Solutions to (some) exercises - Chapter 7 "Estimating the CDF and statistical functionals".

Marco Corneli

Exercise 1. Given $X_1, \dots, X_n \sim F$, i.i.d., the following empirical cdf can be introduced to estimate F

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(X_i). \tag{1}$$

We first compute its expectation

$$\mathbb{E}(\hat{F}_n(x)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbf{1}_{]-\infty,x]}(X_i)) = F(x),$$

since

$$\mathbb{E}(\mathbf{1}_{]-\infty,x]}(X_i)) = \int_{-\infty}^{\infty} \mathbf{1}_{]-\infty,x]}(u)f(u)du = \int_{-\infty}^{x} f(u)du = F(x).$$

Moreover, since $\mathbb{E}(\mathbf{1}^2_{]-\infty,x]}(X_i)) = \mathbb{E}(\mathbf{1}_{]-\infty,x]}(X_i))$ and using that the variance of a random variable Z is

$$\mathbb{V}ar(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2$$

we easily obtain that

$$\mathbb{V}ar(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar(\mathbf{1}_{]-\infty,x]}(X_i)) = \frac{F(x)(1-F(x))}{n}.$$

Since the estimator in Eq. (1) is *unbiased* and its variance is asymptotically null, it converges in probability to its mean $F(x)$.

Exercise 4. This exercise is strongly related to the previous one. By Eq. 1 we see that our estimator is the empirical mean of (indicator) functions of $X_1, \dots, X_n$ that are i.i.d. Moreover, since

$$\mathbb{V}ar(\mathbf{1}_{]-\infty,x]}(X_i)) = F(x)(1-F(x)) \leq 1 < \infty$$

the CLT can be applied and

$$\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \rightsquigarrow \mathcal{N}(0,1).$$

Since the standard deviation, at the denominator, depends on $F(x)$ (unknown) we replace it by its estimate $\hat{F}_n(x)$. Notice that

$$\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}} = \sqrt{n}\frac{\hat{F}_n(x) - \hat{F}_n(x)}{\sqrt{F(x)(1-F(x))}} \frac{\sqrt{F(x)(1-F(x))}}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}$$

where the red term converges <u>in distribution</u> to $\mathcal{N}(0,1)$ and the the blue term converges <u>in probability</u> to 1. Thus the **Slutsky**'s theorem (page 75, point e)) applies and the quantity on the left hand side of the above equality converges to $1 \times \mathcal{N}(0,1)$.

Exercise 2. Not sure there is a unique solution to this exercise. However, this is how I did it. Given $X_1, \ldots, X_n$ i.i.d. following a Bernoulli distribution of parameter p. Now, if we denote by $F_{X_i}(\cdot)$ the cdf of X_i, we have

$$F_{X_i}(u) = \begin{cases} 0 \quad \forall u \in]-\infty, 0[\\ 1-p \quad \forall u \in [0,1[\\ 1 \quad \forall u \in [1, \infty[\end{cases}$$

it means that, for instance, $p = 1 - F_{X_i}(0)$ so that we can estimate p by

$$\hat{p}_n := 1 - \hat{F}_n(0) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,0]}(X_i).$$

This estimator is unbiased

$$\mathbb{E}(\hat{p}_n) = 1 - F_{X_i}(0) = 1 - (1-p) = p$$

and its variance is

$$\mathbb{V}ar(\hat{p}_n) = \mathbb{V}ar(\hat{F}_n(0)) = \frac{F_{X_i}(0)(1-F_{X_i}(0))}{n} = \frac{(1-p)p}{n}.$$

Thus, an estimate of the standard error is

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

The easiest way to build an approximate confidence interval is to rely on the CLT:

$$C_n := [\hat{p}_n - z_{\frac{\alpha}{2}}\hat{se}(\hat{p}_n), \hat{p}_n + z_{\frac{\alpha}{2}}\hat{se}(\hat{p}_n)],$$

where $z_{\frac{\alpha}{2}} := \Phi^{-1}(1 - \frac{\alpha}{2})$ and $\Phi(\cdot)$ denotes de standard normal cdf. So, if you look for a 90% confidence interval (i.e. $1 - \alpha = 0.9$), then

$$\frac{\alpha}{2} = \frac{1-0.9}{2} = 0.05$$

and $z_{0.05} = \Phi^{-1}(0.95) = 1.64$.

