Stat 13 final, Prof. Rick Paik Schoenberg, 9/14/23, 10am-11:30am, via zoom.

1. You may use a calculator, a pencil, and any books and notes you want during the exam, but no internet searching or communicating or use of the computer other than to submit your answers.

2. Final numerical answers have been rounded to 3 significant digits.

3. There are 25 multiple choice questions worth 4 points each.

4. No partial credit is given for multiple choice questions. Choose ONE answer only.

5. Use the exam that matches the FIRST LETTER of your **FIRST NAME**!!!!!!!!!!!
If THE FIRST LETTER OF YOUR FIRST NAME is A through L, then use the exam called finalA.pdf .
If THE FIRST LETTER OF YOUR FIRST NAME is M through Z, then use the exam called finalM.pdf .

6. You have from 10am to 11:30am. By 11:30am you must EMAIL me your answers, to frederic@stat.UCLA.edu . Your email should just contain your answers, like
ADDBC CDAAB BBCCD DAABC DDCCA. You do not need to show work. Make sure your answers are exactly in the correct order!

7. You must zoom in to the usual zoom while taking the exam. If you have a question during the exam, ask it using chat.

_____ 1. The average person drinks 1/2 a gallon of water per day. Researchers study a simple random sample of 1000 residents in a certain city, and ask them about their water consumption as well as several other questions to assess their overall health. They find that those who drink more than 1 gallon of water per day tend to be in significantly better health, on average, than those who drink less water. They conclude that everyone should drink more water if they want to be healthier. Do you agree?
a. Yes, because drinking a lot of water flushes out the negative energy in your system and thus causes better health.
b. No, because those who drink more than 1 gallon of water per day most likely exercise frequently, and the exercise may be what leads to better health, not the water.
c. Yes, because Hydrogen and Oxygen in water are necessary for survival, so the more of these one intakes, the heathier one is.
d. No, because subjects may be lying about how much water they drink, and this lying is likely making them healthier.
e. No, because the experiment was not blinded, so the subjects knew they were receiving more than 1 gallon of water per day and this knowledge would likely contribute to a placebo effect resulting in better health metrics on average.

_____ 2. Researchers investigating the genetic causes of a rare disease take a random sample of 800 patients suffering from the disease, and a random sample of 800 others as controls. The researchers check 500 genes to see if any are statistically significantly related to disease incidence. For each gene, they will do a two-sided z-test to determine if the two groups differ significantly in terms of the proportion with the gene, and consider the gene statistically significantly linked to the disease if the p-value < α. The researchers are not sure what value of α to choose, and they approach you, saying they would like to pick α in order to ensure that, if none of the 500 genes is actually linked to the disease, then there is at most a 5% chance that at least one of the genes would be statistically significantly linked to the disease. What value of α would you recommend?
a. 0.01%. b. 0.05%. c. 0.1%. d. 0.5%. e. 5%.

_____ 3. The management at a hospital would like to predict how long patients are likely to stay in the hospital after a certain surgery, given the length of time taken by the surgery. They take a simple random sample of 100 patients who have had the surgery, and perform linear regression to predict the length of stay using surgery length as an explanatory variable. They find the correlation r = 0.75 and this correlation is statistically significantly different from 0. Should the hospital use surgery length as a predictor of length of stay in the hospital?
a. No, because this is an observational study and there may be many hidden confounding factors.
b. No, because those whose surgeries take longer might be less healthy, on average, than those whose surgeries take less time.
c. No, because surgeries are painful, and it may be this pain, not the actual length of time of the surgery, that is actually causing the hospital stay to be longer.
d. No, because those whose surgeries take longer are more likely to adhere to the hospital stay protocol.
e. Yes, because the hospital is interested in prediction of hospital stay length, so whether surgery length is causally linked to hospital stay length is irrelevant.

For the next 5 problems, suppose a researcher is studying the percentage of students at USC and UCLA who have been a *Primary Responsible Driver in a car accident* (PRD). The researcher takes a simple random sample of 300 USC students and a simple random sample of 500 UCLA students, and checks the driving records of each of the students. The researcher finds that 52 of the sampled USC students have been a PRD, and 40 of the sampled UCLA students have been a PRD.

______ 4. Which of the following is true?
a. Since 52>40, the evidence suggests that going to UCLA is positively associated with being a PRD.
b. Since 52 + 40 < 300, the evidence suggests that there are more USC students than UCLA students who have been PRDs.
c. Since 40/500 < 52/300, the evidence suggests that going to USC is positively associated with being a PRD.
d. Since 52/300 > 40/500, the evidence suggests that being a PRD makes you more likely to go to UCLA than USC.
e. Since 52/300 > 40/500, the evidence suggests that students who have been PRDs are more likely to participate in this study.

