Solutions to selected exercises from Chapter 7 of *Wasserman — All of Statistics*

(2) We have $p = T(F) = \int x dF$ so the plug-in estimator for p is $\hat{p} = \frac{1}{n}\sum X_i = \overline{X_n}$. The standard error for $\overline{X_n}$ is $\sqrt{(p(1-p)/n}$ so the estimated standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$. Our approximate 90% normal-based confidence interval is $T(\widehat{F}_n) \pm z\widehat{\text{se}}$ where $z = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.95) \approx 1.64$. Plugging in, the interval is

$$\hat{p} \pm 1.64\sqrt{\hat{p}(1-\hat{p})/n}.$$

Similarly, we take $\hat{q} = \frac{1}{m}\sum Y_i$. Using Example 7.15, the plug-in estimator for $p - q$ is $\hat{p} - \hat{q}$ and the estimated standard error is

$$\widehat{\text{se}} = \sqrt{\frac{\hat{p}^2(1-\hat{p})^2}{n^2} + \frac{\hat{q}^2(1-\hat{q})^2}{m^2}}.$$

Thus a 90% confidence interval is

$$\hat{p} - \hat{q} \pm 1.64\sqrt{\frac{\hat{p}^2(1-\hat{p})^2}{n^2} + \frac{\hat{q}^2(1-\hat{q})^2}{m^2}}.$$

(3) See the Jupyter Notebook 3.ipynb.

(5) We have $\hat{F}_n(x) = \frac{1}{n}\sum I(X_i \leq x)$ so $\hat{F}_n(x) = \frac{1}{n}\sum Y_i$ where Y_i is the random variable

$$Y_i = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}$$

and $\hat{F}_n(y) = \frac{1}{n}\sum Z_i$ where Z_i is defined similarly, interchanging x with y in the above piecewise definition.
We have $\text{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \frac{1}{n^2}\sum_{i,j} \text{Cov}(Y_i, Z_j)$. If $i \neq j$ then Y_i and Z_j are independent so $\text{Cov}(Y_i, Z_j) = 0$. Otherwise, note that $\text{Cov}(Y_i, Z_i) = \mathbb{E}(Y_i Z_i) - \mathbb{E}(Y_i)\mathbb{E}(Z_i)$ and

$$\mathbb{E}(Y_i) = \int I(z \leq x) dF(z) = \int_{-\infty}^{x} dF(z) = F(x) \text{ and similarly } \mathbb{E}(Z_i) = F(y).$$

Without loss of generality, assume that $x < y$. For $\mathbb{E}(Y_i Z_i)$, note that if $X_i > x$ then $Y_i Z_i = 0$ since $Y_i = 0$. If $X_i \leq x$ then $Y_i Z_i = 1$. So in fact $Y_i Z_i = Y_i$. Thus,

$$\text{Cov}(Y_i, Z_i) = \mathbb{E}(Y_i) - \mathbb{E}(Y_i)\mathbb{E}(Z_i) = F(x) - F(x)F(y).$$

Plugging this into our formula yields

$$\text{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \begin{cases} \frac{F(x)}{n}(1 - F(y)) & x < y \\ \frac{F(y)}{n}(1 - F(x)) & y > x. \end{cases}$$

(6) By definition of the empirical distribution,

$$\hat{\theta} = \frac{1}{n}(\#\{i : X_i \leq b\} - \#\{i : X_i \leq a\}) = \frac{1}{n}\sum_{i=1}^{n} I_{(a,b]}(X_i)$$

where I_* denotes an indicator function. By independence,

$$\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(I_{(a,b]}(X_i)).$$

So we want to find $\mathbb{V}(I_{(a,b]}(X_i))$. We have $I_{(a,b]}(X_i)^2 = I_{(a,b]}(X_i)$ so

$$\mathbb{V}(I_{(a,b])}(X_i)) = \mathbb{E}(I_{(a,b]}(X_i)^2) - \mathbb{E}(I_{(a,b]}(X_i))^2 = \mathbb{E}(I_{(a,b]}(X_i)) - \mathbb{E}(I_{(a,b]}(X_i))^2$$

and $\mathbb{E}(I_{(a,b]}(X_i)) = \int_a^b dF(x) = F(b) - F(a)$. Thus, $\mathbb{V}(I_{(a,b]}(X_i)) = F(b) - F(a) - (F(b) - F(a))^2$ and

$$\mathbb{V}(\hat{\theta}) = \frac{1}{n}\left(F(b) - F(a) - (F(b) - F(a))^2\right).$$

Our estimated standard error would be

$$\widehat{\text{se}}(\hat{\theta}) = \frac{1}{\sqrt{n}}\sqrt{\hat{F}_n(b) - \hat{F}_n(a) - (\hat{F}_n(b) - \hat{F}_n(a))^2}$$

and our estimated $(1-\alpha)$-confidence interval would be the normal-based interval

$$\hat{F}_n(b) - \hat{F}_n(a) \pm z_{\alpha/2}\widehat{\text{se}}(\hat{\theta}).$$

(7) See the Jupyter Notebook 7.ipynb.

(8) See the Jupyter Notebook 8.ipynb.

(9) The plug-in estimators are

$$\hat{p}_1 = 0.9, \quad \hat{p}_2 = 0.85, \quad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = 0.05.$$

To estimate the standard error we first compute sample standard deviations. The sample standard deviation for p_1 is

$$\hat{\sigma}_1 = \sqrt{\frac{1}{100}(90 * (1-0.9)^2 + 10 * (0-0.9)^2)} = 0.3$$

and for $\hat{p}_2$

$$\hat{\sigma}_2 = \sqrt{\frac{1}{100}(85 * (1-0.9)^2 + 15 * (0-0.9)^2)} \approx 0.3606$$

Then the estimated standard error for $\hat{\theta}$ is

$$\widehat{\text{se}} = \sqrt{\frac{1}{100}\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)} \approx 0.0469.$$

For $\alpha = 0.2$, we have $z_{\alpha/2} = \Phi^{-1}(0.9) \approx 1.282$. An 80% confidence interval is therefore

$$0.05 \pm (0.0469)(1.282) = (-0.01, 0.11).$$

For $\alpha = 0.05$, we have $z_{\alpha/2} = \Phi^{-1}(0.975) \approx 1.96$. A 95% confidence interval is therefore

$$0.05 \pm (0.0469)(1.96) = (-0.042, 0.142).$$

(10) See the Jupyter Notebook 10.ipynb.