# IMT573 Problem Set 5: Inference and Monte Carlo Simulations

Your name:

Deadline: Wed, Nov 12th 8pm

## Instructions

This problem set revolves around statistical hypotheses and inference. It contains two tasks: a) test a statistical hypothesis using simulations; and b) test it using t-test.

- Please write clearly! Answer each question in a way that if the code chunks are removed from your document, the result is still readable!

## 1 Are Sons Taller than Fathers?

### 1.1 Descriptive Analysis (20pt)

1. (3pt) load the *fatherson.csv* data. Do the basic descriptive work on it: what is the number of observations? Are there any missings?

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

2. (7pt) 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?

3. (7pt) Lets add a graphical comparison. Create density plots of both heights on the same figure. Comment the density plots. Which distribution do these resemble? Do they agree with the conclusion above that sons are taller?

   Hint: you can do density plots with `stat_density` when using *ggplot*.

4. (3pt) Finally, for the further reference, compute how much taller are sons in average.

### 1.2 Monte Carlo approach (40pt)

If you did the previous part right, you probably saw that sons are taller by $\sim 2.5$ cm. But can it be just a statistical fluke? Just bad luck for the fathers? Let us figure it out through simulations. You will proceed as follows: create two samples of random normals, similar to the data above, using the overall mean and standard deviation in the data. Call one of these samples "fake fathers" and the other "fake sons". What is the difference in their means? Is this close to 2.5 you saw in data?

It is probably not. But maybe this was just an unhappy experiment. So now lets repeat this exercise many times and see how big or how small differences you typically see between fake sons and fake fathers.

1. (5pt) compute the overall mean and standard deviation of pooled fathers' and sons' heights. (I.e combine all heights, and compute just one mean and one standard deviation for this combined data.)

2. (5pt) now create two sets of random normals, both with the same mean and standard deviation that you just computed above. Call one of these "fake fathers" and the others "fake sons".

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

   Example: Compute mean difference for an hypothetic sample of mean 100 and standard deviation 10, and sample size 5:

```
fakefathers <- rnorm(5, mean=100, sd=10)
print(fakefathers)

## [1] 122.87247  88.03228  93.05707  95.87707  90.29327

fakesons <- rnorm(5, mean=100, sd=10)
print(fakesons)

## [1]  90.52720 107.48139  98.83045 101.52658 121.89978
```

```
diff <- mean(fakefathers) - mean(fakesons)
print(diff)

## [1] -6.026646
```

The difference is -6.027. This is much larger than the actual difference of 2.5 (in absolute value). But obviously, we were using wrong values in this example.

3. (8pt) 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. (2pt) What is the mean of the differences? Explain why do you get what you get.

   Hint: it should be close to 0.

5. (2pt) What is the standard deviation of the differences?

6. (2pt) What is the largest value amont the differences (in absolute value)? How does it compare to the actual sons/fathers difference of 2.5?

7. (8pt) find 95% confidence intercals for the differences you computed. Does the actual difference fall inside or outside of the CI?

   Hint: use the R function `quantile` for this.

8. (8pt) What is your conclusion? Can you confidently say that sons are taller than fathers? Why?

   Hint: is the claim: $H_0$: sons and father are equally tall (in average) compatible with the data?

### 1.3 t-test (20pt)

Simulations are great but also cumbersome. t-test comes to help.

1. (5pt) Compute the standard error for the difference of means.

   Hint: read OIS 7.3 (p 267)

2. (5pt) Compute the t-value (OIS denotes it by T). Here we ask to *compute it yourself*, not use any pre-existing t-test functions! What is the t-value you find?

   Hint: read OIS 7.3

   Hint 2: it is large, above 8.

3. (5pt) Look up the t-distribution table. What is the likelihood that such a t value happens just by random chance?

   Hint: read OIS 7.1.3

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

4. (5pt) finally, state clearly your conclusion. Is the actual difference you see compatible with $H_0$ that fathers and sons are of similar height?

## 2 Extra credit: parallel computing (10pt = 1EC credit)

Here your task is to repeat the previous exercise (only MC part of it) using parallel processing:

1. (1pt) Lets return to your MC example. Increase the number of repetitions R substantially, so it runs for at least 10 seconds.

2. (4pt) Conduct the MC analysis using a parallel loop.

   Hint: check out the packages *parallel* and *foreach*

3. (3pt) Time your code. Create a table that shows how the simulation time depends on the number of employed CPU cores.

   Can you get a noticeable speed improvement by running the simulation code in parallel?

4. (2pt) finally, use the fastest approach and increase R as much as your computer can stand. Sure, you can afford to run it for 10 seconds. Maybe you can afford an hour. Report how large R did you use and how long time did it take.

   See how large difference you can find. Can you get anything comparable to the actual difference in the data?