Online S520 Midterm 2

Instructions

- Type your answers in a Word document or Latex and submit through Canvas/Assignment/Midterm2.
- This exam is due Monday 11:59pm, Nov 13th (Pacific Time). **Late submission will not be accepted!**
- **You must NOT discuss this exam with anyone other than the instructor and the TAs until the due date has passed.**
- Write explanations for all your answers. **Answers alone will not get credit.** For questions where you use R, you must give R code, but the code alone is not a sufficient explanation.
- You may use any R functions.
- Round answers sensibly, e.g. to 3 significant figures. Unrounded or inaccurately rounded answers may receive point deductions.
- Give both numerical results (e.g. P-values, confidence intervals) and substantive conclusions. For example:
  - "We reject the null hypothesis." — NOT MANY POINTS
  - "The P-value is 0.005. This means the data gives strong evidence that three-toed sloths have more toes than two-toed sloths." — LOTS OF POINTS

What can I ask the instructor by email?

- If you think there is an error in the exam, notify the instructor immediately.
- General questions about course material or help handling the data. (However, it's easier to talk about these issues during office hours.)

What can ask at the instructor's or TA's office hours?

- General questions about course material.
- Help handling the data.

What can I ask other students?

Nothing.

1. Boxes of cereal are advertised as having a net weight of 8 ounces. The weights of boxes are assumed to be normally distributed. A new cereal box-filling machine is purchased, and we wish to be sure that on average, it puts at least the correct amount of cereal in the boxes. We randomly select 16 boxes, and find they have weights with sample mean 8.10 ounces and sample standard deviation 0.20 ounces.

   (a) (2 points) Suppose your data included all the box weights. How would you check the normal distribution assumption? Explain what would indicate a violation of this assumption. (Note: You do not have to perform the check on the given data.)

   (b) (8 points) Perform a test of the null hypothesis that the average net weight of the boxes of cereal produced by the new machine is less than or equal to 8 ounces, against the alternative that it is greater than 8 ounces. Test at level $\alpha = 0.05$. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion)

2. To test whether my friend's fish Googly had psychic powers, I wrote R code to display two windows. I entered either "Left" or "Right" depending on which way Googly was facing. Then the random number generator in R selected either the left or the right window, with probability 0.5 for each, in which to display a star. Let p be the probability Googly guesses correctly on a given trial (assume this is constant.) In 80 trials, Googly correctly guessed the window with the star 41 times.

   (a) (3 points) Using mathematical notation, write down null and alternative hypotheses for a one-sided test.

   (b) (3 points) If the test statistic is the number of correct guesses (41) in 80 trials, write down R code to find the P-value of a one-sided test.

   (c) (2 points) Even without R, we can see that Googly's success rate was close to its expected value under the null, so the one-tailed P-value will be close to 0.5. State your conclusion about the fish's psychic powers.

   (d) (2 points) Continued from part (b). If you only known that the R code `dbinom(40, 80, 0.5)` gives the number 0.0889, how would you find the exact P-value of the test? (Hint: use the property of a symmetric probabilty distribution.)

3. The file `snoqualmie14.txt` contains the **daily** precipitation (in inches) in Snoqualmie Falls, WA, for a random sample of 365 days. After saving the file to your computer, you can load it into R by entering the command:

   ```
   rainfall = scan(file.choose())
   ```

   and then selecting the file.

   (a) (4 points) Show that the data does not come from a normally distributed population. Include a graph to support your answer.

   (b) (8 points) The mean **annual** rainfall in Seattle is 37.7 inches per year. Test (at level $\alpha = 0.05$) the hypothesis that the mean rainfall in Snoqualmie Falls is *different* from the mean rainfall in Seattle. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion)

4. In one year in the United States, 4.247 million babies were born. Of these, 2.173 million were male and 2.074 million were female. With very few exceptions (e.g. identical twins), the sexes of the babies are independent, so we can use the binomial distribution to model the number of babies that are female. Let p be the probability that a random (future) newborn is female.

   (a) (2 points) What percentage of the babies were female? (To get credit for this question, you must give your answer as a percentage and you must round appropriately.)

   (b) (3 points) Suppose we wish to test the null hypothesis that the probability a baby is female is 50%. Write down null and alternative hypotheses in mathematical notation for this test.

   (c) (10 points) Find the P-value, **using both Binomial probability and Normal approximation**. (Carefully show your work and R codes.)

