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

EXAM 1

Practice B

- 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. **G** Accurate sample
2. **E** Cross-sectional data
3. **I** Element
4. **C** Gambler's fallacy
5. **F** Law of large numbers
6. **D** Hot-hand fallacy
7. **B** Time series data

A. Example: Historic unemployment rates for different countries.
B. Example: Historic unemployment rate for the United States.
C. Example: Lisbeth thinks that she's more likely to have a power outage simply because she hasn't had one in a while.
D. Example: Muhammad believes he won't get rear-ended by another car because he hasn't been rear-ended in years.
E. Example: Unemployment rates in 2012 for different countries.
F. Suggests a running average will tend to change less and less as each additional observation is included in the sample.
G. This means the statistic does not systematically bias the parameter.
H. This means the sample has a lot of observations.
I. What's referenced to determine a variable's value.

1. *A good sample is accurate: it doesn't under or over estimate the population parameter.*
2. *Cross-sectional data concerns the (roughly) same moment in time with the element being anything that can distinguish data at that moment, such as different countries. The variable—unemployment rate—would be composed of unemployment rates in different countries.*
3. *The element determines the variable's value for each observation. It's what the values have in common. For example, a data set with student as the element and the variables as GPA, major, and year, values of 2.56, Business, and freshman would all have one particular student in common.*
4. *Lisbeth believes that simply because she hasn't had a power outage, they are more likely to happen. Power outages are random events and just because you haven't had one for a while doesn't make it more likely you will.*
5. *The LLN states that as the sample size increases, the empirical average approaches the theoretical, or expected, average. If we calculate the average value as observations come in, the average should at first be very wild because each additional observation has a major impact on the average*

for the few observations we have. But as we have many observations, the average should "settle" to the theoretical average and each additional observation, regardless of what it is, shouldn't change the average a noticeable amount.

6. *Muhammad believes he's getting better at driving because he hasn't been rear-ended in a while. But there's very little a driver can do to prevent being rear-ended. He's just been lucky but he thinks that under-representation will persist in the future.*
7. *In a time series, time is the element. This is historic data; the unemployment rate for each observation is based on time.*

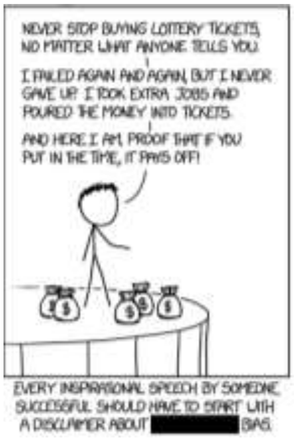
You might have selected A for either 2 or 7. It's kind of both, but it's kind of neither. It's not clear what's the element and what's the variable.

Part II: Multiple Choice. *Circle the best answer to the following.*
4 points each.

8. Imagine the average cost of living grew by 3% during 2015, 2% during 2016, and 4% during 2017. If the average cost of living was $30,000 at the end of 2014, what is the average cost of living at the end of 2017?
    a. $30,918.72
    b. $32,700.00
    c. **$32,778.72**
    d. $32,781.81
    e. None of the above

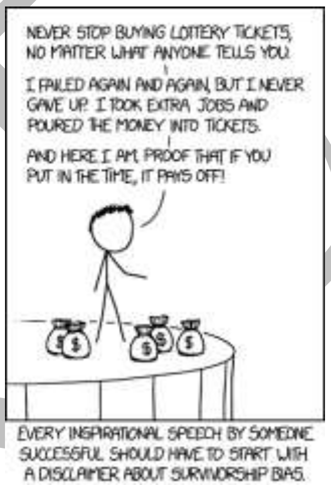
To calculate a growth rate, multiply the starting value by one plus each growth rate (expressed as a decimal). Thus $30,000(1.03)(1.02)(1.04) = $32,778.72. Remember, we add one because we want to include the starting value. A 3% growth rate results in an additional $900 added to the average of cost of living. Thus the new cost of living is $30,000 + $900 = $30,900. An additional 2% growth rate means not just the starting $30,000 increases by 2% but also the $900 from the previous year. Thus cost of living increased by $618; the new cost of living is $31,518. And so on.

9. Consider the following xkcd comic. I've blacked out the kind of bias mentioned in the caption. What kind of bias should go here?
   a. Self-selection
   b. Accuracy
   c. Under-coverage
   d. Precision
   e. None of the above


NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT [redacted] BIAS.

This is about survivorship bias. Many people try to be successful at very difficult things and they fail. They don't give speeches; who wants to hear about someone who started a company that went bust, forcing them to move back into their parents' house? No one, that's who. Inspirational speakers are always survivors but there's never any indication of how common survival is.

