Chapter 7 Solutions

Andrew Wu

Wasserman: All of Statistics

March 14, 2025

Problem 7.1. Prove that at any fixed value of x,

$$\begin{aligned}
\mathbb{E}(\widehat{F}_n(x)) &= F(x), \\
\mathbb{V}(\widehat{F}_n(x)) &= \frac{F(x)(1-F(x))}{n}, \\
\text{mse} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{P} F(x).
\end{aligned}$$

Solution. We have

$$\begin{aligned}
\mathbb{E}(\widehat{F}_n(x)) &= \mathbb{E}\left(\frac{\sum_{i=1}^n I(X_i \le x)}{n}\right) \\
&= \frac{1}{n} \cdot n \cdot \mathbb{E}(I(X_1 \le x)) \\
&= 1 \cdot F(x) + 0 \cdot (1 - F(x)) \\
&= F(x).
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{V}(\widehat{F}_n(x)) &= \mathbb{V}\left(\frac{\sum_{i=1}^n I(X_i \le x)}{n}\right) \\
&= \frac{1}{n^2} \cdot n \cdot \mathbb{V}(I(X_1 \le x)) \\
&= \frac{1}{n}\left[\mathbb{E}((I(X_1 \le x))^2) - \mathbb{E}(I(X_1 \le x))^2\right] \\
&= \frac{1}{n}[F(x) - F(x)^2] \\
&= \frac{F(x)(1-F(x))}{n}.
\end{aligned}$$

We just showed that $\text{bias}(\widehat{F}_n) = 0$, as $\mathbb{E}(\widehat{F}_n(x)) = F(x)$. Thus $\text{mse} = \mathbb{V}(\widehat{F}_n(x))$, and $\text{mse} \to 0$ as $n \to \infty$.

It also follows that as $\text{mse} = \mathbb{E}[(\widehat{F}_n - F)^2] \to 0$, then $\widehat{F}_n(x) \xrightarrow{\text{qm}} F(x)$, so thus $\widehat{F}_n(x) \xrightarrow{P} F(x)$. ☐

Problem 7.2. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $Y_1, \ldots, Y_m \sim \text{Bernoulli}(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90 percent confidence interval for p. Find the plug-in estimator and estimated standard error for $p - q$. Find an approximate 90 percent confidence interval for $p - q$.

Solution. We know that $p = \int x dF(x)$, so the plug-in estimator is $\widehat{p} = \int x d\widehat{F}(x) = \sum_i x_i f(x_i) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n$.

Next, we know that

$$\begin{aligned}
\text{se} &= \sqrt{\mathbb{V}(\widehat{p})} \\
&= \sqrt{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} \\
&= \sqrt{\frac{1}{n^2}\cdot n \cdot \mathbb{V}(X_i)} \\
&= \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

so therefore

$$\widehat{\text{se}} = \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}.$$

We know that an approximate $1-\alpha$ confidence interval for $T(F)$ is $T(\widehat{F}_n) \pm z_{\alpha/2}\widehat{\text{se}}$. Taking $\alpha = 0.1$ and $T(\widehat{F}_n) = \overline{X}_n$, we obtain

$$\left(\overline{X}_n - z_{0.05}\cdot\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{0.05}\cdot\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right).$$

The plug-in estimator for $p-q$ is given by $\widehat{p}-\widehat{q} = \overline{X}_n - \overline{Y}_m$. The standard error is

$$\begin{aligned}
\text{se} &= \sqrt{\mathbb{V}(\overline{X}_n - \overline{Y}_m)} \\
&= \sqrt{\mathbb{V}(\overline{X}_n) + \mathbb{V}(\overline{Y}_m)} \\
&= \sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}
\end{aligned}$$

so therefore

$$\widehat{\text{se}} = \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n} + \frac{\overline{Y}_m(1-\overline{Y}_m)}{m}}$$

and a 90% confidence interval would be

$$\left(\overline{X}_n - \overline{Y}_m - z_{0.05}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n} + \frac{\overline{Y}_m(1-\overline{Y}_m)}{m}}, \overline{X}_n - \overline{Y}_m + z_{0.05}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n} + \frac{\overline{Y}_m(1-\overline{Y}_m)}{m}}\right).$$

□

Problem 7.4. Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}_n(x)$ be the empirical distribution function. For a fixed x, use the central limit theorem to find the limiting distribution of $\widehat{F}_n(x)$.

