7

Estimating the CDF and Statistical Functionals

The first inference problem we will consider is nonparametric estimation of the CDF F. Then we will estimate statistical functionals, which are functions of CDF, such as the mean, the variance, and the correlation. The nonparametric method for estimating functionals is called the plug-in method.

7.1 The Empirical Distribution Function

Let $X_1, \ldots, X_n \sim F$ be an IID sample where F is a distribution function on the real line. We will estimate F with the empirical distribution function, which is defined as follows.

7.1 Definition. *The* **empirical distribution function** $\widehat{F}_n$ *is the* CDF *that puts mass* $1/n$ *at each data point* X_i. *Formally,*

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \tag{7.1}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$

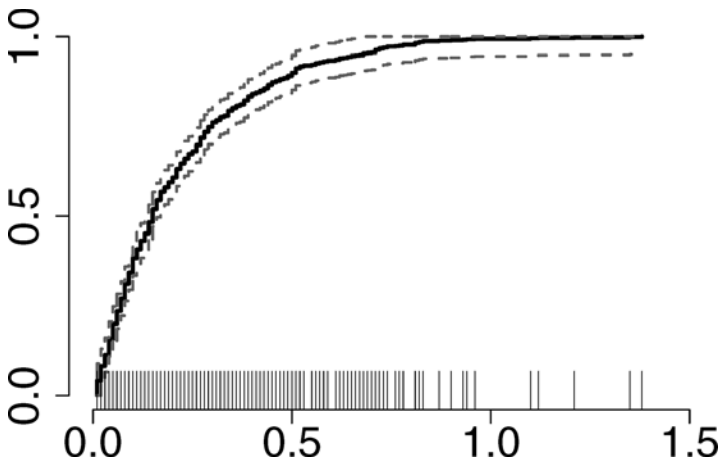

FIGURE 7.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

7.2 Example (Nerve Data). Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. Figure 7.1 shows the empirical CDF $\widehat{F}_n$. The data points are shown as small vertical lines at the bottom of the plot. Suppose we want to estimate the fraction of waiting times between .4 and .6 seconds. The estimate is $\widehat{F}_n(.6) - \widehat{F}_n(.4) = .93 - .84 = .09$. ■

7.3 Theorem. *At any fixed value of* x,

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

7.4 Theorem (The Glivenko-Cantelli Theorem). *Let* $X_1, \ldots, X_n \sim F$. *Then* [1]

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{P}} 0.$$

Now we give an inequality that will be used to construct a confidence band.

[1] More precisely, $\sup_x |\widehat{F}_n(x) - F(x)|$ converges to 0 almost surely.

7.5 Theorem (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality). *Let* $X_1, \ldots, X_n \sim F$. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{7.2}$$

From the DKW inequality, we can construct a confidence set as follows:

A Nonparametric $1-\alpha$ Confidence Band for F

Define,

$$\begin{aligned} L(x) &= \max\{\widehat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\widehat{F}_n(x) + \epsilon_n, 1\} \\ \text{where} \quad \epsilon_n &= \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}. \end{aligned}$$

It follows from (7.2) that for any F,

$$\mathbb{P}\left(L(x) \leq F(x) \leq U(x) \text{ for all } x\right) \geq 1 - \alpha. \tag{7.3}$$

7.6 Example. The dashed lines in Figure 7.1 give a 95 percent confidence band using $\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{.05}\right)} = .048$. ■

7.2 Statistical Functionals

A **statistical functional** $T(F)$ is any function of F. Examples are the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2\, dF(x)$ and the median $m = F^{-1}(1/2)$.

7.7 Definition. *The* **plug-in estimator** *of* $\theta = T(F)$ *is defined by*

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

In other words, just plug in $\widehat{F}_n$ *for the unknown* F.

7.8 Definition. *If* $T(F) = \int r(x) dF(x)$ *for some function* $r(x)$ *then* T *is called a* **linear functional.**

The reason $T(F) = \int r(x)dF(x)$ is called a linear functional is because T satisfies

$$T(aF + bG) = aT(F) + bT(G),$$

hence T is linear in its arguments. Recall that $\int r(x)dF(x)$ is defined to be $\int r(x)f(x)dx$ in the continuous case and $\sum_j r(x_j)f(x_j)$ in the discrete. The empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each X_i. Hence, if $T(F) = \int r(x)dF(x)$ is a linear functional then we have:

7.9 Theorem. *The plug-in estimator for linear functional* $T(F) = \int r(x)dF(x)$ *is:*

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i). \tag{7.4}$$

Sometimes we can find the estimated standard error se of $T(\widehat{F}_n)$ by doing some calculations. However, in other cases it is not obvious how to estimate the standard error. In the next chapter, we will discuss a general method for finding $\widehat{\mathsf{se}}$. For now, let us just assume that somehow we can find $\widehat{\mathsf{se}}$.

