INFO370 Problem Set 4: Inference, Monte Carlo Simulations

Your name:

Deadline: Wed, Nov 6th 2pm

Instructions

This problem set revolves around statistical hypotheses and inference.

1. Please write clearly! Answer each question in a way that if the code chunks are removed from your document, the result is still readable!
2. Don't output irrelevant, or too much of relevant, information. A few figures is helpful. A few thousand figures is only noise.
3. Please keep data file in the same folder as your code, and read these w/o any path like `"data.csv"` (or `"./data.csv"`). This makes the checking your code much easier!

1 Are fathers and sons of different height? (50pt)

Here we analyze the fathers' and sons' height data. You will, see that, in average, sons are taller than fathers. But can this difference just be a result of the small sample?

1. (5pt) load the *fatherson.csv* data. Perform basic description of it: what is the number of observations? Are there any missings?

   Note: *fheight* and *sheight* are fathers' and sons' height, respectively (in cm).

2. (5pt) Comment the measure type (nominal, ordinal, ...) of the variables. Which statistics are appropriate for this kind of measures? What is the expected range and value type (discrete/continuous/...) of these?

3. (5pt) Describe fathers and sons: compute the mean, median, standard deviation, and range of their heights. According to these figures, who are taller: fathers or sons?

4. (10pt) Lets add a graphical comparison. Plot histograms (or even better, density plots) of both heights. Comment the histograms/density plots. How do these look like? What do these suggest in terms of fathers' and sons' relative height?

5. (10pt) But is this difference statistically significant? Let's do a `t`-test. Here I ask you to *compute yourself the* `t`*-value*, do not use any pre-existing functions! What do you find?

   Hint: read OIS 5.3 (2017 version)

6. (10pt) Look up the `t`-distribution table. (Or compute the relevant quantiles). What is the likelihood that such a `t` value happens just by random chance?

   Hint: read OIS 5.1.2.

   Hint 2: what is the *degrees of freedom* in current case?

7. (5pt) finally, state clearly your conclusion.

2 Fathers and sons: the Monte Carlo approach (50pt)

Next, let's re-visit the fathers and sons height, but this time by doing MC analysis on computer. You will proceed as follows: create two samples of random normal numbers as in the data above, using the mean and standard deviation over both fathers and sons. Call one of these samples "fathers" and the other "sons". What is the difference of their means? And now you repeat this exercise many times and see if you can get as big a difference as you got above.

1. (5pt) compute the overall mean and standard deviation of combined fathers' and sons' heights.

2. (10pt) now create two sets of random normals, both with the same mean and standard deviation that you just computed above. Call one of these "fathers" and the others "sons".

   What is the father-son mean difference? Compare the result with that you found in the previous problem.

3. (10pt) Now repeat the previous question a large number of times R (1000 or more). Each time store the difference, so you end up with R different values for the difference.

4. (3pt) What is the mean of the difference values? Explain what do you get.

5. (3pt) What is it standard deviation? Compare it to that you computed in the previous problem for the difference in data (when doing t-test).

6. (4pt) What is the largest difference (in absolute value)?

7. (10pt) find 95pct quantile of (the absolute value) your difference. Compare this number to the actual father-son difference you found in the data.

8. (5pt) finally, increase your number of repetitions R as much as your computer can stand. See how large difference you can find. Can you get anything comparable to the actual difference in the data?

   Hint: if you like coding challenges, try to run this simulation in parallel!

References