Now, if we denote by $\hat{F}_{n,X}$ and $\hat{F}_{m,Y}$ the empirical cdfs of F_X and F_Y, respectively, we can introduce the following estimator of $p - q$

$$\hat{p}_n - \hat{q}_m = 1 - \hat{F}_{n,X}(0) - (1 - \hat{F}_{m,Y}(0)) = \hat{F}_{m,Y}(0) - \hat{F}_{n,X}(0)$$

which is unbiased

$$\mathbb{E}(\hat{p}_n - \hat{q}_m) = 1 - q - (1 - p) = p - q$$

and whose variance is

$$\mathbb{V}ar(\hat{p}_n - \hat{q}_m) = \mathbb{V}ar(\hat{p}_n) + \mathbb{V}ar(\hat{q}_m) = \frac{p(1-p)}{n} + \frac{q(1-q)}{m}.$$

Thus

$$\hat{se}(\hat{p}_n - \hat{q}_m) = \left(\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{\hat{q}_m(1-\hat{q}_m)}{m}\right)^{\frac{1}{2}}.$$

Thus, an asymptotic $1 - \alpha$ confidence interval is

$$C_{n,m} := [(\hat{p}_n - \hat{q}_m) - z_{\frac{\alpha}{2}}\hat{se}(\hat{p}_n - \hat{q}_m), (\hat{p}_n - \hat{q}_m) + z_{\frac{\alpha}{2}}\hat{se}(\hat{p}_n - \hat{q}_m)],$$

Exercise 5. We are asked to compute $Cov(\hat{F}_n(x), \hat{F}_n(y))$. Recalling that

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

and that $\mathbb{E}(\hat{F}_n(x)) = F(x)$ for all x, we have

$$\begin{aligned}
Cov(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{]-\infty,x]}(X_i) - F(x))\right)\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{]-\infty,y]}(X_i) - F(y))\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(\mathbf{1}_{]-\infty,x]}(X_i) - F(x))(\mathbf{1}_{]-\infty,y]}(X_j) - F(y))\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Cov(\mathbf{1}_{]-\infty,x]}(X_i), \mathbf{1}_{]-\infty,y]}(X_j)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}Cov(\mathbf{1}_{]-\infty,x]}(X_i), \mathbf{1}_{]-\infty,y]}(X_i) \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}Cov(\mathbf{1}_{]-\infty,x]}(X_i), \mathbf{1}_{]-\infty,y]}(X_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}Cov(\mathbf{1}_{]-\infty,x]}(X_i), \mathbf{1}_{]-\infty,y]}(X_i))
\end{aligned} \tag{2}$$

where the last equality comes from the independence between X_i and X_j when $i \neq j$. Now, the last term inside the sum is

$$\mathbb{E}\left(\mathbf{1}_{]-\infty,x]}(X_i)\mathbf{1}_{]-\infty,y]}(X_i)\right) - F(x)F(y).$$

If we assume that $y \geq x$, this term reduces to

$$\mathbb{E}\left(\mathbf{1}_{]-\infty,x]}(X_i)\right) - F(x)F(y) = F(x)(1-F(y)).$$

Thus, replacing into Eq. (2), we finally obtain

$$Cov(\hat{F}_n(x), \hat{F}_n(y)) = \frac{F(x)(1-F(y))}{n}.$$

Notice that when $y = x$ we find the variance of $\hat{F}_n(x)$ which makes sense! The case $y \leq x$ can be treated similarly.