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

EXAM 4

Practice A

- There are 110 possible points on this exam. The test is out of 100.
- You have two hours 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. **E** Binomial distribution
2. **J** Central Limit Theorem
3. **G** Median
4. **M** Poisson distribution
5. **A** Population
6. **C** Self-selection bias
7. **H** Standard deviation
8. **B** Survivorship bias
9. **L** t score
10. **K** z score

A. All possible objects of interest
B. Example: Comparing average newspaper revenues in 1990 and average revenues in 2015.
C. Example: Examining average test scores in a country where testing is voluntary
D. Example: Having all students in a remedial math class take a survey that's meant to represent the opinions of all students.
E. Example: Likelihood that 10 out of 50 randomly chosen buses, all in different cities, will be late; assume 400 out of 10,000 buses are late.
F. Example: Probability that one of ten very different companies will perfect a new treatment for cancer this year
G. If an odd number of observations are all either very high or very low, this will be either very high or very low.
H. If an odd number of observations are all either very high or very low, this will be very high.
I. If an odd number of observations are all either very high or very low, this will be somewhere between the observations.
J. Main reason for why we never say an analysis "proves" something
K. Use when you do know σ
L. Use when you don't know σ
M. Used if wondering the probability of ten neighborhood burglar alarms going off in a week.

1. *The presence of a number of successes in the population (400) and the size of the population (10,000) makes it look like hypergeometric but remember the key distinction between hypergeometric and binomial distribution: constant probability. There's no reason to think that the chance of a bus being late in one city will influence the chance of being late in another city. We could easily restate the above information to say there's a 4% chance of a bus being late.*

Normally, F would also be binominal but because the companies are very different, the probability of success is not constant; some firms will surely have a higher chance of developing the treatment than others.

2. *The CLT informs us that any sample average might be unusual due to chance. You can have evidence for something, but you can't prove it.*
3. *Because the sample size is odd, the median will be either very high or very low. If it was an even size, it's possible the two middle numbers will split—one high and one low—thus the middle number would be the average of the two. In the case of a dummy variable, it would be 0.5.*
4. *Any number of alarms could go off in a week; since this is a defined interval (in this case of time), use the Poisson distribution to determine the probability that exactly ten alarms go off.*
5. *The population is what you're studying, what you're trying to learn more about. It could be "all Montgomery College students that are business majors," "all Montgomery College students," "all current U.S. students," "all future and current U.S. students," and so on.*
6. *Like the SAT issue, only the students who think they will do well will take the test, causing the average test score to increase.*
7. *Standard deviation is the spread of the data; if all observations are either high or low, this will be big because no observation is near the mean. Of course, this assumes there's a relatively equal spread distribution between high and low values. If 99% were high and 1% were low, standard deviation would be low.*
8. *Only the newspapers that didn't go out of business between 1990 and 2015 will be counted in the second average. The average could actually go up BECAUSE newspapers have had such a hard time.*
9. *t is used when you don't know sigma, or the population standard deviation.*
10. *z is used when you do.*

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

11. A test that's 100% specific and 0% sensitive:
    a. Will *never* have a false positive
    b. Will *never* have a false negative
    c. Will *never* have a true positive
    d. A & C
    e. B & C

Such a test will always be able to detect a negative state; therefore, all of its negatives will be true. There will be no false positives. Such a test will also never correctly detect a positive state. It will never get a true positive.

In fact, all results will come back negative. It will either correctly identify a negative state, calling it negative (thus no false positives), or it will incorrectly identify a positive state, still calling it negative (thus no true positives). It will always give back negative results. It's a stupid test; it tells you nothing.

12. Which of the following causal claims is better described as reverse causation?
    a. Education causes income
    b. Sales cause advertising
    c. Size of house causes quality of car
    d. B & C
    e. None of the above

Option A makes sense; option B is reverse causation, option C is confounding variable: income is causing both.

13. What does R^2 represent?
    a. The fraction of the variation the regression explains
    b. Regression Sum of Squares divided by Error Sum of Squares
    c. Regression Sum of Squares divided by Total Sum of Squares
    d. A & C
    e. None of the above

Option C is true by definition. Option A is what that means: the proportion of the variation of the data the model explains.

