Problem Set 2

Michael Tang

April 18, 2019

1 Problem 1: Stock & Watson

1.1 Exercise 3.2

(a)

$$\overline{Y} = E\left[Y_1, ..., Y_n\right] = \hat{p} * 1 + (1 - \hat{p}) * 0$$
$$= \hat{p}$$

(b)

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

Therefore, $\hat{p}$ is an unbiased estimator of p.

(c)

$$var\left(\hat{p}\right) = var\left(\overline{Y}\right) = var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var\left(Y_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2, \sigma_Y = p\left(1 - p\right) \text{ for Bernoulli rv.}$$

$$= \frac{p\left(1 - p\right)}{n} \ \Box$$

1.2 Exercise 3.3

(a)

$\hat{p} = 215/400 = .538$

(b)

$SE\left(\hat{p}\right) = \sqrt{var\left(\hat{p}\right)} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.538*.462}{400}} = .025$

(c)

By the Central Limit Theorem, $\left|\frac{\hat{p}-\mu_0}{SE(\hat{p})}\right|$ has approximate distribution $N(0,1)$.

$$p - value = P_{H_0}\left\{\left|\frac{\hat{p}-\mu_0}{SE(\hat{p})}\right| > \left|\frac{\hat{p}^{act}-\mu_0}{SE(\hat{p}^{act})}\right|\right\}$$

$$= 2\Phi\left(-\left|\frac{.538-.5}{.025}\right|\right) = 2\Phi(-1.52) = 2(.0643)$$

$$= .1286$$

(d)

$$p - value = P_{H_0}\left\{\frac{\hat{p}-\mu_0}{SE(\hat{p})} > \frac{\hat{p}^{act}-\mu_0}{SE(\hat{p}^{act})}\right\}$$

$$= \Phi\left(\frac{.038}{.025}\right) = \Phi(-1.52)$$

$$= .0643$$

(e)
p-value of (d) is one half that of (c) because it only counts one direction of inequality and not two.
(f)
Depends on the significance level one wants to take. Assuming a significance level $\alpha = .05$, both p-values show that we do not have significant enough evidence to suggest that the incumbent is ahead of the challenger.

1.3   Exercise 3.4

(a)
95% confidence for $p = \{\hat{p} \pm 1.96 * SE(\hat{p})\} = \{.538 \pm .049\} = \{.489, .587\}$
(b)
99% confidence for $p = \{.538 \pm 2.58 * .025\} = \{.4735, .6025\}$
(c)
The interval in (b) is wider because to be able to capture the population fraction in a larger percentage of samples one has to expand the accepted range of values around the sample mean.
(d)
We cannot reject the null hypothesis at the 5% significance level; see part (f) of Exercise 3.3.

1.4 Exercise 3.15

(a)

$$E\left(\hat{p}_a\right) = \frac{1}{n_a} \sum_{i=1}^{n_a} E\left(Y_i\right) = \frac{1}{n_a} \sum_{i=1}^{n_a} p_a = p_a$$

$$var\left(\hat{p}_a\right) = \frac{1}{n_a} \sum_{i=1}^{n_a} var\left(Y_i\right) = \frac{\sigma_a}{n_a} = \frac{p_a\left(1-p_a\right)}{n_a}$$

$$E\left(\hat{p}_b\right) = \frac{1}{n_b} \sum_{i=1}^{n_b} p_b = p_b$$

$$var\left(\hat{p}_b\right) = \frac{\sigma_b}{n_a} = \frac{p_b\left(1-p_b\right)}{n_b}$$

(b)

Since samples are independent:

$$var\left(\hat{p}_a - \hat{p}_b\right) = var\left(\hat{p}_a\right) - var\left(\hat{p}_b\right) = \frac{p_a\left(1-p_a\right)}{n_a} - \frac{p_b\left(1-p_b\right)}{n_b} \ \square$$

