Gov 2002: Problem Set 7

Spring 2021

Problem Set Instructions

This problem set is due on April 7, 11:59 pm Eastern time. Please upload a PDF of your solutions to gradescope. We will accept hand-written solutions but we strongly advise you to typeset your answers in Rmarkdown. Please list the names of other students you worked with on this problem set.

Question 1

Let $X \sim \text{Unif}(0, 1)$ and $Y|X = x \sim \text{Unif}(x, 1)$ (that is, Y given $X = x$ is uniformly distributed between x and 1).

   a) Find $E[Y]$

   b) Find the PDF of the joint distribution of X and Y, $f_{X,Y}(x, y)$. Remember to specify the domain.

   c) Find $E[Y - X]$ and $E[Y - X|X = x]$. Are $Y - X$ and X independent?

Question 2: Power Analysis

This question will give you some realistic practice conducting *power analysis*, a very useful procedure when planning data collection for research.

Recall the subprime data from previous problem sets. You have been given some money from IQSS to conduct a field experiment to replicate some of the findings in this observational dataset. Your field experiment will take the form of an audit study, where you test whether or not lenders are discriminating on the basis of gender. Such audit studies are a common way of studying discrimination in housing, wage, and many other markets. You will send out loan applications to a variety of lenders, and will randomly assign the gender of the applicant. This random assignment is important because it would allow you to make a causal claim about the relationship between gender and the amount of money applicants will be loaned.

IQSS's money pot is deep, but it has its limits, so you need to know what size experiment you can afford. You decide to conduct a power calculation, so that you can know the likelihood of

"finding effects" (i.e. rejecting the null hypothesis) in your proposed study. For this problem, you will use the same `subprime.csv` dataset we have used in the last few weeks.

(a)

You already know your quantity of interest, $\overline{X}_w - \overline{X}_m$, and your outcome measure, `loan.amount`. What are the factors do we need to consider that will affect the power of your experiment? (Hint: There are four of them.) In what way does each of them affect power, that is, what happens to power as each of these factors increases or decreases in size?

(b)

IQSS wants you to come up with an accurate estimate for the effect size you expect to see. Calculate a reasonable value for the effect size using the population (i.e. the subprime data as a whole).

In what ways might this value be an accurate estimate for the effect you expect to find in your experiment? In what ways might it be inaccurate? If this estimate shows a stronger effect than the real effect size, what will that imply about our power calculation?

(c)

Suppose that you have received a grant of $3,000 for your experiment, and you estimate that your experiment will cost approximately $5 per participant. You plan on assigning half of your applicants to be male and the other half to be female.

Assuming an effect size equal to what you calculated in (b), and using a standard $\alpha = 0.05$, does this budget allow you to run an experiment where your power will be greater than or equal to 0.8?

Hint: For this part, assume that $\sigma_w^2 = \sigma_m^2$, and assume that the variances in the experiment will be the same as the population variances in the subprime dataset.

Question 3[1]

In the case Hazelwood School District v. United States (1977), which went to the Supreme Court, the United States argued that the Hazelwood School District in Missouri was practicing racial discrimination in their hiring of teachers. Data were presented showing that only 15 out of 405 (3.7%) of the teachers hired by the school district were black.

---

[1]Thanks to the 2019 Stat 111 teaching team for this problem

(a)

In the original case, the District Court ruled in favor of the school district, saying "The number of black teachers employed by the Hazelwood district is undeniably meager. Nonetheless, it has kept pace with the small but steadily increasing black enrollment in the district. For the 1970-71 school year the six black teachers hired by the Hazelwood district comprised less than one percent of its total faculty. However, the number of black students enrolled during that period was likewise only one percent of the total district attendance."

