Applied Econometrics, 2021
Assignment 1
Due: September 12th, 2021

Instructions. Do your best to make your arguments rigorous. You may discuss this problem set with your classmates and consult any books or notes, but write out the answers on your own using your own words and show your derivation so that your understanding is transparent from the answers.

Review of Probability

1. *(Random Variable)* Let $S = \mathbf{R}$, and define $X : S \to \mathbf{R}$ as follows: for each $s \in S$,

$$\begin{aligned} X(s) &= 1 \text{ if } 2.2 \leq s \leq 3.1 \text{ or } 3.2 \leq s \leq 3.9, \text{ and} \\ X(s) &= 0 \text{ otherwise.} \end{aligned}$$

Suppose that the probability $\mathbf{P}$ on a field $\mathcal{F}$ of S is given as follows:

$$\begin{aligned} \mathbf{P}(\{0.2\}) &= 1/4, \\ \mathbf{P}(\{1.5\}) &= 1/4, \\ \mathbf{P}(\{2.4\}) &= 1/4, \quad \text{and} \\ \mathbf{P}(\{3.3\}) &= 1/4. \end{aligned}$$

(a) Draw the CDF of X.

(b) Show that X is a discrete random variable.

(c) Compute $\text{Var}(X)$.

2. *(Variance and Covariance)* Suppose that $\{(Y_i, X_i)\}_{i=1}^n$ is a set of i.i.d. random vectors. (Note that this implies that the joint distribution of (Y_1, X_1) and the joint distribution of (Y_2, X_2) are identical, etc., and also (Y_1, X_1) is independent of (Y_2, X_2), etc. This does NOT mean that Y_i and X_i are independent for some or any i.) Then let $Cov(Y_1, X_1)$ denote the (population) covariance between Y_1 and X_1 and let $\widehat{Cov}(Y_1, X_1)$ denote the sample covariance defined by

$$\widehat{Cov}(Y_1, X_1) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)\left(X_i - \bar{X}_n\right), \text{ and } \widehat{Var}(Y_1) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)^2.$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ denote the sample means of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ respectively. Then show the following properties.

(a) For any constants (i.e., nonstochastic numbers) a, b, c, and d, we have

$$\widehat{Cov}(aY_1 + c, bX_1 + d) = ab\widehat{Cov}(Y_1, X_1).$$

(b)

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\mathrm{Var}(Y_1)$$

(c)

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)\left(X_i - \bar{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)X_i = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(X_i - \bar{X}_n\right).$$

Review of Statistics

3. *(Unbiased Estimator vs. Consistent Estimator)* Suppose that we observe a set of i.i.d. random variables $X_1, ..., X_n$ and another set of i.i.d. random variables $Y_1, ..., Y_n$. Let us assume that the correlation between X_i and Y_i is not zero and unknown for any $i = 1, ..., n$, but the correlation between X_i and Y_j for any $i \neq j$ is zero. Suppose that our parameter of interest is as follows:

$$\theta = \mathbf{E}X_1\mathbf{E}Y_1.$$

   (a) Show that an estimator defined by $\hat{\theta} = \bar{X}_n\bar{Y}_n$ for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a consistent estimator of θ. (Hint: apply the Law of Large Numbers and Slutsky Theorem)

   (b) Show that $\hat{\theta} = \bar{X}_n\bar{Y}_n$ is not an unbiased estimator.

   (c) Provide an unbiased estimator of θ. (Hint: what is an unbiased estimator of θ when $n = 2$?)

4. *(Mean Squared Errors and Optimal Weight)* Suppose that we observe i.i.d. random variables $Y_1, ..., Y_n$ with $n = 400$, such that

$$Y_i = \theta + \varepsilon_i,$$

where ε_i's are i.i.d. random variables that are unobserved. It is known that $\mathbf{E}\varepsilon_i = 0$ and $Var(\varepsilon_i) = 0.5$.

   (a) Show that, given any estimator $\tilde{\theta}$ of θ, the Mean Squared Errors defined by $MSE(\tilde{\theta}) = E[(\tilde{\theta} - \theta)^2]$ satisfies

$$MSE(\tilde{\theta}) = \left(\mathrm{Bias}(\tilde{\theta})\right)^2 + \mathrm{Var}(\hat{\theta}),$$

   where $\mathrm{Bias}(\tilde{\theta}) = E[\tilde{\theta}] - \theta$.

