Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

November 2, 2023
34 total points

Problem Set 8 Solutions

---

1. Data are gathered on 16 students. Half of the students are randomly assigned to a new tutoring program and half have their usual schooling experience. A study finds that test scores for the tutoring program students are 10 points higher on average than those for the other students (p=*0.3*). The mean test score overall is 200 with a standard deviation of 40. (**6 points**)

   (a) Is the study's finding statistically significant? Explain why or why not.

   > No, the p-value from a test of differences in means is reported to be 0.3, above most thresholds for statistical significance.

   (b) Is the study's finding practically significant, in your opinion? Explain what practical significance means here.

   > To assess practical significance, we can compare the 10 point average gain from tutoring to the overall standard deviation on the exam of 40, for Cohen's $d = 10/40 = 0.25$. According to the review by Kraft, a 0.25 standard deviation difference is on the large side for educational interventions.

   (c) Would you conclude that the tutoring program is effective based on these results? Ineffective? Briefly explain.

   > No, the results are inconclusive. We cannot rule out large beneficial effects of tutoring, but the estimated impact is statistically insignificant and may have occurred by chance. A larger sample would help us determine whether the effect is real or not.

2. Two studies were commissioned to evaluate an intensive program designed to enhance social and emotional learning (SEL) among adolescents. The index used to measure SEL has a mean of 50 and a standard deviation of 10. Study 1 failed to find a statistically significant improvement in SEL, with a 95% confidence interval for the gain in SEL of $(-7, 17)$. Study 2 also failed to find a statistically significant improvement in SEL, with a 95% confidence interval of $(-2.5, 1.5)$. Which of these two studies (if any) is more valuable to a policymaker, in your opinion, and why? (**4 points**)

In both cases the confidence interval contains zero, so neither provides evidence of a statistically significant impact of the program on SEL. However, the 95% confidence interval in Study 2 is much narrower: $(-2.5, 1.5)$. If the confidence interval provides a range of null hypotheses that our data are consistent with (cannot reject), then Study 2 can largely rule out large positive (or negative) effects of the program on SEL. (At the upper bound, a +1.5 effect would be a $1.5/10 = 0.15$ standard deviations, a modestly large effect). The confidence interval in Study 1 ranges from -7 to 17, encompassing both large negative and large positive effects; it provides little guidance to a policymaker.

3. Simulate the multiple hypothesis testing problem in Stata by doing the following steps: (**24 points**)

See attached Stata output.

(a) Begin by creating a dataset with 1,000 observations. Create a variable called *tstat1* consisting of random draws from the $N(0, 1)$ distribution. Think of *tstat1* as test statistics from a sampling distribution in which H_0 is true. Create a second variable called *pval1* that contains $Pr(z > tstat1)$. (That is, the probability of drawing a test statistic at least as large as *tstat1*). Create a third variable called *reject1* that equals 1 if $pval1 < 0.05$ and 0 otherwise. In what proportion of "samples" does $reject1 = 1$ (a Type I error)? (**4 points**)

(b) Now create four additional variables (*tstat2, tstat3, tstat4,* and *tstat5*) defined in the same way as *tstat1*. Create analogous variables *pval2-pval5* and *reject2-reject5* based on *tstat2-tstat5*. Think of each row in your dataset as 5 independent hypothesis tests conducted using *tstat1-tstat5* in the same sample (where H_0 is true in all cases). In what proportion of "samples" does $reject = 1$ for 1 or more hypothesis tests? (**4 points**)

(c) Use the family-wise error rate formula shown in class to calculate the proportion of samples that one would expect to reject H_0 in one or more of these hypothesis tests. How does this compare to your answer in (b)? (**4 points**)

(d) Create five new *reject* variables that use the Bonferroni-corrected thresholds for rejection. Based on these new variables, in what proportion of "samples" does $reject = 1$ for 1 or more hypothesis tests? (**4 points**)

(e) Use the following syntax to create five new *tstat* variables. These test statistics are still random draws from the N(0,1) distribution, but they are now dependent.

That is, when one *tstat* is higher than average, the others tend to be higher than average (and when one *tstat* is lower than average, the other tend to be lower than average).

```
clear
matrix c=(1, 0.8, 1, 0.8, 0.8, 1, 0.8, 0.8, 0.8, 1, 0.8, 0.8, 0.8, 0.8, 1)
drawnorm tstat1 tstat2 tstat3 tstat4 tstat5, n(1000) corr(c) cstorage(lower)
```

As in parts (a)-(b), create *pval* and *reject* variables based on these new test statistics. In what proportion of "samples" does $reject = 1$ for 1 or more hypothesis tests? Is this proportion higher or lower than that in (b)? What is your intuition about why these differ from b (if they do)? (**4 points**)