Here's the completed comic.


NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.

10. Suppose the average number of pounds of potatoes grown on five acres of land is exactly 50,000 per acre. If an additional acre of land is acquired and it produces exactly 50,000 pounds of potatoes, what happens to the average and standard deviation of potato production for this now six-acre farm?
    a. The average stays the same and the standard deviation decreases.
    b. The average stays the same and the standard deviation stays the same.
    c. The average stays the same and the standard deviation increases.
    d. The average either increases or decreases and the standard deviation stays the same.
    e. It is impossible to tell without knowing how many pounds each of the other acres produces.

Whenever you have an average and you add an observation which has a value equal to the average, the new average doesn't change. Why should it? The central tendency hasn't changed.

But the spread of the data <u>has</u> changed. You added an additional observation that's right on the mean. In the equation for standard deviation, that results in an additional zero in the numerator (because the observation minus the average is zero) which doesn't change the numerator but increases the number of observations by one, which causes the denominator to increase. Thus, the standard deviation falls.

11. In 2012, the average income in D.C. was about $164,000, over three times the U.S. average. But because of how the statistic is calculated, this number is misleading compared to other states. How?

    a. Because a lot of jobs are high-paying government jobs.
    b. Because a lot of people earn income in D.C. but don't live there.
    c. Because the average is thrown off by a few very high income people.
    d. Because there are many embassies and thus that income doesn't count as being earned "in" D.C.
    e. None of the above

Typically, states include major cities as well as much of all of the surrounding suburban area. Where people earn income is usually in the same place as where they live. Thus a state's average income (the total income earned in the state divided by the state's population) makes sense.

But the District is a strange observation compared to other states. It's largely a downtown area and excludes most of the suburbs. A lot of people earn income in D.C. (adding to the numerator) but don't live there (and thus don't add to the denominator), inflating the average.

You might want to have claimed C is the right answer; it's true that as an average, it will be thrown off by outliers. But outliers exist in all states and the question is what about the D.C. average, <u>compared to other states</u>, is misleading.

12. Use the Practice Exam 1 Data Set 2 for this question. It includes data from the Job Openings and Labor Turnover Survey (JOLTS) concerning the total number of hires and separations (divided into layoffs/discharges, quits, and other) per year and by sector (public or private). (See JOLTS Level tab.) Suppose you were curious which sector—public or private—had the most consistent number of people retiring (based on the coefficient of variation). While you can't directly measure it (retirement isn't an explicit variable), perhaps you might think you can get an idea of which is most consistent. What should you conclude?

    a. Private sector is more consistent, based on the quit rates.
    b. Public sector is more consistent, based on the quit rates.

c. **Private sector is more consistent, based on the other separation rates.**
d. Public sector is more consistent, based on the other separation rates.
e. Since "retirements" is not a listed variable, there's no way of even estimating which was most consistent.

This is a hard question because it requires not only how to interpret the coefficient of variation (lower numbers are more consistent) but also recognizing you should look at the definitions tab. Understanding what exactly your data represents is an important part of analysis.

In this case, retirements are not including in quits; they are treated as a part of "other separations." The other things included there—deaths, disability, and transfers—are not that common (with the possible exception of transfers) so it's reasonable that measuring "other separations" is a good estimation of retirements.

There's a little bit of disagreement possible here but that doesn't render E correct. While we might disagree on how good the "other" category is, E is purposely a strong statement; it's hard to argue that there's "no way to even estimate" what you're trying to measure.

13. Use the Practice Exam 1 Data Set 2 for this question. It includes data from the Job Openings and Labor Turnover Survey (JOLTS) concerning the total number of hires and separations (divided into layoffs/discharges, quits, and other) per year and by sector (public or private). (See JOLTS Level tab.) You can determine the Total Separations by adding together Layoffs/Discharges, Quits, and Other Separations. Total Hires – Total Separations = Net Jobs. Which year had the highest number of Net Jobs for the public sector?
    a. 2011
    b. 2014
    c. 2015
    d. 2016
    e. None of the above

This requires you to create a new column called Net Jobs. For 2008, it should be "=C4-(E4+G4+I4)". The highest number is 192, for 2016. If you picked 2011, you probably reversed your subtraction, resulting in Separations – Hires rather than the other way around.