14. Integrated Systems manufactures internet modems. On average, the modems download one MB in 2.2 seconds with a standard deviation of 0.3 seconds. Assume modem performance follows a normal distribution. Suppose Integrated Systems wants to offer a warranty for underperforming modems. The minimum quality should result in replacing no more than 12% of modems. How many seconds should that minimum quality be?
    a. 0.6474 seconds
    b. 1.8474 seconds
    c. 2.5525 seconds
    d. 5.7250 seconds
    e. None of the above

In this example, less is better and more is worse. That means you want the top end of the distribution. Type "=NORM.INV(0.88,2.2,0.3)" and you should get 2.5525. (Note that we rounded up because if we rounded down, we'd be replacing more than 12% of modems.)

15. Jason works quality control in a textiles factory. His job is to reject any fabric with more than one error per yard of fabric (the average under normal conditions). Suppose there's a

1% chance of getting an error. Jason wants to know if the textiles machinery is working properly and wonders how likely it would be to find three or more errors in one yard. What is the chance that a yard of fabric would have three or more errors if the machine is working properly?

a. 0.08
b. 0.18
c. 0.92
d. 0.98
e. None of the above

Poisson is about the number of events over a defined interval. While usually that interval is time, it doesn't have to be. In this case, the interval is a yard of fabric. Note a defining feature of Poisson—the potential number of events is infinite—applies here as well. There's no (inherent) limit to the number of errors on a yard of fabric.

The Excel command =POISSON.DIST(2,1,1) will give you the chance of getting two or fewer errors. 1 minus that is about an 8% chance of getting three or more.

16. It's conventional wisdom that prevention is always better than a cure because it's often cheaper to prevent a bad thing from happening than correcting a bad thing after it happened. But the Learned Hand Rule suggests that's not true. How?
    a. A cost for non-problems must also be incurred; that's why p matters.
    b. Burden only matters when there's negligence.
    c. The expected value is often negative.
    d. Sensitivity is greater than specificity.
    e. None of the above

When people argue how great prevention is, they often forget that the cost of prevention must be incurred even when no problem happens. Consider a train yard. Whenever a train moves, there is a chance that someone is underneath it doing maintenance work.

17. A risk loving person _____.
    a. ***Always*** chooses the riskiest option.
    b. Would ***never*** pick the same option as a risk averse person.
    c. Prefers a 10% chance to win $50 than a five-dollar bill.
    d. A & B
    e. All of the above

This is true by definition; when an expected value of an uncertain payoff equals the value of a certain payoff, a risk loving person prefers the uncertain payoff.

But keep in mind that "risk loving" is a spectrum. If the certain payoff is greater than the expected value of an uncertain payoff (say, it was a 1% chance of winning $50 instead of a 10% chance), a risk loving person may still decide to go with the certain payoff. Only when the expected value equals the certain payoff can we truly distinguish the risk loving from the risk averse and the risk neutral.

18. Suppose you ran a regression and found heteroscedasticity. What should you do first?
    a. Start over with new variables
    b. Drop as few variables as possible to remove the heteroscedasticity
    c. Check how significant your explanatory variables are
    d. A & B
    e. None of the above

Heteroscedasticity doesn't influence the significance too much. If your variables are strongly significant, at this level we can assume you'll still be fine. (At a more advanced level, we could employ robust standard errors to fix that…but Excel isn't capable of that correction to my knowledge.)

19. The ***most important*** difference between a binomial distribution and a hypergeometric distribution is based on what?
    a. If the size of the sample relative to the population is large or not.
    b. If the standard deviation of the population is constant or not.
    c. If the interval between events is constant or not.
    d. If the chance of success is high or not.
    e. None of the above

The most important difference is that in a hypergeometric, the probability of success is not constant while in a binomial distribution, it is constant (or close enough to being constant).

