S350 final

Spring 2020

Due 11:59 pm, Tuesday 5th May

Instructions, warnings, and veiled threats

- You may freely refer to books, notebooks, and the web. However, **you must not discuss this exam with anyone other than the instructor and the AI until the due date has passed**, except on Piazza to the extent discussed below.
- **Include all R code** as an appendix or as a .R or Markdown file. (Do not *just* upload a code file.)
- Include all graphs you're asked to draw in the body of your document.
- Give your answers in sentences, e.g. "The degrees of freedom for the test was 123." Just giving R output is not sufficient for full credit.
- Round answers sensibly. Confidence intervals stated to six significant figures will not get full credit.

What may I post about on Piazza?

- If you think there is an error in the exam, let us know as soon as possible.
- You may for clarification on the questions.
- You may ask general questions about course material.
- You can ask for help handling the data. However, it may be easier to talk about these issues during office hours.

What may I ask at the lecturer's/TAs' office hours?

- General questions about course material.
- Help entering and manipulating the data.
- Questions about Steph Curry, radishes, and the Iowa housing market.

What can I ask other students?

Nothing.

1 Danish kids (10 points)

A survey in Denmark studied the first three children in a random sample of 154,443 families with three or more children. One variable the observed was the number of the children in each family that were girls (out of 3.) The distribution of the number of girls is given in the table below.

Number of girls in the first three children	Number of families with that number of girls
0	23236
1	58529
2	53908
3	18770
Total	154443

Let X be the number of girls in the first three children of a random Denmark family with at least three children. Let μ be the expected value of X.

Questions

(a) Find the mean and variance of the sample. You may use either the plug-in or the sample version of the standard deviation; it doesn't matter.

(b) Find a 95% confidence interval for μ.

(c) Suppose we wish to perform a similar survey in the United States to find the average number of girls in the first three children in families with at least three children. If we want a 95% confidence interval of width 0.02, how many families should we sample?

2 Cashing out (10 points)

In a randomized experiment in Georgia, a treatment group of 592 ex-convicts received cash payments upon being released from prison, while a control group of 154 ex-convicts received no money upon release. In the first year after release, the members of the treatment group averaged 16.8 weeks of paid work, with a standard deviation of 15.9 weeks. The members of the control group averaged 24.3 weeks of paid work, with a standard deviation of 17.3 weeks. The samples were large and right-skewed.

We wish to test the hypothesis that the treatment and control will, on average, result in the same number of weeks worked. We don't have the full data set, but you have enough information to answer the questions below.

Questions

(a) To allow interpretation, the researcher would prefer not to transform the data. Explain why we may do a Welch's t-test even though the samples are right-skewed.

(b) Calculate the test statistic, and give R code to find the P-value. Hint: The correct number of degrees of freedom is about 225.

(c) Calculate a 95% confidence interval for the average difference in weeks worked between the treatment and control.

3 In the heights (15 points)

One of the first uses of linear regression was to predict the heights of children from the heights of their parents. But is this a reasonable thing to do? We'll look at a data set on parent-child pairs from a famous 1903 study by Karl Pearson and Alice Lee, in the file `PearsonLee.txt`. The variables are:

- `parentHeight`: the parent's height in inches.
- `childHeight`: the child's height in inches.
- `parentGender`: "Father" or "Mother."
- `childGender`: "Son" or "Daughter."

Note: Because of rounding, many numerical values appear more than once, e.g. there are 46 mother-daughter pairs where the mother is 62.5 inches tall and the daughter is 63.5 inches tall. You might want (but do not have to) to add random noise to the data for the purpose of plotting, but you should use data *without* random noise added in your regression model.

Questions

(a) Using all the data, find and write down the equation of the regression line to predict a child's height in inches from a parent's height in inches.

(b) The assumptions of linear regression are:

(i) Linearity
(ii) Independence of observations or errors
(iii) Homoskedasticity (equal variance) of errors
(iv) Normality of errors

Which of these assumptions are met, and which are violated? Include graphs and explain your answers. (Hint: `ggplot()` would be useful here.)