The United States appealed this decision, arguing that the relevant population for comparison was the teachers in St. Louis County and St. Louis City, not the students in the Hazelwood district. (St. Louis City borders but is not included in St. Louis County, since the city seceded from the county in 1876.) About 15.4% of the teachers in this population were black, which intuitively seems like a massive disparity compared with the 3.7% statistic in the Hazelwood district. The Appeals Court ruled in favor of the United States. The school district then appealed the Appeals Court decision to the Supreme Court, arguing that St. Louis City should be excluded from the population for comparison, due to the city having very different hiring guidelines than were present in the county. Discuss the principles and considerations you would use in deciding on the population to compare the Hazelwood district statistic with. (You do not need to resolve the question of whether teachers in St. Louis City should be included in the comparison population.)

(b)

The Supreme Court ruled that the comparison should be with St. Louis County, excluding St. Louis City. In St. Louis County, 5.7% of teachers were black. Suppose that, before observing the data for Hazelwood district, we know that $n = 405$ teachers will be hired and that $X \sim Bin(n, p)$ of these teachers are black. What would you use as your null and alternative hypotheses if you would like to show that Hazelwood district discriminates against black teachers in their hiring process? Conduct this test, and find the p-value. (We are just looking at disparities in hiring; of course in reality there are many complications such as who applies for which jobs, how people are recruited, what salaries are offered, etc.). If helpful, you can use R notation such as `qbinom()` and `pbinom` in place of formal mathematical notation.

Please be sure to state the null and alternative hypothesis and the rejection region for a level α test.

(c)

Repeat (b) except with a Normal approximation to the Binomial, i.e., assuming $X \sim \mathcal{N}(np, np(1-p))$. Discuss how good this approximation is by comparing the p-value with the p-value you found in part (b).

Question 4: Propensity scores

In observational studies, we often want to estimate the effect of a binary treatment on some outcome, but we are worried that our treatment and control group are different on potential confounders. For example, suppose we wanted to measure the effect of watching a presidential debate on support for the Democratic candidate. Obviously, people that watch debates are different from those that do not on many dimensions. Thus, it's difficult to know if any difference we find in the outcome are due to the treatment (watching a debate) or to these background characteristics, also called confounders.

One popular way to adjust for potential confounders is to use the **propensity score**, or the probability of treatment given covariates. Let $T = 1$ indicate the treatment group and $T = 0$ be the control group. Suppose we had a vector of background covariates about our respondents such as age, gender, party, political knowledge and so on, and let $\mathbf{X}$ represent the random vector of these covariates. The propensity score for a given value of covariates is:

$$S(\mathbf{x}) = \Pr(T = 1 \mid \mathbf{X} = \mathbf{x}),$$

and the propensity score as a random variable is

$$S = S(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X}).$$

S is a respodent's probability of watching debates conditional on the covariates.

In this problem, we'll show why this quantity is useful in practice. In short, it is because respondents with the same propensity scores will have the same distribution of the covariates, $\mathbf{X}$, will be the same across the treated and control groups. Thus, conditional on the propensity score, any remaining effect on the outcome won't be due to these covariates. You are going to show this is true!

(a)

Using the law of iterated expectations, show that $E[T \mid S] = S$. That is, if we know the propensity score, the best guess (in terms of mean squared error) about the value of T is just the propensity score itself.

Hint: remember by the fundamental bridge, we can write $S = E[T \mid \mathbf{X}]$.

(b)

Use the result above to show that, conditional on the propensity score, T is independent of the covariates $\mathbf{X}$, or $E[T \mid S, \mathbf{X}] = E[T \mid S]$ (which is equivalent to the more formal statement of independence with $\Pr()$ because of the fundamental bridge).

Hint: think about the relationship between $\mathbf{X}$ and S.

(c)

Now, use the facts that you have established and Bayes' rule (or the definition of conditional probability) to show that the covariates are balanced across treatment groups conditional on the propensity score. That is, show that the probability of a particular value of the covariates is independent of treatment conditional on the propensity score, or

$$\Pr(\mathbf{X} = \mathbf{x} \mid S, T = t) = \Pr(\mathbf{X} = \mathbf{x} \mid S),$$

for any t (for simplicity, you can just show this for $t = 1$).