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? 57/117

(b) What proportion of sampled students are social science majors or male? $P(\textit{yes} \text{ or } \textit{male}) = P(\textit{yes}) + P(\textit{male}) - P(\textit{yes} \text{ and } \textit{male}) = 55/117 + 57/117 - 30/117 = 92/117.$

(c) Are Gender and socSciMajor independent variables? The proportion of males in each socSciMajor group appears to be pretty different: $25/55 = 0.45 \neq 35/72 = 0.56$. This would mean the variables are not independent. (What method could we do to check whether independence seems reasonable?)

		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. FALSE. It should be right skewed.

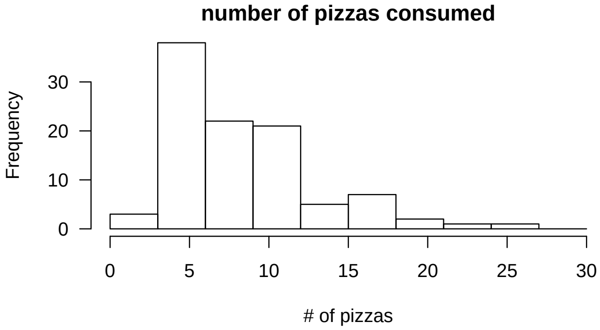
(b) If we take a sample of 100 students and compute the mean number of pizzas consumed, it is reasonable to use the normal approximation to create a confidence interval. TRUE.

(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. FALSE. Since the distribution is right skewed, we would expect the mean to be larger 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. We have a large sample, so we will run a t test even though the data is not normally distributed. $H_0 : \mu = 7, H_A : \mu > 7$. Our test statistic is $t = \frac{8.9-7}{7/\sqrt{100}} = 2.71$ with $df = 100 - 1 = 99$. We would find a (one tail!) p-value to be less than 0.01. This is convincing evidence that the UCLA 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? We cannot compute this probability for a right skewed distribution (with the provided data).

(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? If we were to balance the histogram, it would balance at about 9 (roughly). The median represents the place that splits the data in half, which is roughly 7. The mean should be larger since it is a right skewed distribution.



(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? We have $\mu = 22.1, \sigma = 1.3, x = 20$. Since the distribution is normal, we can use the Z score ($Z = (20 - 22.1)/1.3 = -1.62$) to find the probability from the normal probability table (0.0526). We want 1 minus this: 0.9474.

(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. We will apply the binomial distribution (twice!), which has conditions (i) fixed number of trials, (ii) independent trials, (iii) the response is success/failure, and (iv) fixed probability for a "success" p. All of these conditions are reasonable. To determine P(at least 4 of the next 5 are $>$ 20oz) is the sum of $P(4)$ and $P(5)$. These are each binomial probabilities with $n = 5, p = 0.9474$, and $x = 4$ and $x = 5$, respectively. The sum of these probabilities is $0.212 + 0.763 = 0.975$.

(c) Verify the probability a random burrito weighs between 22 and 23 ounces is 0.29. We first find the Z scores for each of these values ($Z_1 = -0.08, Z_2 = 0.69$), their corresponding lower tail probabilities in the normal probability table (0.468, 0.755), and finally the difference (about 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? Since her burritos are assumed to be independent, the probability is the same as anyone else's burrito being between 22 and 23 ounces: 0.29.

(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.) We need independence, which we already decided was reasonable. Then this scenario means the first four burritos weigh more than 20 ounces and the fifth is under 20 ounces. Using the product rule for independent events, we get $0.9474^4 * 0.0526 = 0.0434$.

(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.) It will be easier to first compute the compliment: the probability that 0 or 1 of the burritos weigh less than 20 ounces. These are each binomial probabilities (verify conditions) and have total probability 0.91. Thus, the probability there will be at least two is $1 - 0.91 = 0.09$.

(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. Since a single burrito will have more variability than the average of five, it is more likely to be far from the mean. Thus, (i) is more likely than (ii).

