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

EXAM 4

Practice #2

- 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. **H** Binominal distribution
2. **G** Correlation coefficient
3. **F** Expected value
4. **L** Hypergeometric distribution
5. **D** Learned Hand formula
6. **K** Margin of error
7. **M** Practically significant
8. **J** Standard deviation
9. **E** Standard normal distribution
10. **B** Statistically significant

A. A stronger correlation means this gets bigger
B. Conclusion if p-value is really small
C. Conclusion if p-value is really big
D. Considers both expected cost and ability to avoid danger to determine negligence
E. Forms the basis of critical z values
F. Mathematically shows that a rare and expensive event equals a common and cheap event
G. Never greater than 1 nor less than -1
H. Probability is constant
I. Probability is sometimes less than zero
J. Sample's value means you use the t distribution; population's value means you use the z distribution
K. Subtracted from and added to the mean
L. The number of successes is bounded by the number of trials
M. When a genuine difference is very large

1. *A constant probability across trials is the defining feature of a binomial distribution. It is **not** that there are only two outcomes (hypergeometric also only has two outcomes).*
2. *The correlation coefficient cannot be outside the boundaries of 1 and -1 as 1 represents perfect positive correlation and -1 represents perfect negative correlation.*
3. *Calculating expected value requires multiplying probability and payoff. Note that if probability is low and payoff is high, that'll have the same result as a higher probability and a proportionally lower payoff. $1,000 times 0.5% equals $10 times 50%.*
4. *There cannot be more successes than trials because a trial represents a chance of success. Note that answer also applies to binominal distributions but since only "probability is constant" applies to binominal distributions, this answer must go with hypergeometric distribution.*
5. *The Learned Hand formula is an application of expected value, or expected cost. By considering both the expected cost of an accident (pL) and how easy it*

is to avoid the accident (B), courts can use the concept as a guide to determine if someone didn't take enough care to avoid the bad thing.

6. *Excel's CONFIDENCE function produces the margin of error, not the confidence interval itself. Recall a confidence interval must be composed of two numbers: an upper bound and a lower bound. To find the interval, add and subtract the margin of error from the mean.*
7. *The word "genuine" is important here. Only statistically significant differences can be practically significant. The vagueness of "large" is also important; what exactly constitutes a practical difference depends on unique circumstances.*
8. *If you have the sample standard deviation (and only the sample standard deviation), then use the t distribution. But if you have population standard deviation, always use the z distribution. The z distribution is more informative.*
9. *With a mean of zero and standard deviation of one, the standard normal distribution is where critical z values come from. A -1 is one standard deviation below the mean of zero. That's why critical (two tailed) values start at about 2, when, by the Empirical Rule, we incorporate 95% of the observations. Anything beyond 2 (technically 1.96) is outside that 95% range and thus statistically significant.*
10. *While calculated t and z values much be large to be considered statistically significant, p-values must be small to be statistically significant. Remember what p-values are: the smallest possible alpha you can claim and still claim statistical significance. The smaller the p-value, the smaller the alpha and thus the larger the confidence level.*

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

11. Students with more psychological issues tend to have poorer nutrition. How should you apply this information?
    a. Eating healthy is never a bad idea.
    b. Better nutrition puts less stresses on your mind
    c. Causation is hard to determine
    d. The average is not the same as any particular observation.
    e. None of the above

This factoid establishes a correlation between mental health and diet. It's tempting to assume that the causation is obvious: if you eat healthier, you're have better mental health. But it's possible that the causation goes the other way: people who have poorer mental health tend to eat less healthy (for example, depression can lead to consuming junk food).

12. Tyrone is looking for a job in information technology. Based on conversations from people in the industry, he estimates that his chance of getting a promotion is 60% and his chance of being hired is 80%. What is the chance that he's promoted assuming that he's hired?
    a. 20%
    b. 48%
    c. 60%
    d. 80%
    e. None of the above

The question asks to determine the conditional probability based on two nonconditionals. Since the chance of being promoted equals the chance of being promoted (assuming he's hired) times the chance of being hired, we can do some simple algebra: 0.6/0.8 = 0.75. If he gets hired, his chance of being promoted in will increase from 60% to 75%.