______ 5. What is the pooled sample percentage who have been PRDs, in both groups combined?
a. 5.05%. b. 11.5%. c. 15.0%. d. 18.5%. e. 20.5%.

______ 6. Using this pooled sample percentage, under the null hypothesis that the two groups have the same percentage of people who have been PRD, what is the standard error for the difference between the two percentages?
a. 2.33%. b. 3.09%. c. 4.23%. d. 4.71%. e. 5.04%.

______ 7. What is the size of the Z statistic for the difference between the two group percentages?
a. 2.51. b. 2.93. c. 3.59. d. 4.01. e. 5.22.

______ 8. What is the most plausible explanation for this large Z statistic?
a. Being a PRD in an accident can lead to concussions or other brain injuries, which can cause one to choose to attend USC instead of UCLA.
b. The difficulty and stress of attending USC makes it more difficult for students to concentrate on driving and therefore more likely to cause a car accident.
c. USC is more expensive than UCLA, and this expense makes it more difficult to afford a safe car and therefore makes it more likely for students to become a PRD.
d. USC students are likely to come from wealthier households on average than UCLA students, and therefore are more likely to have had a car and thus more likely to have been a PRD.
e. Being a PRD is a unique experience that students can use to differentiate themselves on their college applications and therefore be more likely to get admitted to USC.
f. Those who have been a PRD are likely to be healthier, better educated, and more likely to adhere to road signs and traffic regulations than those who have never been a PRD.

For the next 3 problems, suppose a researcher is studying the weights of watermelons to see if weight helps predict the number of seeds a watermelon contains. She takes a simple random sample of 14 watermelons, and counts their seeds. She finds the following.

> weight (kg): mean 5.02, median 7.31, sd 1.53.
> seeds: mean 37.24, median 49.03, sd 10.97.
> r = 0.20.

______ 9. Do the assumptions for a t test, to see if the correlation is significantly different from 0, seem to be satisfied here?
a. No, because the sample size is small and the variables are left skewed and thus not normal.
b. No, because the standard deviations are too large.
c. No, because the two sample means are too different.
d. No, because the effect size is too small.
e. No, because the sample sizes are small and the standard deviations are too large.

______ 10. If one were to do a t test anyway, to see if the correlation is significantly different from 0, what would the t statistic be?
a. 0.362. b. 0.509. c. 0.615. d. 0.707. e. 1.78.

______ 11. How might you use simulations to analyze this data?
a. Randomly shuffle the number of seeds for each melon, recalculate r, repeat, and see for what fraction of the simulations the size of r is 0.20 or more.
b. For each melon, switch its number of seeds with its weight, recalculate r, repeat, and see for what fraction of the simulations the size of r is 0.20 or more.
c. For each melon, flip a coin, and if it is heads, leave the data on that melon alone, but if it is tails, then switch its number of seeds with its weight. Then recalculate r, repeat, and see for what fraction of the simulations the size of r is 0.20 or more.
d. For each melon, choose another melon at random, and switch both their weights and their number of seeds. Repeat for each of the 14 melons, then recalculate r, repeat, and see for what fraction of the simulations the size of r is 0.20 or more.
e. Order the melons from lightest to heaviest, and recalculate r, repeat, and see for what fraction of the simulations the size of r is 0.20 or more.

For the following 5 questions, suppose a researcher is studying sodium consumption (X) and total cholesterol level (Y). She surveys a simple random sample of 1000 American adults and finds their mean daily sodium consumption is 3.2 grams, with a standard deviation of 1.0 grams. She finds their mean total cholesterol level is 186 mg/dL with a SD of 10. Both variables seem approximately normally distributed. The correlation is 0.60. The regression equation predicting Y from X has an estimated intercept of 173.2 and an estimated slope of 6.0 and the associated p-value for this slope, for a 2-sided t-test, is 1.1497%.

______ 12. Using the regression line, what would be the predicted cholesterol level for a respondent whose sodium consumption is 4 grams?
a. 190.7. b. 191.9. c. 193.2. d. 195.1. e. 197.2.

______ 13. How much would such a prediction typically be off by, in mg/dL?
a. 5.50. b. 6.00. c. 6.50. d. 7.50. e. 8.00.

______ 14. Why shouldn't one trust this regression line to predict someone's cholesterol level if their sodium consumption is 8 grams?
a. The sample size is insufficiently large.
b. The value of 8 grams of sodium is too far outside the range of most observations.
c. The correlation of the ANOVA is a t-test confidence interval with statistical significance.
d. The data come from an observational study, so there may be confounding factors.
e. The cholesterol values are heavily right skewed, so the prediction errors are large.