   (d) (2 points) Using the Central Limit Theorem, find a 95% confidence interval for the probability that a birth is female.

   (e) (2 points) Explain what this confidence interval means without using the word "confident."

   (f) (2 points) The P-value for your test in part (c) is basically zero. From this and your confidence interval, write in a sentence your conclusion about the probability that a random newborn is female.

5. The basketball player Steph Curry sometimes shoots free throws with his mouthguard in his mouth, and sometimes shoots free throws with his mouthguard outside of his mouth. His free throw statistics for one season were:

   - Free throws with mouthguard in: 110 completed, 13 missed (89.4%)
   - Free throws with mouthguard out: 198 completed, 16 missed (92.5%)

   His observed free throw completion rate was slightly higher when his mouthguard was outside his mouth. However, we should check whether the difference could be plausibly explained as luck.

   (a) (4 points) Using the Central Limit Theorem, find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *in.* Give a numerical answer.

   (b) (4 points) Using the Central Limit Theorem, find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *out.*

   (c) (3 points) Suppose we wish to test the null hypothesis that Curry's probability of completing a free throw is the same with his mouthguard in as it is with his mouthguard out. The P-value for such a test is 0.33. What does this P-value tell you? Explain.

6. Rosene (1950) studied how quickly hairs on radish roots absorbed water when they were immersed. For each of eleven radishes, she measured the rate of influx of water for a young root hair and an old root hair on that radish. The data is given below.

Radish	Old	Young
A	0.89	2.13
B	0.49	1.16
C	0.91	2.60
D	0.80	1.58
E	0.56	1.53
F	0.79	1.70
G	0.47	2.67
H	0.50	2.64
I	1.08	2.19
J	1.65	2.54
K	1.94	4.46

Table 1: Radish root hair absorption data. Rates are in cubic microns per square micron per minute.

For each pair, the "Young" number is bigger than the "Old" number, so even without a test it's clear that young roots take in water more quickly. But how much more quickly?

(a) (4 points) Explain what test we should consider to use based on the data and context, and why. (Hint: 1-sample or 2-sample test, z-test of t-test, etc)

(b) (4 points) Use a normal probability plot of the differences (old minus young) and the sample size to explain why we should be hesitant do such a test (the one you proposed in the previous question) on these differences (old minus young).

(c) (2 points) Explain why we should not take the logs of these differences (old minus young.)

(d) (8 points) Instead of using the differences, we can look at the **ratio**: old divided by young. This ratio looks to come from a much closer to normal distribution. Write R code to find a **90%** confidence interval for the average value of this ratio.

7. The basketball player Stephen Curry was the NBA's Most Valuable Player for the 2015–16 season. He is known for being very good at shooting (throwing the basketball into the hoop) from long distance. The file `currydist.txt` in the Data folder on Canvas contains data on the 1,598 shots he attempted during the 2015–16 season. Of these shots, 805 of them were successful. For the purpose of this analysis, we treat the data as random independent samples (this isn't quite true but is close enough.)

   The variables in the file are:

   - `distance`: distance of Curry from the hoop, in feet;
   - `venue`: "Home" if the shot was during a home game (in Oakland), "Away" if it was somewhere else;
   - `made`: 1 if the shot was successful, 0 if it was unsuccessful.

   Load the data into R, e.g. using

   ```
   currydist = read.table("currydist.txt", header=TRUE)
   # or
   currydist = read.table(file.choose(), header=TRUE)
   ```

   Then you can extract individual variables from the data frame using the dollar sign, e.g. `currydist$distance` for the distance variable (remember that R is case-sensitive.)

   (a) (4 points) Create two vectors in R: one containing the distances for Curry's shots at home, and one containing the distances for Curry's shots away. Remember that we use square brackets for subsetting in R:

   ```
   currydist$distance[currydist$venue == "Home"]
   ```

   Draw a graph that compares the full distributions of these variables and describe your observations. (Note: Boxplots don't let you see the full distributions.) By "distribution", we mean the PDF or CDF. You should make it visually obvious which distribution is which.

   (b) (8 points) Is there enough evidence to show that the (population) average distance of Curry's shots at home is different from the (population) average distance of Curry's shots away? Perform an appropriate significance test and find an appropriate confidence interval. Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion (i.e. not just "Reject" or "Don't reject.")

   (c) (6 points) The significance test you did requires assumptions. Show that the assumptions are met, or, if they are not met, explain why they can be overlooked.