Estimation of Cumulative Distribution Function

Snehal M. Shekatkar

Department of Scientific Computing, Modeling, and Simulation,
Savitribai Phule Pune University, Pune, India 411007

1 Estimation of CDF

The problem of point estimation and corresponding confidence sets can be applied not just to a few quantities like mean and variance but also to the probability distributions themselves! In other words, instead of asking what is the best estimate of the mean of the underlying distribution, we can even ask what is the best estimate of the shape of the underlying distribution. This is useful in many circumstances, and in fact we have already learned one method of this estimation when we discussed constructing a good histogram from a given data although we did not think in terms of inference then.

Here we would like to proceed in a more principled manner and the main aim of this type of inference won't be visualization. Instead, the emphasis will be on estimating the errors in the estimation or how accurately we could guess the shape of the distribution based on a sample.

When we talked about histogram we were concerned about the shape of the probability density function (for continuous case) or that of probability mass function (for discrete case). However, because these quantities use only local information (e.g. the number of sample points in a given range), they turn out to have certain disadvantages. For example, when the sample size is small, just by chance we may see too many or too few sample points in a particular bin even when the underlying distribution is uniform. That is, histogram is sensitive to the sample size because of noise. Moreover, the estimation depends on the bin-width and hence can't be said to be objective given the data. Hence, the preferred quantity for estimation is the cumulative distribution function. Nevertheless, keep in mind that there is nothing fundamentally wrong with estimating the probability density or mass function, and it is in fact many times used when n is large.

As you know, the *Cumulative Distribution Function* or CDF of a random variable X tells us the probability that X would take value less than or equal to a specific value x:

$$F(x) = \mathbb{P}(X \leq x) \tag{1}$$

Whether X is discrete or continuous, the CDF is given by the above formula only. This is another advantage of talking in terms of CDF.

The **Empirical Distribution Function** or EDF is an estimator of the cumulative distribution function for the IID sample $X_1, X_2, \cdots, X_n$, and is defined as follows:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{2}$$

where

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

The function $I(X_i \leq x)$ is also called as **Indicator function** since it indicates whether X_i is less than a given value or not. We sometimes say that $\widehat{F}_n$ *puts* the mass $1/n$ at each X_i. Here we are basically saying that the probability that X takes value less than or equal to x is well approximated by the fraction of sample points less than or equal to x. This completely avoids the issue of bin sizes and also reduces noise because positive fluctuation in one place tends to get compensated by negative fluctuation in another place (both below x).

At any point x, we can think of $\widehat{F}_n$ as a point estimate of true F. We can then naturally ask whether this estimator possesses properties that any good estimator should possess i.e. whether it is unbiased and consistent. When x is fixed, the probabilities for choice $X_i \leq x$ or $X_i > x$ are governed by the indicator function which is simply a Bernoulli random variable with probabilities $F(x)$ and $1 - F(x)$ respectively. Hence, the EDF $\widehat{F}_n$ is just a sample mean of Bernoulli RVs. We know that for a Bernoulli random variable X with success probability p, the expectation and variance are given by:

$$\mathbb{E}(X) = p \quad \text{and} \quad \mathbb{V}(X) = p(1-p)$$

Hence we have:

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

$$\mathbb{V}(\widehat{F}_n(x)) = \sum_{i=1}^{n}\mathbb{V}\left(\frac{1}{n}I(X_i \leq x)\right) = \sum_{i=1}^{n}\frac{1}{n^2}(F(x)(1-F(x))) = \frac{1}{n}F(x)(1-F(x))$$

From these, it is trivial to verify that $\widehat{F}_n$ is a consistent estimator of F. That is,

$$\widehat{F}_n(x) \xrightarrow{\text{P}} F(x)$$

Now given this, we can in principle construct $1-\alpha$ confidence interval for $F(x)$ at any value of x just like we did in the case of μ. We can actually imagine constructing $1-\alpha$ confidence interval for each value of x. Then all these confidence intervals would form a band around $\widehat{F}_n(x)$ (Here we are talking about all values of x, not one fixed value). This band is another example of a confidence set and is known as **confidence band** for $\widehat{F}_n$.

