Stat 88 Spring 2020 **Final EXAM** **A. Lucas**

NAME (FIRST LAST): ______________________ **SID:** ______________

TIME AND CONDITIONS: You have 2 hours and 50 minutes to complete the exam and 10 minutes to upload your submission to Gradescope. A reference sheet is provided in the end. No other materials are allowed; nor are computers, or the internet.

QUESTIONS AND ANSWERS

- There are 11 questions. Honor code is the first one.
- **Give brief explanations or show calculations in each question** unless the question says this is not required. You may use, without proof, any result proved or used in lecture, the textbook, and homework, unless the question asks for a proof.
- Please leave answers as **unsimplified arithmetic or algebraic expressions including finite sums and the standard normal cdf Φ** unless the question asks for a simplification.
- Please do not leave answers as an infinite sum. **Answers left as an infinite sum will not receive full credit.**
- Calculators are allowed. You will not need calculators unless the question asks you to simplify to numerical answers.

GRADING

- The exam is worth 100 points. Each question is worth 10 points.
- Please commit yourself to a single answer for each question. If you give multiple answers (such as both True and False) then please don't expect credit, even if the right answer is among those that you gave.
- You will have until 2:30pm to write all your answers and upload to Gradescope. **Any submissions after 2:30pm will incur a penalty proportional to how late the submission is.** If you submit your exam n minutes after 2:30pm, you will be deducted 2^{n-2} points.

FORMAT

- **You must answer each question on a separate page for a total of 11 pages.**
- Writing solutions on an iPad or tablet is acceptable
- You must select the correct page associated for each subpart of each question when submitting to Gradescope. If you do not, you will get a zero for that question on the exam.
- Please turn in only your exam, not the reference sheet.

1. Honor Code

Data Science and the entire academic enterprise are based on one quality – integrity. We are all part of a community that doesn't fabricate evidence, doesn't fudge data, doesn't present other people's work as our own, doesn't lie and cheat. You trust that we will treat you fairly and with respect. We trust that you will treat us and your fellow students fairly and with respect. **Please transcribe the UC Berkeley's Honor Code below and sign your name next to it:**

"I certify that all solutions will be entirely my own and that I will not consult or share information with other people during the exam. I promise I will act with honesty, integrity, and respect for others."

2. Halloween will be here before you know it and the children in your neighborhood will come trick-or-treating (that is, they will come to your door and demand candy). Suppose there are 20 children in your neighborhood and 30 houses (one of which is yours). Each child independently chooses 10 houses at random without replacement to visit.

a) What is the probability that a specific child will visit your house?

b) What is the probability that exactly 10 children visit your house?

c) What is the expected number of houses that exactly 10 children visit?

3. Continuing the last question, suppose you have a bunch of candy in your house, ranging from very cheap to very fancy. Suppose that, if you ordered the candy in your house from cheapest to fanciest, the ith candy cost $30i$ cents to buy (so the cheapest candy cost 30 cents, the 10th cheapest cost 300 cents, etc). You want to keep the fancy candy for yourself (the kids would not appreciate the fancy stuff anyway), so when a child comes to your door, you give them the cheapest remaining candy – thus, if 10 children come to your door, you will give out the 10 cheapest pieces of candy that you have.

a) If k children come to your door, what is the total cost of the candy denoted as T_k in terms of cents, that you will give out? (Hint: This uses a summation identity).

b) What is the probability that you will give away more than \$15 worth of candy? Hint: For what k is $T_k > 1500$ and start from there.

4. Multiple Answer. Write down **ALL** correct choices: there may be more than one correct choice, but there is always at least one correct choice for each question. NO partial credit: the set of all correct answers must be checked.

a) Define X to be the sum of 10 indicator random variables, so $X = \sum_{i=1}^{10} I_i$. Which of the following statements that *must* be true for X to be a binomial random variable?

i Each I_i has the same probability of taking the value 1.

ii The indicators are all independent of each other.

iii Each I_i should have probability 1/2 of taking the value 1.

iv None of the above.

correct choices: ______________________________

b) We have two random variables X and Y. Which of the following statements must be true to apply the addition rule for expectations, $E(X+Y) = E(X) + E(Y)$?

i X and Y must be independent.

ii X and Y must be identically distributed.

iii None of the above.

correct choices: ______________________________

c) We have two random variables X and Y. Which of the following statements must be true to apply the addition rule for variance, $Var(X+Y) = Var(X) + Var(Y)$?

i X and Y must be independent.

ii X and Y must be identically distributed.

iii None of the above.

