POLI SCI 231b: Problem Set 2

Spring 2016

University of California, Berkeley

Prof. Thad Dunning/GSI Guadalupe Tuñón

Due Thursday, February 4 by 6 PM

Each group should turn in its problem set solution by email to Guadalupe Tuñón (guadalupe.tunon+231b@berkeley.edu). Please work out the problems on your own, before you meet with your group to agree on solutions. As usual, the assigned readings are very helpful for answering these questions, in addition to lectures and sections. Explain your answers completely and turn in your code (use R Markdown when appropriate).

1. Let X be a variable with average $\overline{X}$. Using the definition of variance, show using math notation that for any constant m, $\text{Var}(\frac{X}{m}) = \frac{\text{Var(X)}}{m^2}$.

2. Draws are being made at random with replacement from a box. The number of draws is getting larger and larger. Say whether each of the following statements is true or false, and explain. (Remember that "converges" means "gets closer and closer.")

   (a) The probability histogram for the sum of the draws (when put in standard units) converges to the standard normal curve.

   (b) The histogram for the numbers in the box (when put in standard units) converges to the standard normal curve.

   (c) The histogram for the numbers drawn (when put in standard units) converges to the standard normal curve.

   (d) The probability histogram for the product of the draws (when put in standard units) converges to the standard normal curve.

   (e) The histogram for the numbers drawn converges to the histogram for the numbers in the box.

(f) The variance of the numbers drawn converges to zero.

(g) The variance of the histogram for the numbers drawn converges to zero.

(h) The variance of the average of the draws converges to zero.

3. Match the phrase in the first column of Table 1 to its synonym in the second column. Here, m in the number of units assigned to the treatment group in an experiment with $N > m$ units in the study group.

(a) The sampling distribution of the treatment group mean	(1) The square root of the variance of the potential outcomes under treatment, divided by the square root of m
(b) The standard error of the mean in the assigned-to-treatment group	(2) The average causal effect of treatment assignment
(c) The estimated standard error of the treatment group mean	(3) the squared s.d. of the sampling distribution of the treatment group mean
(d) The average observed outcome in the assigned-to-treatment group, minus the average observed outcome in the assigned-to-control group	(4) An estimator of the effect of treatment assignment
(e) The sampling variance of the treatment group mean	(5) The standard deviation of observed outcomes in the treatment group, divided by the square root of m
(f) The outcome if every unit were assigned to treatment, minus the outcome if every unit were assigned to control	(6) A histogram showing sample averages in the treatment group, across hypothetical replications of the experiment
(g) The sampling variance of a treatment group "ticket"	(7) The variance of potential outcomes under treatment

Table 1: Match the phrase in the first column to its synonym in the second column.

4. A large college class has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues the following:

   "If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That's what happened to me—chance variation."

   (a) Formulate the chance procedure the TA describes in terms of a box model. Describe the contents of the box.

   (b) Fill in the blanks. The null hypothesis says that the average of the box is ___. The alternative hypothesis says that the average of the box is ___.

(c) Is the TA's defense a good one? Answer yes or no, and explain briefly.

5. A geography test was given to a simple random sample of 250 high school students in a certain large school district. One question involved an outline map of Europe, with the countries identified only by number. The students were asked to pick out Great Britain and France. As it turned out, 65.8% could find France, compared to 70.2% for Great Britain. Is the difference statistically significant? Or can this be determined from the information given? Explain.

6. Consider an experiment with one treatment group and one control group. There are 6 subjects; 3 subjects are assigned to treatment and 3 are assigned to control. Hypothetical potential outcomes under treatment and control are depicted in Table 1.

Subject i	$Y_i(1)$	$Y_i(0)$	τ_i
1	5	4	1
2	7	7	0
3	7	3	4
4	6	7	-1
5	6	3	3
6	5	3	2
Average	**6**	**4.5**	**1.5**

Table 1. Hypothetical potential outcomes for the 6 units in this experiment. Here, $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, while τ_i is the unit causal effect.

