ECON 21020: Econometrics
The University of Chicago, Spring 2022
Instructor: Thomas Wiemann
Problem Set # 2: Review of Statistics
Due: 11:59am on April 20, 2022

Problem 1 10 Points

The following are "True or false?"-questions. If the statement is true, provide a brief proof ($\approx$ 3 lines). If the statement is false, provide a counter example. There are no points awarded for answers without a proof or counter example.

a)

True or false? Let U and X be random variables. If $E[UX] = 0$, then $E[U|X]$.

b)

True or false? Let X be a random variable. Then $E\left[\frac{1}{X}\right] = \frac{1}{E[X]}$.

Problem 2 5 Points

Consider the following statement: "*80% of drivers think they are better than average.*"

Can the drivers be correct? Think carefully and explain briefly.

Problem 3 10 Points

Consider a random sample $X_1, \ldots, X_n \overset{iid}{\sim} X$, and use the sample average $E_n[X]$ as an estimator for $E[X]$.

a)

Suppose that $X \sim N(\mu, \sigma^2)$. What is the the distribution of $E_n[X]$ for $n = 10$?

b)

Suppose now instead that $X \sim \text{Bernoulli}(p)$. Would the sampling distribution derived in part a) still apply for $E_n[X]$ when $n = 10$?

c)

Derive the asymptotic distribution of $E_n[X]$. Does your conclusion depend on whether $X \sim N(\mu, \sigma^2)$ or $X \sim \text{Bernoulli}(p)$?

Problem 4 60 points

Let Y and X be random variables such that $X \sim \text{Bernoulli}(p)$ for $p \in (0, 1)$.

This exercise derives and studies the sample analogue estimator for $E[Y|X = 1]$.[1]

a)

Show that

$$E[Y|X = 1] = \frac{E[XY]}{E[X]}. \tag{1}$$

b)

Use the sample analogue principle to develop an estimator $\hat{\mu}_{Y|1}$ for $E[Y|X = 1]$.

c)

Show that

$$XE[Y|X] = XE[Y|X = 1]. \tag{2}$$

(Hint: For $X : \operatorname{supp} X = \{0, 1\}$ *it holds that* $E[Y|X] = E[Y|X = 1]X + E[Y|X = 0](1 - X)$.*)*

d)

Show that your estimator is unbiased conditional on $\sum_i X_i > 0$.[2] That is, show that

$$E\left[\hat{\mu}_{Y|1} \middle| \sum_{i=1} X_i > 0\right] = E[Y|X = 1]. \tag{3}$$

[1] We first encountered a conditional expectation estimator in Lecture 1: Now we're equipped to study its statistical properties!

[2] Note that without conditioning on $\sum_i X_i > 0$, we may end up attempting to compute a group average of a group from which we don't yet have any observations – that's very difficult!

e)

Show that $\hat{\mu}_{Y|1}$ is consistent for $E[Y|X = 1]$. That is, show that

$$\hat{\mu}_{Y|1} \xrightarrow{p} E[Y|X = 1]. \tag{4}$$

(Hint: Your answer should include four steps similar to those from Examples 13 and 14 in Lecture 3A.)

f)

Show that

$$\sqrt{n}\left(\hat{\mu}_{Y|1} - E[Y|X = 1]\right) = \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1} U_i X_i}{\frac{1}{n}\sum_{i=1} X_i}\right), \tag{5}$$

where $U_i \equiv Y_i - E[Y_i|X_i]$.

g)

Use part e) to show that

$$\sqrt{n}\left(\hat{\mu}_{Y|1} - E[Y|X = 1]\right) \xrightarrow{d} N\left(0, \frac{Var\,(Y|X = 1)}{P(X = 1)}\right). \tag{6}$$

(Hint: Use Slutsky's Theorem.)

h)

Develop a sample analogue estimator $\hat{\sigma}^2_{Y|1}$ for $\sigma^2_{Y|1} \equiv Var\,(Y|X = 1)$ and a sample analogue estimator $\hat{p}_X$ for $P(X = 1)$.

