ECON 21020 PSET 2

Alex Urquhart

I discussed ideas with Nicholas Limon on this assignment.

Problem #1:

(a)

For the estimator $\hat{p}$ to be unbiased, it needs to fulfill the condition $E[\hat{p}] - p = 0$, or $E[\hat{p}] = p$. We can apply the expected value to the formula for $\hat{p}$:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i = 1)$$

$$E[\hat{p}] = E[\frac{1}{n}\sum_{i=1}^{n} 1(Y_i = 1)]$$

$$E[\hat{p}] = \frac{1}{n}E[\sum_{i=1}^{n} 1(Y_i = 1)]$$

The indicator function, $1(Y_i = 1)$, has an expected value of p. We then apply this to the previous line:

$$E[\hat{p}] = \frac{1}{n} * np$$

$$E[\hat{p}] = p$$

Thus we prove that the estimator $\hat{p}$ is unbiased.

(b)

We can apply the same logic as before by finding the variance of both sides of the $\hat{p}$ formula and applying the value of the indicator function:

$$Var[\hat{p}] = Var[\frac{1}{n}\sum_{i=1}^{n} 1(Y_i = 1)]$$

$$Var[\hat{p}] = \frac{1}{n^2}Var[\sum_{i=1}^{n} 1(Y_i = 1)]$$

For Bernoulli distribution, we know that $Var[Y] = p(1-p)$, and since the sample is IID we can apply this to Y_i for all values of i. We can then apply this result to the previous line:

$$Var[\hat{p}] = \frac{1}{n^2} * np(1-p)$$

$$Var[\hat{p}] = \frac{p(1-p)}{n}$$

Problem #2:

$$X \sim N(2000, 200^2)$$

$$H_0 : \mu = 2000$$

$$H_a : \mu > 2000$$

(a)

Here, we are performing an upper one-sided test. We use the test statistic $T_n = \frac{\bar{X}_n - \mu_x}{\frac{\hat{\sigma_x}}{\sqrt{n}}}$, where $\frac{\hat{\sigma_x}}{\sqrt{n}}$ is the standard error. Assuming the null hypothesis is true, a Type I Error will occur when the sample mean is greater than 2100. Thus,

$$P(\text{Type I Error}) = P(\bar{X_n} > 2100) = 1 - P(\bar{X_n} \leq 2100)$$

We can then calculate the test statistic and the probability, using $\mu_x = 2000, \hat{\sigma_x} = 200, n = 100$:

$$P(\bar{X_n} > 2100) = 1 - P(\bar{X_n} \leq 2100)$$

$$P(\bar{X_n} > 2100) = 1 - P(\bar{Z_n} \leq \frac{2100 - \mu_x}{\frac{\hat{\sigma_x}}{\sqrt{n}}})$$

$$P(\bar{X_n} > 2100) = 1 - P(\bar{Z_n} \leq \frac{2100 - 2000}{\frac{200}{\sqrt{100}}})$$

$$P(\bar{X_n} > 2100) = 1 - P(\bar{Z_n} \leq \frac{100}{20})$$

$$P(\bar{X_n} > 2100) = 1 - P(\bar{Z_n} \leq 5)$$

Since we normalized the test statistic, we can use pnorm in R to find the probability:

```r
pnorm(5,lower.tail=FALSE)
```

```
## [1] 0.00000028665
```

The probability of committing a Type I Error is therefore $2.87 * 10^{-7}$.

(b)

If the true mean lifetime is 2150 hours, a Type II Error will occur when the sample mean is less than 2100:

$$P(\text{Type II Error}) = P(\bar{X}_n \leq 2100)$$

We can use the same test statistic as before, but setting $\mu_x = 2150$:

$$P(\bar{X}_n \leq 2100) = P(\bar{Z}_n \leq \frac{2100 - 2150}{\frac{200}{\sqrt{100}}})$$

$$P(\bar{X}_n \leq 2100) = P(\bar{Z}_n \leq \frac{-50}{20})$$

$$P(\bar{X}_n \leq 2100) = P(\bar{Z}_n \leq -2.5)$$

We then use pnorm in R to find the probability:

```r
pnorm(-2.5)
```

```
## [1] 0.0062097
```

The probability of committing a Type II error is approximately 0.0062.

(c)

You could use a one-sided test to test the null at a 5% significance level. This is possible because our sample size validates the Central Limit Theorem ($n = 100 > 30$), and the selections in our sample are independent. We use the following steps:

$$\text{Choose a value c such that } \phi(c_{1-\alpha}) = 1 - \alpha.$$

$$\text{We can rewrite this to } c_{1-a} = \text{qnorm}(1 - \alpha).$$

The test statistic remains the same:

$$T_n = \frac{\bar{X} - \mu_x}{\frac{\hat{\sigma}_x}{\sqrt{n}}}$$

We reject the null hypothesis if $T_n > c_{1-\alpha}$.

For $\alpha = 0.05$, we would reject the null if $T_n > c_{0.95} = 1.64$, and fail to reject it otherwise.