20. Use the Practice Final Exam Data Set 1 to answer this question. It is hypothetical data on a hypothetical grocery store chain called The Happy Spud, with each observation a particular location. Using the coefficient of variance, determine which variable is most consistent.
    a. Annual profit
    b. Square feet
    c. Inventory value
    d. Advertising spent
    e. None of the above

You might be tempted to claim Square feet is the most consistent because its standard deviation is the lowest. But its average is also the lowest, too. You need to adjust for that average using the coefficient of variation: standard deviation divided by

average, times 100. With just 37.01, Advertising spent is the most consistent variable.

	Annual profit	*Square feet*	*Inventory value*	*Advertising spent*	*Number of competing stores*	*Number of families in sales area*
Average	*286.57*	*3.33*	*386.56*	*8.36*	*7.81*	*9.69*
Std Dev	*192.06*	*2.01*	*159.42*	*3.09*	*4.39*	*5.14*
CoV	*67.02*	*60.47*	*41.24*	*37.01*	*56.23*	*53.03*

21. Use the Practice Final Exam Data Set 1 to answer this question. Create a dummy variable called East? to indicate which region the store is in (1=East, 0=West). Then run a regression with East? and Number of competing stores in district predicting Annual profit. At 95% confidence, what is the result of the East? variable?
    a. It's statistically significant, with being located in the East reducing profits by 89.2%.
    b. It's statistically significant, with being located in the East increasing profits by 89.2%.
    c. It's statistically significant, with each additional dollar in profit reduces the chances of being located in the East by 89.2%.
    d. It's not statistically significant.
    e. None the above

First, make a dummy variable in Column F, because it'll need to be next to the other independent variable, Number of competing stores in district. A regression should result in a coefficient of about -89.2 with a p-value of 0.028. The low p-value (lower than 0.05) indicates it's statistically significant at the 95% level.

A is close to the right answer but it's not a percent. The second half the punchline should be in terms of the whatever the dependent variable is: thousands of dollars in profit. Being located in the East (a 1 rather than a 0) reduces profits by $89.2 thousand, not 89.2%.

22. Ursula works for CBS. Suppose CBS changed its time slot of a long running show, *The Amazing Race*, to Fridays. Suppose in previous seasons, the show average 8 million viewers with a population standard deviation of two million (based on over 200 episodes). The new time slot for the latest season—based on nine episodes—averaged 6.5 million episodes. Does changing the time slot result in a statistically significant change in viewership?
    a. At the 95% level, yes.
    b. At the 99% level, yes.

c. A & B
d. There is not enough information to determine an answer.
e. None the above; it's not statistically significant at all.

*This is a test for significance and with so many episodes, it's probably a z-test; in other words, we know the population standard deviation. (And note that the standard deviation of the sample was not given and that two million was explicitly referred as the **population** standard deviation.)*

This is also a two-tailed test—ratings could be higher or lower—so our critical values are 1.96 (95%) and 2.576 (99%). Now we calculate:

$$z = \left|\frac{6.5 - 8}{2/\sqrt{9}}\right| = \left|\frac{-1.5}{2/3}\right| = |-2.25| = 2.25$$

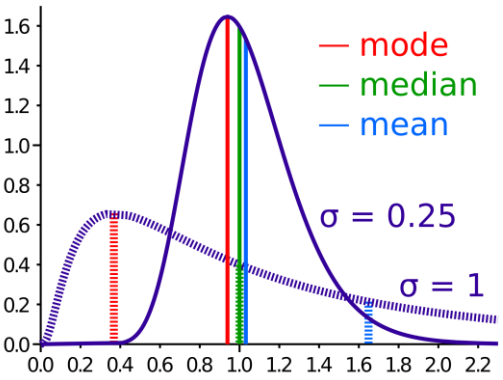
This value is greater than the critical value at 95% confidence but not at 99% confidence.

23. George works in a deli slicing meats using a special machine. This machine can slice meat very thinly (to maximize the flavor). Too thick and it wastes food. But it can never be too thin; thinner is always better. Suppose George wants to test if the machine is working properly. Normally it slices meat 0.5 millimeters thick. What is George's null hypothesis?
    a. $\mu < 0.5$ mm
    b. $\mu \leq 0.5$ mm
    c. $\mu = 0.5$ mm
    d. $\mu \geq 0.5$ mm
    e. $\mu > 0.5$ mm

Though tests for machinery are usually two-tailed tests, this is a special case. The machine can only make one kind of error: slice it too thick. Remember, thinner is always better. Therefore, this is a one-tailed test. Usual performance would be 0.5 mm or less.