   (b) Find out the least squares estimator of θ, i.e., find out $\hat{\theta}$ such that

$$\hat{\theta} = \arg\min_{b}\sum_{i=1}^{n}(Y_i - b)^2.$$

(c) Show that $\hat{\theta}$ is an unbiased estimator of θ.

(d) Show that $\hat{\theta}$ is a consistent estimator of θ.

(e) Compute the MSE of the least squares estimator $\hat{\theta}$ you obtained in (b). Does the MSE of $\hat{\theta}$ converges to zero as $n \to \infty$?

(f) Suppose that we consider the following form of an estimator for $\mathbf{E}Y_1$:

$$\tilde{\theta}(a,b) = \sum_{i=1}^{n} Y_i a_i + b,$$

where a_i's and b are constants to be chosen, and $a = (a_i)_{i=1}^n$, which satisfy the following constraint:

$$\sum_{i=1}^{n} a_i = 1.$$

Find $a = (a_i)_{i=1}^n$ and b which minimize the MSE of $\tilde{\theta}(a,b)$. Let the optimal estimator be denoted by $\tilde{\theta}(a^*, b^*)$, where a^* and b^* are the minimizers. Is the estimator $\tilde{\theta}(a^*, b^*)$ unbiased?

(g) Now, we drop the assumption that Y_i's are i.i.d. We assume instead that Y_i's are independent across i's, and $\mathbf{E}Y_i$'s are identical across i, but $Var(Y_i) = \sigma_i^2$ which are positive and all different across i's, and are known. (That is, Y_i's are heteroskedastic.) Then, compute a and b which minimizes the MSE of $\tilde{\theta}(a,b)$ in this set-up. Let the minimizer be $\tilde{a}$ and $\tilde{b}$. Is the estimator $\tilde{\theta}(\tilde{a}, \tilde{b})$ unbiased?

5. *(Central Limit Theorem, Hypothesis Test, and Power of Test)* Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. random vectors, which means that (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) are mutually independent random vectors and the random vectors are identically distributed. (This does not mean that X_i and Y_i are independent.) The null hypothesis of interest is

$$H_0 : \mathbf{E}X_i = \mathbf{E}Y_i$$

and the alternative hypothesis is

$$H_1 : \mathbf{E}X_i > \mathbf{E}Y_i.$$

Suppose for simplicity that $Var(X_i - Y_i)$ is known to be 2.

(a) Show that

$$\frac{\sqrt{n}(\bar{X}_n - \bar{Y}_n - (\mathbf{E}X_i - \mathbf{E}Y_i))}{\sqrt{2}} \to_d N(0,1),$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. (HINT: Use the Central Limit Theorem.)

(b) Using (a), find a test statistic τ and critical value $c \in \mathbf{R}$ such that under the null hypothesis, as $n \to \infty$,

$$P\{\tau > c\} \to 0.05.$$

(c) Show that under the alternative hypothesis such that

$$\mathbf{E}X_i = \mathbf{E}Y_i + 1 > \mathbf{E}Y_i, \tag{1}$$

the power of the test (τ, c) converges to 1 as $n \to \infty$. (HINT. For this, it is sufficient to show that for any small $\varepsilon > 0$, there exists a sufficiently large n_0 such that for all $n \geq n_0$,

$$P\{\tau > c\} > 1 - \varepsilon,$$

under the alternative hypothesis in (1).)

6. (Stock and Watson E.3.4) A survey of 1000 registered voters is conducted, and the voters are asked to choose between candidate A and candidate B. Let p denote the fraction of voters in the population who prefer candidate A, and let p_n denote the fraction of voters in the sample who prefer candidate A.

(a) You are interested in the competing hypotheses $H_0 : p = 0.4$ v.s. $H_1 : p \neq 0.4$. Suppose that you decide to reject H_0 if $|\hat{p} - 0.4| > 0.01$.

i. What is the size of the test?

ii. Compute the power of the test if $p = 0.45$.