(a) Does the treatment have a positive effect for every unit? On average, does it have a positive effect for all units? Are these quantities parameters or estimators? What is another term for the average τ_i, that is, the number 1.5 found in the bottom-right cell?

(b) How many possible random assignments are there in which 3 units are assigned to treatment and 3 to control?

(c) Create a table showing, for each possible random assignment, the average response in the treatment group, the average response in the control group, and the difference of means across the treatment and control groups.

(d) What is the average of the difference of means across the treatment and control groups, across all these possible assignments? What do you conclude about the difference of means, as an estimator for the average causal effect? Give a statistical rationale for your conclusion.

(e) Calculate the variance of the average responses in the treatment group, across all possible assignments. (That is, compute the mean squared deviation of each treatment group mean from the overall average of the treatment group means, across all assignments).

(f) Compare your answer in part (e) to the analytic formula for the sampling variance of the treatment group mean, as presented in class. (Remember that you need a finite-sample correction factor: $\frac{N-m}{N-1}$, where N is the size of the population, and m is the size of the sample).

N.B. To keep track of your calculations and avoid error, it may be useful to work on items (c-f) in R. Please turn in all of your work, including any code you write.

7. Using the data from Dunning and Harrison (2010),

   (a) Carefully define the sharp (strict) null hypothesis involving the "coethnic cousin" treatment condition and the "non-coethnic, non-cousin" control condition. (Be clear about the population for which the null hypothesis is defined).

   (b) Use randomization inference to test the sharp null hypothesis.

   (c) Carefully define the weak null hypothesis involving the "coethnic cousin" treatment condition and the "non-coethnic, non-cousin" control condition. (Be clear about the population for which the null hypothesis is defined).

   (d) Building on the code you have used in section, write a *t*-test function in R that takes the treatment and the outcome data and returns the difference of means, its standard error and the p-value. Use this function to test the weak null hypothesis.

   (e) Compare your *p*-values from the (b) and (d). In which case (or neither, or both) would you reject the null hypothesis? If your *p*-values are very different, why are they different? If they are very similar, why are they similar?

   (N.B. For guidance in answering this question, see Gerber and Green 2012: Chapter 3 and Dunning 2012: Sections 6.1 and 6.3, as well as lecture and section notes).

8. Is the standard deviation of the sample an unbiased estimator of the standard deviation in the population? Does your answer depend on whether you are sampling with or without replacement? Use R to write simulations to answer these questions. Note that R uses (n-1) for the denominator of sd( ).

9. For this question, you will compare the true standard error of $\widehat{\text{ATE}}$ to the "conservative" standard error.

   (a) First, consider the following R code:

```
set.seed(1234567)
N <- 60
m <- 30
y0 <- rnorm(N, 2, 3)
y1 <- y0 + rnorm(N, 1, 2)
# y0 and y1 are potential outcomes; the ATE is about 1
data <- data.frame(y0, y1)
```

(b) Suppose that `m` units are assigned at random to treatment, with `N-m` assigned to control. Is the difference between the average `y1` in the treatment group and the average `y0` in the control group a random variable?

(c) What is the true standard error of this difference of means? What is the "conservative" standard error? (Note: in both cases we are asking about standard errors defined by parameters—not sample quantities).

(d) Now, complete the code above to build a simulation in which there are 10,000 replicates. In each replicate, `m` of the units are assigned at random to treatment. Save the conservative $\widehat{\text{SE}}(\widehat{\text{ATE}})$ for each replicate. Plot the distribution of the conservative $\widehat{\text{SE}}(\widehat{\text{ATE}})$s across the 10,000 replicates. Add a vertical line to your plot at the value of the true standard error.

(e) Referring to your plot from part (d), explain why the estimated standard error is called "conservative"?

(f) Under what conditions would the true standard error equal the conservative standard error in part (a)? Modify the code excerpted above so that this equality holds.