Problem Sets for Intro to Empirical Methods with Solutions

David A Siegel

June 15, 2021

1 Problems

1.1 Lecture 1

1. State a puzzle or research question that motivates you in no more than three sentences. This may be your proposed paper idea.

2. Specify a theory that you think explains your question or an aspect of your question. This theory must contain a dependent and at least one (but possibly more) independently variables. It must also specify a causal mechanism that answers the question: How does(do) change in the independent variable(s) cause change in the dependent variable.

3. Specify at least one research hypothesis from this theory. It should operationalize both dependent and independent variables.

4. State the null hypothesis for each research hypothesis.

5. Consider at least one threat to causal inference we discussed in class and state in no more than a paragraph how you are going to deal with it.

6. The following seven numbers represent a *sample* taken from a population of interest: 4, 6, 10, 12, 16, 18, 6, 12, 6. Compute the Mean, Median, Mode, Variance, and Standard Deviation.

1.2 Lecture 2

1. Calculate a χ^2 for this table. Does it show a significant association?

	Star Wars	Harry Potter
< 40	30	70
≥ 40	60	40

2. A survey of satisfaction in college was taken at the end of the school year. Though it was intended to get equal numbers of each class, it produced 101 juniors but only 51 seniors. The mean level of satisfaction of the juniors was 3 on a 5-point scale, with a standard deviation of 0.5, while the mean level of satisfaction of the seniors was 4, with a standard deviation of 0.3. Is the measured difference between juniors and seniors statistically significant at conventional levels?

3. The following table presents the results of an O.L.S. regression on fake data over some alternative universe set of 50 U.S. States. The dependent variable is the average rate of charitable donations as a proportion of individual income in each state. The independent variable is a self-reported measure of the average proportion of hours spent watching Sesame Street during childhood, relative to hours spent watching all TV, during the ages 3-7 and conditional on watching at least 10 hours/week of TV, also in each state. (These imaginary people have excellent memories.) A table of critical values can be found at the end of the question.

Estimate	Coefficient	Standard Error
Constant	0.01	0.055
Sesame Street	0.05	0.02

   (a) What is the null hypothesis, expressed in terms of the value of the coefficient on the independent variable?

   (b) In words, describe what the estimate on the coefficient on the independent variable means substantively.

   (c) Calculate and report the t-statistic for the independent variable.

(d) What is the critical value at the .05 level for the independent variable? Use the t-table below and write the value given there *exactly*.

(e) Does the number of hours spent watching Sesame Street have a statistically significant effect on charitable donations?

(f) What would be your prediction as to the rate of charitable donations for someone who exclusively watched Sesame Street growing up (i.e., has a value of the independent variable of 1)?

(g) How would your answers to the previous four questions have changed, if at all, if instead the coefficient estimate had been .01?

Critical Values of t

d.f.	.1	0.05
45	1.679	2.014
46	1.679	2.013
47	1.678	2.012
48	1.677	2.011
49	1.677	2.010
50	1.676	2.009

1.3 Lecture 3

1. The following table presents the results of an O.L.S. regression on fake data capturing 1000 survey respondents in some alternative universe. The dependent variable is the perceived percentage probability that Republican candidate Pat Armstrong would be competitive were they to run for office. The independent variables are the respondent's education (in years), income (in thousands of dollars), and an ordinal measure of partisanship from 0 (Strong Democrat) to 6 (Strong Republican).

Estimate	Coefficient	Standard Error
Constant	60	12.7
Education	-3.5	1.75
Income	0.05	0.02
Partisanship	5.5	1.1

(a) How many null hypotheses can be tested here? List them.

(b) In words, describe what the estimate on the coefficient on each independent variable means substantively.

(c) Calculate and report the t-statistic for each independent variable.

(d) What is the critical value at the .05 level for the independent variables?

(e) Which, if any, variables have a statistically significant effect on Armstrong's support?

(f) What would be your prediction as to Armstrong's perceived competitiveness for an individual with 10 years of education and an income of $40,000 who was an independent (3 on the partisanship scale)?

1.4 Lecture 4

1. Write a model in which the focus is on the manner by which education affects one's likelihood of running for office (assume a linear probability model), but both gender and race condition the effect of education on the likelihood of running for office (but do not condition each other's effects).

2. Return to part (e) of the first problem from Lecture 3. Explain how the use of dummy variables for partisanship might produce a different result.

2 Solutions

2.1 Lecture 1

1. State a puzzle or research question that motivates you in no more than three sentences. This may be your proposed paper idea.

