LPO 8800: Statistical Methods in Education Research
Vanderbilt University
Final Exam
December 9, 2021

Name: ____

By signing below, I agree to the terms of Vanderbilt University's honor code. I attest that I have not collaborated with, or received any external assistance from other individuals on this at-home exam.

Signature: ____

Instructions: Read each question carefully and provide clear, concise responses in a separate document. Be sure to complete every part of every question. Partial credit will be given where appropriate. If you make any assumptions to answer a question, please state those explicitly. You may use lecture notes or other reference materials for this exam, but you must complete the exam on your own. Please submit your responses via email to `sean.corcoran@vanderbilt.edu` before **5:00 p.m. on Tuesday, December 14**. Good luck!

Question 1 - Requires Stata You are using the National Educational Longitudinal Study of 1988 (NELS-88) to analyze educational outcomes of middle and high school students in the U.S. Data were collected on students in 8th, 10th, and 12th grade. You have created a dataset for this purpose named *NELS_subset.dta* that can be found on Github. Use this dataset to answer the following questions. (**25 points**)

(a) Conduct a significance test for the difference in mean reading test scores in 12th grade between students who attended a Catholic high school and those who didn't. In your answer be sure to include the following elements: (**5 points**)

- Write down the null and alternative hypotheses for this test.
- Report the t-statistic and p-value for the test. Provide an interpretation of the p-value in words.
- Using $\alpha = 0.05$, report the conclusion of your test.
- Write down the Stata command you used for this part.

(b) Consider the difference in mean test scores you found in part (a). Regardless of whether this difference is statistically significant or not, would you say this difference is *practically* or *educationally* significant? Briefly explain how you came to this conclusion. (**4 points**)

(c) Fit a least squares regression with math test scores in 12th grade as the outcome and family income in thousands of dollars (*famincimp*) as the explanatory variable. Note *famincimp* is an estimate of family income based on the ordinal variable *faminc*. Write down the prediction equation you obtained, and provide a written interpretation of the least squares intercept and slope. (**5 points**)

(d) Based on your regression in part (c), how large of a difference in 12th grade math scores would you predict between a child in a household earning $25,000 and a child in a household earning $50,000? (**3 points**)

(e) Provide a written interpretation of the R-squared you obtained in part (c). (**3 points**)

(f) This part is unrelated to (a)-(e). Suppose you and your research team have the opportunity to experimentally assign 200 rising 9th grade students to either attend a private high school or a public high school. (You will provide a tuition voucher to half of these students selected at random; assume students comply with the random assignment). At the end of high school, you will administer a test and standardize the scores to have a mean of 0 and standard deviation of 1. Assume 1 is the "known" standard deviation on the test. Use Stata to calculate the statistical power of the test of $H_0 : \mu_1 - \mu_0 = 0$ versus the alternative $H_1 : \mu_1 - \mu_0 > 0.2$. ($\mu_1$ is the mean score of the private school group and μ_0 is the mean score of the public school group). Write down the Stata command you used, and carefully explain what this number represents. (**5 points**)

Question 2. Everybody loves a winner—or do they? The table below shows the average annual home attendance (in thousands) for major league baseball teams in 2002, with and without a winning record, and the standard deviation for each. For example, winning teams drew an average of 484 thousand fans in 2002, with a standard deviation of 155 thousand. (**25 points—5 points each**)

	Teams with a winning record	Teams with a losing record
n	14	16
$\bar{x}$	484	392
s	155	141

(a) Construct a 95% confidence interval for $\Delta = \mu_w - \mu_\ell$, the difference in average home attendance for winning and losing teams. Assume equal population variances, and assume you can use data from 2002 as a random sample from the population of all home games.

(b) Using a t-test, test the null hypothesis that $\Delta = 0$, that there is no difference in average home attendance between winning and losing teams. Briefly explain your answer.

(c) Does your answer in part (b) change if you use a 90% confidence level? Explain why or why not.

(d) Which of the two tests—part (b) or part (c)—has higher statistical power? Briefly explain.

(e) In the context of part (c), explain what a Type II error would be in this test.

Question 3. The Stata output below comes from a study of the relationship between average class size (the student-teacher ratio, or *str*) and reading achievement (*read_scr*) in California schools. Use these results to answer the following questions. (**25 points—5 points each**)

```
Source    |       SS           df       MS         Number of obs   =       420
----------+----------------------------------      F(1, 418)       =     27.06
Model     |  10301.7839         1  10301.7839      Prob > F        =    0.0000
Residual  |  159112.847       418  380.652743      R-squared       =    0.0608
----------+----------------------------------      Adj R-squared   =    0.0586
Total     |  169414.631       419  404.330861      Root MSE        =     19.51

------------------------------------------------------------------------------
read_scr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
str       |  -2.621026   .5038248    -5.20   0.000    -3.611372    -1.63068
_cons     |   706.4485   9.941023    71.06   0.000     686.9079    725.9892
------------------------------------------------------------------------------
```

(a) Provide a written interpretation of the standard error of the *str* coefficient. What assumption(s), if any, must hold for this standard error calculation to be correct?

(b) What is the Pearson correlation coefficient between the student-teacher ratio and reading achievement?

(c) The mean reading score in the sample is 654.97. What is the mean student teacher ratio?

(d) True or false? The intercept and slope coefficient shown above minimize the Root MSE, given the sample data. (Explain your answer).

(e) True or false? Reading achievement explains about 6% of the variation in student-teacher ratios across California schools. (Explain your answer).

Question 4. 66 boys and 83 girls—all in 11th grade—were randomly sampled from a local school district. The following table shows the count of each who were enrolled in an Advanced Placement (AP) course. Use this information to answer the following questions. (**15 points**)

	Taking an AP course?		
	No	Yes	Total
Boys	24	42	66
Girls	38	45	83

(a) Should the two samples in this problem (boys and girls) be considered independent or dependent samples? Briefly explain your answer. (**4 points**)

(b) Conduct a two-sided t-test of the null hypothesis that the population proportion of boys and girls taking AP courses is the same ($\pi_b = \pi_g$). Use $\alpha = 0.05$. In your answer, be sure to report the standard error, test statistic, and p-value. (**6 points**)

(c) For tests like the one you conducted in part (b), many researchers report the difference in sample proportions and indicate whether or not the difference is statistically significant. (They may also report a p-value). True or false: having reported this information, there would be little added value to also reporting a 95% confidence interval. Explain your answer. (**5 points**)