13. Consider the previous question. If promotion in information technology is automatic—100% of hired people get promoted—then what should his chances of the nonconditional promotion?
    a. 20%
    b. 48%
    c. 60%
    d. 80%
    e. None of the above

*If promotion is a guarantee, then that means the chance of getting a promotion equals the chance of getting a job. 0.8*1 = 0.8.*

14. People who had a job as a teenager are more likely to have a job as an adult. This correlation could be explained by a confounding variable. How?
    a. People who get a job as an adult cause them to get a job as a teenager.
    b. People with a lot of motivation are more likely to get a job, regardless of how old they are.
    c. People who get a job as a teenager get experience and that additional experience causes them to get a job as an adult.
    d. People who get a job as a teenager tend to be low income, which makes it easier for them to get a job as an adult.
    e. None of the above

Remember, a confounding variable causes the other two variables independently. Motivation causes teenage employment and that same motivate causes employment as an adult.

15. If the coefficient of an independent variable in a regression analysis is -4.8 with a p-value of 0.094, what should you conclude at 95% confidence?

f. The coefficient is not statistically significant.
g. The coefficient is statistically significant but not practically significant.
h. The coefficient is statistically significant and is practically significant.
i. The coefficient is statistically significant; practical significance depends on what the dependent and independent variables are.
j. It is impossible to conclude anything based on the information provided.

At 95% confidence, the relevant alpha is 0.05. Since 0.094 is greater than 0.05, the coefficient is not statistically significant. Had the p-value been less than 0.05, D would have been the best answer; you can't determine practical significance unless you know what you're dealing with.

16. True or false: For two variables with an equal number of observations, if one variable has a larger range, that variable must also have a larger standard deviation.
    a. True because both range and standard deviation are measures of dispersion. It makes sense that they move in concert.
    b. True because the number of observations are equal; it would not be true otherwise.
    c. False because ranges are always larger than standard deviations.
    d. It is true, but not for a reason listed.
    e. It is false but not for a reason listed.

Range and standard deviation are calculated in two completely different ways. Range simply looks at the maximum and minimum values. Standard deviation looks at every value. It's easy to imagine a variable with one very low observation and one very high observation but all other observations are right on the mean. This standard deviation would naturally be lower than a variable where half the observations are at a slightly higher minimum and the other half are at a slightly higher maximum.

17. Which of the following is an example of someone being risk averse?
    a. Randy spending $60 on a coin flip that has a 50% chance of winning $120 and a 50% chance of winning zero.
    b. Amir learning to juggle using very sharp knives.
    c. Chloe purchasing a lottery ticket.
    d. B & C
    e. None of the above

A is an example of risk neutral; the expected payoff equals the cost Randy paid to play the game. B & C are risk-loving strategies.

18. Which of the following is an example of a question requiring a Poisson distribution?

a. If a major storm hits the DC area, and an average of ten trees typically fall in a neighborhood, what is the chance that four trees will fall in a particular neighborhood?
b. If there's an average of 1.3 flaws per 50 feet of rope, what are the chances that there will be two flaws on a particular 50 feet of rope?
c. If Simon selects four people at random to be his bodyguards, what are the chances that two of his bodyguards will be traitors if there's a five percent chance of any one bodyguard being a traitor?
d. A & B
e. None of the above

C is clearly a binominal distribution: it has a constant probability and note there's a natural limit to the number of traitor bodyguards if Simon only hires four.

B is clearly Poisson. There's no reason to think the average number of flaws will change from 50' segment to 50' segment and there's no natural limit to the number of flaws you can have.

It's tempting to think the answer here is D, since A sounds like Poisson as well. But neighborhood size differs widely. It's crazy to think a large neighborhood should have the same number of downed trees as a small one. Poisson can't be used here.

19. Which of the following ***clearly*** would be independent to the chance of a corn crop failing on Wu's small farm?
    a. The chance of a drought.
    b. The chance a corn crop failing on Aziz's nearby farm.
    c. The chance that the price of corn will fall next year.
    d. B & C
    e. None of the above

Independent probabilities means the outcome of one event doesn't affect the probability of the other event. If there's a drought, that makes it more likely the corn crop failing; those events are not independent.

You can make a pretty good argument B is independent. Just because Aziz's crop fails doesn't mean Wu's crop will fail. But it does depend as to why Aziz's crop failed. Remember, this is a nearby farm. Was there a disease or an infestation, something that could carry over to Wu's crop? Or a drought? Or bad seeds? The chance that these occurring would be part of the chance of crop failure.

