The Empirical Distribution Function and Plug-in Principle

Mingu Qiu

January 4, 2018

1 Introduction

- Problems of statistical inference often involve estimating some aspect of a probability distribution F on the basis of a random sample drawn from F.
- The *empirical distribution function*, which we will call $\hat{F}$, is a simple estimate of the entire distribution F.
- An obvious way to estimate some interesting aspect of F, like its mean or median or correlation, is to use the corresponding aspect of F. This is the *"plug-in principle."*

2 The Empirical Distribution Function

Definition 2.1. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d}}{\sim} F$. The *empirical distribution function* is definied as

$$\hat{F}(x) = \frac{\text{number of elements in the sample} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}},$$

where $\mathbb{1}$ is the indicator function.

In other words, the value of the empirical distribution function at a given point x is obtained by:

1. counting the number of observations that are less than or equal to x;
2. dividing the number thus obtained by the total number of observations, so as to obtain the proportion of observations that is less than or equal to x.

Example 2.1. Suppose we observe a sample made of 4 observations: $x_1 = 3, x_2 = 2, x_3 = 5, x_4 = 2$. What is the value of the empirical distribution function of the sample at the point $x = 4$?

According to the definition above, it is

$$
\begin{aligned}
\hat{F}(3) &= \frac{1}{4}\sum_{i=1}^{4} \mathbb{1}_{\{x_i \le 3\}} \\
&= \frac{1}{4}\Big(\mathbb{1}_{\{x_1 \le 3\}} + \mathbb{1}_{\{x_2 \le 3\}} + \mathbb{1}_{\{x_3 \le 3\}} + \mathbb{1}_{\{x_4 \le 3\}}\Big) \\
&= \frac{1}{4}(1 + 1 + 0 + 1) \\
&= \frac{3}{4}
\end{aligned}
$$

Note. It's simply the distribution function of a discrete random variable that places mass $1/n$ in the points $X_1, \cdots, X_n$ (provided all these are distinct). Namely, the p.m.f. of the empirical distribution is:

$$
\Pr(x) = \begin{cases} \frac{1}{n} & \text{if } x = x_{(1)}, \\ \frac{1}{n} & \text{if } x = x_{(2)}, \\ \vdots & \\ \frac{1}{n} & \text{if } x = x_{(n)}, \\ 0 & \text{othwewise.} \end{cases}
$$

where $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ are the sample observations ordered from the smallest to the largest. Then it is easy to see that the empirical distribution function can be written as

$$
\hat{F}(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ \frac{1}{n} & \text{if } x_{(1)} \le x < x_{(2)}, \\ \frac{2}{n} & \text{if } x_{(2)} \le x < x_{(3)}, \\ \vdots & \\ \frac{n-1}{n} & \text{if } x_{(n-1)} \le x < x_{(n)}, \\ 1 & \text{if } x \ge x_{(n)}. \end{cases}
$$

Example 2.2. The table below shows a random sample of $n = 100$ rools of a die: $x_1 = 6, x_2 = 3, \cdots, x_{100} = 6$. The empirical distribution function $\hat{F}$ put probability $1/100$ on each of the 100 outcomes. In cases like this, where there are repeated values, we can express $\hat{F}$ as the vector of observed frequencies $\hat{f}_k, k = 1, 2, \cdots, 6$,

$$
\hat{f}_k = \#\{x_i = k\}/n.
$$

So the empirical distribution is $(.13, .19, .10, .17, .14, .27)$.

6	3	2	4	6	6	6	5	3	6	2	2	6	2	3	1	5	1
6	6	4	1	5	3	6	6	4	1	4	2	5	6	6	5	5	3
6	2	6	6	1	4	1	5	6	1	6	3	3	2	2	2	5	2
2	4	1	4	5	6	6	6	2	2	4	6	1	2	2	2	5	1
5	3	5	4	2	1	4	6	6	5	6	4	6	4	3	6	4	1
4	5	4	4	2	3	2	1	4	6								

R provides the very useful function `ecdf()` for working with the empirical distribution function.

```
> data <- read.delim("die.txt")
> Fhat <- ecdf(data$outcome)
> Fhat(1)
[1] 0.13
> Fhat(6)
[1] 1
> summary(Fhat)
Empirical CDF:   6 unique values with summary
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    2.25    3.50    3.50    4.75    6.00
> plot(Fhat, verticals = T, do.points = F)
```