(c)

$$95\% \text{ confidence for } p_a - p_b = \left(\hat{p}_a - \hat{p}_b\right) \pm 1.96 * SE\left(\hat{p}_a - \hat{p}_b\right)$$

$$= \left(\hat{p}_a - \hat{p}_b\right) \pm 1.96 \sqrt{SE\left(\hat{p}_a\right)^2 - SE\left(\hat{p}_b\right)^2}$$

$$= \left(\hat{p}_a - \hat{p}_b\right) \pm 1.96 \sqrt{\frac{\hat{p}_a\left(1-\hat{p}_a\right)}{n_a} + \frac{\hat{p}_b\left(1-\hat{p}_b\right)}{n_b}}$$

For 90% confidence critical value = 2.58, so interval:
$\left(\hat{p}_a - \hat{p}_b\right) \pm 2.58 \sqrt{\frac{\hat{p}_a\left(1-\hat{p}_a\right)}{n_a} + \frac{\hat{p}_b\left(1-\hat{p}_b\right)}{n_b}}$

(d)

$\left(.859 - .374\right) \pm 1.96 \sqrt{\frac{.859*.141}{5801} + \frac{.374*.626}{4249}} = [.468, .502]$

1.5 Exercise 3.16

(a)
$\mu_Y = 1000$ is the mean score for the population, and $\overline{Y} = 1013$ is the sample mean.

$$SE\left(\overline{Y}\right) = \frac{s}{\sqrt{n}} = \frac{108}{\sqrt{453}} = 5.074$$

$$95\% \text{ interval} = \overline{Y} \pm 1.96 * SE\left(\left(\overline{Y}\right)\right)$$

$$= 1013 \pm 1.96 * 5.074 = [1003.054, 1022.946]$$

(b)
There is statistically significant evidence of Florida students performing differently than other US students

at 5% significance, because the population mean of 1000 is outside the 95% confidence interval.

(c)

   i. $(\overline{Y}_p - \overline{Y}) \pm 1.96 * SE(\overline{Y}_p - \overline{Y}) = (1019 - 1013) \pm 1.96 * \sqrt{95^2/503 + 108^2/453} = [-6.955, 18.955]$

   ii. Not at 5% significance, because a difference of 0 between prepped and non-prepped students is within the 95% confidence interval.

(d)

   i. $9 \pm 1.96 * (60/\sqrt{453}) = [3.475, 14.525]$

   ii. Yes at 5% significance, because difference of 0 is not within the 95% confidence interval.

   iii. We already have the statistics for when prep is administered and students do not take the test twice. Now, to compare to that effect and also to the two effects combined, isolate the effect of taking twice by having another large sample group take the test twice with no prep.

1.6 Exercise 3.17

(a)

$(25.30 - 24.83) \pm 1.96\sqrt{12.09^2/2004 + 10.85^2/1594} = [-.281, 1.221]$

(b)

$(21.50 - 21.39) \pm 1.96\sqrt{9.99^2/1951 + 8.39^2/1368} = [-.518, .738]$

(c)

$(3.80 - 3.44) \pm 1.96\sqrt{.35^2 + .35^2} = .36 \pm .97 = [-.61, 1.33]$

1.7 Exercise 3.18

(a)

$$
\begin{aligned}
E\left[\left(Y_i - \overline{Y}\right)^2\right] &= var\left(Y_i - \overline{Y}\right) - E\left[Y_i - \overline{Y}\right]^2 \\
&= var\left(Y_i\right) - 2cov\left(Y_i, \overline{Y}\right) + var\left(\overline{Y}\right) - 0^2 \\
&= var\left(Y_i\right) - 2cov\left(Y_i, \overline{Y}\right) + var\left(\overline{Y}\right) \ \square
\end{aligned}
$$

(b)

$$
\begin{aligned}
cov\left(\overline{Y}, Y_i\right) &= cov\left(\frac{1}{n}\sum_{j=1}^{n}, Y_i\right) \\
&= cov\left(\frac{Y_i}{n}, Y_i\right) \\
&= \frac{1}{n}cov\left(Y_i, Y_i\right) = \frac{\sigma_Y^2}{n}
\end{aligned}
$$

