Review for Final

Stat 10

(1) The table below shows data for a sample of students from UCLA.

(a) What percent of the sampled students are male?

(b) What proportion of sampled students are social science majors or male?

(c) Are Gender and socSciMajor independent variables?

		socSciMajor		
		yes	no	TOTAL
Gender	female	25	35	60
	male	30	27	57
	TOTAL	55	62	117

(2) The distribution of the number of pizzas consumed by each UCLA student in their freshman year is very right skewed. One UCLA residential coordinator claims the mean is 7 pizzas per year. The first three parts are True/False.

(a) If we take a sample of 100 students and plot the number of pizzas they consumed, it will closely resemble the normal distribution.

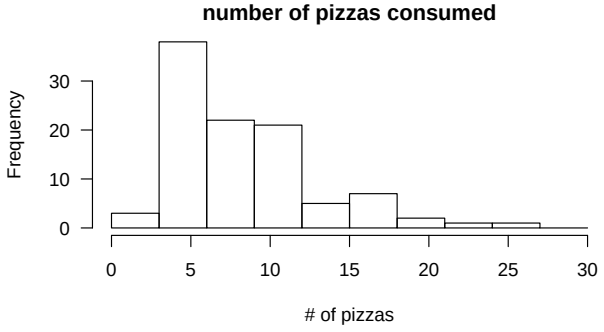
(b) If we take a sample of 100 students and compute the mean number of pizzas consumed, it is reasonable to use the normal approximation to create a confidence interval.

(c) If we took a sample of 100 students and computed the median and the mean, we would probably find the mean is less than the median.

(d) We suspect the the mean number of pizzas consumed is actually more than 7 and take a sample of 100 students. We obtain a mean of 8.9 and SD of 7. Test whether this data provides convincing evidence that the UCLA residential coordinator underestimated the mean.

(e) Suppose the mean is actually 9 and the SD is actually 7. Can you compute the probability a student eats at least five pizzas in his/her freshman year? If so, do it. If not, why not?

(3) In exercise (2), the distribution of pizzas eaten by freshman was said to be right skewed. Suppose we sample 100 freshman after their first year and their distribution is given by the histogram below. Estimate the mean and the median. Explain how you estimated each. Should your estimate for the mean or median be larger?



(4) Chipotle burritos are said to have a mean of 22.1 ounces, a st. dev. of 1.3 ounces, and they follow a normal distribution. Additionally, the weight was found to be independent of the type of burrito ordered. Chipotle advertises their burritos as weighing 20 ounces.

(a) What proportion of burritos weigh at least 20 ounces?

(b) What is the chance at least 4 of your next 5 burritos from Chipotle will weigh more than 20 ounces? Write out any assumptions you need for this computation and evaluate whether the assumptions are reasonable. Even if you decide one or more assumptions are not reasonable, go ahead with your computation anyways.

(c) Verify the probability a random burrito weighs between 22 and 23 ounces is 0.29.

(d) Your friend says that the last 3 times she has been at Chipotle, her burrito weighed between 22 and 23 ounces. What is the chance it happens again the next time she goes?

(e) Ten friends go to Chipotle and bring a scale. What is the probability the fifth burrito they weigh is the first one that weighs under 20 ounces? (Check any conditions necessary.)

(f) What is the probability at least two of the ten burritos will weigh less than 20 ounces? (Check any conditions necessary for your computations.)

(g) Without doing any computations, which of the following is more likely? (i) A randomly sampled burrito will have a mean of at least 23 ounces. (ii) The sample mean of 5 burritos will be at least 23 ounces. Explain your choice. Your explanation should be entirely based on reasoning without any computations (!). Pictures are okay if you think it would be helpful.

(h) Which of the following is more likely? (i) A randomly sampled burrito will have a mean of at least 20 ounces. (ii) The sample mean of 5 burritos will be at least 20 ounces. Explain your choice. Your explanation again does not need any computations.

(i) Verify your conclusion in (h) by computing each probability.

(j) If the distribution of burrito weight is actually right skewed, could you have compute the probabilities in (i) with the information provided? Why or why not?

(5) Below is a probability distribution for the number of bags that get checked by individuals at American Airlines. No one checks more than 4 bags.

number of bags checked	0	1	2	3	4
probability	0.42	0.29	0.22	0.05	0.02

(a) Picking out a passenger at random, what is the chance s/he will check more than 2 bags?

