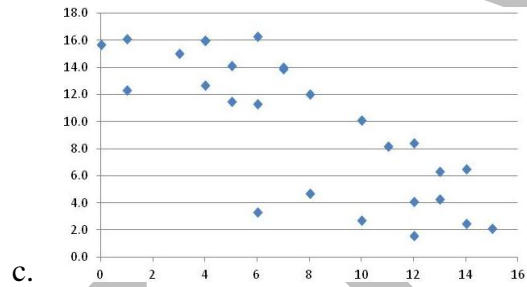
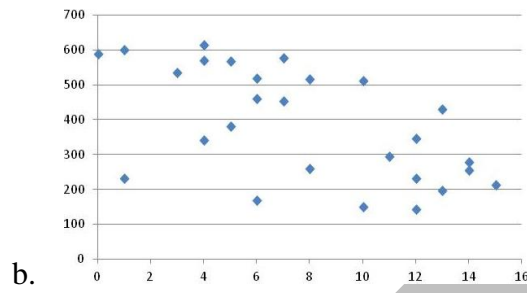
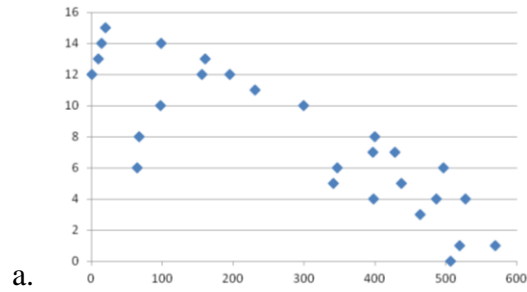
It's notable that the first option is also probably the most expensive thanks to the prize. But this prize incentivizes moderates to take the survey. (This is also why I give extra credit if you fill out an evaluation.) The metro also has a better chance of capturing a more accurate cross section of 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.

14. 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.

15. Use the Practice Exam 1 Dataset for this question. 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?



e. None of the above

Correct Answer

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.)

16. The percent of families who own their own home is positively correlated with average income across different U.S. states. Suppose a governor attempts to increase average income by subsidizing homeownership. If this is a mistake, what is the most likely reason?
- Reverse causation: home ownership is the result of high incomes.**
 - Reverse causation: something else is causing both variables.

- c. Confounding variable: home ownership is the result of high incomes.
- d. Confounding variable: something else is causing both variables.
- e. None of the above / The governor has the correct interpretation

Options (B) and (C) are interpreting reverse causation and confounding variables incorrectly. While (D) is possible, it is unlikely...what would that underlying variable possibly be? Interesting, E isn't a terrible choice—perhaps homeownership allows people to borrow more because the house is collateral and that might mean they go to college which means, later, they get more income. Maybe, but there's a lot of ifs, maybes, and conditions in that statement. Option (A) is the most likely answer.

17. 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.

18. Which of the following pair of variables is positively correlated?

- a. **Value of a car and the owner's income**
- b. Time spent taking a shower and the number of forks that person owns
- c. Frequency of rain and how often people play outside
- d. A & C
- e. None of the above

One would expect cheaper cars would be owned by people with low incomes and expensive cars would be owned by people with high incomes; that is positive correlation. Option C is an example of negative correlation: as it rains more, people will play outside less. Option B is an example of zero correlation. It's hard to imagine the number of forks someone owns would

influence or be influenced, directly or indirectly, by how long that person spends in the shower.

19. Suppose time spent playing video games and non-violent criminal activity are negatively correlated. While there might be a causation story (video games offer a safe outlet for criminal urges), a confounding variable could also be an explanation. Which of the following is the most likely reasonable confounding variable?
- a. Frequency of police patrols
 - b. Frequency of new video game releases
 - c. Weather**
 - d. A & B
 - e. None of the above

While you could argue all these of these are confounding variables, only option C is really strong. Both A and B suggest criminals might stop thieving in order to play games—thus the negative correlation—but that relies on the same people switching tasks.

Weather, however, doesn't require that assumption. If the weather is good, fewer people will play video games and more criminals will engage in illegal activity but these don't have to be the same people. The people could be shopping, surfing, going to bars, etc. And because the criminals are outside, they are not sleeping or watching TV and, yes, playing video games. Because C doesn't require that the gamers are the same folk as the criminals, C is the best option.

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 Dataset):

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

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. If you have a scientific calculator, this can be done by taking 3.2301 to the power of 1/8. But you can also do it with a basic calculator.

Recall that $x^{1/8} = ((x^{1/2})^{1/2})^{1/2}$. In other words, press the square root button three times. You should get about 1.1579. Subtract one and convert back to percent: 15.79%.

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

<i>Month (2014)</i>	<i>Microsoft (Start of the Month)</i>	<i>Apple (Start of the Month)</i>
January	\$35.99	\$555.68
February	\$37.74	\$502.61
March	\$37.92	\$523.42
April	\$41.15	\$537.76

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.

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.

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

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

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

Now, determine the standard deviation.

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

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

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.

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

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

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

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:

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

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.