   Ans: There are many possible answers here, but yours should put forth a clear, coherent, answerable question. For example, How would anti-poverty programs affect likely turnout levels?

2. Specify a theory that you think explains your question or an aspect of your question. This theory must contain a dependent and at least one (but possibly more) independently variables. It must also specify a causal mechanism that answers the question: How does(do) change in the independent variable(s) cause change in the dependent variable.

   Ans: There are many possible answers here, but yours should specify a causal mechanism that links one or more independent variables to one or more dependent variables. You need not have clear measures for the variables at this stage. For example, Turnout is driven in part by a combination of direct and indirect economic costs. Direct costs alter the immediate opportunity cost for voting, while indirect costs alter one's perceived benefits from voting. Lack of education produces both direct and indirect costs.

3. Specify at least one research hypothesis from this theory. It should operationalize both dependent and independent variables.

   Ans: There are many possible answers here, but yours should state a very clear expected relationship between variation in the independent variables and variation in the dependent variable, with all variables operationalized (i.e., tied to a measure of them). For example, Higher levels of education, as measured by years in school, should produce higher levels of turnout.

4. State the null hypothesis for each research hypothesis.

   Ans: There are many possible answers here, but yours should simply restate your research hypothesis in terms of observing no effect. For example, Higher levels of education, as measured by years in school, are unrelated to levels of turnout.

5. Consider at least one threat to causal inference we discussed in class and state in no more than a paragraph how you are going to deal with it.

   Ans: There are many possible answers here, but yours should identify a problem that could come up in the research that would make it

difficult to isolate the effect of your independent variable(s) on your dependent variable. For example, in your available data, education might be strongly correlated with geographic location, and you believe location is causally related to turnout too. This would make it difficult to disentangle the effect of each. We'll talk more about this particular issue in the context of multiple regression.

6. The following seven numbers represent a *sample* taken from a population of interest: 4, 6, 10, 12, 16, 18, 6, 12, 6. Compute the Mean, Median, Mode, Variance, and Standard Deviation.

   Ans:

   (a) Mean: 10

   (b) Median: 10

   (c) Mode: 6

   (d) Variance: 24

2.2 Lecture 2

1. Calculate a χ^2 for this table. Does it show a significant association?

	Star Wars	Harry Potter
< 40	30	70
≥ 40	60	40

   Ans: The expected table, with row and column marginals, looks like:

	Star Wars	Harry Potter	
< 40	$\frac{(90)(100)}{200} = 45$	$\frac{(110)(100)}{200} = 55$	100
≥ 40	$\frac{(90)(100)}{200} = 45$	$\frac{(110)(100)}{200} = 55$	100
	90	110	200

   The $\chi^2 = \frac{(30-45)^2}{45} + \frac{(70-55)^2}{55} + \frac{(60-45)^2}{45} + \frac{(40-55)^2}{55} = 18.18$. This exceeds the critical value of 3.84 for one row and one column at the 95% confidence level, so we reject the null hypothesis of no association.

2. A survey of satisfaction in college was taken at the end of the school year. Though it was intended to get equal numbers of each class, it produced 101 juniors but only 51 seniors. The mean level of satisfaction of the juniors was 3 on a 5-point scale, with a standard deviation of 0.5, while the mean level of satisfaction of the seniors was 4, with a standard deviation of 0.3. Is the measured difference between juniors and seniors statistically significant at conventional levels?

   Ans: We apply a difference of means test. $\bar{X}_1 - \bar{X}_2 = -1$, which is the numerator. The denominator is the pooled estimate of the standard deviation, which is: $\sqrt{\frac{(.5)^2}{100} + \frac{(.3)^2}{50}} = 0.066$. This implies the t-statistic is 15.25, which exceeds the critical value of 1.96 for this sample size at the 95% or the 99% confidence levels. Thus the difference between groups is statistically significant.

3. The following table presents the results of an O.L.S. regression on fake data over some alternative universe set of 50 U.S. States. The dependent variable is the average rate of charitable donations as a proportion of individual income in each state. The independent variable is a self-reported measure of the average proportion of hours spent watching Sesame Street during childhood, relative to hours spent watching all TV, during the ages 3-7 and conditional on watching at least 10 hours/week of TV, also in each state. (These imaginary people have excellent memories.) A table of critical values can be found at the end of the question.

Estimate	Coefficient	Standard Error
Constant	0.01	0.055
Sesame Street	0.05	0.02

   (a) What is the null hypothesis, expressed in terms of the value of the coefficient on the independent variable?

   Ans: $\beta = 0$.

   (b) In words, describe what the estimate on the coefficient on the independent variable means substantively.