(c)

$$E\left(s_y^2\right) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}E\left[\left(Y_i - \overline{Y}\right)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(var\left(Y_i\right) - 2cov\left(Y_i, \overline{Y}\right) + var\left(\overline{Y}\right)\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma_Y^2 - 2*\frac{\sigma_Y^2}{n} + \frac{\sigma_Y^2}{n}\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{n-1}{n}\sigma_Y^2\right)$$

$$= \frac{1}{n-1}*n*\left(\frac{n-1}{n}\sigma_Y^2\right) = \sigma_Y^2$$

2 Problem 2

(a)
By the law of iterated expectations:
$E[U] = E[E[U|X]] = E[1] = 1$
(b)
Again by law of iterated expectations:
$E[XU] = E[E[XU|X]] = E[XE[U|X]] = E[X] = 2$
(c)
$cov[X, U] = \sigma_{XU} = E[XU] - E[X]E[U] = 2 - 2 = 0$
(d)
Yes, because $E[U] = E[U|X]$.
(e)
Yes, because their covariance is 0.
(f)
No, I do not have enough information to determine independence. I do not have the probability distributions of either random variable.

3 Problem 3

(a)

$\hat{\theta}_n$ unbiased estimator of $E[X] \Rightarrow E[\hat{\theta}_n] = E[X]$

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

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

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

(b)

$$var[\hat{\theta}_n] = \sum_{i=1}^{n} a_i^2 var[X_i] \text{ Eqn (2.31)}$$

$$= var[X] \sum_{i=1}^{n} a_i^2 \; \square$$

(c)
Since $var[X]$ is not what we are adjusting here, we want to find:

$$\min_{a_1,...,a_n} \sum_{i}^{n} a_i^2 \text{ such that } \sum_{i}^{n} a_i = 1$$

$$\min_{a_1,...,a_n} \mathcal{L} = a_1^2 + ... + a_n^2 + \lambda \left(1 - a_1 - ... - a_n\right)$$

FOCs:

$[a_i] 2a_i - \lambda = 0$

$[\lambda] 1 - a_1 - ... - a_n = 0 \Rightarrow na_i = 1 \Rightarrow a_i = 1/n$

Setting all $a_i = 1/n$ is optimal.

4 Problem 4

(a)

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n}(a+bX_i)$$

$$= a + \frac{b}{n}\sum_{i=1}^{n}X_i \qquad\qquad\qquad = a + b\overline{X}_n \;\square$$

$$\hat{\sigma}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(a+bX_i - a+b\overline{X}_n\right)^2$$

$$= \frac{b^2}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

$$= b^2\hat{\sigma}_x^2 \;\square$$

(b)
Recall that $\overline{X}_n$ is an unbiased estimator of X.
$E\left[\overline{Z}_n\right] = E\left[a+b\overline{X}_n\right] = a+bE\left[\overline{X}_n\right] = a+bE[X] = E[a+bX] = E[Z] \;\square$
(c)
By the Law of Large Numbers, $\overline{X}_n$ is a consistent estimator of $E[X]$.
By the Continuous Mapping Theorem:
$\overline{Z}_n = a+b\overline{X}_n \Rightarrow a+bE[X] = E[a+bX] = E[Z]$ where $\Rightarrow$ here means convergence in probability.
$\therefore \overline{Z}_n$ is a consistent estimator of $E[Z]$.

5 Problem 5

(a)
$E\left[\overline{X}_n\overline{Y}_n\right] = E\left[\overline{X}_n\right]E\left[\overline{Y}_n\right] + cov\left(\overline{X}_n,\overline{Y}_n\right) = E[X]E[Y] + cov\left(\overline{X}_n,\overline{Y}_n\right)$
From this, estimator is unbiased iff $cov\left(\overline{X}_n,\overline{Y}_n\right) = 0$ which is only true when $cov(X,Y) = 0$ which we do not know. We cannot conclude whether $\overline{X}_n\overline{Y}_n$ is an unbiased estimator of $E[X]E[Y]$.
(b)
By the Law of Large Numbers, $\overline{X}_n$ converges probabilistically to $E[X]$ and $\overline{Y}_n$ likewise to $E[Y]$. Then, by the Continuous Mapping Theorem, $\overline{X}_n\overline{Y}_n$ converges probabilistically to $E[X]E[Y]$ and is a consistent estimator of it.