(b) American Airlines charges \$20 for the first checked bag, \$30 for the second checked bag, and \$100 for each additional bag. How much revenue does American Airlines expect for baggage costs for a single passenger? (Note: the cost of 2 bags is \$20+\$30 = \$50).

(c) How much revenue would AA expect from 20 random passengers?

(d) 18 months ago the fees were \$15, \$25, and \$50 for the first, second, and each additional bag, respectively. Recompute the expected revenue per passenger with these prices instead of the current ones and compare the average revenue per passenger.

(6) 26% of Americans smoke.

(a) If we pick one person (specifically, an American) at random, what is the probability s/he does not smoke?

(b) We ask 500 people to report if they and their closest friend smokes. The results are below. Does it appear that the smoking choice of people and their closest friends are independent? Explain.

	friend smokes	friend does not	TOTAL
smoke	67	68	135
does not	53	312	365
TOTAL	120	380	500

(7) A drug (sulphinpyrazone) is used in a double blind experiment to attempt to reduce deaths in heart attack patients. The results from the study are below:

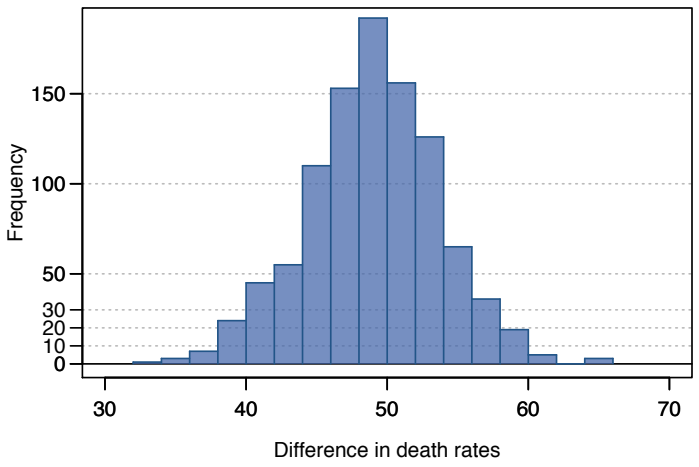
		`outcome`		
		lived	died	Total
`trmt`	drug	693	40	733
	placebo	683	59	742

(a) Setup a hypothesis test to check the effectiveness of the drug.

(b) Run a two-proportion hypothesis test. List and verify any assumptions you need to complete this test.

(c) Does your test suggest `trmt` affects `outcome`? (And why can we use causal language here?)

(d) We could also test the hypotheses in (b) by scrambling the results many times and see if our result is unusual under scrambling. Using the number of deaths in the drug group as our test statistic, estimate the p-value from the plot of 1000 scramble results below. How does it compare to your answer in part (b)? Do you come to the same conclusion?



(8) *For the record, these numbers are completely fabricated.* The probability a student can successfully create and effectively use a tree diagram at the end of an introductory statistics course is 0.67. Of those students who could create a tree diagram successfully, 97% of them passed the course. Of those who could not construct a tree diagram, only 81% passed. We select one student at random (call him Jon).

(a) What is the probability Jon passes introductory statistics course?

(b) If Jon passed the course, what is the probability he can construct a tree diagram?

(c) If Jon did NOT pass the course, what is the probability he canNOT construct a tree diagram?

(d) Stephen knows about these probabilities and decides to spend all his time studying tree diagrams, thinking he will then have a 97% chance of passing the final. What is the flaw in his reasoning?

(9) It is known that 80% of people like peanut butter, 89% like jelly, and 78% like both.

(a) If we pick one person at random, what is the chance s/he likes peanut butter or jelly?

(b) How many people like either peanut butter or jelly, but not both?

(c) Suppose you pick out 8 people at random, what is the chance that exactly 1 of the 8 likes peanut butter but not jelly? Verify any assumptions you need to solve this problem.

(d) Are "liking peanut butter" and "liking jelly" disjoint outcomes?

(10) There are currently 120 students still enrolled in our lecture. I assumed that each student will attend the review session with a probability of 0.5.

(a) If 0.5 is an accurate estimate of the probability a single student will attend, how many students should I expect to show up?

(b) If the review session room seats 71 people (59% of the class), what is the chance that not everyone will get a seat who attends the review session? Again assume 0.5 is an accurate probability.

(c) Suppose 61% of the class shows up. Setup and run a hypothesis to check whether the 0.5 guess still seems reasonable.

