7 Estimating the Distribution Function and Statistical Functionals

The first inference problem we will consider is nonparametric estimation of the cdf F. Actually, we are rarely interested in estimating F but estimating the cdf is a useful first step towards estimating other quantities.

7.1 The Distribution Function

Let $X_1, \ldots, X_n \sim F$ where F is a distribution function on the real line. The *empirical distribution function* $\widehat{F}_n$ is the cdf that puts mass $1/n$ at each data point X_i. Formally, we define $\widehat{F}_n$ as follows. Let

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

Then

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I\{X_i \leq x\}}{n} = \frac{\text{number of observations less than or equal to } x}{n}.$$

We can also write $\widehat{F}_n(x) = n^{-1} \sum_{i=1}^n \delta_{X_i}(x)$ where $\delta_{X_i}(x)$ is a point mass at X_i.

We propose to use $\widehat{F}_n$ as an estimate of F.

EXAMPLE 7.1 *The first plot in Figure 7.1.1 shows the true cdf for a N(0,1). I generated 100 observations from a N(0,1). The empirical cdf is shown in the second plot. The third plot overlays the two cdf's. The R code to generate these plots is as follows.*

```
par(mfrow=c(2,2))
grid <- seq(-3,3,length=1000)
cdf <- pnorm(grid)
plot(grid,cdf,type="l",xlab="x",ylab="cdf",sub="True cdf")
n <- 100
x <- rnorm(n)
x <- sort(x)
cdf.hat <- (1:n)/n
plot(x,cdf.hat,type="s",xlab="x",ylab="cdf",sub="Empirical cdf",xlim=c(-3,3))
```

```
### type = "s"  means, draw a step-function
plot(grid,cdf,type="l",xlab="x",ylab="cdf")
lines(x,cdf.hat,lty=3,col=3,lwd=3,type="s")
```

The following two theorems give some properties of $\widehat{F}_n(x)$.

THEOREM 7.1 *At any fixed value of x,*

$$E\left(\widehat{F}_n(x)\right) = F(x) \quad \text{and} \quad \text{Var}\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n}.$$

Thus,

$$\text{MSE} = \frac{F(x)(1-F(x))}{n} \to 0$$

so $\widehat{F}_n(x) \stackrel{q.m.}{\to} F(x)$ and $\widehat{F}_n(x) \stackrel{p}{\to} F(x)$.

PROOF. Homework.

We now know that $\widehat{F}_n(x) \stackrel{p}{\to} F(x)$ at each x. But that doesn't imply that $\sup_x |\widehat{F}_n(x) - F(x)| \stackrel{p}{\to} 0$. The following theorem shows that this stronger convergence does indeed hold.

THEOREM 7.2 (Glivenko-Cantelli Theorem) [7] *Let $X_1, \ldots, X_n \sim F$. Then*

$$\sup_x |\widehat{F}_n(x) - F(x)| \stackrel{p}{\to} 0.$$

Remark. If you are unfamiliar with *sup*, just think of it as *max*.

The next theorem tells us how close $\widehat{F}_n$ is to $F(x)$.

THEOREM 7.3 (Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.) *Let $X_1, \ldots, X_n$ be iid from F. Then, for any $\epsilon > 0$,*

$$\text{P}(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}. \tag{3}$$

[7] Actually, we have a stronger convergence called almost sure convergence.

