Econometrics (ECON 21020) – Winter 2020
Problem Set #2 Solutions

SW 3.2

(a) Note that $Y = 1$ denotes success and $Y = 0$ denotes failure. Thus, by summing over Y_i, we count the number of successes. Thus, $\hat{p}$ is the sample mean of Y_i's.

$$\begin{aligned}\hat{p} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i=1\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \because Y_i \in \{0,1\} \\ &= \bar{Y}.\end{aligned}$$

(b)

$$\mathbf{E}\left[\hat{p}\right] = \mathbf{E}\left[\bar{Y}\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}\cdot np = p.$$

(c)

$$\begin{aligned}\mathrm{Var}\left(\hat{p}\right) &= \mathbf{E}\left[\left(\hat{p} - \mathbf{E}\left[\hat{p}\right]\right)^2\right] = \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - p\right)^2\right] \\ &= \mathbf{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(Y_i - p\right)^2 + \frac{2}{n^2}\sum_{i\neq j}\left(Y_i - p\right)\left(Y_j - p\right)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\left[\left(Y_i - p\right)^2\right] + \frac{2}{n^2}\sum_{i\neq j}\mathbf{E}\left[(Y_i - p)(Y_j - p)\right] \\ &= \frac{n}{n^2}\cdot p(1-p) + \frac{2n(n-1)}{n^2}\cdot 0 = \frac{p(1-p)}{n}\end{aligned}$$

SW 3.3

(a) Let us use the fraction of survey respondents who preferred the incumbent as an estimator for the fraction of all likely voters who preferred the incumbent.

$$\hat{p} = \frac{215}{400} = \frac{43}{80} = 0.5375.$$

Our estimate is 0.5375.

(b)

$$\frac{\hat{p}\left(1-\hat{p}\right)}{n} = \frac{1}{400}\frac{43}{80}\frac{37}{80} = \frac{1591}{2560000} \approx 0.0006215.$$

The standard error is 0.02493.

(c) By the CLT and Slutksy's Theorem,

$$\frac{\sqrt{n}\,(\hat{p}-p)}{\sqrt{\hat{p}\,(1-\hat{p})}} \xrightarrow{d} N(0,1).$$

Given the null hypothesis and the alternative hypothesis, our test is to reject the null if

$$\left|\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}}\right| \geq c$$

with some constant c. Consider some random variable $Z \sim N(0,1)$. The p-value is

$$\begin{aligned}\Pr\left\{|Z| \geq \left|\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}}\right|\right\} &= \Pr\left\{|Z| \geq \left|\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}}\right|\right\} \\ &\approx \Pr\left\{|Z| \geq \frac{0.0375}{0.02493}\right\} \\ &\approx \Pr\left\{|Z| \geq 1.5042\right\} \approx 0.1325.\end{aligned}$$

(d) Note, the test is to reject the null if

$$\frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}} \geq c$$

with some constant c. Consider some random variable $Z \sim N(0,1)$. The p-value is

$$\begin{aligned}\Pr\left\{Z \geq \frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}}\right\} &= \Pr\left\{Z \geq \frac{\sqrt{n}\,(\hat{p}-0.5)}{\sqrt{\hat{p}\,(1-\hat{p})}}\right\} \\ &\approx \Pr\left\{Z \geq \frac{0.0375}{0.02493}\right\} \\ &\approx \Pr\left\{Z \geq 1.5042\right\} \approx 0.0663.\end{aligned}$$

(e) The test in **c.** is a two-sided test while the test in **d.** is a one-sided test. Since the one-sided test has only one side of the critical region, the p-value is smaller, (by the symmetry of the normal distribution, the p-value is actually exactly half).

(f) With the significance level of 0.95, the both tests do not reject the null. We do not have statistically significant evidence.