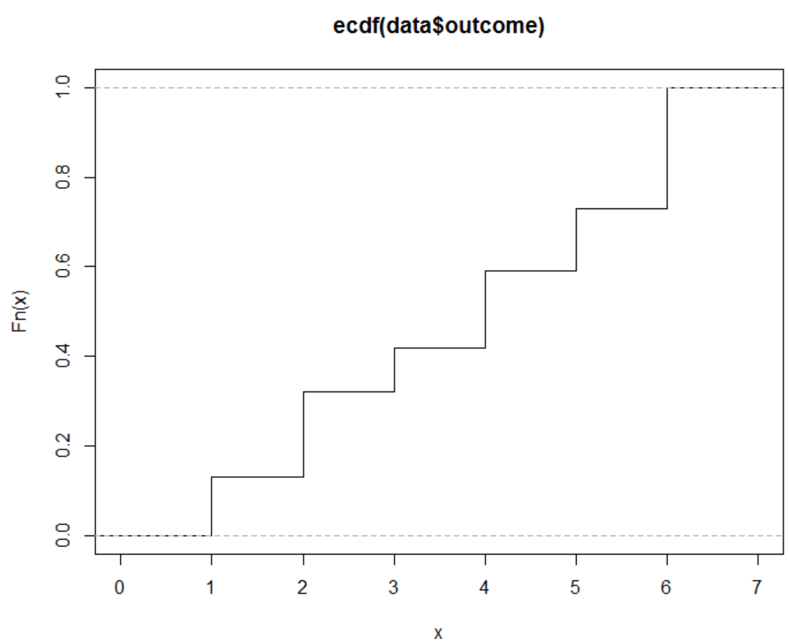

Figure 1: Empirical Distribution Function

Proposition 2.1. *For a fixed (but arbitrary) point $x \in \mathbb{R}$ we have that $n\hat{F}(x) \sim B(n, F(x))$ and*

$$\mathbb{E}(\hat{F}(x)) = F(x) \quad and \quad Var(\hat{F}(x)) = \frac{F(x)(1-F(x))}{n}.$$

Proof. Since $\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$ is the sum of n independent Bernoulli random variables with success probability $p = F(x)$, therefore

$$n\hat{F}(x) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \sim B(n, F(x)).$$

Hence

$$\mathbb{E}(\hat{F}(x)) = \frac{\mathbb{E}(n\hat{F}(x))}{n} = \frac{nF(x)}{n} = F(x)$$

and

$$Var(\hat{F}(x)) = \frac{Var(n\hat{F}(x))}{n^2} = \frac{nF(x)(1-F(x))}{n^2} = \frac{F(x)(1-F(x))}{n}.$$

□

This implies that:

Proposition 2.2. *For a fixed (but arbitrary) point $x \in \mathbb{R}$,*

a. $\hat{F}(x) \xrightarrow{\text{P}} F(x) \quad as \quad n \to \infty;$

b. $\hat{F}(x) \xrightarrow{\text{a.s.}} F(x) \quad as \quad n \to \infty;$

c. $\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{\text{d}} N(0, F(x)(1-F(x)) \quad as \quad n \to \infty.$

Proof. The first statement of the proposition follows simply by Chebyshev's inequality: for any $\epsilon > 0$

$$\Pr\{|\hat{F}(x) - F(x)| \geq \epsilon\} \leq \frac{F(x)(1-F(x))}{n\epsilon^2}.$$

The second statement follows by the strong law of large numbers. Since $\mathbb{1}_{\{X_i \leq x\}}$ is Bernoulli random variable with success probability p=F(x), then

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \xrightarrow{\text{a.s.}} \mu = F(x).$$

where n is the sample size.

The last statement follows by the central limit theorem. □

Note. The above results were all about *pointwise* convergence. That is, we examined what happens to $\hat{F}(x)$ for a fixed point $x \in \mathbb{R}$.

There is a stronger result than, called the Glivenko-Cantelli theorem, which states that the convergence in fact happens *uniformly* over $\mathbb{R}$:

Theorem 2.1 (Glivenko-Cantelli Theorem)**.** [1] *The empirical distribution function $\hat{F}(x)$ converges uniformly to $F(x)$, namely*

$$\sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x)| \xrightarrow{\text{a.s.}} 0,$$

as $n \to \infty$.

3 Parameter and Statistic

- A *parameter* is a function of the probability distribution F.
- A *statistic* is a function of the sample x.

Thus f_k is a parameter of F in the die example, while $\hat{f}_k$ is a statistic, $k = 1, 2, \cdots, 6$.

We will sometimes write parameters directly as fnctions of F as follow :

$$\theta = t(F).$$

For example, if F is a probability diatribution in the real line, the expectation can be thought of as the parameter

$$\theta = t(F) = E_F(x).$$

For a given distribution F such as $B(n, p)$, we can evaluate $E_F(x) = t(F) = np$.

4 Plug-in Principle

The plug-in principle is a simple method of estimating parameters from samples.