correct choices: ______________________________

d) We have two random variables X and Y. Which of the following statements is true (there can be more than one):

i $E(Y|X)$ is a function of X.

ii $E(Y|X)$ is a function of Y.

iii $E(E(Y|X))$ is a non constant random variable.

iv $E(E(E(Y|X))) = E(Y)$.

correct choices: ______________________________

e) You flip a fair coin N times where N is a random variable, $N \sim$ Poisson(5). What is the expected number of heads you will get?

i 5

ii 5/2

iii N/2

iv None of the above

correct choices: ______________________________

5. Your family is trying to decide sleeping arrangements in the house. There are 10 young people staying in the house who will be divided into two rooms. You will choose a set of 5 people to sleep in the first room and the rest will sleep in the second room. A sleeping arrangement is acceptable if it meets the following two criteria:

- Your two youngest cousins (say, 7 and 8 years old) always fight, so exactly one of them must be in the first room.
- There are 4 people over 18 years old, and exactly 2 of them must be in the first room to supervise the others.

You are too lazy to work out the arrangement by hand, so you decide to do this randomly by choosing a set of 5 people at random to sleep in the first room, and repeat this until you get an acceptable arrangement.

a) For a single draw of 5 people, what is the probability that you will draw an acceptable sleeping arrangement?

b) What is the expected number of draws of 5 people until you get an acceptable arrangement? (Hint: Each drawing of 5 people is independent of every other drawing of 5 people.)

6. You are using a telescope to measure the speed at which the planet Saturn crosses the night sky. To do this you draw two lines on your lens, and measure the time it takes for Saturn to cross between the two lines. However, your time measurement is noisy, so you will conduct this observation several times and average their results.

Let X_i represent the time measurement from the ith observation. Your measurements are well calibrated, so for each i, $E(X_i) = \mu_X$, where μ_X is the true time it takes Saturn to cross between the lines. Each measurement also has standard deviation $SD(X_i) = 0.03$ seconds.

a) You will take n measurements, $X_1, \ldots, X_n$, using the same procedure, and use the sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ_X. In terms of n, what is $SD(\bar{X})$?

b) What is the smallest number of measurements you will need to take so that your estimate $\bar{X}$ has at most a $\frac{1}{25}$ probability of falling outside the interval $\mu_X \pm 0.003$ seconds? (Hint: Chebyshev)

7. You are interested in knowing the preference at Cal between two statistical software packages R and Python. In a SRS of 250 students, 80% prefer R over Python, and the rest (20%) prefer Python over R.

a) Construct a 95% CI for the percentage of students who prefer R over Python. Is there a 95% chance that the true percentage lies in your interval?

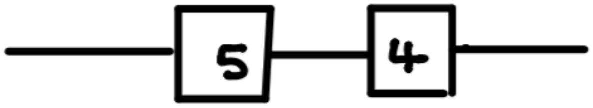
b) Cal Data Science believes that 70% of Cal students prefer R over Python. Based on the data from your SRS, would you agree with Cal Data Science's belief?

c) Your high school friend, who studies Data Science at Stanford, finds from a SRS of 250 students, that only 75% of Stanford Data Science students prefer R over Python and the rest (25%) prefer Python over R. Do a hypothesis test (at a 5% level of significance) to decide whether this difference in R preferences is due to chance (make a one sided alternative). Please sate your null hypothesis, alternative hypothesis, test statistics and p value.

8. Let T_1 and T_2 be independent Exponential random variables with rates λ_1 and λ_2.

a) Show that the CDF of $T = \max(T_1, T_2)$ is $F(t) = (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t})$.

b) An electric circuit consists of 2 components in the following diagram.



The lifetimes of the components, measured in days, have independent exponential distributions with means given in the diagram. Let T be the lifetime of the circuit. Find the CDF of T. (Hint: Is T the max or the min of the two components lifetimes?)

9. Let $X_1, X_2, \ldots, X_n$ be iid Uniform$(0, 2\theta)$ for an unknown parameter 2θ. Let $M = \max(X_1, X_2, ..., X_n)$.

a) What is the expected distance of X_1 from zero?

b) Find the expectation of M

c) Find an unbiased estimator of 2θ in terms of M.