In many cases, it turns out that

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2). \tag{7.5}$$

By equation (6.11), an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$T(\widehat{F}_n) \pm z_{\alpha/2}\,\widehat{\mathsf{se}}. \tag{7.6}$$

We will call this the **Normal-based interval.** For a 95 percent confidence interval, $z_{\alpha/2} = z_{.05/2} = 1.96 \approx 2$ so the interval is

$$T(\widehat{F}_n) \pm 2\,\widehat{\mathsf{se}}.$$

7.10 Example (The mean)**.** Let $\mu = T(F) = \int x\,dF(x)$. The plug-in estimator is $\widehat{\mu} = \int x\,d\widehat{F}_n(x) = \overline{X}_n$. The standard error is $\mathsf{se} = \sqrt{\mathbb{V}(\overline{X}_n)} = \sigma/\sqrt{n}$. If $\widehat{\sigma}$ denotes an estimate of σ, then the estimated standard error is $\widehat{\sigma}/\sqrt{n}$. (In the next example, we shall see how to estimate σ.) A Normal-based confidence interval for μ is $\overline{X}_n \pm z_{\alpha/2}\,\widehat{\mathsf{se}}$. ∎

7.11 Example (The Variance)**.** Let $\sigma^2 = T(F) = \mathbb{V}(X) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$. The plug-in estimator is

$$\widehat{\sigma}^2 = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2$$

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

Another reasonable estimator of σ^2 is the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

In practice, there is little difference between $\widehat{\sigma}^2$ and S_n^2 and you can use either one. Returning to the last example, we now see that the estimated standard error of the estimate of the mean is $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. ■

7.12 Example (The Skewness). Let μ and σ^2 denote the mean and variance of a random variable X. The skewness is defined to be

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}.$$

The skewness measures the lack of symmetry of a distribution. To find the plug-in estimate, first recall that $\widehat{\mu} = n^{-1}\sum_i X_i$ and $\widehat{\sigma}^2 = n^{-1}\sum_i (X_i - \widehat{\mu})^2$. The plug-in estimate of κ is

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_i (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}. \quad ■$$

7.13 Example (Correlation). Let $Z = (X, Y)$ and let $\rho = T(F) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)/(\sigma_x \sigma_y)$ denote the correlation between X and Y, where $F(x, y)$ is bivariate. We can write

$$T(F) = a(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$$

where

$$T_1(F) = \int x\, dF(z), \quad T_2(F) = \int y\, dF(z), \quad T_3(F) = \int xy\, dF(z),$$
$$T_4(F) = \int x^2\, dF(z), \quad T_5(F) = \int y^2\, dF(z),$$

and

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}.$$

Replace F with $\widehat{F}_n$ in $T_1(F), \ldots, T_5(F)$, and take

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)).$$

We get

$$\widehat{\rho} = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_i (X_i - \overline{X}_n)^2}\sqrt{\sum_i (Y_i - \overline{Y}_n)^2}}$$

which is called the **sample correlation.** ■

7.14 Example (Quantiles)**.** Let F be strictly increasing with density f. For $0 < p < 1$, the p^{th} quantile is defined by $T(F) = F^{-1}(p)$. The estimate if $T(F)$ is $\widehat{F}_n^{-1}(p)$. We have to be a bit careful since $\widehat{F}_n$ is not invertible. To avoid ambiguity we define

$$\widehat{F}_n^{-1}(p) = \inf\{x : \ \widehat{F}_n(x) \geq p\}.$$

We call $T(\widehat{F}_n) = \widehat{F}_n^{-1}(p)$ the p^{th} **sample quantile.** ■

Only in the first example did we compute a standard error or a confidence interval. How shall we handle the other examples? When we discuss parametric methods, we will develop formulas for standard errors and confidence intervals. But in our nonparametric setting we need something else. In the next chapter, we will introduce the bootstrap for getting standard errors and confidence intervals.

7.15 Example (Plasma Cholesterol)**.** Figure 7.2 shows histograms for plasma cholesterol (in mg/dl) for 371 patients with chest pain (Scott et al. (1978)). The histograms show the percentage of patients in 10 bins. The first histogram is for 51 patients who had no evidence of heart disease while the second histogram is for 320 patients who had narrowing of the arteries. Is the mean cholesterol different in the two groups? Let us regard these data as samples from two distributions F_1 and F_2. Let $\mu_1 = \int x dF_1(x)$ and $\mu_2 = \int x dF_2(x)$ denote the means of the two populations. The plug-in estimates are $\widehat{\mu}_1 = \int x d\widehat{F}_{n,1}(x) = \overline{X}_{n,1} = 195.27$ and $\widehat{\mu}_2 = \int x d\widehat{F}_{n,2}(x) = \overline{X}_{n,2} = 216.19$. Recall that the standard error of the sample mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is

$$\mathsf{se}(\widehat{\mu}) = \sqrt{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i)} = \sqrt{\frac{n\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$$

which we estimate by

$$\widehat{\mathsf{se}}(\widehat{\mu}) = \frac{\widehat{\sigma}}{\sqrt{n}}$$

where

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}.$$

For the two groups this yields $\widehat{\mathsf{se}}(\widehat{\mu}_1) = 5.0$ and $\widehat{\mathsf{se}}(\widehat{\mu}_2) = 2.4$. Approximate 95 percent confidence intervals for μ_1 and μ_2 are $\widehat{\mu}_1 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_1) = (185, 205)$ and $\widehat{\mu}_2 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_2) = (211, 221)$.