14. Use the Practice Exam 1 Data Set 2 for this question. It includes data from the Job Openings and Labor Turnover Survey (JOLTS) concerning the total number of hires and separations (divided into layoffs/discharges, quits, and other) per year and by sector (public or private). (See JOLTS Level tab.) What's the average number of <u>total</u> hires from 2008 to 2017?
    a. About 51.9 thousand
    b. About 55.6 thousand
    c. About 51.9 million

d. **About 55.6 million**
e. None of the above

Adding the private and public hires together gets you the total number of hires: 54,763 for 2008, 46,191 for 2009, etc. The average is 55,566. But keep in mind that these values are in the thousands. The average number of hires is about 55,566,000, or about 55.6 million.

15. In 1970, 65% of primary school age girls worldwide were enrolled in school. In 2015, that number was 90%. Suppose you wanted to really show how dramatic that change is by truncating the axis. How might you do that?
    a. Set the x axis to 50%
    b. Set the y axis to 50%
    c. Set the x axis to 0%
    d. Set the y axis to 0%
    e. Set the y axis to -65%

If you're looking to highlight change over a period of time, you adjust the y-axis by having it start at a higher number, cutting of (or truncating) numbers between zero and the new starting number. By having it start at 50%, the graph will show a more dramatic change over time. You could even start it at 65% to show an even more dramatic change but then it would look like your number's coming from zero to the casual observer; that might be pushing the realm of honesty. Regardless, you should always clearly label your axis.

Note that setting an even lower number than the default (a negative number) would make the changes look less dramatic. That would be blatantly dishonest here if you were trying to argue that enrollment was basically flat. Why? Because you'd be including negative numbers in the graph and, of course, you can't have negative enrollment.

16. Which of the following is/are true?
    a. A weighted average is a normal average but with equal weights.
    b. The sample's standard deviation can be greater than the sample's mean.
    c. The geometric mean of a series of growth rates will overstate the true growth rate.
    d. A & B
    e. All of the above are true

A is true because the sum of observations divided by n is the same as 1/n multiplied by each observation and then summed. Since 1/n will never be larger than one and 1/n added together n times equals 1, you can think of each 1/n as the weight for each observation. Since 1/n is a constant, all observations have equal weight. In other words,

$$\frac{\sum_i^n (x_i)}{n} = (\frac{1}{n}) \sum_i^n (x_i) = \sum_i^n \left((\frac{1}{n})(x_i) \right)$$

To understand why B is correct, imagine any two numbers. The average of the two is naturally halfway between them. If you increase the larger number by X and, at same time, decrease the smaller number by X, the average hasn't changed. The number of observations haven't changed. But the distance between the observation and the average has increased, thus standard deviation increased. Since there's no limit to X, there's no limit to how big standard deviation can be, holding average constant. Thus, standard deviation can be larger than the mean.

C, however, is not true. The geometric mean is exactly what you should use for determining the average growth rate. Using the arithmetic mean will result in overstating the average growth rate.

17. According to the Pew Research Center, adults who don't participate in elections (nonvoters) tend to be poorer, younger, more racial diverse, and less educated compared to adults in general. If we think of the people who voted as a sample of what the adult population wants, is there anything wrong with this sample? Why or why not?
    a. Yes; it is not precise.
    b. Yes; it is not accurate.
    c. Yes; it is neither precise nor accurate.
    d. No; though it's over-represented in some areas, it's under-represented in others and the effects cancel out.
    e. No; these factors don't matter.

This sample is systematically different from the population it's supposed to represent. Age, income, education, and racial background clearly play a role in how people feel about politics and so the people who are making the decisions (the voters) are not reflective of who they are, in essence, representing.

You could, in theory, argue that (e) is correct. If I gave you data that nonvoters were, say, more likely to be redheads and liked spaghetti, then that would be a reasonable answer. It's hard to see how either of these differences would impact electoral preferences.

By the way, this is an example of self-selection bias.

18. "Freemium" software is free to use but you can pay to have an enhanced experience. Most users pay nothing and a few users pay a lot. Describe the median and the mean revenue per user.
    a. The median is zero and the mean is more than zero.

b. The mean is zero and the median is more than zero.
c. The mean is higher than the median; neither are zero.
d. The median is higher than the mean; neither are zero.
e. It is impossible to tell given the information provided.

If most users pay nothing, then your typical user, you median user, pays nothing. The median is zero. But the mean would be greater than zero because it'll be influenced by the minority of users who pay. In fact, given how much they pay, it might be much larger than zero.