SW 3.4

(a) From the CLT and Slutsky's theorem, we have

$$\begin{aligned}0.95 &\approx \Pr\left\{-1.96 \leq \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}\,(1-\hat{p})}} \leq 1.96\right\} \\ &= \Pr\left\{\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right\}.\end{aligned}$$

By plugging in $\hat{p} = 0.5375$, the 95% confidence interval is

$$[0.4886, 0.5864].$$

(b) We have

$$\begin{aligned} 0.99 &\approx \Pr\left\{-2.58 \leq \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}\,(1-\hat{p})}} \leq 2.58\right\} \\ &= \Pr\left\{\hat{p} - 2.58\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + 2.58\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right\}. \end{aligned}$$

By plugging in $\hat{p} = 0.5375$, the 99% confidence interval is

$$[0.4732, 0.6018].$$

(c) The interval in **b.** is wider than the interval in **a.** since we need to increase the width to increase the frequency of capturing p.

(d) The null is not rejected since 0.5 is included in the 95% confidence interval.

SW 3.15

(a) Refer to **SW 3.2**.

(b) This follows by part (a) and independence of the two samples, since then we have

$$\mathrm{Var}[\hat{p}_a - \hat{p}_b] = \mathrm{Var}[\hat{p}_a] + \mathrm{Var}[\hat{p}_b]\ .$$

(c) We claimed in class that for a two sample test of $E[X] - E[Y]$,

$$\left(\mathrm{Var}[\bar{X}_n] + \mathrm{Var}[\bar{Y}_m]\right)^{-1/2}\left(\bar{X}_n - \bar{Y}_m - (E[X] - E[Y])\right) \approx N(0,1).$$

Thus, using the notation in our problem

$$\begin{aligned} 0.95 &\approx \Pr\left\{\left|\left(\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}\right)^{-1/2}(\hat{p}_a - \hat{p}_b - (p_a - p_b))\right| \leq 1.96\right\} \\ &= \Pr\left\{\hat{p}_a - \hat{p}_b - 1.96 \cdot \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}\right. \\ &\qquad \left. \leq p_a - p_b \leq \hat{p}_a - \hat{p}_b + 1.96 \cdot \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}\right\}. \end{aligned}$$

For 90% confidence interval,

$$0.9 \approx \Pr\left\{\left|\left(\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}\right)^{-1/2}(\hat{p}_a-\hat{p}_b-(p_a-p_b))\right| \le 1.645\right\}$$
$$= \Pr\left\{\hat{p}_a-\hat{p}_b-1.645\cdot\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}\right.$$
$$\left.\le p_a-p_b \le \hat{p}_a-\hat{p}_b+1.645\cdot\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}\right\}.$$

The 90% confidence interval is

$$\hat{p}_a-\hat{p}_b \pm 1.645\cdot\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}.$$

(d) We have

$$\begin{aligned}\hat{p}_a &= 0.859\\ \hat{p}_b &= 0.374\\ n_a &= 5801\\ n_b &= 4249.\end{aligned}$$

The 95% confidence interval is

$$0.859-0.374 \pm 1.96\cdot\sqrt{\frac{1}{5801}\cdot 0.859\cdot 0.141+\frac{1}{4249}\cdot 0.374\cdot 0.626}$$
$$\Leftrightarrow [0.4679, 0.5021].$$

SW 3.16

(a) By the CLT,

$$\frac{\sqrt{n}(\bar{X}-\mathbf{E}[X])}{\sqrt{\mathrm{Var}(X)}} \xrightarrow{d} N(0,1).$$

Since

$$s^2_{X,n} \xrightarrow{p} \mathrm{Var}(X),$$

we have by Slutsky's theorem that

$$\frac{\sqrt{n}(\bar{X}-\mathbf{E}[X])}{s_X} \xrightarrow{d} N(0,1).$$

The 95% confidence interval is

$$\left[1013-1.96\cdot\frac{108}{\sqrt{453}}, 1013+1.96\cdot\frac{108}{\sqrt{453}}\right] = [1003.0544, 1022.9456].$$

(b) Since the confidence interval does not include 1000, we have a statistically significant evidence that the Florida mean is different that the national mean.

(c) We can apply the same procedure as in **SW 3.15**.

i. The 95% confidence interval is

$$\left[1019 - 1013 - 1.96 \cdot \sqrt{\frac{108^2}{453} + \frac{95^2}{503}}, 1019 - 1013 + 1.96 \cdot \sqrt{\frac{108^2}{453} + \frac{95^2}{503}}\right]$$

$$= [-6.9554, 18.9554]\,.$$

ii. Since the confidence interval includes 0, we do not have a statistically significant evidence that the average test score for students with the prep course is higher than students without the prep course.