(11) Obama's approval rating is at about 51% according to Gallup. Of course, this is only a statistic that tries to measure the population parameter, the *true* approval rating based on all US citizens (we denote this proportion by p).

(a) If the sample consisted of 1000 people, determine a 95% confidence interval for his approval.

(b) Interpret your confidence interval.

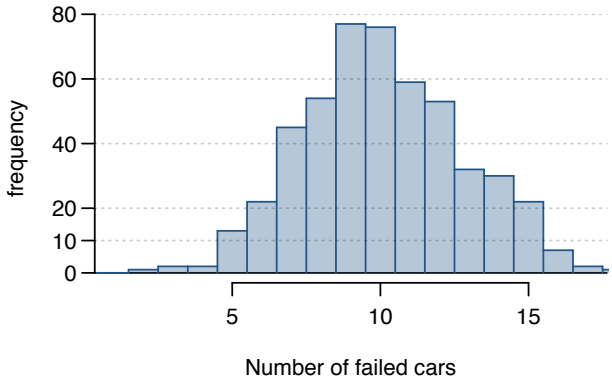
(c) What does 95% confidence mean in this context?

(d) If Rush Limbaugh said Obama's approval rating is below 50%, could you confidently reject this claim based on the Gallup poll?

(12) A company wants to know if more than 20% of their 500 vehicles would fail emissions tests. They take a sample of 50 cars and 14 cars fail the test.

(a) Verify any assumptions necessary and run an appropriate hypothesis test.

(b) The simulation results below represent the number of failed cars we would expect to see if 20% of the cars in the fleet actually failed. Use this picture to setup and run a hypothesis test. You should specifically estimate the p-value based on the simulation results.



(c) How do your results from parts (a) and (b) compare?

(d) Which test procedure is more statistically sound, the one from part (a) or the one from part (b)? (To my knowledge, we haven't really discussed this in class. That is, make your best guess at which is more statistically sound.)

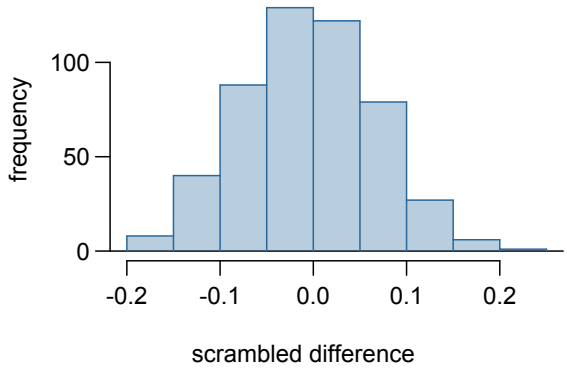
(13) Stephen claims that *Casino Royale* was a significantly better movie than *Quantum of Solace*. Jon disagrees. To settle this disagreement, they decide to test which movie is better by examining whether *Casino Royale* was more favorably reviewed on Amazon than *Quantum of Solace*.

(a) Setup the hypotheses to check whether *Casino Royale* is actually rated more favorably than *Quantum of Solace*.

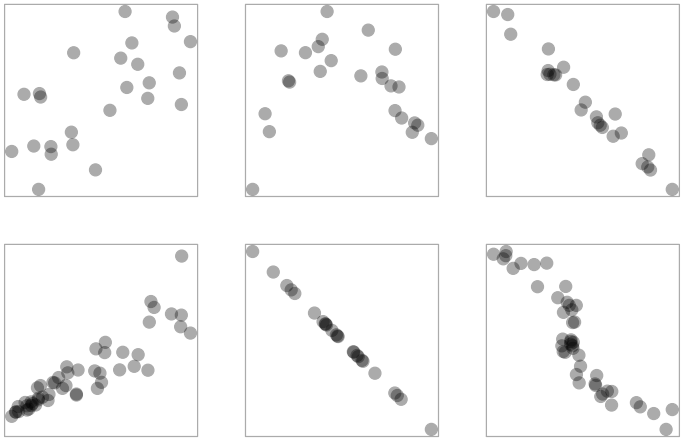
(b) The true difference in the mean ratings is 0.81. Stephen and Jon decide to scramble the results, compute the difference under this randomization, and see how it compares to 0.81. What hypothesis does scrambling represent?

(c) Approximately what difference would they expect to see in the scrambled results?