Solution. Note that for some fixed x, we have that

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}.$$

But $I(X_i \leq x)$ is just a Bernoulli random variable; it takes on 1 with probability $F(x)$ and 0 with probability $1-F(x)$. Thus $\widehat{F}_n(x)$ is the sum of n Bernoulli($F(x)$) random variables then divided by n, all of which have mean $F(x)$ and variance $F(x)(1-F(x))$.

Then, by the Central Limit Theorem, we know that

$$\widehat{F}_n(x) \approx N\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

and we are done. □

Problem 7.5. Let x and y be two distinct points. Find $\mathrm{Cov}(\widehat{F}_n(x), \widehat{F}_n(y))$.

Solution. Assume without loss of generality that $x > y$. We have $\mathrm{Cov}(\widehat{F}_n(x), \widehat{F}_n(y)) = \mathbb{E}(\widehat{F}_n(x)\widehat{F}_n(y)) - \mathbb{E}(\widehat{F}_n(x))\mathbb{E}(\widehat{F}_n(y))$. As we know that $\mathbb{E}(\widehat{F}_n(x)) = F(x)$ and $\mathbb{E}(\widehat{F}_n(y)) = F(y)$, we need only compute $\mathbb{E}(\widehat{F}_n(x)\widehat{F}_n(y))$.

We have

$$\begin{aligned}
\mathbb{E}(\widehat{F}_n(x)\widehat{F}_n(y)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \cdot \frac{1}{n}\sum_{i=1}^{n} I(X_i \le y)\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} I(X_i \le x) \cdot \sum_{i=1}^{n} I(X_i \le y)\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} I(X_i \le x)I(X_i \le y) + \sum_{1\le i,j\le n, i\ne j} I(X_i \le x)I(X_j \le y)\right).
\end{aligned}$$

We can split this using linearity of expectation. Note that

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} I(X_i \le x)I(X_i \le y)\right) &= \mathbb{E}\left(\sum_{i=1}^{n} I(X_i \le y)\right) \\
&= \sum_{i=1}^{n} \mathbb{E}(I(X_i \le y)) \\
&= nF(y)
\end{aligned}$$

where we use the assumption that $x > y$. Next, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{1\le i,j\le n, i\ne j} I(X_i \le x)I(X_j \le y)\right) &= n(n-1)\mathbb{E}(I(X_1 \le x)I(X_2 \le y)) \\
&= n(n-1)F(x)F(y).
\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathbb{E}(\widehat{F}_n(x)\widehat{F}_n(y)) &= \frac{1}{n^2}(nF(y) + n(n-1)F(x)F(y)) \\
&= \frac{F(y)}{n} + \frac{n-1}{n}F(x)F(y) \\
&= \frac{F(y) + (n-1)F(x)F(y)}{n}
\end{aligned}$$

and so

$$\begin{aligned}
\mathrm{Cov}(\widehat{F}_n(x), \widehat{F}_n(y)) &= \frac{F(y) + (n-1)F(x)F(y)}{n} - F(x)F(y) \\
&= \frac{F(y) - F(x)F(y)}{n}.
\end{aligned}$$

□

Problem 7.6. Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}$ be the empirical distribution function. Let $a < b$ be fixed numbers and define $\theta = T(F) = F(b) - F(a)$. Let $\widehat{\theta} = T(\widehat{F}_n) = \widehat{F}_n(b) - \widehat{F}_n(a)$. Find the estimated standard error of $\widehat{\theta}$. Find an expression for an approximate $1 - \alpha$ confidence interval for θ.