24. If a distribution is positively skewed, which of the following is true about the central tendency?
    a. The median is higher than the mean
    b. The mean is higher than the median
    c. The mode is higher than the median
    d. B & C
    e. None of the above

A positive skew means there is a long tail on the right: a few extreme values on the high side. This brings up the average compared to the median. The mode is actually lower than the median when the distribution is skewed. Imagine you have a normal distribution that's symmetric. Now imagine you add several high values. That brings up the mean and it brings up the median, but the mode is unaffected. Here's an illustration with two right-skewed distributions:



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

25. Yolanda Jade is testing if any of her employees might sell company secrets to a competitor. She knows from industry analysis and historical evidence that 0.3% of employees will betray the company for money (we'll call them "traitors"). Suppose Yolanda develops a series of questions to test for treachery. It is 99% sensitive and 98% specific. If a result comes back positive, what is the chance that the employee is actually a traitor?

    Here, we use Bayes' Theorem:

$$P(T|+) = \frac{0.99 * 0.003}{0.99 * 0.003 + 0.02 * 0.997} = \frac{0.00297}{0.00297 + 0.01994} = \frac{0.00297}{0.02291} = 0.1296$$

    There is a 12.96% chance of a positive result is actually traitor.

26. Alfonso works for a fruit company. He's in charge of quality control for bananas. It's too expensive to test every banana bunch in a crate so he requires his fellow workers to select a sample. Suppose he has them selects three banana bunches from each crate containing six banana bunches. Suppose, in one instance, the chosen box has three bad banana bunches. What is the probability that the sample from that crate will have exactly two bad banana bunches? Be sure to include any commands you put into Excel.

Here, "success" is defined as getting a bad bunch. Since the chance of success changes with each trial (three trials but only six bananas), we use hypergeometric. Let's make a list of what we know.

N = 6 (bunches in the crate)
R = 3 (bad bunches)
n = 3 (number of bunches being pulled)

x = 2 (successes in question)

Here's what you should have put in Excel:

=HYPGEOM.DIST(2,3,3,6)

For a result of 0.45

If you want to write out the equation by hand, here's what you would have wrote (we didn't get into the actual equations behind the Excel commands but just so you appreciate all that the program is doing for you):

$$\mathrm{P}(2,3) = \frac{\left(\frac{(6-3)!}{((6-3)-(3-2))!\,(3-2)!}\right)\left(\frac{3!}{(3-2)!\,2!}\right)}{\left(\frac{6!}{(6-3)!\,3!}\right)} = \frac{\left(\frac{3!}{(3-1)!\,1!}\right)\left(\frac{3!}{1!\,2!}\right)}{\left(\frac{6!}{3!\,3!}\right)} = \frac{\left(\frac{3!}{2!}\right)\left(\frac{3!}{2!}\right)}{\left(\frac{6!}{3!\,3!}\right)}$$

$$= \frac{(3)(3)}{\left(\frac{(6)(5)(4)}{(3)(2)(1)}\right)} = \frac{9}{(2)(5)(2)} = \frac{9}{20} = 0.45$$

Regardless, you have a 45% of getting exactly two bad bunches.

27. Justin is venture capitalist. He's thinking about backing a new video game with a radically unusual approach. Suppose there is a 60% chance of getting the game to work correctly. If it doesn't work, Justin would lose the $10 million he invested. If it works, the profit Justin can make on this investment is determined by how successful the game is (see table). Calculate Justin's expected value of this deal. Remember to show all work.

Market Performance	*Probability*	*Profit*
Good	50%	$20 million
Average	30%	$8 million
Bad	20%	$3 million

We begin by recognizing that Justin has a 40% chance of losing $10 million, or -$4 million. All other probabilities are multiplied by 60%.

