Statistics for the Behavioral and Social Sciences 1
New York University Fall 2017
Prof. Sean P. Corcoran

Final Exam

Instructions: Question 1 (using Stata) is required. Choose *any three* of the remaining questions to answer. It *must* be clear which questions (of #2-5) you want graded, or the first three will be scored. Answer each question in its entirety. Your answers should be clear, complete, concise, and legible, and *show your work* where applicable. Good luck!

---

1. (**25 points - *requires Stata***) You work in the institutional research office for Major State University (MSU). For your first project, you have conducted a survey of 141 MSU undergraduate students to learn more about their life at MSU, and predictors of their academic success. Type this link into a web browser to save the survey data to your local drive, and then open it in Stata: `https://goo.gl/WIokyn`.

   (a) (**6 points**) For each of the following variables, report the sample mean ($\bar{x}$) and a 99% confidence interval for the population mean (μ): Write down the Stata command(s) you used to obtain these.

      i. ACT: college admissions test score

      ii. *hsGPA*: grade point average in high school

   (b) (**3 points**) Report the sample proportion ($\hat{\pi}$) of students who currently work at a job for 20 or more hours per week (*job20*), and a 95% confidence interval for the population proportion (π). Write down the Stata command(s) you used to obtain these.

(c) (**2 points**) Which of the following two variables, each measured before college admission, is more strongly correlated with students' college GPA ($colGPA$): ACT or $hsGPA$? Use and report the Pearson correlation coefficient for both.

(d) (**3 points**) Calculate the intercept and slope coefficient for the least squares regression line using ACT as the explanatory variable and $colGPA$ as the response. Report the results below by writing out the prediction equation and write down the Stata command(s) you used to obtain these

(e) (**3 points**) Carefully interpret—in words—the estimated slope coefficient from part (d).

(f) (**3 points**) Based on your results in part (d), what is your best prediction of the college GPA for a student entering MSU with an ACT score of 27?

(g) (**5 points**) Finally, conduct an independent samples t-test for the difference in mean college GPA for two groups: students who have a personal computer at school (PC=1) and those who do not (PC=0). Use the null hypothesis that there is no difference in these groups' mean GPA (and a two-sided alternative). Report your test statistic, p-value, and conclusion, and write down the Stata command(s) you used to get your answer.

This page ironically left blank.

2. (**25 points**) The height of male high school seniors in the population is believed to be normally distributed with a mean (μ) of 70 inches and a variance (σ^2) of 9 inches. Use this information to answer the following questions.

   (a) (**5 points**) The armed services have strict height and weight requirements for new recruits. The United States Naval Academy (USNA), for example, will only consider male applicants who are 64 to 76 inches in height. Given the information provided above, what fraction of male high school seniors would **not** meet the USNA height requirements? Provide a numeric answer, and draw a sketch if it helps).

   (b) (**5 points**) To you, male high school seniors in New York seem unusually tall. You decide to take a random sample of 16 students and calculate their mean height ($\bar{x}$) to assess whether male seniors in New York are taller on average than the general population. If the distribution of male high school seniors in New York is not different from the general population, what would you expect the distribution of $\bar{x}$ to look like? That is, what would be its mean, standard error, and distributional shape (if this can be determined)? Briefly explain your answer.

(c) (**5 points**) Given your sample of students, you calculate a mean height of 72 inches. If the distribution of male high school seniors in New York is no different from the general population, what is the probability of drawing a random sample with a mean of 72 inches or taller? Briefly justify your answer.

(d) (**5 points**) If the variance of height in the population (σ^2) were **12** inches rather than 9, would your sample mean of 72 inches be more or less likely than in part (c)? Briefly explain your answer.

(e) (**5 points**) If your sample were of **25** students rather than 16 (and the variance remained $\sigma^2 = 9$, would your sample mean of 72 inches be more or less likely than in part (c)? Briefly explain your answer.

3. (**25 points**) The body mass index (BMI) of men and women in the Framingham heart study was compared to see whether BMI differs substantially for men and women. Random samples of 200 men and 200 women were drawn from the larger study. Stata output for an independent samples *t*-test is provided below. (Be sure to note which group is being subtracted from the other to obtain the "diff" in means).

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
Group    |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  men    |     200    26.30545    .2459713    3.478559    25.82041    26.79049
women    |     200     25.1796    .2995249    4.235922    24.58895    25.77025
---------+--------------------------------------------------------------------
combined |     400    25.74253    .1955872    3.911743    25.35801    26.12704
---------+--------------------------------------------------------------------
diff     |             1.12585    .3875784                .3638932    1.887807
------------------------------------------------------------------------------
diff = mean(men) - mean(women)                                  t =   2.9048
Ho: diff = 0                                     degrees of freedom =      398

Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 0.9981         Pr(|T| > |t|) = 0.0039          Pr(T > t) = 0.0019
```

(a) (**4 points**) Write down the null and alternative hypotheses for a two-sided hypothesis test for a significant difference in the two population means.

(b) (**4 points**) Next, write down two assumptions that are required to conduct this test. Is the *t*-test for independent samples appropriate in this particular case? Briefly explain why or why not.

(c) (**4 points**) Carefully provide an interpretation—in words—of the 95% confidence interval shown in the table above on the line labeled “diff”.

(d) (**3 points**) Using the information in the Stata output, make a conclusion for the test described in part (a). Provide a p-value for this test, and use $\alpha = 0.05$. Briefly explain how you obtained your answer.

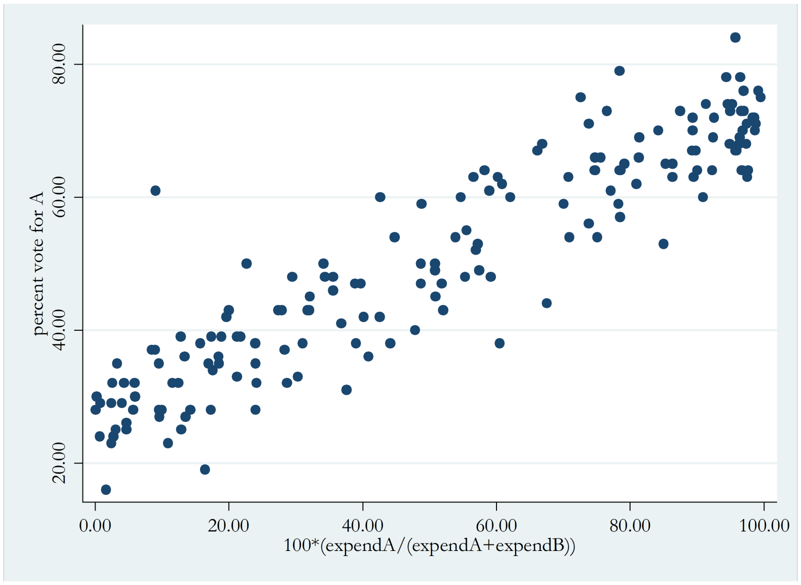
(e) (**4 points**) Will a 99% confidence interval for the difference in population means be *narrower* or *wider* than the one shown in the Stata output? Briefly explain your answer.

(f) (**3 points**) Finally, use the information in the Stata output to construct a 99% confidence interval for the difference in means.

This page left blank for comic effect.

4. (**25 points**) You have collected data on campaign expenditures and election results for 173 U.S. Congressional races in 1988 and 1990, to examine the relationship between campaign spending and election outcomes.

   (a) (**3 points**) Using Stata, you create the following scatterplot between *voteA*, the percent voting for Candidate "A" (1 of the 2 major candidates) and *shareA*, the percentage of campaign spending (between the two candidates) that was spent by Candidate "A." Briefly assess/describe the association you see. Would the Pearson correlation coefficient be appropriate to use here? Why or why not?



(b) (**6 points**) Using the same variables, you estimate a simple linear regression, and obtain the following results. In words, carefully interpret the estimated intercept and slope coefficient. Be sure to use the correct units in your description.

```
. regress voteA shareA

  Source     |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  1,   171) = 1017.66
   Model     |  41486.2307     1  41486.2307           Prob > F      =  0.0000
Residual     |  6971.01783   171  40.7661862           R-squared     =  0.8561
-------------+------------------------------           Adj R-squared =  0.8553
   Total     |  48457.2486   172  281.728189           Root MSE      =  6.3848

------------------------------------------------------------------------------
   voteA     |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  shareA     |   .4638269   .0145397    31.90   0.000     .4351266    .4925272
   _cons     |   26.81221   .8872146    30.22   0.000     25.06091    28.56352
------------------------------------------------------------------------------
```

(c) (**4 points**) The following are descriptive statistics for the two variables used in part (b). Using this information and/or that from part (b), what is the Pearson correlation coefficient between *voteA* and *shareA*? Show your work.

```
. tabstat voteA shareA, stat(mean sd n) col(stat)

variable     |      mean        sd         N
-------------+------------------------------
   voteA     |  50.50289  16.78476       173
  shareA     |  51.07654  33.48358       173
--------------------------------------------
```

(d) (**4 points**) Consider a Congressional candidate whose campaign spending represents 85 percent of all spending in his race. Using your regression results, compute a best prediction for their election vote share. Do you predict a win for this candidate, with a majority of the votes?

(e) (**4 points**) In your own words, carefully interpret the “R-squared” shown in the regression results.

(f) (**4 points**) Finally, should the slope coefficient in part (b) be interpreted as the *causal* impact of campaign spending on election outcomes in U.S. Congressional races? Why or why not? Briefly explain, and ignore the fact that the data is old.

5. (**25 points**) Answer parts (a)-(e) below. Note that parts (d)-(e) are unrelated to parts (a)-(c).

   (a) (**6 points**) A recent survey asked married respondents, "Did you live with your husband/wife before you got married?" Of 172 couples who called themselves politically liberal, 57 responded "yes" to this question. Of 283 couples who called themselves politically conservative, 45 responded "yes." Construct a 95% confidence interval for the population proportion of liberal couples who lived together before marriage. Show your work.

   (b) (**5 points**) Can you reject the null hypothesis that the population proportion of liberal couples who lived together before marriage is 0.45 (or 45%)? Explain why or why not (and use $\alpha = 0.05$).

(c) (**5 points**) Now test the null hypothesis that the population proportions of liberal and conservative couples who live together before marriage is the same. Show your work (and continue to use $\alpha = 0.05$).

(d) (**5 points**) A pharmaceutical company has developed a new drug that it believes will reduce cholesterol levels in some patients. In a preliminary study of 22 subjects, the company finds that cholesterol levels for these subjects fell an average of 18 points ($\bar{x} = -18$) with a sample standard deviation of $s = 12$. Construct a 95% confidence interval for the population mean cholesterol reduction.

(e) (**4 points**) Suppose—based on your finding in part (d)—that the pharmaceutical company decides to go ahead and market this drug because of its statistically significant effects on cholesterol levels. Explain, in your own words, what a Type I Error would be in this case. (This does not require a numeric answer, just an explanation).