Biostat 561: Homework 4

Instructor: Amy Willis, Biostatistics, UW

25 April, 2019

Homework due May 1, 3:15pm

Link to Homework 4 submission: `https://classroom.github.com/a/zEjWAL5-`

If are doing a simulation study for a research project and would like to convert your current simulation into a "simulator" Simulation to receive credit for this homework in place of the below question, please e-mail me with details of the proposed simulation study and report.

Suppose we are interested in comparing the prevalence of a quality across two groups. For example, we may wish to test the null hypothesis that the proportion of incoming Biostatistics PhD students with R experience is the same as the proportion of incoming Statistics PhD students with R experience.

If we sample n_i individuals from the ith group, and $\hat{p}_i = x_i/n_i$ of them have the quality of interest, the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is normally distributed under the null hypothesis that $p_1 = p_2$, if individuals are sampled independently. Note

$$\hat{p} = (x_1 + x_2)/(n_1 + n_2).$$

However, this test may not be reliable, especially when the sample sizes are small and the true prevalence is low.

Design a simulation to investigate the Type 1 and Type 2 error rate of this test. What factors contribute to the error rate? Refer to your results in your answer.

Some things to note

- Of course, you must use the `simulator`
  - The `simulator` vignettes provide clear examples of the structure of the simulator
- The `simulator` provide an excellent mechanism for producing reports via the `writeup.Rmd` file; your report should be completed using this infrastructure
- Type 1 error: Rather than evaluating the Type 1 error rate for a specific α, it is much better to look over all possible choices of α using qq-plots. `qqplotr` is a nice package for qqplots in `R`.
- Type 2 error: Typically Type 2 error rate (1 - power) is evaluated as a function of effect size.
- A nice paper that includes investigation of error rates of a proposed hypothesis test is by Bryan Martin (UW Statistics). If you are not sure how to conduct a sensible simulation investigating error rates, I recommend taking a look at his preprint: https://arxiv.org/pdf/1902.02776.pdf. He also made his code available through github, and he used the simulator, so with some digging (also a useful research skill!) you will find the code he used to produce his figures.

You are invited but not obligated to investigate other tests for the null hypothesis, including permutation-based approaches!

This is the only question on this week's homework, so I expect a very high quality report! This includes a clearly described set of conclusions and clearly labelled figures.

Submit the `simulator` code along with your report, but do not submit the data generated by the `simulator`.