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

October 31, 2024
46 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. The table below summarizes the number of hours spent in housework per week by gender, based on a 2002 survey. (**10 points**)

		Housework Hours	
Gender	Sample size	Mean	SD
Men	292	8.4	9.5
Women	391	12.8	11.6

NOTE: you can use the Stata *t*-test calculator to check your answer to the questions below, but please show your work. (I need to see that you understand the calculation).

(a) What is the estimated difference in the mean hours spent in housework per week between men and women, and what is its standard error? Assume equal variances and use the pooled variance estimator. Provide a written interpretation of the standard error. (**5 points**)

The estimated difference between the two population means (women - men) is: $\bar{x}_w - \bar{x}_m = 12.8 - 8.4 = \mathbf{4.4}$ hours. The pooled variance estimator and standard error of the difference in means is below. The standard error is a measure of how much the difference between two sample means will vary across many repeated samples.

$$s_p^2 = \frac{(n_m - 1)s_m^2 + (n_w - 1)s_w^2}{(n_m - 1) + (n_w - 1)}$$

$$s_p^2 = \frac{(291)9.5^2 + (390)11.6^2}{(291) + (390)} = 115.626$$

$$se_{\bar{x}_w - \bar{x}_m} = \sqrt{s_p^2 \left(\frac{1}{n_m} + \frac{1}{n_w} \right)}$$

$$se_{\bar{x}_w - \bar{x}_m} = \sqrt{115.626 \left(\frac{1}{292} + \frac{1}{391} \right)} = 0.832$$

(b) Find the 99% confidence interval for the population difference in mean hours spent in housework per week. (Use $n_1 + n_2 - 2$ for the degrees of freedom). (**3 points**)

A 99% confidence interval for the difference in two population means is below. The t-statistic 2.583 is the value of t for which there is a probability $\alpha/2 = 0.005$ of exceeding, using degrees of freedom $n_m + n_w - 2 = 292 + 391 - 2 = 681$:

$$(\bar{x}_w - \bar{x}_m) \pm t_{\alpha/2} se_{\bar{x}_w - \bar{x}_m}$$

$$4.4 \pm 2.583 * 0.832 = (2.25, 6.55)$$

(c) Using the information in part (b), test the null hypothesis that women and men in the population on average spend an equal number of hours per week doing housework. (Use the 1% significance level). Briefly explain your answer. (**2 pts**)

The 99% confidence interval will contain the true population mean in 99% of random samples. The null hypothesis of no difference in housework hours ($\mu_w - \mu_m = 0$) is not contained in this confidence interval. This suggests that the null hypothesis is probably not true, so it is **rejected** (at the 1% level).

Stata output using `ttesti` is shown below, corresponding to parts (a)-(c).

```
. ttesti 391 12.8 11.6 292 8.4 9.5, level(99)

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     391        12.8    .5866372        11.6    11.28149    14.31851
       y |     292         8.4    .5559454         9.5    6.958528    9.841472
---------+--------------------------------------------------------------------
combined |     683    10.91889    .4195122    10.96364    9.835263    12.00251
---------+--------------------------------------------------------------------
    diff |                 4.4    .8316831                2.251706    6.548294
------------------------------------------------------------------------------
```

```
    diff = mean(x) - mean(y)                                        t =   5.2905
Ho: diff = 0                                       degrees of freedom =      681

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

4. Men are considered overweight if their body mass index is greater than 27.8. In the 1980 *National Health and Nutrition Examination Survey*, 130 of 750 randomly surveyed men aged 20-24 were found to be overweight, while in the 1994 version of the survey, 160 of the 700 randomly surveyed men were overweight. Test the hypothesis that the proportion overweight is the same in 1994 as it was in 1980. (**5 points**)

   NOTE: you can use the Stata t-test calculator to check your answer to the question above, but please show your work. (I need to see that you understand the calculation).

$$H_0 : \pi_{1994} - \pi_{1980} = 0$$
$$H_1 : \pi_{1994} - \pi_{1980} \neq 0$$

The point estimates for the population proportions in 1980 and 1994 are $\hat{\pi}_{1980} = 130/750 = 0.1733$ and $\hat{\pi}_{1994} = 160/700 = 0.2286$. Under H_0, the population proportions were the same in 1980 and 1994. The standard error calculation for the difference in proportions assumes H_0 is true, and thus we use the *pooled* estimate of π:

$$\pi = (130 + 160)/(700 + 750) = 0.2$$

(You could equivalently take the weighted average of $\hat{\pi}_{1980}$ and $\hat{\pi}_{1994}$—you would get the same answer). Using this to calculate the standard error for the difference in proportions:

$$\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$\sqrt{0.20(1-0.20)\left(\frac{1}{700}+\frac{1}{750}\right)} = 0.021$$

Therefore the estimated difference in proportions is (0.2286 - 0.1733) / 0.021 = 2.63 standard errors above the null difference of zero. From the standard normal table (since the sample sizes are large), the p-value for this test statistic is very small (0.004*2 = 0.008), so we can safely reject H_0. The proportion of men in the population who are overweight has changed between 1980 and 1994.

Stata output using `prtesti` is shown below.