Problem #3:

(a)

We are given the formula for $\hat{p}$ in Problem #1:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i = 1)$$

We are still dealing with Bernoulli distribution, where we assume that incumbent voters are $P(Y = 1)$ and challenger voters are $P(Y = 0)$. Thus, the indicator function is only activated for $\frac{215}{400}$ respondents:

$$\hat{p} = \frac{1}{400}\sum_{i=1}^{400} 1(Y_i = 1)$$

$$\hat{p} = \frac{215}{400}$$

(b)

For a proportion test, the standard error is equal to the standard deviation, $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. We then calculate the standard error, using the value of p from (a):

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE = \sqrt{\frac{\frac{215}{400}(1-\frac{215}{400})}{400}}$$

$$SE = \sqrt{0.000621}$$

$$SE = 0.0249$$

(c)

We start by calculating the test statistic, $T_n = \frac{\hat{p}-p_0}{SE}$:

$$T_n = \frac{\frac{215}{400} - 0.5}{0.0249}$$

$$T_n = 1.506$$

We can then calculate the P-Value using pnorm. Since this is a two-sided test, we need to multiply it by 2 to account for both tails:

```r
2*pnorm(1.506,lower.tail=FALSE)
```

```
## [1] 0.13207
```

The P-Value of the two-sided test is 0.1321.

(d)

Here, we use the same test statistic as before, and calculate the P-Value using pnorm. Since this is an upper one-sided test, we don't need to multiply pnorm by 2:

```r
pnorm(1.506, lower.tail=FALSE)
```

```
## [1] 0.066034
```

The P-Value of the upper one-sided test is 0.0660.

(e)

The results from (c) and (d) differ because we are using two different tests (two-sided v.s. upper one-sided), and therefore two different rejection regions. In the case of (c), the rejection region is twice as large, which explains the higher P-Value.

(f)

At a 5% significance level, we would fail to reject the null in either (c) or (d), since both P-Values are greater than 0.05. This implies that there was not statistically significant evidence of the incumbent being ahead.

(g)

$$CI = [\hat{p} - SE * c_{1-\frac{0.05}{2}}, \hat{p} + SE * c_{1-\frac{0.05}{2}}]$$

Let $\hat{p} = \frac{215}{400}, SE = 0.0249, c_{1-\frac{0.05}{2}} = 1.96$.

$$CI = [\frac{215}{400} - 0.0249 * 1.96, \frac{215}{400} + 0.0249 * 1.96]$$

$$CI = [0.489, 0.586]$$

(h)

Same formula as before, but let $\hat{p} = \frac{215}{400}, SE = 0.0249, c_{1-\frac{0.01}{2}} = 2.57$.

$$CI = [\frac{215}{400} - 0.0249 * 2.57, \frac{215}{400} + 0.0249 * 2.57]$$

$$CI = [0.474, 0.601]$$

(i)

The 99% CI is wider than the 95% CI because of the Z-Score used in calculating the Confidence Interval; $c_{0.005} > c_{0.025}$. It also makes intuitive sense, as a higher-confidence interval would need to span more potential sample parameters.

(j)

Since 0.6 falls within the 99% CI but not the 95% CI, we reject the null at the 5% level but fail to reject it at the 1% level.

Problem #4:

(a)

For $\hat{\theta}$ to be an unbiased estimator of $E[X]$, it must fulfill the equation $E[\hat{\theta}] = E[X]$. We can rewrite the equation using the given formula for $\hat{\theta}$:

$$E[\hat{\theta}] = E[\sum_{i=1}^{n} a_i X_i]$$

Since our sample is IID from X, we know that all values of X_i have a distribution of X, and thus $E[X_i] = E[X]$ for all values of i:

$$E[\hat{\theta}] = \sum_{i=1}^{n} a_i E[X]$$

We can then substitute our original equation, $E[\hat{\theta}] = E[X]$, to prove that $\sum_{i=1}^{n} a_i = 1$:

$$E[\hat{\theta}] = \sum_{i=1}^{n} a_i E[\hat{\theta}]$$

$$\frac{E[\hat{\theta}]}{E[\hat{\theta}]} = \sum_{i=1}^{n} a_i$$

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

(b)

Using our result from (a), we can apply the variance to both sides of the equation:

$$Var(\hat{\theta}) = Var(\sum_{i=1}^{n} a_i X_i)$$

We now apply the variance property $Var(aX) = a^2 Var(X)$:

$$Var(\hat{\theta}) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

Finally, since we know from (a) that the distribution of X_i is equal for all values of i, the variance will also be identical for all values of i:

$$Var(\hat{\theta}) = Var(X) \sum_{i=1}^{n} a_i^2$$

(c)

We want to minimize $Var[\hat{\theta_n}]$ subject to $1 = \sum_{i=1}^{n} a_i$. We can use a Lagrangian for the minimization problem:

$$L = Var[\hat{\theta_n}] + \lambda(1 - \sum_{i=1}^{n} a_i)$$

$$L = \sum_{i=1}^{n} a_i^2 Var(X_i) + \lambda(1 - \sum_{i=1}^{n} a_i)$$

We then take the first-order partial derivatives of λ and a_i, and set them to 0:

$$L_{a_i} = 2a_i - \lambda = 0 \rightarrow a_i = \frac{\lambda}{2}$$

$$L_\lambda = 1 - \sum_{i=1}^{n} a_i = 0 \rightarrow 1 = \sum_{i=1}^{n} a_i$$

We can then plug the first equation into the second:

$$1 = \sum_{i=1}^{n} \frac{\lambda}{2}$$

$$1 = n\frac{\lambda}{2}$$

$$1 = na_i$$

$$a_i = \frac{1}{n}$$

This is our minimization solution for all values of i.