(d) i. The 95% confidence interval is

$$\left[9 - 1.96 \cdot \frac{60}{\sqrt{453}}, 9 + 1.96 \cdot \frac{60}{\sqrt{453}}\right] = [3.4747, 14.5253].$$

ii. Since the confidence interval does not include 0, we have a statistically significant evidence that the average test score for students increases when they retake the test after taking the prep course.

iii. Randomly select n students who have taken the test only one time. Randomly select one half of these students and have them take the prep course. Administer the test again to all of the n students. Compare the gain in performance of the prep-course second-time test takers to the non-prep-course second-time test takers.

Problem 2

(a)

$$\mathbf{E}[U] = \mathbf{E}[\mathbf{E}[U|X]] = \mathbf{E}[1] = 1.$$

(b)

$$\mathbf{E}[XU] = \mathbf{E}[\mathbf{E}[XU|X]] = \mathbf{E}[X\mathbf{E}[U|X]] = \mathbf{E}[X \cdot 1] = \mathbf{E}[X] = 2.$$

(c)

$$\mathrm{Cov}(X, U) = \mathbf{E}[XU] - \mathbf{E}[X]\mathbf{E}[U] = 2 - 1 \cdot 2 = 0.$$

(d) U is mean independent of X. $\mathbf{E}[U|X]$ is a constant.

(e) U is uncorrelated with X. The covariance of X and U is zero.

(f) We do not have enough information to determine whether X is independent of U. To say X and U are independent, we need to know the joint distribution. Here we provide an example where we do not have independence. Suppose X and U are discrete, where $X \in \{1, 3\}$, $U \in \{-1, 0, 2, 3\}$. The joint p.m.f is given below:

$$p_{X,U}(x, u) = \begin{cases} \frac{1}{4}, & \text{if } (x, u) = (1, 0), (1, 2), (3, -1), (3, 3) \\ 0, & \text{otherwise} \end{cases}$$

Note that

$$\mathbf{E}[X] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 3 = 2$$
$$\mathbf{E}[U|X] = \begin{cases} \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 2 = 1, & \text{if } X = 1 \\ \frac{1}{2} \cdot -1 + \frac{1}{2} \cdot 3 = 1, & \text{if } X = 3 \end{cases}$$

but

$$\mathbf{E}[X|U = 0] = 1 \neq 3 = \mathbf{E}[X|U = 3].$$

Problem 3

(a)

$$\begin{aligned} \mathbf{E}\left[\hat{\theta}_n\right] = \mathbf{E}[X] &\Leftrightarrow \mathbf{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \mathbf{E}[X] \\ &\Leftrightarrow \sum_{i=1}^{n} a_i \mathbf{E}[X_i] = \mathbf{E}[X] \\ &\Leftrightarrow \sum_{i=1}^{n} a_i \mathbf{E}[X] = \mathbf{E}[X] \qquad \because X_1, \ldots, X_n \text{ are i.i.d.} \\ &\Leftrightarrow \sum_{i=1}^{n} a_i = 1. \end{aligned}$$

(b)

$$\begin{aligned} \mathrm{Var}[\hat{\theta}_n] &= \mathrm{Var}[\sum_{i=1}^{n} a_i X_i] \\ &= \sum_{i=1}^{n} \mathrm{Var}[a_i X_i] \qquad \because X_1, \ldots, X_n \text{ are independent} \\ &= \sum_{i=1}^{n} a_i^2 \, \mathrm{Var}[X_i] \\ &= \sum_{i=1}^{n} {a_i}^2 \, \mathrm{Var}[X] \qquad \because X_1, \cdots, X_n \text{ are identically distributed} \end{aligned}$$

(c) The constraint that $\hat{\theta}_n$ is an unbiased estimator of $\mathbf{E}[X]$ is

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

Thus, the minimization problem we are facing is

$$\min \sum_{i=1}^{n} {a_i}^2 \mathrm{Var}(X) \text{ s.t. } \sum_{i=1}^{n} a_i = 1.$$