```
. prtesti 700 0.2286 750 0.1733

Two-sample test of proportions                    x: Number of obs =      700
                                                  y: Number of obs =      750
------------------------------------------------------------------------------
             |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |      .2286   .0158719                      .1974916    .2597084
           y |      .1733   .0138211                      .1462111    .2003889
-------------+----------------------------------------------------------------
        diff |      .0553   .0210461                      .0140503    .0965497
             |  under Ho:   .0210214     2.63   0.009
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =   2.6307
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.9957         Pr(|Z| > |z|) = 0.0085          Pr(Z > z) = 0.0043
```

5. In Stata, open the NELS.dta dataset from class. As you know, this extract from the larger National Education Longitudinal Study of 1988 contains data for 500 students followed from 8th through 12th grade. For this problem you will be asking the following question: Among the population of college-bound students, do students whose families owned a computer in 8th grade (*computer*) score differently in 12th grade math (*achmat12*), on average, than those whose families did not own a computer? Use the variable *edexpect* to select the subset of students who are college-bound. (**16 points**)

   (a) Are the two samples being compared here independent or dependent? Briefly explain your answer. (**1 point**)

   NELS is designed to be a representative sample of 8th graders, and there is no link between students whose families owned a computer and students whose families did not. Thus the samples are **independent**. (Note that NELS is a longitudinal study, which often implies a dependent sample. However, in this problem we are not comparing the same students at different points in time, but rather two groups of students at one point in time).

   (b) For this hypothesis test to be valid, does the distribution of math achievement in these two populations have to be normal? Briefly explain why or why not. (**1 point**)

No. We will be assuming the sampling distribution for the difference in means ($\bar{x}_2 - \bar{x}_1$) is normal (or approximately normal), an assumption that should hold if the samples sizes are large enough. As shown below, both samples are quite large (220+).

```
. table computer if edexpect>=2,row

----------------------
computer  |
owned by  |
family in |
eighth    |
grade?    |      Freq.
----------+-----------
       no |        229
      yes |        223
          |
    Total |        452
----------------------
```

(c) Write down the null and alternative hypotheses for a t-test to determine whether 12th grade math achievement for students whose families owned a computer in 8th grade differs, on average, from that of students whose families did not own a computer. (**2 points**)

$$H_0 : \mu_c - \mu_{nc} = 0$$
$$H_1 : \mu_c - \mu_{nc} \neq 0$$

(d) Using the appropriate Stata command, what is the test statistic and p-value associated with this test? (**2 points**)

As shown below, the t-statistic is **-3.2693** and the p-value for a two-tailed test is **0.0012**. (These values would not differ much if we assumed unequal variances and used the `unequal` option in the `ttest` command.)

```
. ttest achmat12 if edexpect>=2, by(computer)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      no |     229    56.70223     .451125     6.82676    55.81332    57.59113
     yes |     223    58.93758    .5152514    7.694344    57.92217    59.95299
---------+--------------------------------------------------------------------
combined |     452    57.80507     .345497    7.345368    57.12608    58.48405
---------+--------------------------------------------------------------------
    diff |           -2.235351    .6837501               -3.579091   -.8916117
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =  -3.2693
Ho: diff = 0                                     degrees of freedom =      450
```