(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. Here keeping the weight near the mean is good (in that the probability is higher). Thus, the mean of five burritos is more likely to be above 20 ounces.

(i) [IGNORE THIS PROBLEM.] Verify your conclusion in (h) by computing each probability. The first probability is $1 - 0.0526 = 0.9474$ while the second probability is $1 - 0.0002 = 0.9998$ (we use the normal model for the sample mean because (1) we know the SD and do not need to estimate it from the data and (2) the data is normally distributed).

(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? No we could not. computing the probability of a single burrito would not have been possible. Additionally, since 5 is a small sample size, we could not safely apply the Central Limit Theorem to compute the second probability.

(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? This corresponds to 3 or 4 bags, which has probability $0.05 + 0.02 = 0.07$.

(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). We want to compute an expected value. Note that 0 bags costs \$0, 1 bag costs \$20, 2 bags cost \$50, 3 bags cost \$150, and 4 bags costs \$250. Then the expected value is $\$0 * P(0) + \$20 * P(\$20) + \$50 * P(\$50) + \$150 * P(\$150) + \$250 * P(\$250) = 0 + 20 * 0.29 + 50 * 0.22 + 150 * 0.05 + 250 * 0.02 = \$29.30$.

(c) How much revenue would AA expect from 20 random passengers? About \$29.30 per person, which is \$586 total.

(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. I will omit the details. The solution is \$20.45, about \$9 less than their current expected 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? If 26% smoke, then 74% do not smoke. (0.74)

(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. It does not. 50% (67/135) of the smoking group has "friend smokes" while only about 14% (53/365) of non-smokers has "friend smokes".

	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:

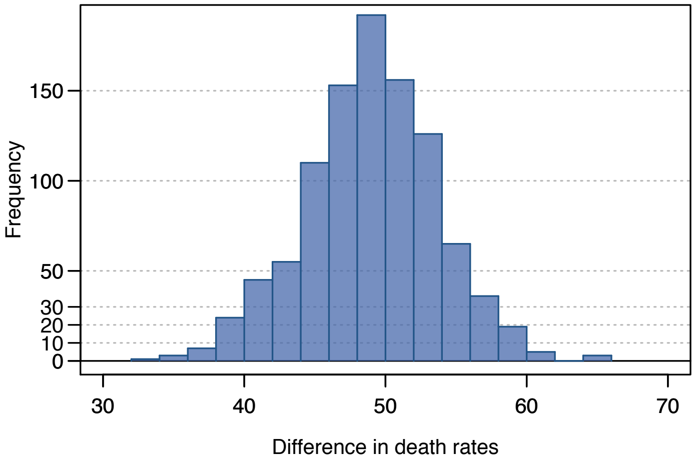
		`outcome`		
		lived	died	Total
`trmt`	drug	693	40	733
	placebo	683	59	742

(a) Setup a hypothesis test to check the effectiveness of the drug. Let p_d represent the drug proportion who die and p_p represent the proportion who die who got the placebo. Then $H_0 : p_p - p_d = 0$ and $H_A : p_p - p_d > 0$, where the alternative corresponds to the drug effectively reducing deaths.

(b) Run a two-proportion hypothesis test. List and verify any assumptions you need to complete this test. (i) The success/failure condition is okay. If we pool the proportions, we get $p_c = \frac{40+59}{733+742} = 0.067$ and $n_p p_c \approx 49$ and $n_d p_c \approx 50$ are each greater than 10. (ii) It is reasonable to expect our patients to be independent in this study. (iii) We certainly have less than 10% of the population. All conditions are met. Then we compute our pooled Z test statistic (pooled since $H_0 : p_1 - p_2 = 0$) as $Z = 1.92$, which has a p-value of 0.028. This is sufficiently small to reject H_0, so we conclude the drug is effective at reducing heart attack deaths.

(c) Does your test suggest `trmt` affects `outcome`? (And why can we use causal language here?) Yes! It is an experiment.