- *(0.60)(0.50)($20 million) = (0.3)($20 million) = $6 million*
- *(0.60)(0.30)($8 million) = (0.18)($8 million) = $1.44 million*
- *(0.60)(0.20)($3 million) = (0.12)($3 million) = $0.36 million*

So -$4 + $6 + $1.44 + $0.36 = $3.8 million.

28. Pepper's testing a new workout routine he's developed to see if it's better than the standard one. The standard workout routine results in 5.1 pounds lost on average over the course of a week. It has a population standard deviation of 1.7 pounds. Using a sample of 25 people, Pepper finds that his results in 5.9 pounds of weight loss over a week with the same number of hours per day. Is his results statistically significant at the 95%, 99%, and/or 99.9% level?

    If you want to stretch yourself, determine the p-value, to four decimal places. Make sure to show your work and report the relevant critical z-score and calculated z-score and how you found the p-value. (HINT: You will need Excel to find the p-value.)

First, note this is a one-tailed test. This is also a z-test because we know the population standard deviation.

So we do some math:

$$z = \left|\frac{5.9 - 5.1}{1.7/\sqrt{25}}\right| = \left|\frac{0.8}{0.34}\right| = 2.35$$

The critical value for a one-tailed test at 99% confidence is 2.326; at 99.9%, it's 3.09. It is significant at 99% (and thus 95% and 90%), but not at 99.9% because 2.35 > 2.326.

To determine the p-value, we use the standard normal distribution function: NORM.S.DIST such that =NORM.S.DIST(2.352941) is put into Excel. This tells us the area under the curve at just over 2.35 standard deviations. This will give us about 0.99068. To get the tail, we subtract this value from 1: about 0.00932. This is the p-value.

Note would also would have gotten the p-value by putting in =NORM.S.DIST(-2.352941) as it would be the same distance but in the opposite direction.

Note that this p-value is less than 0.01 but greater than 0.001, as predicted.

Exam 4 Equation and Information Reference

Function	*Output*
ABS	The absolute value of an input
AVERAGE	Arithmetic mean of a dataset
BINOM.DIST	Binominal distribution for x number of successes
CONFIDENCE.NORM	Determines the margin of error to make a confidence interval (known σ)
CONFIDENCE.T	Determines the margin of error to make a confidence interval (unknown σ)
CORREL	Correlation coefficient of two variables
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)
HYPGEOM.DIST	Hypergeometric distribution for x number of successes
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
NORM.DIST	Returns the normal distribution for a specified mean and standard deviation.
NORM.INV	Returns the inverse of the normal cumulative distribution for a specified mean and standard deviation.
NORM.S.DIST	Returns the standard normal distribution. Can also be used to find the critical z scores.
NORM.S.INV	Returns the inverse of the standard normal cumulative distribution. Useful for finding critical z scores.
POISSON.DIST	Poisson distribution for x number of successes
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
T.INV	Finds area under a t distribution; useful for finding one-tailed critical t scores.
T.INV.2T	Finds area under a t distribution; useful for finding two-tailed critical t scores.
T.TEST	Various two population tests which use a t score.

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

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

Confidence interval for proportion

$$\widehat{CI}_{\bar{p}} = \bar{p} \mp z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|{\sim}A)P({\sim}A)}$$

Learned Hand Formula

$$B < pL$$

Binominal Distribution

$$\mu = np$$

Hypergeometric Distribution

$$\mu = \frac{nR}{N}$$

Poisson

$$\mu = \lambda, \sigma = \sqrt{\lambda}$$

z-test

$$z_{\bar{x}} = \left|\frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}}\right|$$

t-test

$$t_{\bar{x}} = \left|\frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}}\right|$$

z-test (proportion)

$$z_p = \left|\frac{\bar{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right|$$

Critical z scores

Use =NORM.S.INV command

Confidence	*α*	$z_{\alpha/2}$	z_{α}
95%	5%	1.960	1.645
99%	1%	2.576	2.326
99.9%	0.1%	3.291	3.090

Critical t scores

Use T.INV or T.INV.2T commands

Adjusted R^2

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$