Exercises 4 · Uncertainty

Due Monday, February 22, 2016

(1) Bootstrapping

Complete the "Creatinine, revisited" walkthrough on the class website.[1] This will introduce you to the idea of bootstrapping as a way to approximate a sampling distribution when you cannot simulate samples from the population (as you did in the previous question). Make sure you also read up through page 116 in Chapter 5 of the course packet.

[1] http://jgscott.github.io/teaching/r/creatinine/creatinine_bootstrap.html

Once you've done this:

(A) Return to the data set "ut2000.csv" on SAT scores from UT students across all 10 undergraduate colleges. Calculate an approximate 95% confidence interval for the difference in mean SAT math (SAT.Q) scores between students in the colleges of architecture and liberal arts. I can think of at least two ways you could accomplish this, so make sure you describe precisely what you did and why, and report the interval.

(B) Fit a regression model for graduating GPA in terms of SAT combined score (SAT.C) and College (with no interaction term), and provide a 95% confidence interval for the slope of the SAT score.

(C) In your own words, briefly describe the idea of bootstrapping (both what we do and why we do it).

(2) Bootstrapped prediction intervals

For this problem, use the data set "shocks.csv." This data was taken by Monroe Shocks and Struts, a company that manufactures high-performance shock absorbers for top-end cars. Monroe offers a range of shock absorbers for cars of various sizes. These different shocks are distinguished from one another by their "rebound," a number which describes how aggressively the vibrations from the road are absorbed by the shock. Having an accurate understanding of a shock's rebound is important for safety; you don't want to put shocks designed for an SUV on a small car, or vice versa.

As part of its manufacturing process, Monroe tests each shock absorber to make sure it performs to the required rebound specification. They have one very accurate test of the shock's rebound, but this test is expensive. They also have a cheaper test, but this is less accurate.

In "shocks.csv," you have rebound readings on 35 different shock absorbers for both the expensive test and the cheap test. If the cheap test can accurately predict the result of the expensive test with minimal uncertainty, then it's OK to use the cheap test. But if it can't, then the expensive test must be used instead.

(A) Suppose the company is willing to use the cheap test as long it can predict at least 90% of the total variation in the readings given by the expensive test. In light of this data, should they use the cheap test? Why or why not?

(B) Now suppose the company adopts a more specific standard, and decides it is willing to use the cheap test if both of the following criteria are met under the assumptions of the normal linear regression model. First, the slope of the regression line for the expensive test, given the cheap test, is close to 1, as measured by a 95% confidence interval. Second, the 95% prediction interval for the value of the expensive test, given the cheap test, is no wider than 16.5 units of rebound, as measured from center to endpoint. (Or, measured from endpoint to endpoint, the interval can be no wider than 33 units of rebound.) This criterion must be met for readings of the cheap test (x) in the low (510), middle (550), and high (590) end of the rebound scale. That is, if the prediction interval for y is too wide at any of these three different x values, then the cheap test is not precise enough and cannot be used.

    In light of the data and these criteria, should the company use the cheap test? If not, what criterion was missed and how? Use bootstrapping to account for your parameter and prediction uncertainty, and describe your methodology and results in a careful write-up. Make sure you use enough bootstrapped samples so that you can address the questions without Monte Carlo error substantially affecting your results.

(3) The PREDIMED trial: a first look

For this problem, we'll revisit the PREDIMED trial, described in the course packet. For details, see this paper. The data is in predimed.csv from the course website.

The main goal of the trial was to understand the relationship between a Mediterranean diet and the likelihood of experiencing a major cardiovascular event (stroke, heart attack, or death from heart-related causes). Trial participants were assigned to one of three treatment arms,

described in the paper as: "a Mediterranean diet supplemented with extra-virgin olive oil, a Mediterranean diet supplemented with mixed nuts, or a control diet (advice to reduce dietary fat)."

The predimed.csv file has data on many variables on each trial participant; we'll focus only on two:

- group: which treatment arm the person was assigned to
- event: yes or no, did the person experience a cardiac event during the study period

If you look at a contingency table for these two categorical variables, you get the following.

```
> xtabs(~event + group, data=predimed)
     group
event Control MedDiet + Nuts MedDiet + VOO
  No     1945           2030          2097
  Yes      97             70            85
```

Thus there is a hint that cardiact events happened at a slightly higher rate among participants in the control group.

Your task is to use a permutation test to assess whether this difference in event rates across the dietary categories could be explained due to chance. Note: you've seen a walkthrough of this kind of thing for a 2x2 table, but this is a 3x2 table with three levels of the predictor. You will have to define your own test statistic that collapses the association across categories to a single number. You have considerable freedom to choose a test statistic here; just make sure you are clear about what you are doing and why.

Note: we'll revisit this question and this data set later in the semester using more sophisticated techniques that take into account not just whether an event happened, but how long it took to happen.