(d) Stephen and Jon scramble the results 500 times and plot a histogram of the observed *chance* differences, which is shown below. Based on these results, would you reject or not reject your null hypothesis from part (a)? Who was right, Stephen or Jon?



(14) Examine each plot below.



(a) Would you be comfortable using an LSR line to model the relationship between the explanatory and response variables in each plot using the techniques learned in this course?

(b) If we did fit a linear model for each scatterplot, what would the residual plot look like for each case?

(c) For those plots where you would not apply our methods, explain why not. (Hint: only three are alright with our methods discussed in class.)

(d) Identify the approximate correlation *for those plots where we could apply our methods* from the following options (i) -1, (ii) -0.98, (iii) -0.60, (iv) 0, (v) 0.65, (vi) 0.90, and (vii) 1.

(15) p301, #25 (modified). You are about to take the road test for your driver's license. You hear that only 34% of candidates pass the test the first time, but the percentage rises to 72% for subsequent retests.

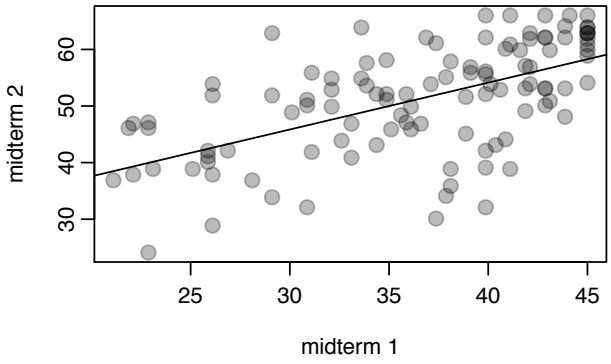
(a) Describe how you would simulate one person's experience in obtaining her license.

(b) Run your simulation ten times, i.e. run your simulation for ten people. Some random numbers:

73867	54439	92138	01549	38302	08879
80786	81483	75366	64652	71227	48755
28088	09478	76440	75881	94643	84652

(c) Based on your simulation, estimate the average number of attempts it takes people to pass the test.

(16) Below are scores from the first and second exam. (Each score has been randomly moved a bit and only students with both moved scores above 20 were included to ensure anonymity.)



(a) From the plot, does a linear model seem reasonable?

(b) If we use only the data shown, we have the following statistics:

	midterm 1	midterm 2	correlation (R)
$\bar{x}$	36.4	51.1	
s	6.8	9.7	0.58

Determine the equation for the least squares regression line based on these statistics.

(c) Interpret both the y-intercept and the slope of the regression line.

(d) If a random student scored a 27 on the first exam, what would you predict she (or he) scored on the second exam?

(e) If that same student scored a 47 on midterm 2, did s/he have a positive or negative residual?

(f) Would you rather be a positive or negative residual?

(17) Lightning round. For (c) and (d), change the sentence to be true in the cases it is false. (Note: There is typically more than one way to make a false statement true.)

(a) Increasing the confidence level affects a confidence interval how?
slimmer wider no effect

(b) Increasing the sample size affects a confidence interval how?
slimmer wider no effect

(c) True or False: If a distribution is skewed to the right with mean 50 and standard deviation 10, then half of the observations will be above 50.

(d) True or False: If a distribution is skewed to the right and we take a sample of 10, the sample mean is normally distributed.

(e) You are given the regression equation $\hat{y} = 2.5 + 0.12 * x$ where y represents GPA and x represents how much spinach a person eats. Suppose we have an observation ($x = 5 ounces/week, y = 2.0$). Can we conclude that eating more spinach would cause an increase in this person's GPA?

(f) Researchers collected data on two variables: sunscreen use and skin cancer. They found a positive association between sunscreen use and cancer. Why is this finding not surprising? What might be really going on here?

(g) In the Franken-Coleman dispute (MN Senate race in 2008 that was disputed for many months), one "expert" appeared on TV and argued that the support for each candidate was so close that their support was statistically indistinguishable, and we should conduct another election for this race. What is wrong with his reasoning?

(h) In Exercise 11, you computed a confidence interval to capture President Obama's approval rating. True or False: There is a 95% chance that the true proportion is in this interval.

(i) True or False: In your confidence interval in Exercise 11, you confirmed that 95% of all sample proportions would fall between (0.48 and 0.54).

(j) Researchers ran an experiment and rejected H_0 using a test level of 0.05. Would they make the same conclusion if they were using a testing level of 0.10? How about 0.01?