(c) Instead of fitting one regression on all the data, it might be more appropriate to fit separate models for fathers and mothers. (Of course, you could also fit separate models for sons and daughters, but we won't do that here.)

Fit separate linear regression models to:

(i) Predict a child's height from a father's height
(ii) Predict a child's height from a mother's height

Give the regression equations. Are the assumptions of linear regression above more or less reasonable when you fit two separate models?

4 Flowers, birds, and nickels (10 points)

Does money affect people's judgments of how often something has happened?

A randomized experiment was carried out on 77 German psychology students. Each student was shown a sequence of 57 flower pictures and 57 bird pictures in random order. Afterwards, each student flipped a coin. If the coin was heads, the student was told they would receive five cents for each flower picture they saw. If the coin was tails, the student was told they would receive five cents for each bird picture they saw. They were then asked to guess the number of flower pictures they were shown and the number of bird pictures they were shown.

The file `flowersandbirds.txt` on Canvas contains three columns:

- `Group`: The result of the coin toss ("Heads" or "Tails.")
- `FlowerGuess`: The student's guess for how many flower pictures they saw.
- `BirdGuess`: The student's guess for how many bird pictures they saw.

Download the data to your computer and read it into R, e.g. by downloading the file to your working directory and running the code

```
FlowerBird <- read.table("flowersandbirds.txt", header = TRUE)
```

Questions

For parts (a) and (b), consider the data to be two measurements (flower guess and bird guess) on one sample from one population of German psychology students.

(a) Is the population mean for the flower guesses equal to the correct answer, 57? If not, what can you say about the value of the population mean? Perform a hypothesis test. State hypotheses, the specific type of test, the test statistic, P-value, and a substantive conclusion.

Note: For this question we want to make direct statements about the population mean, so avoid transformations.

(b) Draw a well-labeled graph to illustrate the relationship between individuals' flower and bird guesses. Label your graph clearly. Also write a summary of what the graph tells you about the relationship, including the form of the relationship (e.g. linear or curved), its direction, and its strength.

For part (c), consider the data to be two samples from two (theoretical) populations: students for whom the coin came up "Heads" and students for whom the coin came up "Tails."

(c) We can use the difference between the two guesses (flower guess minus bird guess) as a quantitative variable. Test the hypothesis that the expected value of this variable is the *same* for both the Heads population and the Tails population. State the specific type of test, the test statistic, P-value, degrees of freedom, and a substantive conclusion.

5 Moving to Iowa (15 points)

The file `IowaHouses.txt` contains data of 29 sales of houses in Ames, IA. The two variables are:

- `Price`: the sale price of the house in dollars. This is seriously right-skewed, so for the following analysis, use the natural log of this variable.
- `Condition`: the condition of the house rated on a 1 to 9 scale, where 9 indicates the best condition.

Suppose we want to know: Does the typical house price depend on the condition of the house?

Because the sample size is small, we must choose between imperfect methods. Two possibilities are:

- Linear regression using `Condition` as a numeric variable
- ANOVA using `Condition` as a categorical variable

You can use the R function `factor()` to make a variable categorical and the function `as.numeric()` to make a variable numeric. You'll need to be careful you have the `Condition` variable in the right format for the questions below.

Questions

(a) Fit a linear regression using `Condition` as a numeric predictor and `log(Price)` as the response variable. Does this analysis provide evidence that the typical house price depends on condition? Explain.

(b) Perform an analysis of variance using `Condition` as a categorical variable and `log(Price)` as the response variable. Does this analysis provide evidence that the typical house price depends on condition? Explain.

(c) What are the problems with doing the linear regression in (a)? What are the problems with doing the ANOVA in (b)? If you had to choose one of these two analyses, which would you choose, and why? (Think about things like assumptions and power.)