Ans: The estimate on the coefficient on the independent variable represents the marginal effect of the independent variable on the dependent variable. In other words, it represents the amount that the dependent variable would change if the independent variable were to increase by one unit, assuming a linear relationship and normally distributed errors.

(c) Calculate and report the t-statistic for the independent variable.

Ans: 0.05/0.02=2.5

(d) What is the critical value at the .05 level for the independent variable? Use the t-table below and write the value given there *exactly*.

Ans: 2.011

(e) Does the number of hours spent watching Sesame Street have a statistically significant effect on charitable donations?

Ans: yes

(f) What would be your prediction as to the rate of charitable donations for someone who exclusively watched Sesame Street growing up (i.e., has a value of the independent variable of 1)?

Ans: Rate of Charitable Donations is $0.01 + 0.05 * 1 = 0.06$.

(g) How would your answers to the previous four questions have changed, if at all, if instead the coefficient estimate had been .01?

Ans: The t-statistic for the independent variable would instead be 0.5 which is below the critical value, suggesting that watching Sesame Street is not a statistically significant predictor of charitable donations. The predicted rate of charitable donations would now be: $0.01 + 0.01 * 1 = 0.02$.

Critical Values of t

d.f.	.1	0.05
45	1.679	2.014
46	1.679	2.013
47	1.678	2.012
48	1.677	2.011
49	1.677	2.010
50	1.676	2.009

2.3 Lecture 3

1. The following table presents the results of an O.L.S. regression on fake data capturing 1000 survey respondents in some alternative universe. The dependent variable is the perceived percentage probability that Republican candidate Pat Armstrong would be competitive were they to run for office. The independent variables are the respondent's education (in years), income (in thousands of dollars), and an ordinal measure of partisanship from 0 (Strong Democrat) to 6 (Strong Republican).

Estimate	Coefficient	Standard Error
Constant	60	12.7
Education	-3.5	1.75
Income	0.05	0.02
Partisanship	5.5	1.1

   (a) How many null hypotheses can be tested here? List them.

   Ans: There are three independent variables (besides the constant) and three null hypotheses. These are: 1) There is no effect of education on Armstrong's perceived competitiveness; 2) There is no effect of income on Armstrong's perceived competitiveness; 3) There is no effect of partisanship on Armstrong's perceived competitiveness.

   (b) In words, describe what the estimate on the coefficient on each independent variable means substantively.

Ans: 1) Each year of education decreases Armstrong's perceived competitiveness by 3.5 on a 100-point scale. 2) Each thousand dollars of income increases Armstrong's perceived competitiveness by 0.05 on a 100-point scale. 3) Moving one step toward Strong Republican on the partisanship scale increases Armstrong's perceived competitiveness by 5.5 on a 100-point scale.

(c) Calculate and report the t-statistic for each independent variable.

Ans: 1) -2.0 for Education, 2) 2.5 for Income, 3) 5 for partisanship.

(d) What is the critical value at the .05 level for the independent variables?

Ans: 1.96, as there are 996 degrees of freedom.

(e) Which, if any, variables have a statistically significant effect on Armstrong's support?

Ans: All three variables have statistically significant effects.

(f) What would be your prediction as to Armstrong's perceived competitiveness for an individual with 10 years of education and an income of $40,000 who was an independent (3 on the partisanship scale)?

Ans: Perceived Competitiveness is $60-3.5*10+0.05*40+5.5*3 = 60-35+2+16.5 = 43.5$.

2.4 Lecture 4

1. Write a model in which the focus is on the manner by which education affects one's likelihood of running for office (assume a linear probability model), but both gender and ethnicity condition the effect of education on the likelihood of running for office (but do not condition each other's effects).

Ans: If X_1 is education, gender is X_2, and ethnicity is X_3, then a model that satisfies the requirements in the problem is: $Y = \beta_1 X_1 +$

$\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$. Remember that the constituent terms for gender and ethnicity must both be included.

2. Return to part (e) of the first problem from Lecture 3. Explain how the use of dummy variables for partisanship might produce a different result.

   Ans: Using the ordinal measure of partisanship assumes that there is a constant effect of partisanship on Armstrong's perceived competitiveness for all steps on the scale. But we may think, for example, that Democrats might have very different standards for the competitiveness of a Republican candidate than Republicans do, so that moving from a Strong Democrat to an Intermediate Democrat (0 to 1) has a very different effect than moving from a Weak Republican to an Intermediate Republican (4 to 5). The effects may even have opposite signs! Turning the ordinal variable into multiple dummies may help to parse out those different effects, some of which may prove to be significant.