```
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.9994
```

(e) Use the p-value to determine whether or not H_0 can be rejected in favor of the alternative. Use a significance level of $\alpha = 0.05$. (**2 points**)

> The p-value here for a two-tailed test is very small (0.0012), leading us to **reject** H_0 in favor of the alternative.

(f) Provide a 95% confidence interval for the mean difference in 12th grade math achievement between those whose families owned a computer in 8th grade and those whose families did not. (**2 points**)

> Using the Stata output above, the 95% confidence interval is **(-3.579, -0.892)**.

(g) Use the confidence interval found in part (f) to conduct the test in parts (c)-(e). Are the results consistent? (**2 points**)

> Since zero lies outside the 95% confidence interval in part (f), we can **reject** H_0 at the 0.05 level (the same conclusion).

(h) Now provide a 99% confidence interval for the mean difference in 12th grade math achievement. Does your conclusion change at the $\alpha = 0.01$ level of significance? (**2 points**)

> The easiest way to obtain a 99% confidence interval is to change the `level` option in Stata, below. The 99% confidence interval is (-4.004, -0.467). Because zero lies outside this interval as well, we can **reject** H_0 at the 0.01 level.

```
. ttest achmat12 if edexpect>=2, by(computer) level(99)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
      no |     229    56.70223     .451125     6.82676     55.5304    57.87405
     yes |     223    58.93758    .5152514    7.694344    57.59888    60.27628
---------+--------------------------------------------------------------------
combined |     452    57.80507     .345497    7.345368    56.91134    58.69879
---------+--------------------------------------------------------------------
    diff |           -2.235351    .6837501               -4.004075   -.4666275
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =  -3.2693
Ho: diff = 0                                     degrees of freedom =      450
```

```
      Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.9994
```

(i) Finally, calculate the Cohen's d as a measure of effect size. Would you consider the observed effect practically significant? Explain why or why not. (**2 points**)

Cohen's d is the difference in the sample means divided by the overall standard deviation for these students. Using the output above, $d = -2.235/7.345 =$**0.304**, a practically significant and large effect size. (Put another way, the observed difference in math achievement between students whose families owned a computer and that of students whose families did not (-2.235 points) is quite large, relative to the overall standard deviation in achievement).

Notice we used the SD for this particular sample of interest (students who are college-bound), since that is the relevant benchmark population.

The Stata command `esize` will also calculate Cohen's d. However, the standard deviation it uses is s_p (below), which is very close but not exactly the same as the overall standard deviation.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$$

```
. esize twosample achmat12 if edexpect>=2, by(computer)

Effect size based on mean comparison

                               Obs per group:
                                          no =        229
                                         yes =        223
---------------------------------------------------------
        Effect Size |   Estimate     [95% Conf. Interval]
--------------------+------------------------------------
          Cohen's d |  -.3075725    -.4928887   -.1219183
         Hedges's g |  -.3070595    -.4920667    -.121715
---------------------------------------------------------
```

6. Consider again Question #2 from Problem Set 8. In that problem, you were interested in the effect of charter school attendance on student math achievement. Suppose now you have the opportunity to randomly assign a group of students to either attend a charter or traditional public school. At the end of the year, you will administer a math test to your study students. Based on prior evidence, you still believe the standard deviation in math achievement is 84. You consider a meaningful difference in math

achievement to be +21 points. How many students (in total) will you need to randomly assign in order to correctly reject the null hypothesis of no effect in 80% of random samples? Use $\alpha = 0.05$ and assume that students will be assigned in equal numbers to the treatment and control groups. Hint: use Stata to answer this question. (**5 points**)

In Stata, use the Power and Sample Size Analysis tool for two independent sample means. Set the control and experimental means to 420 and 441 (reflecting an desired minimum effect size of 21) and the common standard deviation of 84. (Assume a known standard deviation). Let $\alpha = 0.05$, Power=0.80, and use one-sided test and an equal allocation ratio. The Stata results are below, indicating a minimum total sample size of **396**, or **198 per group**.

```
. power twomeans 420 441, sd(84) knownsds onesided

Estimated sample sizes for a two-sample means test
z test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   21.0000
           m1 =  420.0000
           m2 =  441.0000
           sd =   84.0000

Estimated sample sizes:

            N =       396
  N per group =       198
```

Note the levels of the two means (m1 and m2) are not important, only the difference of 21. In the example below I used 0 and 21 for the two group means and produced the same result.