The plug-in estimate of a parameter $\theta = t(F)$ is defined to be

$$\hat{\theta} = t(\hat{F}),$$

obtained by replacing the distribution function F with the empirical distribution function $\hat{F}$.

Example 4.1 (the mean)**.** Let $\mu = \mathbb{E}_F(X) = \sum_{i=1}^{n} x_i p(x_i)$ be the mean of the distribution F. Then the plug-in estimator of μ is

$$\hat{\mu} = E_{\hat{F}}(X) = \sum_{i=1}^{n} X_i \hat{p}(X_i) = \sum_{i=1}^{n} X_i \frac{1}{n} = \bar{X},$$

where $p(X)$ is the pmf of F and $\hat{p}(X_i) = 1/n, i = 1, 2, \cdots, n$.

[1] This theorem originates with Valery Glivenko and Francesco Cantelli in 1933.

Example 4.2 (the variance)**.** Let $\sigma^2 = Var_F(X) = \mathbb{E}_F(X^2) - (\mathbb{E}_F(X))^2$ denote the variance of X. The plug-in estimator for σ^2 is

$$\hat{\sigma}^2 = Var_{\hat{F}}(X) = \mathbb{E}_{\hat{F}}(X^2) - (\mathbb{E}_{\hat{F}}(X))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example 4.3 (the median)**.** Define $F^{-1}(y) = \inf\{x : F(x) \geq y\}$ and $F^{-1}(y+) = \inf\{x : F(x) > y\}$.Let $\theta = F^{-1}(1/2)$. The median of distribution F can be denoted by

$$\theta = \frac{F^{-1}(1/2) + F^{-1}(1/2+)}{2},$$

then the plug-in estimator of the median is

$$\hat{\theta} = \frac{\hat{F}^{-1}(1/2) + \hat{F}^{-1}(1/2+)}{2} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)} & \text{if } n \text{ is even} \end{cases}.$$

Example 4.4. The law school population F can be written as $F = (f_1, f_2, \cdots, f_{82})$. The population correation coefficient can be written as

$$corr(y, z) = \frac{\sum_{j=1}^{82} f_j(Y_j - \mu_y)(Z_j - \mu_z)}{\left[\sum_{j=1}^{82} f_j(Y_j - \mu_y)^2 \sum_{j=1}^{82} f_j(Z_j - \mu_z)^2\right]^{1/2}} \tag{1}$$

where

$$\mu_y = \sum_{j=1}^{82} f_j Y_j, \mu_z = \sum_{j=1}^{82} f_j Z_j. \tag{2}$$

Now for the sample of 1, $\hat{f}_1 = 0, \hat{f}_2 = 0, \hat{f}_3 = 0, \hat{f}_4 = 1/15$ etc. Plugging these values $\hat{f}_j$ into (1) and (2) gives $\hat{\mu}_y$, $\hat{\mu}_z$ and $\hat{corr}(y, z)$ respectively. That is, $\hat{\mu}_y$,$\hat{\mu}_z$ and $\hat{corr}(y, z)$ are *plug -in* estimates of μ_y, μ_z and $corr(y, z)$.

5 How good is the plug-in principle?

- It is usually quite good, if the only available information about F comes from the sample x.
- However, the plug-in principle is less good in situations where there is information about F other than provided by the sample x.

A Appendix

A1. Chebyshev's inequality

Let X be a random variable and c be a positive constant, then

$$\Pr\{|X - \mu| \geq c\sigma\} \leq \frac{1}{c^2},$$

where $\mu = \mathbb{E}(X)$ and $\sigma^2 = Var(X)$.

School	LAST	GPA	School	LAST	GPA
1	576	3.39	9	651	3.36
2	635	3.30	10	605	3.13
3	558	2.81	11	653	3.12
4	578	3.03	12	575	2.74
5	666	3.44	13	545	2.76
6	580	3.07	14	572	2.88
7	555	3.00	15	594	2.96
8	661	3.43			

Table 1: *The law school data. A random sample of size n=15 was taken from the collection of N=82 American law schools participating in a large study of admission practices. Two measurements were made on the entering classes of each school in 1973:LAST, the average score for the class on a national law test, and GPA, the average undergraduate grade-point average for the class.*

A2. The strong law of large numbers

Assume that $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d random variables with $\mathbb{E}(X_n) = \mu \leq \infty$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, then

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

A3. The central limit theorem

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with common mean μ and common variance $\sigma^2 > 0$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ and $Y_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, then

$$Y_n \xrightarrow{d} Z,$$

where $Z \sim N(0, 1)$.

A4. Converge in Probbility

A sequence of random variables $\{X\}_{n=1}^{\infty}$ is said to converge in probability to a random variable X, denote by $X_n \xrightarrow{\mathrm{P}} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|X_n - X| \geq \epsilon\} = 0.$$