Final Exam

Online S520

Instructions

- Type your answers in a Word document or Latex and submit through Canvas/Assignment/Final Exam.
- This exam is due Wednesday 11:59pm, Dec 13 (Pacific Time). **Late submission will NOT be accepted!**
- **You must not discuss this exam with anyone other than the instructor and the TA 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 should be able to answer all questions using the methods we covered this semester. **Don't search and use any approach/tests/R functions not covered in this course!**
- 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 NFL

In a National Football League (NFL) regular season, each team plays 16 games. Let “team wins” be the number of regular season wins by a team in a particular season (taking a tie as half a win.) Since on average teams win half their games, the distribution of team wins has mean 8. Assume the distribution of team wins stays about the same from year to year.

There is a positive correlation between a team’s wins one year and their wins the next ($r = 0.327$.) Because of this, we can use regression to predict a team’s win one year by using their wins the previous year.

1. (3 points) Find the regression line to predict a team’s wins one year from their wins the previous year. (Hint: You do not need to know the standard deviations, but if you cannot work out how to do the problem without standard deviations, make a reasonable guess.)
2. (3 points) In 2013, Houston had 2 wins, while in 2014 they had 9 wins. In 2013, Dallas had 8 wins, while in 2014 they had 12 wins. Use regression to predict 2014 wins for Houston and Dallas based on their 2013 wins. Which team exceeded their prediction by a larger margin?
3. (2 points) A cable sports analyst who does not know statistics suggests a different prediction system — simply predict a team will win as many games one year as they did the previous year. Explain convincingly to the analyst why in the long run, this prediction system will not be as accurate as a regression line. (Note: This will be graded fairly strictly.)

2 Citation

The file `citations.txt` contains the number of citations for a random sample of 1000 journal articles published in 1981. (The data is from the ISI.) After saving the file to your computer, you can load it into R by entering the command:

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

and then selecting the file.

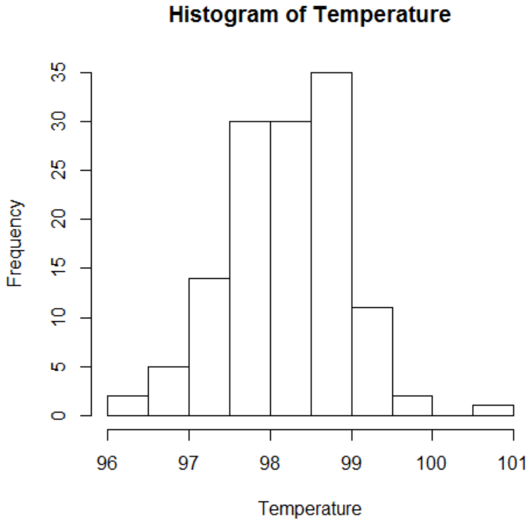
1. (2 points) Draw an appropriate graph of the data, and briefly describe (in words) the shape of the distribution.
2. (4 points) Find an approximate 95% confidence interval for the mean number of citations.
3. (5 points) The command

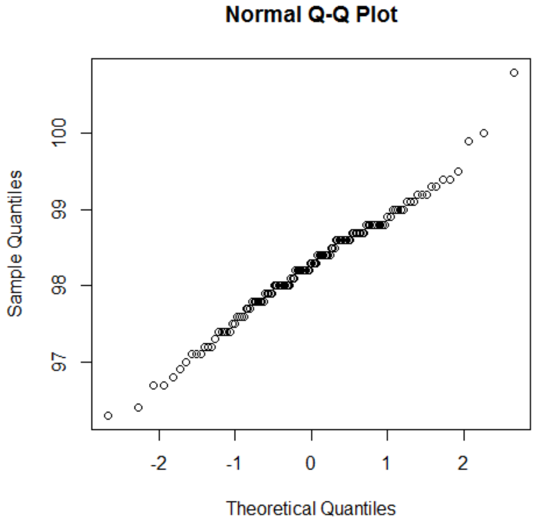
   ```
   sum(citations==0)
   ```

   will tell you how many of the 1000 journal articles had no citations. Use this statistic to find an approximate 95% confidence interval for the proportion of journal articles with no citations.

3 Body Temperature

It has long been asserted that the average body temperature was 98.6 degrees Fahrenheit. A 1992 study aimed to test this hypothesis. (The data presented here is fictionalized but similar to the study data.) The body temperatures of a sample of 130 adults were taken to one decimal place. The mean temperature of the sample was 98.5 degrees, the median was 98.3 degrees, and the standard deviation was 0.73 degrees. The figures below show a histogram of the data and a normal quantile plot of the data.





1. (2 points) From the information provided, does it seem like the distribution of body temperatures is (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? Explain your choice.

2. (4 points) Let μ be the population mean body temperature (not the median!) We wish to test $H_0 : \mu = 98.6$ against $H_1 : \mu \neq 98.6$. Assuming this is a random sample, calculate a test statistic and give R code for the P-value of this test. (Only use R code for the P-value!)

