Mathematical Statistics II

Statistical Computing Activity: Module 6

One purpose of these Statistical Computing Activities (SCAs) is to give you a chance to explore statistics using the computer. Another purpose is to give you more skills in thinking about the randomness that is life.

Usually, like here, these SCAs will have a theme and several problems dealing with that theme or purpose. The reason for that extra layer of complexity is to tie what we do in the class with what we can use these techniques for in our lives as statisticians and/or consultants and/or full members of a democratic society.

The purpose of this particular activity is to give you some practice in collecting and analyzing data. Since the data are two numeric variables and since we are interested in the relationship between those two variables, we will use linear regression to perform the analysis.

⋆⋆⋆

The Procedure

The research question is "What is the relationship between the number of iterations and the time it takes to perform a simulation exercise?"

Overview. This is an interesting activity because it tries to get at the fact computers take time — sometimes a lot of it. In this course, I have focused on providing code that is educationally helpful. Each line in the code mirrors the thought process to achieve the end. However, such code is often rather slow. Code can be modified to make it faster, but those modifications frequently make the code difficult to understand.

In this activity, you will run simulations with different numbers of iterations and estimate the time it will take to do a million iterations.

To make things more interesting, the three of you will have different simulations to run.

My Turn: Sample Means

In this section, I determine how long it takes to simulate the distribution of a sample mean. I will generate a sample of size $n = 500$ from a $X \sim \mathcal{N}(0, 1)$, calculate the mean, save that mean, then repeat B times. The parameter B will take on 20 values of my choice. I will record the time it takes to perform the simulation for each of those B values, then estimate the time it takes to do $B = 1 \times 10^7$ iterations.

Along the way, I will determine if a linear model is appropriate or if I need to do some transformations to the variables.

Here is my code for the simulation

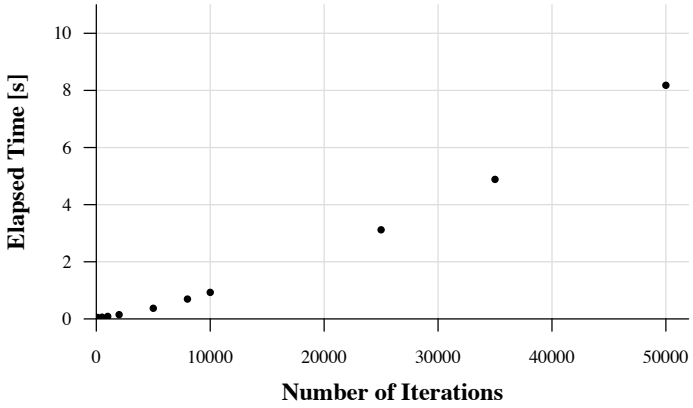
```
start = Sys.time()

B = 1e4
ts = numeric()
for(i in 1:B) {
  x = rnorm(500)
  ts[i] = mean(x)
}

end = Sys.time()
end-start
```

You should know — or be able to guess — what each line does in this code.

By changing the value of B, recording the final value (elapsed time), I created the following ugly scatter plot.



Next, I will model the elapsed time using this data, checking the assumptions, transforming the variables (if needed) to ensure the model is appropriate, etc.

Finally, I will create a prediction interval for the time it takes to do $B = 1,000,000$ and $B = 1,000,000,000$ iterations. The first will help you make sure your model is working well. The second is what I am after.

Your Turn!

Now that you see what I did, here are your assigned simulations:

(1) Distribution of sample means with $n = 10,000$ and X comes from a standard Cauchy.
(2) Distribution of the sample variances with $n = 10,000$ and X comes from a standard Cauchy.
(3) Distribution of the p-values from the Shapiro-Wilk test when the data are from a t distribution with $\nu = 2$ degrees of freedom ($n = 1000$).
(4) Distribution of the p-values from Fisher's chi-square variance test (this is implemented as `onevar.test` in the class's supplemental functions) when the data are from a t distribution with $\nu = 3$ degrees of freedom, the sample size is $n = 1000$, and the hypothesized variance is $\sigma^2 = 3$.

Remember you need to choose 20 values of B that will help you best estimate the relationship between the two variables (number of iterations and elapsed time). Also, remember you are predicting the elapsed time for two specific, and large, values of B.

Step 1: Do the Experiment. Do the experiment. Make sure you write down your elapsed times for each number of iterations. In fact, you may want to create the data table in your notes. [I fully expect this step to take about 1800 seconds.]

Step 2: Check the Requirements. You know that there are (at least) six assumptions that must be met for ordinary least squares (OLS) regression to be mathematically sound. Check them. If any of the assumptions is not met, transform the dependent or independent variable appropriately. Then, retest.

Continue this until you have an appropriate model.

Step 3: Predict the Times. Predict the time it takes to run a million iterations. [Now, run those million iterations to see if your observation is actually within the prediction interval. Do this if you have time, however.]

Finally, predict the time it will take to perform a billion (1×10^9) iterations. [Why did I need to define 'billion' in this step?]