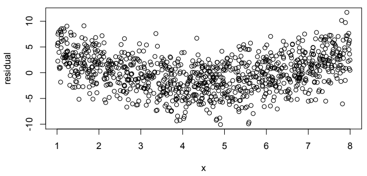
______ 15. How should one interpret the estimated slope of 6.0?
a. Each extra gram of sodium you consume daily causes your cholesterol to increase by 6 mg/dL.
b. Each extra unit of cholesterol you consume causes your sodium level to increase by 6 grams.
c. For each extra gram of sodium you consume daily, your predicted cholesterol increases by 6 mg/dL.
d. The Z-score corresponding to the correlation between sodium and cholesterol is 6.
e. The proportion of variance in cholesterol levels explained by the regression equation is 6 percent.

______ 16. Does the fact that the correlation between X and Y is significant provide strong evidence that sodium intake causes an increase in cholesterol level?
a. Yes, because correlation is proof of causation.
b. No, because those who eat high sodium foods may also eat other foods that raise cholesterol levels.
c. No, because the sample size is too small.
d. No, because the data are heteroskedastic.
e. No, because high cholesterol levels make people crave sodium.

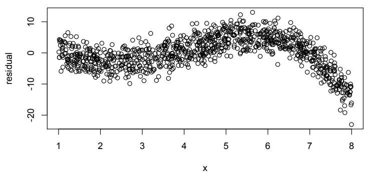
______ 17. What is a residual in a linear regression?
a. An observation with an x-value very far from most of the other observed x-values.
b. The difference between an observed y-value and the observed x-value.
c. An observation where both x and y are very far from most of the other points on the scatterplot.
d. The difference between an observed y-value and the y-value predicted using the regression line.
e. The difference between an observed y-value and the nearest other observed y-value.

______ 18. The four plots below, labeled a,b,c, and d above them, each represent a residual plot from a linear regression. In which case do the usual assumptions for linear regression appear to hold?

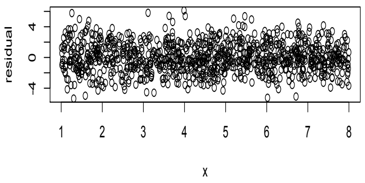
a



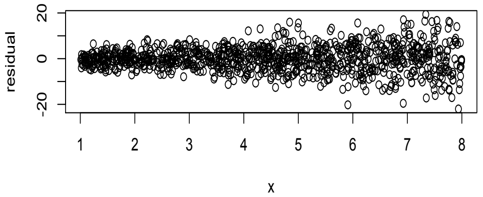
b



c



d



______ 19. In the portacaval shunt example, why did the studies with historical controls find that the portacaval shunt seemed to be associated with lower death rates?
a. Those getting the shunt smoked more.
b. Those getting the shunt were healthier.
c. Those getting the shunt were genetically predisposed to die younger.
d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.
e. None of the above.

______ 20. Suppose there is no such thing as extra-sensory perception (ESP), yet the majority of published studies find that psychics predict events statistically significantly better than non-psychics. What might be a plausible explanation?
a. The psychics in these studies are unusually talented.
b. Most psychic studies that do not find statistical significance do not get published.
c. The studies finding statistical significance suffered from the placebo effect.
d. The studies finding statistical significance had confounding factors.
e. The studies finding statistical significance did not use simple random sampling.

For the next 5 problems, suppose in a random sample of 10,000 edible chicken eggs :

the mean number of calories per egg is 58.2,
the SD is 11.1 calories,
the IQR is 23.0 calories,
the median is 50.1 calories,
the 25th percentile is 40.0 calories,
and the range is [22.4, 87.8] calories.

______ 21. Which of the following is true?
a. The data are symmetric. b. The data are right skewed.
c. The data are normally distributed. d. The data are left skewed.

______ 22. Which of the following is an interval containing the number of calories of exactly 25% of the eggs in the sample?
(a) [40.0, 60.0]. (b) [40.0, 63.0]. (c) [50.1, 73.1]. (d) [50.1, 63.0]. (e) [22.4, 87.8].

______ 23. What is the standard error, in calories, for the mean number of calories per egg?
a. 0.111. b. 0.235. c. 0.353. d. 0.412. e. 0.509.

______ 24. If we were to do a one-sample, two-sided Z-test at significance level 5% to see if the population mean number of calories per egg might really be 60.0 calories, what would be the size of the Z statistic?
a. 3.52. b. 4.44. c. 8.23. d. 12.9. e. 16.2.

______ 25. Continuing the previous problem, what would the conclusion of the test be?
a. We fail to reject the null hypothesis that the population mean is greater than 60.0 calories.
b. We fail to reject the alternative hypothesis that the sample mean is 60.0 calories.
c. We fail to reject the null hypothesis that the population mean is 60.0 calories.
d. We reject the null hypothesis that the population mean is 60.0 calories.
e. We fail to reject the null hypothesis that the population mean is different from 60.0 calories.