3. (4 points) Construct an approximate 95% confidence interval for the population mean body temperature. (Show how you calculate it and give a numerical answer — apply the Central Limit Theorem if necessary.) Summarize the evidence for or against the null hypothesis.

4 Exam and Anxiety

The file `examanxiety.txt` on Canvas contains information on a number of variables measured on a sample of 103 students taking a math exam:

- `Code`: a label for the individual in the sample (not scientifically interesting.)
- `Revise`: hours spent revising for the math exam.
- `Exam`: score on a math exam on a scale from 0 to 100.
- `Anxiety`: "math anxiety" on a scale from 0 to 100 (100 is most anxious.)
- `Gender`: female or male.

Assume the data is a random sample from a larger population of students.

1. (5 points) Is there a significant difference between average anxiety for the population of male students and the population of female students? Perform an appropriate significance test, stating hypotheses, a P-value, and a substantive conclusion.

2. (3 points) Let anxiety be your x-variable and exam score be your y-variable. Find the regression line to predict exam score from anxiety. Carefully explain (in words or using math) what your regression line means — do not just paste R output.

3. (5 points) Let anxiety be your x-variable and exam score be your y-variable. Which of the following regression assumptions are met? Make arguments and/or show graphs to support your answers.

   (a) *Linearity*

   (b) *Independence*

   (c) *Equal variance (homoskedasticity)*

   (d) *Normality of errors*

   (e) *Bivariate normality*

5 Student ID

Take a random sample from a bivariate normal data based on your Student ID number as follows:

```
set.seed(StudentID) # use the numerical value of your Student ID
x = rnorm(500)
y = 2 * x + rnorm(500)
```

Note: DO NOT print out the data. You should not have the same data as anyone else.

1. (3 points) For your sample, use R to find the mean of x, the mean of y, the standard deviation of x, the standard deviation of y, and the correlation between x and y. (You must give R code for credit.)
2. (4 points) Find the equation of the regression line to predict y from x.
3. (4 points) We select a point (x_i, y_i) from the parent population of your data. Suppose $x_i = 1$. What is the probability that y_i is greater than 3? (You may find this either by using theory or based on your data.)
4. (4 points) Find a 95% confidence interval for the slope coefficient of the regression line predicting y from x.

6 Singer

The file `singer.txt` contains the heights of 235 opera singers. Load the file into R using `read.table()`. We will consider four groups of singers: the 66 sopranos, 62 altos, 42 tenors, and 65 basses. (We will ignore any difference between, for example, "Soprano 1" and "Soprano 2".) The data is close enough to normal and homoskedastic to perform an analysis of variance.

1. (2 points) Suppose we wish to test the hypothesis that sopranos, altos, tenors, and basses all have the same average height. Construct an ANOVA table to test this hypothesis. Carefully write down the hypotheses and give a conclusion.
2. (3 points) Two more interesting null hypotheses to test are:

   (a) Sopranos and altos have the same average height

   (b) Tenors and basses have the same average height

   Test each of these null hypotheses at level 0.025, giving P-values and conclusions.

7 College and high school earnings

A researcher performs a survey of a random sample of young people aged 25 to 32. The sample of such people included 340 individuals with four-year college degrees and 260 individuals with high school degrees but no college. (Individuals with community college degrees or no high school degree are excluded here, as are individuals who are not working; you should also exclude them from your analysis.) The results are given in two data files (posted in the Data folder of the Files section of Canvas):

- `college.txt` contains the annual earnings of the 340 individuals with college degrees.
- `highschool.txt` contains the annual earnings of the 260 individuals with high school degrees but no college.

Note: The data is fictitious but realistic. Save this data to your computer and read it into R, e.g. by using `scan(file.choose())`.

1. (2 points.) Suppose we wish to estimate the PDFs of (i) the earnings of young people with college degrees, and (ii) the earnings of young people with only high school degrees. For each of these populations, draw ONE graph that shows an estimate of the PDF (so two graphs in total. These should be the only graphs you include in your submission.) For each graph, explain in words what it tells you about the shape of the underlying distribution. (You must include these explanations to get credit.)

2. (2 points.) Test the hypothesis that young people with college degrees have the same **mean** earnings as young people with only high school degrees.

3. (2 points.) Find a 95% confidence interval for the difference between the mean earnings of young people with college degrees and the mean earnings of young people with only high school degrees.

4. (2 points.) Calculations show that the approximate 95% confidence interval for the *median* earnings of young people with only high school degrees is ($28,500, $36,300), and the first quartile of the college degrees earnings is $22,890. This ($22,890) is below the lower bounds of the confidence interval for the median earnings of young people with only high school degrees. Based on this, a commentator draws the following conclusion: "At least a quarter of college students would have probably had higher earnings if they had not gone to college." Convince the commentator that this conclusion is not proven by the data.