Chapter 7: Estimating the CDF and Statistical Functionals

All of Statistics, Wasserman

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

Solution. Let F be a Bernoulli(p) distribution. Then $p = \int x\, dF(x)$ so the plug-in estimator is

$$\widehat{p} = \int x\, d\widehat{F}_n(x) = \sum_{i=1}^{n} X_i \cdot \frac{1}{n} = \overline{X}_n.$$

Calculating directly,

$$\text{se}(\widehat{p}) = \text{se}_F(\widehat{p}) = \sqrt{V_F(\overline{X}_n)} = \sqrt{\frac{V_F(X_1)}{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

Therefore, the estimated standard error for p (equivalently, the plug-in estimator of $T(F) = \text{se}_F(\widehat{p})$) is

$$\widehat{\text{se}}\,(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

This coincides with the general plug-in formula for the standard error: $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ (cf. Example 7.11) since for Bernoulli samples, $\widehat{p}(1-\widehat{p})$ equals the sample variance. It's possible to avoid the general formula in this case since $\text{se}(\widehat{p})$ is a function p. By the CLT, $\overline{X}_n \approx N(p, (\widehat{\text{se}}(\widehat{p}))^2)$. Therefore,

$$\overline{X}_n \pm z_{0.05}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is a (normal-based) 90 percent confidence interval. The following is cynically hand-wavy for multiple reasons[1] since Wasserman doesn't provide the tools for a rigurous solution. The plug-in estimator of $p - q$ is (cf. Example 7.15)

$$\widehat{p}_n - \widehat{q}_m = \overline{X}_n - \overline{Y}_m.$$

Moreover,

$$\text{se}(\widehat{p}_n - \widehat{q}_m) = \sqrt{V(\overline{X}_n - \overline{Y}_m)} = \sqrt{V(\overline{X}_n) + V(\overline{Y}_m)} = \sqrt{(\text{se}(\widehat{p}))^2 + (\text{se}(\widehat{q}))^2}$$

Therefore,

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

Assuming[2] $\widehat{p}_n - \widehat{q}_m \approx N(p - q, (\widehat{\text{se}}(\widehat{p}_n - \widehat{q}_m))^2)$,

$$(\overline{X}_n - \overline{Y}_m) \pm z_{0.05}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{q}(1-\widehat{q})}{m}}$$

is a (normal-based) 90 percent confidence interval. ◊

[1](1) Is the plug-in estimator of a statistical functional of two distributions $T(F, G)$ defined as $T(\widehat{F}_n, \widehat{G}_m)$? (2) With respect to which distribution is $V(\overline{X}_n - \overline{Y}_m)$ being calculated? (GPT's ans: It is taken with respect to the joint distribution $(F^n \otimes G^m)$ of the two independent samples. Since the samples are independent, the cross-covariance term vanishes, yielding $V(\overline{X}_n - \overline{Y}_m) = V(\overline{X}_n) + V(\overline{Y}_m)$.)

[2]More hand-waving. Notice that the this would be a double limit: $n, m \to \infty$.

4. Let $X_1, \ldots, X_n \sim F$. For a fixed x, use the CLT to find the limiting distribution of $\widehat{F}_n(x)$.

Solution. Let x be fixed and let $Y_i = I(X_i \leq x)$. Note that $Y_i \sim \text{Bernoulli}(F(x))$ are iid and $\widehat{F}_n(x) = \overline{Y}_n$. Therefore, by the CLT

$$\widehat{F}_n(x) = \overline{Y}_n \approx N\left(E(Y_1), \frac{V(Y_1)}{n}\right) = N\left(F(x), \frac{F(x)(1-F(x))}{n}\right).$$

$\lozenge$

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

Solution. Calculating directly,

$$\begin{aligned}
\text{Cov}(\widehat{F}_n(x), \widehat{F}_n(y)) &= \text{Cov}\left(\frac{1}{n}\sum_{i=1}^n I(X_i \leq x),\ \frac{1}{n}\sum_{j=1}^n I(X_j \leq y)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \text{Cov}\left(I(X_i \leq x), I(X_j \leq y)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \text{Cov}\left(I(X_i \leq x), I(X_i \leq y)\right)
\end{aligned}$$

where the last equality is because $X_i \perp X_j$ for $i \neq j$. Moreover,

$$\begin{aligned}
\text{Cov}\left(I(X_i \leq x), I(X_i \leq y)\right) &= E\left(I(X_i \leq x) \cdot I(X_i \leq y)\right) - E\left(I(X_i \leq x)\right) \cdot E\left(I(X_i \leq y)\right) \\
&= F(\min\{x, y\}) - F(x)F(y)
\end{aligned}$$

since

$$I(X_i \leq x) \cdot I(X_i \leq y) = \begin{cases} 1 & \text{if } X_i \leq \min\{x, y\}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\text{Cov}(\widehat{F}_n(x), \widehat{F}_n(y)) = \frac{1}{n}\left(F(\min\{x, y\}) - F(x)F(y)\right).$$

$\lozenge$

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

Solution. Calculating directly,

$$\begin{aligned}
\left(\text{se}(\widehat{\theta})\right)^2 &= V(\widehat{F}_n(b) - \widehat{F}_n(a)) \\
&= V(\widehat{F}_n(a)) + V(\widehat{F}_n(b)) - 2\text{Cov}(\widehat{F}_n(a), \widehat{F}_n(b)) \\
&= \frac{1}{n}\left(F(a)(1-F(a)) + F(b)(1-F(b)) - 2(F(a) - F(a)F(b))\right) \\
&= \frac{\theta(1-\theta)}{n}.
\end{aligned}$$

Therefore, the estimated standard error of $\widehat{\theta}$ is

$$\widehat{\text{se}}(\widehat{\theta}) = \sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}$$

and $\widehat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}$ is an approximate $1 - \alpha$ CI for θ. Alternatively, let $W_i = I(a < X_i \leq b)$ for $i = 1, \ldots, n$. Then $P(W_i = 1) = P(a < X_i \leq b) = F(b) - F(a) = \theta$ and $\overline{W}_n = \widehat{F}_n(b) - \widehat{F}_n(a) = \widehat{\theta}$. In words, $\widehat{\theta}$ is the sample mean of iid Bernoulli(θ) variables. The desired results follow immediately. $\lozenge$

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 people 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. Notice this is a particular case of Exercise 2[3]. The plug-in estimate of θ is

$$\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2 = 0.9 - 0.85 = 0.05.$$

The estimated standard error is

$$\widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} = \sqrt{\frac{0.9(1-0.9)}{100} + \frac{0.85(1-0.85)}{100}} = 0.0466.$$

A $1-\alpha$ confidence interval for θ is[4]

$$0.05 \pm z_{\alpha/2} \cdot 0.0466.$$

◊

[3]We are implicitly assuming two independent iid samples: $X_1, \ldots, X_{100} \sim \text{Bernoulli}(p_1)$ and $Y_1, \ldots, Y_{100} \sim \text{Bernoulli}(p_2)$.
[4]GPT says normal approximation is fine here since $n_k\widehat{p}_k$ and $n_k(1-\widehat{p}_k)$ are all ≥ 10.