ECON 626: Empirical Microeconomics

Multiple Testing

Department of Economics
University of Maryland
Fall 2019

1. **Implementing Benjamini-Hochberg adjusted q-values.**

   (a) Generate a list of $k = 20$ p-values, including several below 0.05 that are quite close together and would normally be rejected (absent adjustment). How many hypotheses would be rejected if you did not impose any sort of multiple test correction?

   (b) Calculate Bonferroni-adjusted p-values — how many would be rejected now?

   (c) Calculate Benjamini-Hochberg q-values as follows:

      i. Sort p-values and calculate ranks (smallest to largest)

      ii. Calculate preliminary adjusted q-values by multiplying p by rank/k

      iii. Whenever the q-value for hypothesis j is above that of hypothesis $j + 1$ (etc.), adjust accordingly by setting the q-value for hypothesis j equal to the q-value for hypothesis $j + 1$

   (d) How many hypotheses are rejected under each of the three approaches (no adjustments, Bonferroni, and Benjamini-Hochberg)?

   (e) Compare the q-values you calculated to those you obtain using Michael Anderson's do file (`anderson_qvalues.do`) — are they similar?

2. **Comparing approaches.**

(a) Generate a data set with 1000 observations and 100 outcome variables. Generate a treatment dummy that is equal to one for half the observations. Suggested Stata code appears below.

```
forvalues i = 1/100 {
    gen y‘i’ = rnormal()
}
gen treatment = (_n<=500)
```

(b) Calculate p-values from 100 regressions of y_i on t and a constant.

(c) Calculate Bonferroni-adjusted p-values.

(d) Calculate Benjamini-Hochberg q-values.

(e) Calculate Romano-Wolf adjusted p-values using Stata’s `rwolf` command.

(f) Plot the four sets of adjusted p-values. How do they compare (under the null)? How many (true) hypotheses are rejected under each approach?

(g) Repeat steps (a) through (f) in a new `do` file that imposes a treatment effect. Specifically, for $i = 1, \ldots, 100$, replace y_i with $y_i + \lambda \times$ `treatment` where $\lambda > 0$ is a scale factor that you choose so that you detect the effect of treatment 80 percent of the time (when p-values are not adjusted for multiple testing).