(b) In the survey, $p_n = 0.44$.

i. Test $H_0 : p = 0.4$ v.s. $H_1 : p \neq 0.4$ using a 10% significance level.

ii. Test $H_0 : p = 0.4$ v.s. $H_1 : p < 0.4$ using a 10% significance level.

iii. Construct a 90% confidence interval for p.

iv. Construct a 99% confidence interval for p.

v. Construct a 60% confidence interval for p.

(c) Suppose that the survey is carried out 30 times, using independently selected voters in each survey. For each of these 30 surveys, a 90% confidence interval for p is constructed.

i. What is the probability that the true value of p is contained in all 30 of these confidence intervals?

ii. How many of these confidence intervals do you expect to contain the true value of p?

(d) In survey jargon, the "margin of error" is $1.96 * SE(p_n)$, that is, it is halft he length of the 95% confidence interval. Suppose you want to design a survey that has a margin of error of at most 0.5%. That is, you want $P(|p_n - p| > 0.005 \leq 0.005)$. How large should n be if the survey uses simple random sampling?

Programming

7. For this question, include your code with the output in the submission.

(a) *(Discrete Random Variables)* Sampling Suppose you are the lottery fairy in a weekly lottery, where 5 out of 36 unique numbers are drawn. Instructions: Draw the winning numbers for this week.
(Hints: Start with `set.seed(123)` to set the seed. You may use the function `sample()` to draw random numbers.)

(b) Consider a random variable X with probability density function (PDF) $f_X(x) = \frac{x}{4}e^{-x^2/8}, x \geq 0$. Define the PDF from above as a function and check whether the function you have defined is indeed a PDF. (Hints: You may use the function integrate to check the integration of a function.)

(c) *(Normal Distribution)* Let $Y \sim N(3, 10)$, compute the 99%-th quantile of the given distribution.

8. Consider the following alternative estimator for μ_Y, the mean of the Y_i,

$$\tilde{Y_i} = \frac{1}{n-1}\sum_{i=1}^{n} Y_i.$$

(a) In this exercise we will illustrate that this estimator is a biased estimator for Y_i.
Instructions: 1. Define a function `Y_tilde` that implements the estimator above.
2. Randomly draw 5 observations from the `N(10,25)` distribution and compute an estimate using `Y_tilde()`. Repeat this procedure 10000 times and store the results in `est_biased`.
3. Plot a histogram of `est_biased`. Add a red vertical line at using the function `abline()`.

(b) do the same procedure as in part (a). Increase the number of observations to draw from 5 to 1000 this time. What do you notice? What can you say about this estimator?

9. (Stock and Watson Empirical Exercise.3.4) On the text website[1], you will find the data file CPS96_15, which contains an extended version of the data set used in Table 3.1 of the text for the years 1996 and 2015. It contains data on full-time workers, ages 25–34, with a high school diploma or a B.A./B.S. as their highest degree. A detailed description is given in CPS96_15_Description, available on the website. Use these data to complete the following.

(a) i. Compute the sample mean for average hourly earnings (AHE) in 1996 and 2015.

ii. Compute the sample standard deviation for AHE in 1996 and 2015.

iii. Construct a 95% confidence interval for the population means of AHE in 1996 and 2015.

iv. Construct a 95% confidence interval for the change in the population means of AHE between 1996 and 2015.

(b) In 2015, the value of the Consumer Price Index (CPI) was 237.0. In 1996, the value of the CPI was 156.9. Repeat (a), but use AHE measured in real 2015 dollars ($2015); that is, adjust the 1996 data for the price inflation that occurred between 1996 and 2015.

[1] `http://www.pearsonglobaleditions.com`

(c) If you were interested in the change in workers' purchasing power from1 996 to 2015, would you use the results from (a) or (b)? Explain.

(d) Using the data for 2015:

  i. Construct a 95% confidence interval for the mean of AHE for high school graduates.

  ii. Construct a 95% confidence interval for the mean of AHE for workers with a college degree.

  iii. Construct a 95% confidence interval for the difference between the two means.

(e) Repeat (d) using the 1996 data expressed in \$2015.

(f) Using appropriate estimates, confidence intervals, and test statistics, answer the following questions:

  i. Did real (inflation-adjusted) wages of high school graduates increase from 1996 to 2015?

  ii. Did real wages of college graduates increase?

  iii. Did the gap between earnings of college and high school graduates increase? Explain.