(Hint: Recall that $Var\,(Y|X = 1) = E[Y^2|X = 1] - (E[Y|X = 1])^2$ and use part a).)

i)

Show that

$$\hat{\sigma}^2_{Y|1} \xrightarrow{p} \sigma^2_{Y|1}, \tag{7}$$

and that

$$\hat{p}_X \xrightarrow{p} P(X = 1). \tag{8}$$

j)

Show that

$$\sqrt{\frac{\hat{\sigma}^2_{Y|1}}{\hat{p}_X}} \xrightarrow{p} \sqrt{\frac{\sigma^2_{Y|1}}{P(X = 1)}}. \tag{9}$$

k)

Show that

$$\frac{\sqrt{n}\left(\hat{\mu}_{Y|1} - \mu_{Y|1}\right)}{\sqrt{\frac{\hat{\sigma}^2_{Y|1}}{\hat{p}_X}}} \xrightarrow{d} N(0, 1). \tag{10}$$

l)

Part k) shows that $se(\hat{\mu}_{Y|1}) \equiv \frac{1}{\sqrt{n}}\sqrt{\frac{\hat{\sigma}^2_{Y|1}}{\hat{p}_X}}$. Use this to construct a symmetric two-sided confidence interval for $E[Y|X = 1]$ with significance level $\alpha = 0.05$.

m)

Suppose now that $\hat{\mu}_{Y|1} = 10$ and $se(\hat{\mu}_{Y|1}) = 3$. Consider testing $H_0 : E[Y|X = 1] = 4$ against $H_1 : E[Y|X = 1] \neq 4$. Do you reject H_0 at a 5% significance level?

Problem 5 15 Points

This exercise uses the data of Angrist and Krueger (1991) to put our analysis of Problem 4 to practice.[3]

A cleaned version of the data is posted to Canvas (see the file `ak91.csv`). It contains 329,509 observations of American men born between 1930 and 1939. The variables we focus on in this problem set are:

[3] Angrist and Krueger (1991) is one of the most highly cited studies on the returns to education with more than 3,400 citations on Google Scholar.

- `YRS_EDUC` $\equiv$ years of education;
- `WKLY_WAGE` $\equiv$ the weekly wage.

Once downloaded, you can load the data into R using the following code:

```
# Load the ak91.csv data
df <- read.csv("data/ak91.csv")

# Store years of education and the weekly wage in separate variables
yrs_educ <- df$YRS_EDUC
wkly_wage <- df$WKLY_WAGE
```

We focus on the individuals in the sample who have completed 16 years of education, which we view as synonymous with having obtained a college degree. We may find the observations associated with 16 of education via the following code:

```
# Find college graduates
has_college_degree <- yrs_educ == 16
```

Note: This exercise must be completed in base R. That is, don't load any dependencies.

If you upload your solutions to a GitHub repository and share the link in your homework solutions, you earn an extra credit of 5 percentage points on this problem set.

a)

Let Y and W be two random variables, denoting the weekly wage and the years of education, respectively. Consider a sample $(Y_1, W_1), \ldots, (Y_n, W_n) \overset{iid}{\sim} (Y, W)$, and suppose that the data of Angrist and Krueger (1991) is a realization of this sample for $n = 329,509$.

Define $X \equiv \mathbb{1}\{W = 16\}$. Hence, X_i is 1 if the ith individual has obtained a college degree, and 0 otherwise.