(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? The p-value is the sum of the bins to the left of 40, which is about 35. Since there were 1000 simulations, the estimated p-value is 0.035, which is pretty close to our result when we use the normal approximation and we 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? First, construct a tree diagram, breaking by the treeDiagram variable first and then passClass second. Total the probabilities where Jon passes: $0.65 + 0.27 = 0.92$.

(b) If Jon passed the course, what is the probability he can construct a tree diagram? $0.65/0.92 = 0.71$.

(c) If Jon did NOT pass the course, what is the probability he canNOT construct a tree diagram? First find the probability he did not pass the course (0.08), then the probability he cannot construct a tree diagram AND did not pass (0.13), and take the ratio: $0.06/0.08 = 0.75$.

(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? This is only observational data and not necessarily causal. In fact, it seems like he will guarantee he does poorly if he only understands one topic!

(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? Construct a Venn diagram to use in each part. Then we can compute this probability as 0.91.

(b) How many people like either peanut butter or jelly, but not both? 13%

(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. This is a binomial model problem. Refer to problem (4b) for conditions. Then identify $n = 8, x = 1, p = 0.02$ and the probability is 0.1389.

(d) Are "liking peanut butter" and "liking jelly" disjoint outcomes? No! 78% of people like both.

(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? About half, so 60 students.

(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. We can use the normal approximation. We first compute $Z = \frac{0.59-0.5}{\sqrt{0.5*0.5/120}} = 1.97$, which corresponds to an upper tail of 0.024.

(c) Suppose 61% of the class shows up. Setup and run a hypothesis to check whether the 0.5 guess still seems reasonable. $H_0 : p = 0.5, H_A : p \neq 0.5$. The test is two-sided since we did not conjecture beforehand whether the outcome would over or undershoot 50%. The test statistic is $Z = \frac{0.61-0.5}{\sqrt{0.5*0.5/120}} = 2.41$ and the p-value is $2 * 0.008 = 0.016$. We reject the notion that 0.5 is a reasonable probability estimate.

(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. The estimate is 0.51, and the $SE = \sqrt{.51 * .49/1000} = 0.016$. Then using $Z = 1.96$, we get the confidence interval $estimate \pm Z * SE$, i.e. (0.48, 0.54).

(b) Interpret your confidence interval. We are 95% confident that Obama's true approval rating is between 48% and 54%.

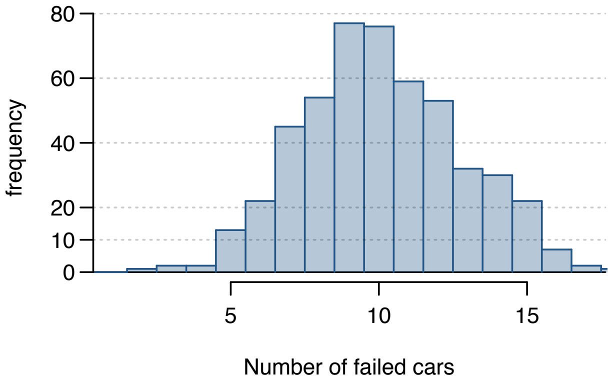
(c) What does 95% confidence mean in this context? About 95% of 95% confidence intervals from all such samples will capture the true approval rating. Most importantly, it is trying to capture the *true* approval rating for the entire population and this is not a probability!

(d) If Rush Limbaugh said Obama's approval rating is below 50%, could you confidently reject this claim based on the Gallup poll? No. Values below 50% (e.g. 49.9%) are in the interval.

(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. You should verify the success/failure condition, that this is less than (or equal to) 10% of the fleet, and that the cars are reasonable independent (which is okay if the sample is random). $H_0 : p = 0.2, H_A : p > 0.2$. Find the test statistic: $Z = \frac{0.28-0.2}{\sqrt{0.2*0.8/50}} = 1.41$. Then find the p-value: 0.08. Since the p-value is greater than 0.05, we do not reject H_0. That is, we do not have significant evidence against the claim that 20% or fewer cars fail the emissions 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. Hypothesis setup as before. We get the p-value from the picture as about $60/500 = 0.12$. Same conclusion as before.



