PLSC 502: "Statistical Methods for Political Research"

Exercise Six
October 20, 2016

Part I

We're again going to use simulations to examine the properties of the techniques we've been learning. As I noted in class, a key difference between Student's original version of the t-test and the Welch version in use today is that the latter performs better when the variances of Y are different for different values of X. In this part of the exercise, you're going to show (in a rudimentary way) that fact.

Your task is to demonstrate (via simulation) the practical differences between Student's and Welch's version of the t-test. To make things a bit more concrete, use the following general guidelines:

- For all simulations, assume that Y is Normally distributed, with mean -2.0 and variance σ_0^2 for cases with $X = 0$ and mean 2.0 and variance σ_1^2 for cases with $X = 1$. Thus, the "population" value of $\mu_1 - \mu_0$ will always $= 4$.
- Draw samples with n_0 observations with $X_i = 0$ and n_1 observations with $X_i = 1$.
- Welch's version of the t-test is implemented via (e.g.):

  ```
  test <- t.test(Y~X)
  ```

  while Student's (common-variance) version is:

  ```
  test <- t.test(Y~X, var.equal=TRUE)
  ```

Your assignment is the following:

1. Begin with samples comprising $n_0 = n_1 = 20$ (so $N = 40$), and where $\sigma_0^2 = \sigma_1^2 = 2$ (i.e., common variances). Calculate both Student and Welch t-tests for these data, and compare their t values, P-values, and associated confidence intervals. Talk briefly about what you find.
2. Next, relax the assumption of common variance in Y, but keep the sample sizes the same. More specifically, simulate data with $n_0 = n_1 = 20$ but where $\sigma_0^2 = 1$ but $\sigma_1^2 = 10$. Again: Examine t-statistics, P-values, and confidence intervals, and briefly discuss what you found when you relaxed the common-variance assumption.
3. Finally, while keeping the variances different (that is, for $\sigma_0^2 = 1$ and $\sigma_1^2 = 10$), change the sample sizes:

   (a) First, so that $n_0 = 20$ but $n_1 = 200$,

   (b) And then the reverse, so that $n_0 = 200$ but $n_1 = 20$.

   For each case, discuss briefly what you learn.

Finally: Note that when I say "calculate tests... and compare them," that does *not* mean to do so for a single random sample / simulation.

Part II

We might call this part of the homework exercise "A Half-Dozen Questions," because you're going to be asked to answer six questions about counties in the U.S., using a (simple random) sample ($N \approx 400$) of the data we used for Exercise Three. For each question, report the appropriate estimates, test statistic(s), confidence intervals, etc. along with whatever relevant inferential information you want to present, and write a 1-2 sentence answer to the question.

More specifically, choose **six** of the following hypotheses, and use your sample data to test them:

1. $\overline{\texttt{LandArea}} = 1000$.
2. $\overline{\texttt{JanTemp}} = 32.9$ degrees F.
3. $\overline{\texttt{JulHumid}} = 59$.
4. $\overline{\texttt{PCPersIncome05}} = \$27{,}367$.
5. $\overline{\texttt{MedianAge05}} = 40$.
6. $\overline{\texttt{MHPSACty07}} = 0.14$.
7. $\overline{\texttt{UnEmpRate05}} = 5.9$ percent.
8. $\overline{\texttt{WarmHumid}} = 0.14$.
9. $\overline{\texttt{PopGr00-05}} = 2$ percent.
10. $\overline{\texttt{AmenityScale}} = 0.05$.
11. $\overline{\texttt{PropTaxPerCap02}} = \$825$.
12. $\overline{\texttt{CrimeRate04}} = 2770$.
13. Finally, compare each of your conclusions / inferences to the actual population means in the data for Exercise Three. (You don't have to write anything up on this; I just thought you might be curious).

This exercise is due by 5:00 p.m. EST on Friday, October 28, 2016, and is worth 50 possible points.