```
. power twomeans 0 21, sd(84) knownsds onesided

Estimated sample sizes for a two-sample means test
z test assuming sd1 = sd2 = sd
```

```
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   21.0000
           m1 =    0.0000
           m2 =   21.0000
           sd =   84.0000

Estimated sample sizes:

            N =       396
  N per group =       198
```

```
.
. // *********************************************************************
. // LPO.8800 Problem Set 8
. // Last updated: October 31, 2024
. // *********************************************************************
.
. cd "$pset"
C:\Users\corcorsp\Dropbox\_TEACHING\Statistics I - PhD\Problem sets\Problem set
> 8

.
. // *************
. // Question 3
. // *************
.
. ttesti 391 12.8 11.6 292 8.4 9.5, level(99)

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. err.   Std. dev.   [99% conf. interval]
---------+--------------------------------------------------------------------
       x |     391        12.8    .5866372        11.6    11.28149    14.31851
       y |     292         8.4    .5559454         9.5    6.958528    9.841472
---------+--------------------------------------------------------------------
Combined |     683    10.91889    .4195122    10.96364    9.835263    12.00251
---------+--------------------------------------------------------------------
    diff |                 4.4    .8316831                2.251706    6.548294
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   5.2905
H0: diff = 0                                     Degrees of freedom =      681

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

.
.
. // *************
. // Question 4
. // *************
.
. prtesti 700 0.2286 750 0.1733

Two-sample test of proportions                     x: Number of obs =      700
                                                   y: Number of obs =      750
------------------------------------------------------------------------------
             |       Mean   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |      .2286   .0158719                      .1974916    .2597084
           y |      .1733   .0138211                      .1462111    .2003889
-------------+----------------------------------------------------------------
        diff |      .0553   .0210461                      .0140503    .0965497
             |  under H0:   .0210214     2.63   0.009
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =   2.6307
    H0: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.9957         Pr(|Z| > |z|) = 0.0085          Pr(Z > z) = 0.0043
```

```
.
.
. // *************
. // Question 5
. // *************
.
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/nels.dta, clear

.
. // ********
. // Part b
. // ********
. tabulate computer if edexpect>=2

   computer |
   owned by |
  family in |
     eighth |
     grade? |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |        229       50.66       50.66
        yes |        223       49.34      100.00
------------+-----------------------------------
      Total |        452      100.00

.
. // ********
. // Part d-g
. // ********
. ttest achmat12 if edexpect>=2, by(computer)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
      no |     229    56.70223     .451125     6.82676    55.81332    57.59113
     yes |     223    58.93758    .5152514    7.694344    57.92217    59.95299
---------+--------------------------------------------------------------------
Combined |     452    57.80507     .345497    7.345368    57.12608    58.48405
---------+--------------------------------------------------------------------
    diff |           -2.235351    .6837501               -3.579091   -.8916117
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =  -3.2693
H0: diff = 0                                     Degrees of freedom =      450

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.9994

.
. // ********
. // Part h-i
. // ********
. ttest achmat12 if edexpect>=2, by(computer) level(99)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [99% conf. interval]
---------+--------------------------------------------------------------------
      no |     229    56.70223     .451125     6.82676     55.5304    57.87405
     yes |     223    58.93758    .5152514    7.694344    57.59888    60.27628
---------+--------------------------------------------------------------------
Combined |     452    57.80507     .345497    7.345368    56.91134    58.69879
---------+--------------------------------------------------------------------
    diff |           -2.235351    .6837501               -4.004075   -.4666275
------------------------------------------------------------------------------
    diff = mean(no) - mean(yes)                                   t =  -3.2693
H0: diff = 0                                     Degrees of freedom =      450

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.9994
```

```
.
. esize twosample achmat12 if edexpect>=2, by(computer)

Effect size based on mean comparison

                               Obs per group:
                                          no =        229
                                         yes =        223
---------------------------------------------------------
        Effect size |   Estimate     [95% conf. interval]
--------------------+------------------------------------
          Cohen's d |  -.3075725    -.4928887   -.1219183
         Hedges's g |  -.3070595    -.4920667    -.121715
---------------------------------------------------------

.
.
. // *************
. // Question 6
. // *************
. power twomeans 420 441, sd(84) knownsds onesided

Estimated sample sizes for a two-sample means test
z test assuming sd1 = sd2 = sd
H0: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   21.0000
           m1 =  420.0000
           m2 =  441.0000
           sd =   84.0000

Estimated sample sizes:

            N =       396
  N per group =       198

.
.
.
.
. capture log close
```