(f) If you apply the Bonferroni correction to the variables created in part (e), how often will you reject 1 or more of these tests? (**4 points**)

```
.
. // ************************************************************************
. // LPO.8800 Problem Set 8 - Solution to Question 3
. // Last updated: November 1, 2023
. // ************************************************************************
.
. cd "$pset"
C:\Users\corcorsp\Dropbox\_TEACHING\Statistics I - PhD\Problem sets\Problem set
> 8

.
. // *************
. // Question 3
. // *************
.
. // *************
. // Parts (a)-(b)
. // *************
.
. // Create dataset
. // tstat1-5 are independent draws from N(0,1)
. // pval1-5 is the one tail probability z>tstat
. // reject1-5 equals one if pval<0.05
.
. clear

. set seed 3121

. set obs 1000
number of observations (_N) was 0, now 1,000

. gen tstat1=rnormal(0,1)

. gen tstat2=rnormal(0,1)

. gen tstat3=rnormal(0,1)

. gen tstat4=rnormal(0,1)

. gen tstat5=rnormal(0,1)

. gen pval1=1-normal(tstat1)

. gen pval2=1-normal(tstat2)

. gen pval3=1-normal(tstat3)

. gen pval4=1-normal(tstat4)

. gen pval5=1-normal(tstat5)

. gen reject1=(pval1<0.05)

. gen reject2=(pval2<0.05)

. gen reject3=(pval3<0.05)

. gen reject4=(pval4<0.05)
```

```
. gen reject5=(pval5<0.05)

.
. // How often is reject1==1? 5.9% in my case (close to expected)
.
. tabulate reject1

    reject1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        941       94.10       94.10
          1 |         59        5.90      100.00
------------+-----------------------------------
      Total |      1,000      100.00

.
. // How often is at least one test rejected? 23.9% in my case
.
. egen rejectany=rowmax(reject*)

. tabulate rejectany

  rejectany |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        761       76.10       76.10
          1 |        239       23.90      100.00
------------+-----------------------------------
      Total |      1,000      100.00

.
.
. // *************
. // Part (c)
. // *************
. // FWER with 5 independent tests (and alpha=0.05 on each test) = 0.226
.
. display 1 - (1 - 0.05)^5
.22621906

.
.
. // *************
. // Part (d)
. // *************
. // The Bonferroni correction is the desired FWER 0.05/k. Create new
. // "reject" variables using this new threshold.
.
. gen reject1b=(pval1<0.05/5)

. gen reject2b=(pval2<0.05/5)

. gen reject3b=(pval3<0.05/5)

. gen reject4b=(pval4<0.05/5)

. gen reject5b=(pval5<0.05/5)

. egen rejectanyb=rowmax(reject*b)

.
. // Now how often is at least one test rejected? 4.4% in my case
```

```
. tabulate rejectanyb

 rejectanyb |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        956       95.60       95.60
          1 |         44        4.40      100.00
------------+-----------------------------------
      Total |      1,000      100.00

.
.
. // *************
. // Part (e)
. // *************
. // Correlated data example--new tstats that are dependent (correlated),
. // along with new pvals and reject variables
.
. clear

. matrix c=(1, 0.8, 1, 0.8, 0.8, 1, 0.8, 0.8, 0.8, 1, 0.8, 0.8, 0.8, 0.8, 1)

. drawnorm tstat1 tstat2 tstat3 tstat4 tstat5, n(1000) corr(c) cstorage(lower)
(obs 1,000)

. corr tstat*
(obs=1,000)

             |   tstat1   tstat2   tstat3   tstat4   tstat5
-------------+-----------------------------------------------
      tstat1 |   1.0000
      tstat2 |   0.7980   1.0000
      tstat3 |   0.7971   0.8106   1.0000
      tstat4 |   0.8144   0.8329   0.8202   1.0000
      tstat5 |   0.8083   0.8233   0.8238   0.8251   1.0000


.
. gen pval1=1-normal(tstat1)

. gen pval2=1-normal(tstat2)

. gen pval3=1-normal(tstat3)

. gen pval4=1-normal(tstat4)

. gen pval5=1-normal(tstat5)

. gen reject1=(pval1<0.05)

. gen reject2=(pval2<0.05)

. gen reject3=(pval3<0.05)

. gen reject4=(pval4<0.05)

. gen reject5=(pval5<0.05)

. egen rejectany=rowmax(reject*)

.
```

```
. // How often do we reject one or more tests in this case? 11.7% here.
. // In the independent samples case, all it took was one extreme value out of
. // 5 to reject in at least one case. That is the case here too, but now the
. // tstats are correlated. When tstat1 is not extreme, the other tstats2-5
. // are also likely to be not extreme. In a sense, there are fewer
. // "opportunities" for extreme values that would lead to rejection. To
. // take this to the limit, suppose the 5 tests were perfectly correlated.
. // In this case we would only reject in 5% of samples. The "effective"
. // number of tests in this limiting case is only 1.
.
. tabulate rejectany

  rejectany |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        883       88.30       88.30
          1 |        117       11.70      100.00
------------+-----------------------------------
      Total |      1,000      100.00

.
. // *************
. // Part (f)
. // *************
.
. // Bonferroni correction
. gen reject1b=(pval1<0.05/5)

. gen reject2b=(pval2<0.05/5)

. gen reject3b=(pval3<0.05/5)

. gen reject4b=(pval4<0.05/5)

. gen reject5b=(pval5<0.05/5)

. egen rejectanyb=rowmax(reject*b)

.
. // How often do we reject one or more tests in this case? Only 3.8% here.
. // The correction is overly conservative when the tests are dependent.
.
. tabulate rejectanyb

 rejectanyb |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        962       96.20       96.20
          1 |         38        3.80      100.00
------------+-----------------------------------
      Total |      1,000      100.00

.
. capture log close
```