But C is clearly independent. The price of corn next year can't influence Wu's crop success now (since the price of corn is in the future). At the same time, if Wu's crop fails, that can't influence next year's corn prices. Not only is it a year away, Wu's

farm is small. Corn is a big market; Wu's crop won't affect this year's price, let alone next year's.

20. Use Practice Final Exam Data Set 2 for this question. It's made up of hypothetical data of fictional cities. What's the difference between the average Net Water for cities on the coast and the average Net Water for cities not on the coast?
    a. 2.1
    b. 4.8
    c. 5.9
    d. 7.0
    e. None of the above

The first step should be use the sort function with the Coast? variable so cities are grouped into coast and non-coast cities. Net Water is 5.9 for non-coast cities and for coast cities, it's -1.1; the difference is 7.

21. Armadillo Business Consulting has 40 associates who are eligible for promotion; seven of those associates are accounting majors. If eight associates are promoted, what is the chance that at least two of those will be accounting majors? (Use hypergeometric distribution to determine the answer.)
    a. 0.1282
    b. 0.3024
    c. **0.4306**
    d. 0.5694
    e. 0.8718

0.3024 is the chance that exactly two will be accounting. The question asks for <u>at least</u> two.

0.8718 is the chance that two or fewer will be accounting. 0.1282 is the chance that more than two will be accounting. (Note these two added together equals one.)

0.5694 is the chance one or fewer will be accounting; thus one minus that—0.4306—is the chance that two or more will be accounting. In Excel, it equals 1 - HYPGEOM.DIST(1,8,7,40,1)

22. Kali's investing in three companies. Each company has an 80% chance of failing. What's the chance that exactly one company will succeed?
    a. 12.8%
    b. 20.0%
    c. 38.4%
    d. 48.8%
    e. There is not enough information to determine the answer.

You can solve this in two ways. 1) You can note that this is a binominal distribution (a constant probability of success—0.2—predicting exactly one success). BINOM.DIST(1,3,0.2,0)=0.384, or 38.4%.

*You can also note that if one company succeeds, then the other two must fail. The probability of that happening is 0.8*0.8*0.2=0.128. Since this can happen in two other ways (0.8*0.2*0.8 and 0.2*0.8*0.8), multiply the result by three to get 0.384 or 38.4%.*

23. Allen's testing a new user interface to understand if people who are color blind like it more than the previous user interface. Color blind users rated the old interface at 6.7 out of 10. Based on a sample of 12 color blind users, with a population standard deviation of 1.8 and a sample standard deviation of 2.9, the new average rating was 8.4 out of 10. Is the new user interface a statistically significant improvement for color blind users?
    a. At the 99% level, yes.
    b. At the 99.9% level, yes.
    c. A & B
    d. None of the above, but it would at the 95% level.
    e. None the above; it's not statistically significant at all.

Note that this is a z test, not a t test. Though the question indicates your sample standard deviation, it also indicates your population standard deviation. If you have the population standard deviation, you use a z test.

This is also a one-tailed test—Allen is looking for an improvement—so our critical values are 2.33 (99%) and 3.09 (99.9%). Now we calculate:

$$z = \left|\frac{8.4 - 6.7}{1.8/\sqrt{12}}\right| = \left|\frac{-1.7}{1.8/3.46}\right| = |-3.27| = 3.27$$

This value is greater than the critical value at the 99% confidence level and at the 99.9% confidence level.

Note that B would never be an answer for this kind of question; if it's significant at the 99.9% level, it must be significant at the 99% level, too.

24. D'angelo runs an assembly line and wants to give his fastest employees a bonus but he can't afford to give a bonus to anymore than 12% of his employees. Speed of assembly follows a normal distribution with a mean of 7 minutes and a standard deviation of 45 seconds. What would be an appropriate cut off time for giving a bonus?
    a. 6 minutes
    b. 6 minutes, 15 seconds

c. 7 minutes, 45 seconds
d. 8 minutes
e. None of the above

Because he's rewarding his fastest employees, lower numbers are better. NORM.INV(0.12,7,0.75) results in 6.11. 6.11 minutes is a strange amount of time so it seems appropriate to round it to the nearest 15-second increment. Rounding up to 6.25 (option B) would result in more than 12% of employees getting a bonus. A cutoff of 6 minutes even would be more appropriate.