19. The chance that any one driver will be in an accident is very hard to predict. But insurance companies can reasonably predict how many drivers they cover will be in an accident. What's this an example of?
    a. Central tendency
    b. Range
    c. Sampling bias
    d. Sampling accuracy
    e. None of the above

This is an example of the law of large numbers. A single random event is hard to predict, much like the spin of a roulette wheel. But you can reasonably predict, based on 1,000 spins, how many times the result will be black or red or a particular number. Similarly, insurance companies have reasonable estimates on how many drivers will get into accidents, how many policy holders' homes will be robbed, and so on. As the number of observations increase, the observed mean approaches the theoretical (or expected) means.

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

20. Elona works at a major bank with many kinds of loans. She wants to know what the average rate of return the bank has across all kinds of loans. Here's the total assets the bank has in each type of loan, the percent of their total assets in each type of loan, and the rate of return for each loan type. Determine the average rate of return for the whole bank.

Loan Type	*Total Assets in this Loan Type ($B)*	*% of Total Assets in this Loan Type*	*Rate of Return of this Loan Type*
Mortgages	72	36%	3.5%
Consumer durables	24	12%	5.1%
Education	10	5%	5.8%
Small businesses	16	8%	5.6%
Big businesses	28	14%	3.1%
Government	50	25%	1.2%

This is a little tricky because you have two percent value: rate of return and percent of total assets. Which is your weight? What is your value the weight will be multiplied by?

Remember, Elona's trying to determine the average rate of return. Much like if she was calculating average grade, she'd look at individual grades, she will look the rates of return. Despite it being a percent, these are not the weights. These are what she will multiply the weights by.

If she found the simple average, she'd find the average rate of return is 4.1%. But that value treats all rates of return the same. That's odd because while education has the highest rate of return (5.8%), it's the smallest part of the company's portfolio (5%). In other words, the rates of return should be weighed by how much the company has invested in each kind of loan.

You can find this two ways. The first is to multiply each % of Total Asset value by the corresponding rate of return and then adding the products up. (SUMPRODUCT is helpful for this.) You'd get 3.344% (we have to also divide by the total of all weights but that's 1 so nothing changes).

% of Total Assets in this Loan Type	*Rate of Return of this Loan Type*	*Product*
36%	3.5%	1.260%
12%	5.1%	0.612%
5%	5.8%	0.290%
8%	5.6%	0.448%
14%	3.1%	0.434%
25%	1.2%	0.300%
100%	**<--------Sum of values ---------->**	**3.344%**
	3.344%/1=	**3.344%**

Or you can multiply the rates of return by the corresponding amount of money in each loan type, adding up the values, and then divided by the sum of the assets. The numerator is 6.688, the denominator is 200 and 6.688/200 = 0.03344, or 3.344%.

Total Assets in this Loan Type ($B)	*Rate of Return of this Loan Type*	*Product*
72	3.5%	252.000%
24	5.1%	122.400%
10	5.8%	58.000%

16	5.6%	89.600%
28	3.1%	86.800%
50	1.2%	60.000%
200	**<--------Sum of values ---------->**	**668.800%**
	668.800%/200=	**3.344%**

21. Use the Practice Exam 1 Data Set 2 for this question. It includes data from the Job Openings and Labor Turnover Survey (JOLTS) concerning the growth of hires and separations (divided into layoffs/discharges, quits, and other) from the previous year by sector (public or private). (See JOLTS Growth tab.) Determine the average growth rate for private sector layoffs/discharges. Express your answer as a percent to two decimal places (e.g. 45.62%).

Since this is a growth rate, it's best to not use the arithmetic mean but the geometric mean. For column D, add one to every observation, then use the GEOMEAN function then subtract one from the result. You should get -1.82%.

Remember, you can use SUMPRODUCT to combine the adding one step and the GEOMEAN step as so: =SUMPRODUCT(GEOMEAN(1+D4:D12))-1

In case you are curious, here are the average growth rates for all variables.

Hires		**Layoffs/Discharges**		**Quits**		**Other Separations**	
Private	*Public*	*Private*	*Public*	*Private*	*Public*	*Private*	*Public*
1.97%	1.87%	-1.82%	1.72%	2.56%	2.65%	0.73%	1.82%

22. A metropolitan statistical area (MSA) is an economically integrated geographic definition with an urban core. The ten largest MSAs in the U.S. are indicated in the table below, along with the population density (people per square miles), average income, and population. (Because of the awkwardness of how the Census defines L.A.'s MSA, I've left out its population density.)

Use weighted average to determine the average density and average income for the top ten MSAs in the U.S.