Now, consider the functional $\theta = T(F_2) - T(F_1)$ whose plug-in estimate is $\widehat{\theta} = \widehat{\mu}_2 - \widehat{\mu}_1 = 216.19 - 195.27 = 20.92$. The standard error of $\widehat{\theta}$ is

$$\mathsf{se} = \sqrt{\mathbb{V}(\widehat{\mu}_2 - \widehat{\mu}_1)} = \sqrt{\mathbb{V}(\widehat{\mu}_2) + \mathbb{V}(\widehat{\mu}_1)} = \sqrt{(\mathsf{se}(\widehat{\mu}_1))^2 + (\mathsf{se}(\widehat{\mu}_2))^2}$$

and we estimate this by

$$\widehat{\mathsf{se}} = \sqrt{(\widehat{\mathsf{se}}(\widehat{\mu}_1))^2 + (\widehat{\mathsf{se}}(\widehat{\mu}_2))^2} = 5.55.$$

An approximate 95 percent confidence interval for θ is $\widehat{\theta} \pm 2\,\widehat{\mathsf{se}}(\widehat{\theta}_n) = (9.8, 32.0)$. This suggests that cholesterol is higher among those with narrowed arteries. We should not jump to the conclusion (from these data) that cholesterol causes heart disease. The leap from statistical evidence to causation is very subtle and is discussed in Chapter 16. ■

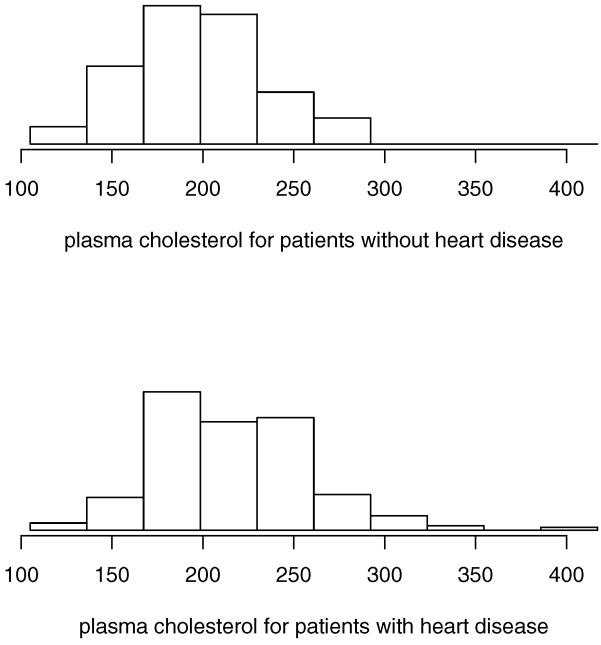

FIGURE 7.2. Plasma cholesterol for 51 patients with no heart disease and 320 patients with narrowing of the arteries.

7.3 Bibliographic Remarks

The Glivenko-Cantelli theorem is the tip of the iceberg. The theory of distribution functions is a special case of what are called empirical processes which underlie much of modern statistical theory. Some references on empirical processes are Shorack and Wellner (1986) and van der Vaart and Wellner (1996).

7.4 Exercises

1. Prove Theorem 7.3.

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 $p - q$.

3. (Computer Experiment.) Generate 100 observations from a N(0,1) distribution. Compute a 95 percent confidence band for the CDF F (as described in the appendix). Repeat this 1000 times and see how often the confidence band contains the true distribution function. Repeat using data from a Cauchy distribution.

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 $\widehat{F}_n(x)$.

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

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.

7. Data on the magnitudes of earthquakes near Fiji are available on the website for this book. Estimate the CDF $F(x)$. Compute and plot a 95 percent confidence envelope for F (as described in the appendix). Find an approximate 95 percent confidence interval for $F(4.9) - F(4.3)$.

8. Get the data on eruption times and waiting times between eruptions of the Old Faithful geyser from the website. Estimate the mean waiting time and give a standard error for the estimate. Also, give a 90 percent confidence interval for the mean waiting time. Now estimate the median waiting time. In the next chapter we will see how to get the standard error for the median.

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.

10. In 1975, an experiment was conducted to see if cloud seeding produced rainfall. 26 clouds were seeded with silver nitrate and 26 were not. The decision to seed or not was made at random. Get the data from

    http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html

    Let θ be the difference in the mean precipitation from the two groups. Estimate θ. Estimate the standard error of the estimate and produce a 95 percent confidence interval.