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

25. Using Practice Final Exam Data Set 2, run a regression with Net Water as the dependent variable and Coast?, Pollution, Precipitation, Average Income, and Population Density as the independent variables. Then answer the following:
    k. Which variables are statistically significant at the 95% level?
    l. There is multicollinearity in this model; how should you fix the multicollinearity?
    m. Re-run the regression with the fix you suggest in part B. This is a better model than the original; what about the regression output shows you that?

a) When you run the regression, you should notice that nothing is statistically significant. The closest value is Precipitation with a p-value of 0.12. Not even good enough for 90%!

b) A correlation 0.87 for density and income indicate multicollinearity (the high correlations for income and coast and pollution and coast are nearly multicollinearity but technically fall short of our 0.8 threshold). To fix the multicollinearity, drop the income variable. (You can instead drop the density variable but dropping the income variable has the virtue of getting rid of the high correlation pair of income and coast. Dropping income over density creates a slightly better model but dropping either would be acceptable answers.)

c) A regression with the dropped variable makes a better regression, evidenced by the improved F stat and the higher adjusted R^2. (Note that the unadjusted R^2 is higher in the first model, which is what we'd expect because, all other things being equal, more explanatory variables always leads to a higher R^2.)

26. Trinity is a lawyer attempting to determine which expert she wants to use in a trial. The expert must be able to explain the clearly explain the material to a jury while also being personable, cooperative, trustworthy, and able to withstand cross examination. Trinity does not have time to interview these candidates in depth so she relies on an agency's recommendation. Suppose 10% of potential experts are good in a trial. The agency's

recommendation is 85% sensitive and 75% specific. What's the chance that the expert would be good at trial (G) if the agency recommended him (R)?

Here, we use Bayes' Theorem:

$$\mathrm{P}(G|R) = \frac{P(R|G) * P(G)}{P(R|G) * P(G) + P(R|\sim G) * P(\sim G)} = \frac{0.85 * 0.1}{0.85 * 0.1 + 0.25 * 0.9}$$
$$= \frac{0.085}{0.085 + 0.225} = \frac{0.085}{0.310} = 0.2742$$

There is a 27.42% chance that the recommendation is a good expert.0

27. Jaya's company is thinking about investing in new blood testing technology. There's a 90% chance it will not work and Jaya's company will lose $1 million from the investment. But there's a 10% chance it will work and the profits for the company depend on market performance, as indicated by the table. What's the expected value of this investment?

Market Performance	*Probability*	*Profit*
Excellent	5%	$25 million
Good	15%	$15 million
Fair	50%	$6 million
Terrible	30%	$2 million

First, recognized that this is not a sure thing; there's a 90% chance that the firm will lose $1 million and each of these probabilities are conditional on the technology working. So we do the following calculations:

- *(0.90)(-$1 million) = -$0.9 million*
- *(0.10)(0.05)($25 million) = $0.125 million*
- *(0.10)(0.15)($15 million) = $0.225 million*
- *(0.10)(0.50)($6 million) = $0.3 million*
- *(0.10)(0.30)($2 million) = $0.06 million*

$0.71 million – $0.9 million = -$0.19 million; this is not a good investment.

28. Highland Farms grows potatoes with an average yield of 30,900 pounds per acre. On 13 acres, they try a new fertilizer to try to increase yields. On these 13 acres, the average yield of 32,500 pounds and a sample standard deviation of 1300 pounds. Is this a statistically significant difference or not at 99.9% confidence? Remember to show your work and justify your answer.

First, note this is a one-tailed test. This is also a t-test because we do not know the population standard deviation.

So we do some math:

$$t = \left|\frac{32{,}500 - 30{,}900}{1300/\sqrt{13}}\right| = \left|\frac{1600}{360.6}\right| = 4.44$$

Now we determine the critical t score using T.INV(0.001,12)=3.93

It is statistically significant because the calculated value is more than the critical value.

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.

Critical z scores

Confidence	α	$z_{\alpha/2}$	z_{α}
90%	0.1	1.645	1.280
95%	0.05	1.960	1.645
99%	0.01	2.576	2.330
99.9%	0.001	3.291	3.090

Critical t scores

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

Coefficient of Variation

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

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

$$\mu = np,\ \sigma = \sqrt{npq}$$

Hypergeometric Distribution

$$\mu = \frac{nR}{N},\ \sigma = \sqrt{\frac{nR(N-R)}{N^2}}\sqrt{\frac{N-n}{N-1}}$$

Poisson

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

Optimal Sample Size

$$n = \left(\frac{z_{\alpha/2}\sigma}{ME}\right)^2$$

z-test

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

Proportion

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

t-test

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

Adjusted R^2

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