(c) How do your results from parts (a) and (b) compare? The conclusions are the same, however, the p-values are pretty different.

(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.) Generally the randomization/simulation technique is more statistically sound. The normal model is only a rough approximation for samples this small and, while it is good, it is inferior to the other technique. (It will *always* be inferior, even when the sample size is very large.)

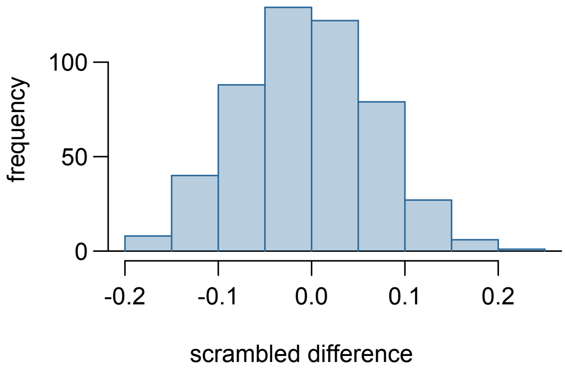
(13) Stephen claims that *Casino Royale* was a significantly better movie than *Quantum of Solace*. Jon disagrees. To settle this disagreement, they decide to test which movie is better by examining whether *Casino Royale* was more favorably reviewed on Amazon than *Quantum of Solace*.

(a) Setup the hypotheses to check whether *Casino Royale* is actually rated more favorably than *Quantum of Solace*. H_0 : *Casino Royale* and *Quantum of Solace* are equally well-accepted. H_A : *Casino Royale* was better accepted than *Quantum of Solace*.

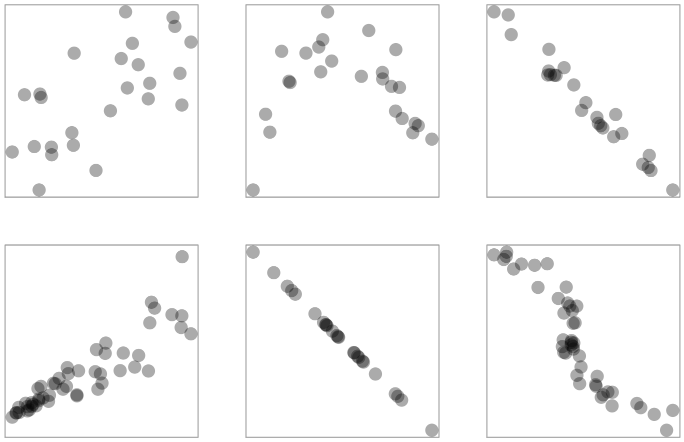
(b) The true difference in the mean ratings is 0.81. Stephen and Jon decide to scramble the results, compute the difference under this randomization, and see how it compares to 0.81. What hypothesis does scrambling represent? It represents the null hypothesis since the scrambled results have an expected difference between the ratings of 0.

(c) Approximately what difference would they expect to see in the scrambled results? Values close to 0.

(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? There were NO observations anywhere remotely close to 0.81, so this suggests *Casino Royale* is rated more favorably and Stephen is, according to our test, right.



(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? I will read across rows for when a linear model is appropriate: yes, no, yes, yes, yes, no. But notice that on the fourth plot (bottom left) that the variability about the line changes, so we cannot move forward with the methods learned in our class.

(b) If we did fit a linear model for each scatterplot, what would the residual plot look like for each case? Completely random, curved shape, completely random, completely random but with more variability for larger values of x, no variability, a "wave".

(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.) The second plot does not represent a linear fit. The fourth plot has different amounts of variability about the line in different spots. The last plot shows a nonlinear fit would be appropriate.