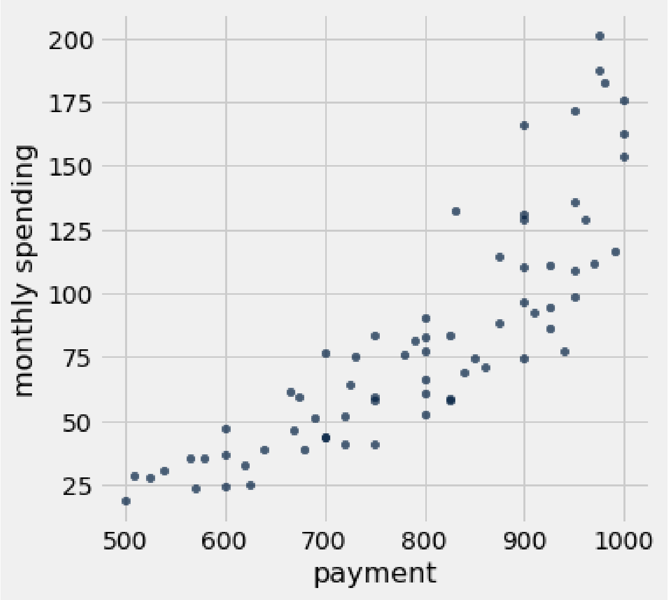
10. You run a study at Thanksgiving. Let X be how many pounds of turkey each person ate, and Y be how many hours past midnight they slept. You find that $r(X, Y) = 0.6$.

a) Qualitatively, what sort of pattern does this correlation indicate?

b) You rerun the study, this time defining X as grams of turkey consumed, and Y as minutes slept past midnight. Does $r(X, Y)$ change?

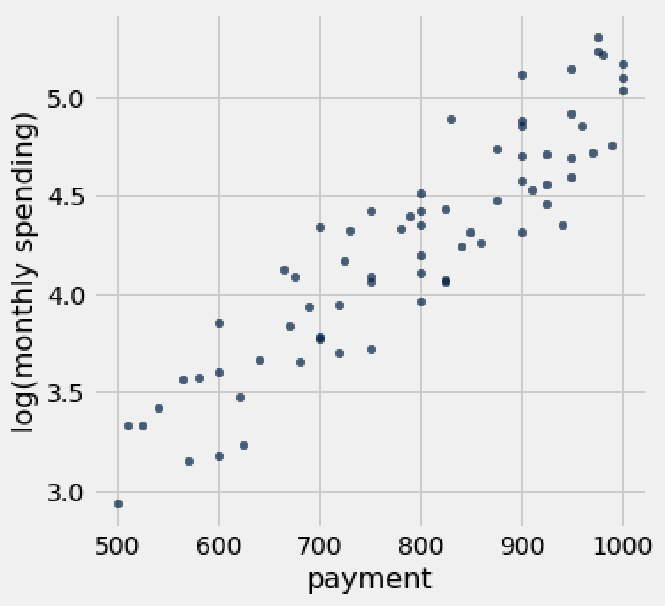
c) Now you define Z how much money each person saves by attending Black Friday sales, which start (in your town) at midnight. You estimate that every minute spent shopping (i.e. not sleeping) after midnight is worth 2 dollars in savings. What is $r(X, Z)$, or the correlation between turkey eaten and dollars saved?

11. As a statistical intern for the US Treasury you are studying the effect of the stimulus payment (in dollars) on monthly household spending (in dollars) during Covid-19. You decide to make a linear regression model of monthly household spending as a function of stimulus payment size. You create a scatter plot based on a random sample of 70 households.

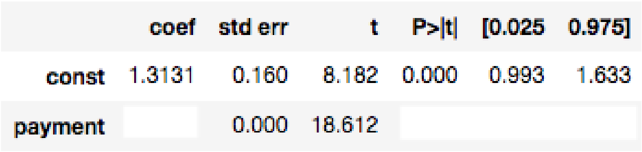


a) Does it look like a regression line is a good fit for the data? Explain your answer in terms of the assumptions of the Simple Linear Regression Model.

You determine the regression line after taking the natural log (base e) of all the observations for the response variable. That is to say, for data points $i = 1, 2, 3, \ldots,$ the response is assumed to be $\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$.



You now run a linear regression on this transformed data and print out a summary of the output, but some of the entries are removed for national security reasons.



	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3131	0.160	8.182	0.000	0.993	1.633
payment		0.000	18.612			

You also know, $\mu_x = 792.8$ and $\sigma_x = 140.1$
$\sigma_{\log(Y)} = 0.6$ and r=.91
where X is the payment and $\log(Y)$ is the log of monthly spending.

b) Find the equation of the regression line.

c) What is the predicted monthly spending, in dollars, for a household receiving a stimulus payment of \$900?

d) Find the average log(monthly spending), $\mu_{\log Y}$.

e) Conduct a hypothesis test $H_0 : \beta_1 = 0$ versus alternative $H_A : \beta_1 \neq 0$. What can you conclude about the effect of payment size and household spending? (Hint: calculator might be needed)