Compute your sample analogue estimator $\hat{p}_X$ for $P(X = 1)$ from Problem 4 part h). Report the estimate in your solutions.

b)

Compute your sample analogue estimator $\hat{\mu}_{Y|1}$ for $E[Y|X = 1]$ from Problem 4 part b). Store the value in a variable called `mu_college` and report it in your solutions.

c)

Compute the associated standard errors $se(\hat{\mu}_{Y|1})$ from Problem 4 part k). Store the value in a variable called `se_college` and report it in your solutions.

d)

Compute a symmetric two-sided confidence interval at a 5% significance level using your solution to Problem 4 part l). Report the confidence interval in your solutions.

e)

Consider testing $H_0 : E[Y|X = 1] = 600$ against $H_1 : E[Y|X = 1] \neq 600$. Do you reject H_0 at a 5% significance level? Give an economic interpretation of your result.

f)

Consider testing $H_0 : E[Y|X = 1] = 595$ against $H_1 : E[Y|X = 1] \neq 600$. Do you reject H_0 at a 5% significance level? Give an economic interpretation of your result.

Problem 6 15 Points (Extra credit)

This is an optional extra credit exercise.

This exercise must be completed in base R. That is, don't load any dependencies.

a)

Write a function `my_confint` that takes three scalars: 1) an estimate `mu_hat` $\in \mathbb{R}$, 2) the associated standard error `se` ≥ 0, and 3) a significance level `alpha` $\in (0, 1)$, and that returns a bivariate vector `confint` $\in \mathbb{R}^2$ that gives the bounds of a symmetric two-sided confidence interval with significance level `alpha`.

Confirm your function matches the confidence interval values reported in the code below.

```
# Define a custom function that returns a two-sided confidence interval
my_confint <- function(mu_hat, se, alpha) {
  # Compute and return the confidence interval
  confint <- # [INSERT YOUR CODE HERE]
  return(confint)
}#MY_CONFINT

# Test the function
```

```
my_confint(mu_college, se_college, 0.01) # [1] 588.6420 600.3312
```

b)

Write a function `my_testrejects` that takes bivariate vector `confint`$\in \mathbb{R}^2$ and a scalar `mu_0`$\in \mathbb{R}$, and returns `TRUE` if `mu_0` $\notin$ `confint` and `FALSE` otherwise.

Confirm your function matches the results reported in the code below.

```
# Define a custom function that returns TRUE if mu_0 is not in confint
my_testrejects <- function(confint, mu_0) {
  # Check whether mu_0 is in confint
  is_in_confint <- #  [INSERT YOUR CODE HERE]
  # If mu_0 is in confit, don't reject. Else, reject.
  is_rejected <- #  [INSERT YOUR CODE HERE]
  # Return boolean
  return(is_rejected)
}#MY_TESTREJECTS

# Check whether the test rejects on 1\% significance level
confint_01 <- my_confint(mu_college, se_college, 0.01)
my_testrejects(confint_01, 600) # [1] FALSE

# Check whether the test rejects on 10\% significance level
confint_10 <- my_confint(mu_college, se_college, 0.1)
my_testrejects(confint_10, 600) # [1] TRUE
```

c)

Write a function `my_twosidedtest` that takes the same arguments as your function `my_confint` (i.e., `mu_hat`, `se`, and `alpha`) as well as a scalar `mu_0`. The function should print a message informing the user of whether the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ rejects at a `alpha`% significance level or not.

Your function `my_twosidedtest` must call both `my_confint` and `my_testrejects`. (Don't copy paste code you wrote for part a) or b) for your solution to part c)!).

```
# Define a custom function for a two-sided test
my_twosidedtest <- function(mu_hat, se, alpha, mu_0) {
  # Compute the confidence interval w/ significance level alpha
  # [INSERT YOUR CODE HERE]
  # Check whether mu_0 is in the confidence interval
  is_rejected <- # [INSERT YOUR CODE HERE]
```

```
  # Construct test message
  if (is_rejected) {
    message <- # [INSERT YOUR CODE HERE]
  } else {
    message <- # [INSERT YOUR CODE HERE]
  }#IF
  # Print the message
  print(message)
}#MY_TWOSIDEDTEST

# Check whether the test rejects on 1\% significance level
my_twosidedtest(mu_college, se_college, 0.01, 600) # Should not reject

# Check whether the test rejects on 10\% significance level
my_twosidedtest(mu_college, se_college, 0.10, 600) # Should reject
```

References

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.