(d) Identify the approximate correlation *for those plots where we could apply our methods* from the following options (i) -1 5, (ii) -0.98 3, (iii) -0.60, (iv) 0, (v) 0.65 1, (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. We use two random digits for each attempt. For their first attempt, assign the random numbers 00-33 for PASS and the rest for FAIL. If the person failed the first exam, then assign 00-71 for PASS and the rest for FAIL. Continue until they PASS and record the number of attempts.

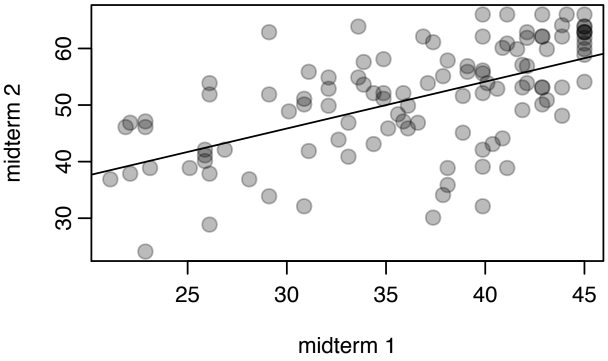
(b) Run your simulation ten times, i.e. run your simulation for ten people. Some random numbers:

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

Trials: 4 (73, 86, 75, **44**), 3 (39, 92, **13**), 2 (80, **15**), 2 (49, **38**), 1 (**30**), 1 (**20**), 5 (88, 79, 80, 78, **68**), 1 (**14**), 3 (83, 75, **36**), 2 (66, **46**).

(c) Based on your simulation, estimate the average number of attempts it takes people to pass the test. The average of our ten trials is 2.4, which estimates the true value (about 2.75).

(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? Yes. There are no obvious signs of curvature. The values at the far top of the class appear to deviate slightly, but not so strongly that we should be alarmed.

(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. First find the slope: $b_1 = r\frac{s_y}{s_x} = 0.58 * 9.7/6.8 = 0.827$. Next find the intercept: $\bar{y} = b_0 + 0.827 * \bar{x}$ (plug in the means and solve for $b_0 = 21$). The equation is $\widehat{midterm}2 = 21 + 0.827 * (midterm1)$.

(c) Interpret both the y-intercept and the slope of the regression line. If person 1 scored a point higher than person 2 on exam 1, we would expect (predict) the first person's score to be about 0.827 points higher than the second person on exam 2. The y-intercept describes our predicted score if someone got a 0 on the first exam. We should be cautious about the meaning of this intercept since our line doesn't include any values there.

(d) If a random student scored a 27 on the first exam, what would you predict she (or he) scored on the second exam? Plug 27 in for midterm1 in the equation and get $\widehat{midterm2} = 43.3$.

(e) If that same student scored a 47 on midterm 2, did s/he have a positive or negative residual? Positive since she scored higher than we predicted (her score falls *above* the regression line).

(f) Would you rather be a positive or negative residual? Positive – that means you outperformed your predicted score.

(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. False. Change "half" to "less than half" or change "skewed to the right" to be "symmetric" or "normal".

(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. False. We need more observations before we can safely assume the normal model. Change "10" to "100" (or another large number) or "skewed to the right" to "normal".

(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 $(x = 5ounces/week, y = 2.0)$. Can we conclude that eating more spinach would cause an increase in this person's GPA? No. Regression lines only represent causal relationships when in the context of a proper experiment.

(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? It might be that sun exposure is influencing both variables! Sun exposure is a lurking variable.

(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? An election represents a measure on the entire (voting) population. It is not a sample.

(h) In Exercise 12, 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. False! Confidence levels do not correspond to probabilities.

(i) True or False: In your confidence interval in Exercise 12, you confirmed that 95% of all sample proportions would fall between (0.48 and 0.54). False! The confidence interval only tries to capture the population parameter, in this case the true proportion.

(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? Yes since p-value < 0.05. They still reject H_0. How about 0.01? Unknown since we don't know if the p-value is smaller or larger than 0.01. False!