Construct the Lagrangian:

$$\mathcal{L}(a_1, \cdots, a_n, \lambda) = \sum_{i=1}^{n} {a_i}^2 \mathrm{Var}(X) + \lambda \left(1 - \sum_{i=1}^{n} a_i\right).$$

The first order condition is

$$2a_i \mathrm{Var}(X) - \lambda = 0 \qquad i = 1, \cdots, n.$$

and the second order condition is

$$\frac{\partial^2}{\partial (a_1, \cdots, a_n)^2} \mathcal{L} = \mathrm{Var}(X) \mathbb{I}_n,$$

which is positive definite. By solving the linear equations of the first order condition,

$$a_i = \frac{\lambda}{2\mathrm{Var}(X)} \quad \text{for all } i \text{ and} \quad \sum_{i=1}^{n} a_i = \frac{n\lambda}{2\mathrm{Var}(X)} = 1.$$

The solution is

$$\lambda^* = \frac{2\mathrm{Var}(X)}{n}$$
$$a_1^* = \cdots = a_n^* = \frac{1}{n}.$$

Since each of $X_1, \cdots, X_n$ have the same information, it is optimal to put equal weights.

Problem 4

(a)

$$\begin{aligned}
\bar{Z}_n &= \frac{1}{n} \sum_{i=1}^{n} Z_i = \frac{1}{n} \sum_{i=1}^{n} (a + bX_i) \\
&= \frac{1}{n} \cdot na + \frac{1}{n} \sum_{i=1}^{n} bX_i = a + b\bar{X}_n \\
\hat{\sigma}_{Z,n}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(Z_i - \bar{Z}_n\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(a + bX_i - a - b\bar{X}_n\right)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left(b(X_i - \bar{X}_n)\right)^2 = b^2 \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 = b^2 \hat{\sigma}_{X,n}^2.
\end{aligned}$$

(b) Note that

$$\mathbf{E}[Z] = \mathbf{E}[a + bX] = a + b\mathbf{E}[X].$$

Thus,

$$\mathbf{E}\left[\bar{Z}_n\right] = \mathbf{E}\left[a + b\bar{X}_n\right] = a + b\mathbf{E}\left[\bar{X}_n\right] = a + b\mathbf{E}[X] = \mathbf{E}[Z].$$

(c) From the WLLN,

$$\bar{X}_n \xrightarrow{p} \mathbf{E}[X] \text{ as } n \to \infty.$$

Consider a function $g(t) = a + bt$. g is continuous on $\mathbb{R}$. By applying the CMT,

$$\bar{Z}_n = a + b\bar{X}_n = g\left(\bar{X}\right) \xrightarrow{p} g\left(\mathbf{E}[X]\right) = a + b\mathbf{E}[X] = \mathbf{E}[Z].$$

Problem 5

(a) We do not have enough information about distribution of (X, Y). For example, if X and Y are independent, we have

$$\mathbf{E}\left[\bar{X}_n\bar{Y}_n\right] = \mathbf{E}\left[\bar{X}_n\right]\mathbf{E}\left[\bar{Y}_n\right] = \mathbf{E}[X]\mathbf{E}[Y].$$

However, we can also construct a counterexample:

$$\begin{aligned} &n = 1 \\ &X \sim \text{Bernoulli}(0.5) \\ &Y = X \end{aligned}$$

Then,

$$\begin{aligned} \mathbf{E}\left[\bar{X}_n\bar{Y}_n\right] = \mathbf{E}\left[X_1Y_1\right] = \mathbf{E}\left[{X_1}^2\right] &= \frac{1}{2}\cdot 1^2 + \frac{1}{2}\cdot 0^2 = \frac{1}{2} \\ &\neq \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} = \mathbf{E}[X]\mathbf{E}[Y]. \end{aligned}$$

(b) By the WLLN, $\bar{X}_n \xrightarrow{p} E[X]$ and $\bar{Y}_n \xrightarrow{p} E[Y]$. Consider the function $g(t, s) = ts$. Note that g is continuous on $\mathbb{R}^2$. By the CMT,

$$\bar{X}_n\bar{Y}_n = g\left(\bar{X}_n, \bar{Y}_n\right) \xrightarrow{p} g\left(\mathbf{E}[X], \mathbf{E}[Y]\right) = \mathbf{E}[X]\mathbf{E}[Y].$$