Problem #5:

(a)

We can start by defining the relationship between Z_i and $\bar{Z_n}$:

$$E[Z_i] = \frac{1}{n}\sum_{i=1}^{n} Z_i = \bar{Z_n}$$

We then apply this relationship to the given formula for Z_i:

$$E[Z_i] = E[a + bX_i]$$

$$\bar{Z_n} = a + bE[X_i]$$

$$\bar{Z_n} = a + b(\frac{1}{n}\sum_{i=1}^{n} X_i)$$

$$\bar{Z_n} = a + b\bar{X_n}$$

We can then define the variance estimation of X as $\hat{\sigma_X}^2 = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$. Note that, because the sample is IID from X, this result will hold for all values of i:

$$Z_i = a + bX_i$$

$$Var(Z_i) = Var(a + bX_i)$$

$$Var(Z_i) = b^2 Var(X_i)$$

$$\frac{1}{n^2}\sum_{i=1}^{n} Var(Z_i) = b^2 \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$\hat{\sigma_Z}^2 = b^2 \hat{\sigma_X}^2$$

(b)

For $\bar{Z}_n$ to be an unbiased estimator of $E[Z]$, it must fulfill the equivalence $E[\bar{Z}_n] = E[Z]$. We can start by applying the expected value to both sides of the equation

$$E[Z] = E[a + bX]$$

$$E[Z] = a + bE[X]$$

We can then repeat that process on the equation from (a):

$$E[\bar{Z}_n] = E[a + b\bar{X}_n]$$

$$E[\bar{Z}_n] = a + bE[\bar{X}_n]$$

Since $X_1, ..., X_n$ is IID from X and $Var(X) < \infty$, we can use the Weak LLN to prove that $\bar{X}_n$ converges in probability to $E[X]$:

$$E[\bar{Z}] = a + bE[E[X]]$$

$$E[\bar{Z}] = a + bE[X]$$

$$E[\bar{Z}] = E[Z]$$

And we prove that $\bar{Z}_n$ is an unbiased estimator of $E[Z]$.

(c)

For $\bar{Z}_n$ to be a consistent estimator of $E[Z]$, it must converge to $E[Z]$ as $n \to \infty$. We start by making $\bar{Z}_n$ a function of $\bar{X}_n$:

$$\bar{Z}_n = a + b\bar{X}_n = g(\bar{X}_n)$$

We know from the Weak LLN that $\bar{X}_n$ converges in probability to $E[X]$. Additionally, since both $Z = a + bX$ and $\bar{Z}_n = a + b\bar{X}_n$ are continuous, we know from the CMT that $E[X]$ and $E[\bar{X}_n]$ will be spanned by those functions as well. Thus, as $n \to \infty$,

$$g(\bar{X}_n) \to g(E[X])$$

$$g(E[X]) = a + bE[X]$$

$$g(E[X]) = E[Z]$$

We thus prove that $\bar{Z}_n$ is a consistent estimator of $E[Z]$.

Problem #6:

(a)

For $\bar{X_n}\bar{Y_n}$ to be an unbiased estimator of $E[X]E[Y]$, the following equivalence would need to hold:

$$E[\bar{X_n}\bar{Y_n}] = E[X]E[Y]$$

However, this is not always true. Consider the case where $X_n = [a, b] = Y_n$, with probability of $\frac{1}{2}$. Then, we would have

$$E[\bar{X_n}\bar{Y_n}] = E[\bar{X_n}^2] = \frac{a^2 + b^2}{2}$$

.

But, if we calculate

$$E[X]E[Y] = \frac{a+b}{2} * \frac{a+b}{2}$$

$$E[X]E[Y] = \frac{(a+b)^2}{4}$$

We see that they are not equal, since $\frac{a^2+b^2}{2} \neq \frac{(a+b)^2}{4}$. Thus, we cannot prove that the estimator is unbiased.

(b)

For the estimator to be consistent, $\bar{X_n}\bar{Y_n}$ must converge to $E[X]E[Y]$ as $n \to \infty$.

We know from the Weak LLN that $\bar{X_n} \to E[X]$ and $\bar{Y_n} \to E[Y]$ as $n \to \infty$.

We can then assign a function $g(\bar{X_n}, \bar{Y_n}) = \bar{X_n}\bar{Y_n}$, and use the CMT to prove that it converges to $g(E[X], E[Y]) = E[X]E[Y]$.

Problem #7:

Find my repository at https://github.com/aurquhart23/ECON-21020-Problem-Repository.