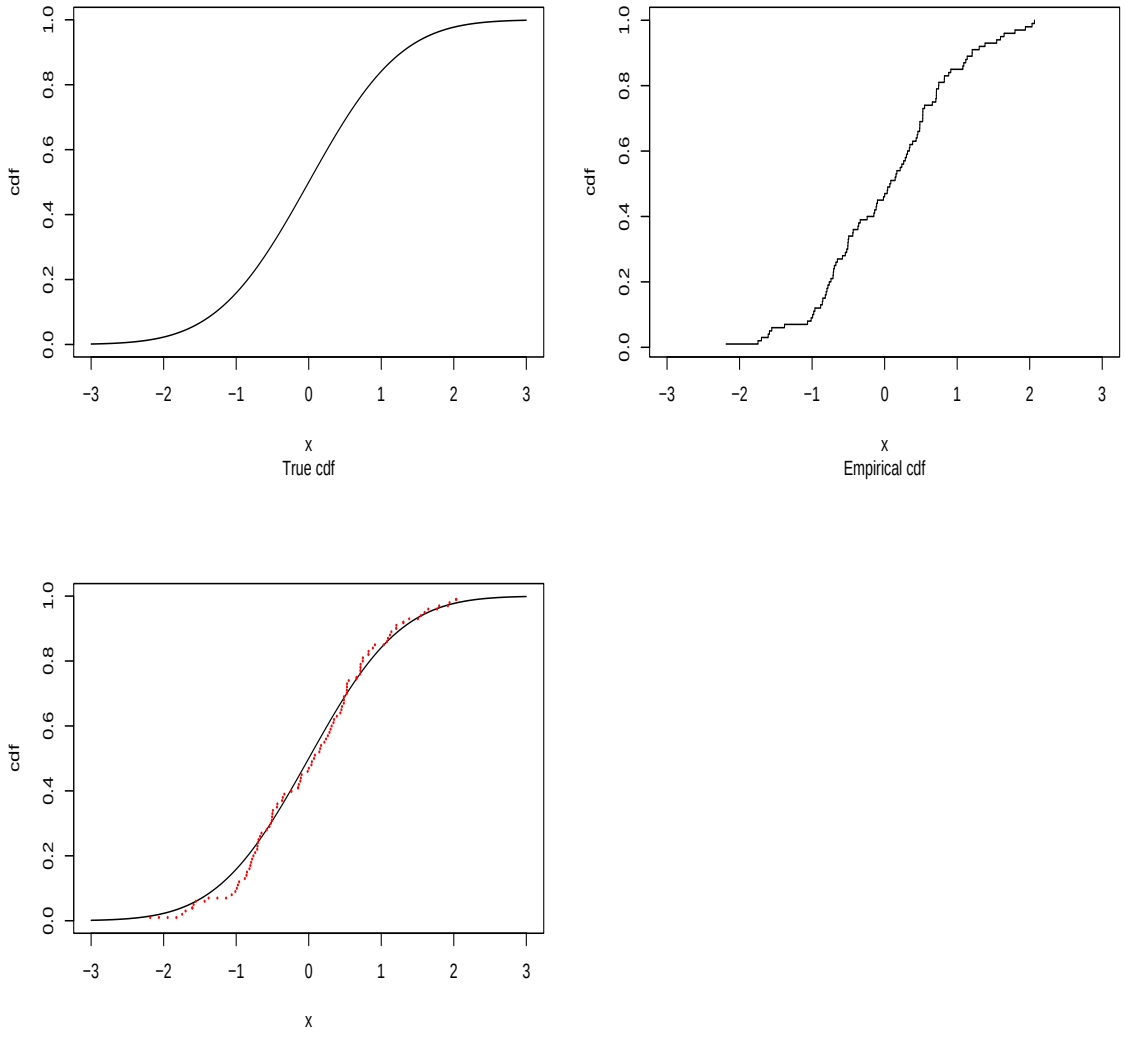

Figure 7.1.1. cdf of N(0,1) and empirical cdf from 100 observations.

From the DKW inequality, we can construct a confidence set. Let $\epsilon_n^2 = \log(2/\alpha)/(2n)$, $L(x) = \max\{\widehat{F}_n(x) - \epsilon_n, 0\}$ and $U(x) = \min\{\widehat{F}_n(x) + \epsilon_n, 1\}$. It follows from (3) that, no matter what the true F is,

$$P(F \in C_n) \geq 1 - \alpha$$

where $\mathcal{F}$ is the set of all distribution Thus, C_n is a nonparametric $1-\alpha$ confidence set for F. A better name for C_n is a *confidence band*. To summarize: a $1-\alpha$ nonparametric confidence interval for F is $(L(x),\ U(x))$ where

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

EXAMPLE 7.2 *Figure 7.1.2 shows the true cdf, the empirical cdf, and the 95 per cent confidence band using 100 observations from a N(0,1). The extra R code to compute the confidence band is:*

```
alpha <- .05
eps   <- sqrt(log(2/alpha)/(2*n))
l     <- pmax(cdf.hat - eps, 0)
u     <- pmin(cdf.hat + eps, 1)
### type help(pmin) or help(pmax) to see what these functions do
plot(grid,cdf,type="l",xlab="x",ylab="cdf")
lines(x,l,lty=2,col=2,type="s")
lines(x,u,lty=2,col=2,type="s")
```

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 $m = F^{-1}(1/2)$. We shall also refer to statistical functionals as *parameters* although that's an abuse of terminology. Another example of a functional is $\int r(x)dF(x)$ where $r(x)$ is a any function of x. The mean is of this form with $r(x) = x$. A functional of the form $\int r(x)dF(x)$ is called a linear functional.