Solution. To compute $\widehat{\text{se}}$, we want to begin by finding $\sqrt{\mathbb{V}(\widehat{F}_n(b) - \widehat{F}_n(a))}$. We have

$$\begin{aligned}
\widehat{F}_n(b) - \widehat{F}_n(a) &= \frac{\sum_{i=1}^n [I(X_i \leq b) - I(X_i \leq a)]}{n} \\
&= \frac{\sum_{i=1}^n I(a < X_i \leq b)}{n}
\end{aligned}$$

so thus

$$\begin{aligned}
\mathbb{V}(\widehat{F}_n(b) - \widehat{F}_n(a)) &= \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^n I(a < X_i \leq b)\right) \\
&= \frac{1}{n^2} \cdot n \cdot \mathbb{V}(I(a < X_1 \leq b)) \\
&= \frac{1}{n}(F(b) - F(a))(1 - F(b) + F(a)).
\end{aligned}$$

It follows that the standard error is

$$\text{se} = \frac{\sqrt{(F(b) - F(a))(1 - F(b) + F(a))}}{\sqrt{n}}$$

and that thus the estimated standard error is

$$\widehat{\text{se}} = \frac{\sqrt{(\widehat{F}_n(b) - \widehat{F}_n(a))(1 - \widehat{F}_n(b) + \widehat{F}_n(a))}}{\sqrt{n}}.$$

A $1 - \alpha$ confidence interval would be

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

where $\widehat{\text{se}}$ is the value we just computed. □

Problem 7.9. 100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 recover. Let p_1 be the probability of recovery under the standard treatment and let p_2 be the probability of recovery under the new treatment. We are interested in estimating $\theta = p_1 - p_2$. Provide an estimate, standard error, an 80 percent confidence interval, and a 95 percent confidence interval for θ.

Solution. We can model the data as $X_1, \ldots, X_{100} \sim \text{Bernoulli}(p_1)$ and $Y_1, \ldots, Y_{100} \sim \text{Bernoulli}(p_2)$, where X_i and Y_i represent people getting the standard and new antibiotic, and take on values 1 for recovery and 0 for non-recovery.

A good estimate $\widehat{\theta}$ for θ would be $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$, where $\widehat{p}_1$ and $\widehat{p}_2$ are estimates for the probability of recovery under the standard and new treatments, respectively.

We can set $\widehat{p}_1 = \frac{90}{100} = 0.9$ and $\widehat{p}_2 = \frac{85}{100} = 0.85$, so $\widehat{\theta} = 0.05$.

The standard error would be $\text{se} = \sqrt{\mathbb{V}(\widehat{\theta})} = \sqrt{\mathbb{V}(\widehat{p}_1) + \mathbb{V}(\widehat{p}_2)}$. Note that $\widehat{p}_1 = \frac{X_1 + \ldots + X_{100}}{100}$, so

$$\begin{aligned}
\mathbb{V}(\widehat{p}_1) &= \mathbb{V}\left(\frac{X_1 + \ldots + X_{100}}{100}\right) \\
&= \frac{1}{10000} \cdot 100 \cdot \mathbb{V}(X_1) \\
&= \frac{p_1(1 - p_1)}{100}
\end{aligned}$$

and thus $\text{se} = \sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{100}}$ and $\widehat{\text{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)+\widehat{p}_2(1-\widehat{p}_2)}{100}} = 0.0466$.

Then an 80 percent confidence interval is given by

$$0.05 \pm z_{0.2/2}\widehat{\text{se}} = (-0.0097, 0.1097).$$

A 95 percent confidence interval is given by

$$0.05 \pm z_{0.05/2}\widehat{\text{se}} = (-0.0413, 0.1413).$$

Note that we can find $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ with qnorm$(1-\alpha/2)$ in R. That is, $z_{0.2/2} = \phi^{-1}(1-0.2/2) =$ qnorm(0.9). $\square$