Metro Statistical Area	*Density (sq mi)*	*Average Income*	*2010 Census*
New York-Newark-Jersey City	1,781.30	68,525	19,567,410
Los Angeles-Long Beach-Anaheim		59,441	12,828,837
Chicago-Naperville-Elgin	1,318.00	56,423	9,461,105
Dallas-Fort Worth-Arlington	634.00	58,744	6,426,214

Houston-The Woodlands-Sugar Land	630.30	67,746	5,920,416
Washington-Arlington-Alexandria	1,084.00	76,712	5,636,232
Miami-Fort Lauderdale-West Palm Beach	1,004.00	45,041	5,564,635
Philadelphia-Camden-Wilmington	2,746.32	58,463	5,965,343
Atlanta-Sandy Springs-Roswell	631.18	51,993	5,286,728
Boston-Cambridge-Newton	1,375.00	72,494	4,552,402

A lot's going on here; let's begin by understand why we're using a weighted average. If we were to take a simple average—adding all the incomes and then dividing by ten—we'd treat New York high income with the same weight as Dallas-Fort Worth's lower income. That's strange because there are three times as many people in New York. We shouldn't give these values equal weight; New York should have more weight because it has more of what matters: people in this case.

So one easy first step is to determine the percent of top ten MSA dwellers live in each MSA. Adding up all the listed MSAs results in 81.2 million people. New York's population divided by 81.2 million results in about 0.241, or 24.1%; 24.1% of people who live in one of the top ten MSAs live in New York. This gives us our weight. We can repeat this for each MSA, as so:

Metro Statistical Area	*Density (sq mi)*	*Average Income*	*2010 Census*	*Pop %*
New York-Newark-Jersey City	1,781.30	68,525	19,567,410	24.1%
Los Angeles-Long Beach-Anaheim		59,441	12,828,837	15.8%
Chicago-Naperville-Elgin	1,318.00	56,423	9,461,105	11.7%
Dallas-Fort Worth-Arlington	634.00	58,744	6,426,214	7.9%
Houston-The Woodlands-Sugar Land	630.30	67,746	5,920,416	7.3%
Washington-Arlington-Alexandria	1,084.00	76,712	5,636,232	6.9%
Miami-Fort Lauderdale-West Palm Beach	1,004.00	45,041	5,564,635	6.9%
Philadelphia-Camden-Wilmington	2,746.32	58,463	5,965,343	7.3%

Atlanta-Sandy Springs-Roswell	631.18	51,993	5,286,728	6.5%
Boston-Cambridge-Newton	1,375.00	72,494	4,552,402	5.6%

Now it’s a simple matter of multiplying each percent value (the weight) by average income and then adding these values together. Note that the weights add up to one so no further division is needed.

If you put the Pop % column in Column E, =SUMPRODUCT(C2:C11,E2:E11) results in the weighted average income of $62,215.59

Density’s a little trickier; I’ve excluded L.A.’s density because the Census doesn’t include the greater LA area here (Greater Los Angeles is actually a combination of MSAs). (Yes, if I excluded density because the MSA is incomplete, I should have also excluded average income for the same reason but this is just for practice; I wanted one variable that had all the values and one that didn’t.)

Let’s start by doing the same thing as before but now using the density column: =SUMPRODUCT(B2:B11,E2:E11) which results in 1,142.81. However, we don’t have the full picture because LA is not included. We’re not working with 100% of the data we care about; we’re only working with 84.2% of the observations. The remaining 15.8% are in LA.

We thus divide by the sum of the weights which is not one but 0.842. Using =SUMPRODUCT(B2:B11,E2:E11)/(E2+SUM(E4:E11)) results in an average population density of 1,357.21 people per square mile.

Equation and Information Sheet

Function or Command	*Result*
ABS	Returns the absolute value of an input
AVERAGE	Arithmetic mean of an array
CTRL + `	Show formulas
CTRL + F	Find
CTRL + P	Print
CTRL + X	Cut highlighted area
CTRL + C	Copy highlighted area
CTRL + V	Past highlighted area
CTRL + Z	Undo
F4	Makes cell reference absolute
GEOMEAN	Geometric mean of an array (adjustments must be added manually)

LARGE	Larger values of an array (k=1 is largest, k=2 is second largest, k=3 is third largest…)
MAX	Maximum value of an array
MEDIAN	Median of an array
MIN	Minimum value of an array
MODE	Mode of an array
QUARTILE	Returns the 0th to 4th quartile of an array
SQRT	Finds the square root of the value in question.
SMALL	Smaller values of an array (k=1 is smallest, k=2 is second smallest, k=3 is third smallest…)
STDEV.S	Standard deviation of a sample
SUMPRODUCT	The summed product of two or more arrays.

Geometric Mean

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

Weighted Average

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

Standard deviation of a sample

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Coefficient of Variation

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