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

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

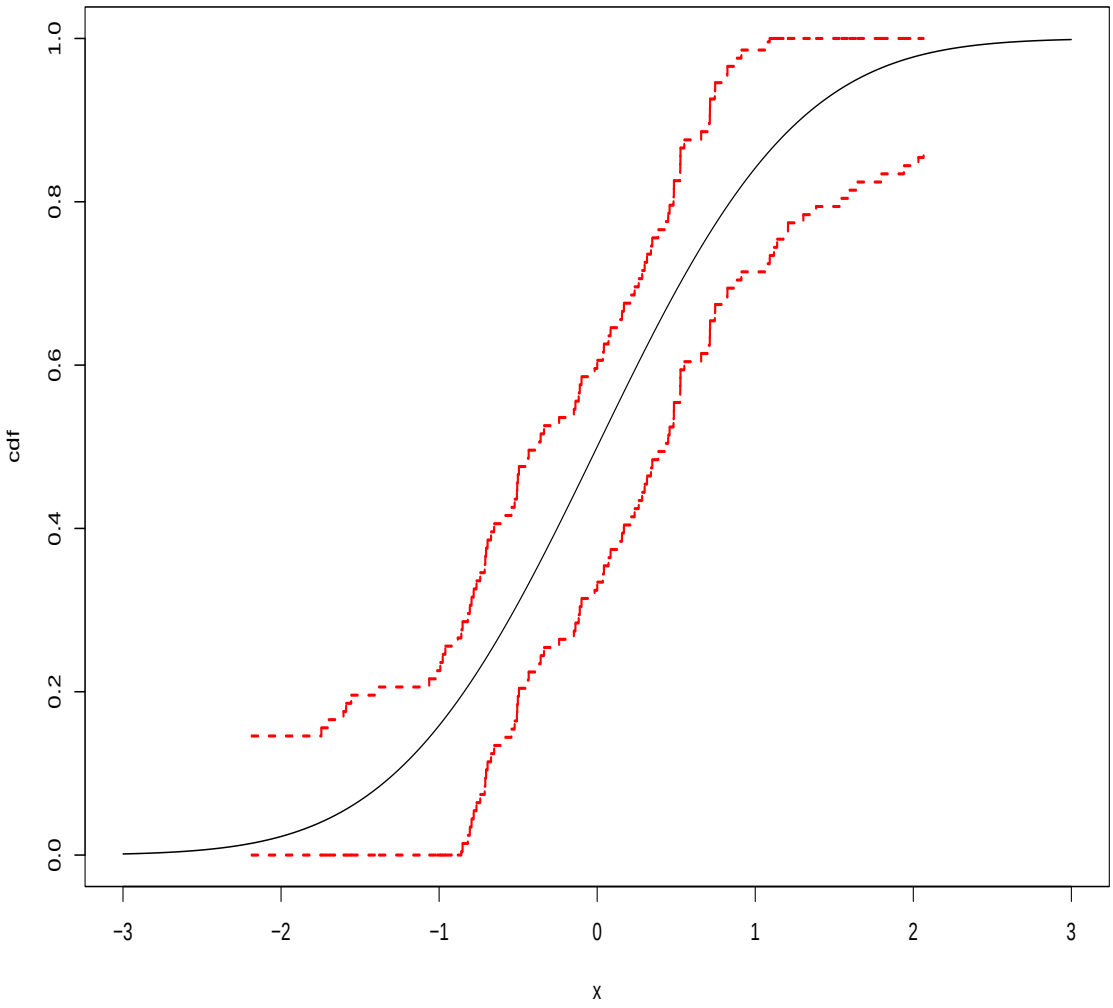

Figure 7.1.2. cdf of N(0,1) and confidence band from 100 observations.

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

Before proceeding, let me remind you of about some notation. Let $r(x)$ be a function of x. The quantity $\int r(x)dF(x)$ is to be interpreted as $\int r(x)f(x)dx$ in the continuous case and $\sum_j r(x_j)f(x_j)$ in the discrete. Now, the empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each X_i. Hence, $\int r(x)d\widehat{F}_n(x) = n^{-1}\sum_i r(X_i)$. So the plug-in estimator of $T(F) = \int r(x)dF(x)$ is $\int r(x)d\widehat{F}_n(x) = n^{-1}\sum_i r(X_i)$.

EXAMPLE 7.3 (The mean.) *Let $\mu = T(F) = \int x\,dF(x)$. The plug-in estimator is $\widehat{\mu} = \int x\,d\widehat{F}_n(x) = \overline{X}_n$. We can compute the standard error in this case:* $\text{se} = \sqrt{\text{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. *If $\widehat{\sigma}$ is 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 $\overline{x} \pm z_{\alpha/2}\,\text{se}^2$.*

EXAMPLE 7.4 (The Variance) *Let $\sigma^2 = T(F) = \text{Var}(X) = \int x^2 dF(x) - (\int x dF(x))^2$. The plug-in estimator is*

$$\begin{aligned}\widehat{\sigma}^2 &= \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.\end{aligned}$$

Some statistics texts use a different estimator, namely,

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

which is called the sample variance. *The reason they prefer this estimator is that it is unbiased, $E(S_n^2) = \sigma^2$. In practice, there is little difference between $\widehat{\sigma}^2$ and S_n^2 and we shall use both. Returning the our last example, we now see that the estimated standard error of the estimate of the mean is $\widehat{se} = \widehat{\sigma}/\sqrt{n}$.*

EXAMPLE 7.5 (The Skewness) *Let μ and σ^2 denote the mean and variance of a random variable X. The skewness is defined to be*

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

The skewness measure 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*

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

EXAMPLE 7.6 (Correlation.) *Let* $Z = (X, Y)$ *and let* $\rho = T(F) = E(X - \mu_X)(Y - \mu_Y)/(\sigma_x \sigma_y)$ *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*

$$\begin{array}{lll} T_1(F) = \int x\, dF(z) & T_2(F) = \int y\, dF(z) & T_3(F) = \int xy\, dF(z) \\ T_4(F) = \int x^2\, dF(z) & T_5(F) = \int y^2\, dF(z) & \end{array}$$

and

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\{(t_4 - t_1^2)(t_5 - t_2^2)\}^{1/2}}.$$

If you 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*

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

EXAMPLE 7.7 (Quantiles.) *Let* F *be strictly increasing with density* f. *The* $T(F) = F^{-1}(p)$ *be the* p^{th} *quantile. 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* $\widehat{F}_n^{-1}(p)$ *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 formulae for standard errors and confidence intervals. But in our nonparametric setting we need something else. In the next section, we will introduce two methods – the jackknife and the bootstrap – for getting standard errors and confidence intervals.

EXAMPLE 7.8 (The Mice Data.) *Below are data on the survival times in days of mice after surgery for a control group and treatment group.*

```
Treatment Group
94 197 16 38 99 141 23

Control Group
52 104 146 10 50 31 40 27 46
```

Let $\mu_X = \int x dF_X(x)$ *denote the mean survival time of treated mice and let* $\mu_Y = \int y dF_Y(y)$ *denote the mean survival time of untreated mice. The plug-in estimates are* $\widehat{\mu}_X = \int x d\widehat{F}_{X,n}(x) = \overline{X}_n = 86.86$ *and* $\widehat{\mu}_Y = \int y d\widehat{F}_{Y,n}(y) = \overline{Y}_n = 56.22$. *The standard error of* $\overline{X}_n$ *is* $se_X = se(\overline{X}_n) = \sqrt{Var(\overline{X}_n)} = \sigma_X/\sqrt{n} = \sigma_X/\sqrt{7}$ *which we estimate by* $\widehat{se}_X = s_X/\sqrt{n}$ *where* s_X^2 *is an estimate of* σ^2 *such as* $\sum_i (X_i - \overline{X})^2/n$ *or* $\sum_i (X_i - \overline{X})^2/(n-1)$. *Using the latter, we get* $\widehat{se}_X = 25.44$ *and* $\widehat{se}_Y = 14.14$. *The Normal-based 95 per cent confidence intervals are* $86.86 \pm 2(25.24) = (36.38, 137.34)$ *and* $56.22 \pm 2(14.14) = (27.94, 84.50)$.

However, we are probably more interested in the difference of the means $\delta = \mu_X - \mu_Y$. *The estimate of this is* $\widehat{\delta} = \overline{X} - \overline{Y} = 30.64$ *which suggests that the treated mice live, on average, about a month longer. The standard error of* $\widehat{\delta}$ *is*

$$\begin{aligned} se(\widehat{\delta}) &= \sqrt{Var(\overline{X} - \overline{Y})} \\ &= \sqrt{Var(\overline{X}) + Var(\overline{Y})} \\ &= \left\{ \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} \right\}^{1/2} \\ &= \left\{ se^2(\overline{X}) + se^2(\overline{Y}) \right\}^{1/2}. \end{aligned}$$

We estimate the standard error by

$$\begin{aligned} \widehat{se}(\widehat{\delta}) &= \left\{ \widehat{se}^2(\overline{X}) + \widehat{se}^2(\overline{Y}) \right\}^{1/2} \\ &= \left\{ 25.44^2 + 14.14^2 \right\}^{1/2} = 28.93. \end{aligned}$$

Finally, a 95 per cent confidence interval for δ *is* $30.64 \pm 2(28.93) = (-27.22, 88.50)$. *This confidence interval is huge. We have much uncertainty about* δ. *Looking only at the point estimate is very misleading.*