However, constructing a confidence band in this fashion is not that straightforward because it needs the knowledge of the variance of $\widehat{F}_n(x)$ which itself is a function of $F(x)$. The situation is similar to the one we encountered during the estimation of $1-\alpha$ confidence interval for μ, and there we had to substitute the estimate of the variance in place of actual variance. Here too we can do the same by using $\widehat{F}_n(x)$ in place of $F(x)$ in the formula for $\mathbb{V}(x)$. This is probably fine as far as the sample size n is large. However, just like the case of confidence interval for μ, when n is small we will need to worry about the actual coverage of the confidence band constructed this way. Turns out that there is a much straightforward solution to this problem if we take into account a theorem called **DKW inequality** given below.

Let $X_1, X_2, \cdots, X_n$ be a IID sample with cumulative distribution function F. Then, given $\epsilon > 0$,

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

In a crude sense this is telling us that the probability that the maximum difference between the true $F(x)$ and its estimate $\widehat{F}_n(x)$ (for any value x) is greater than some specified positive value ϵ is not larger than $2e^{-2n\epsilon^2}$. For $1-\alpha$ confidence band, we want this probability at most to equal α, the level of significance. Thus, equating the two, we get the value of ϵ that satisfies our requirement as:

$$\epsilon_n = \sqrt{\frac{1}{2n}\log_e\left(\frac{2}{\alpha}\right)} \tag{3}$$

Note that this is the maximum value when we take into consideration all possible x values. Hence for some values the coverage of the band may even be greater than $1-\alpha$. Thus sometimes subtracting ϵ_n from $\widehat{F}_n(x)$ may result into negative value or adding ϵ_n to $\widehat{F}_n(x)$ may result into value greater than 1. Since $F(x)$ is bounded between $[0,1]$ for all x, we need to fix it. Thus we define a nonparametric $1-\alpha$ confidence band for $F(x)$ as follows.
Let,

$$L(x) = \max\{\widehat{F}_n(x) - \epsilon_n, 0\}$$

$$U(x) = \min\{\widehat{F}_n(x) + \epsilon_n, 1\}$$

where ϵ_n is as given in (3). Then for all x,

$$\mathbb{P}\left(L(x) \leq F(x) \leq U(x)\right) \geq 1-\alpha \tag{4}$$

Fig. 1 shows the Empirical Distribution Function and the corresponding 95% confidence band for the *sepal widths* from the famous *iris* dataset collected by Ronald Fisher.

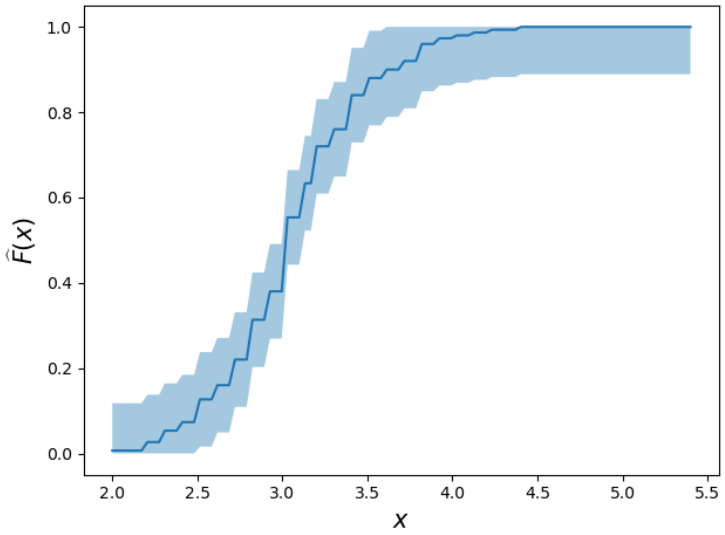

Figure 1: Empirical Distribution Function (EDF) and the corresponding 95% confidence band for the *sepal widths* from the classic *iris* dataset.

2 Statistical Functionals and Plug-in Estimators

A statistical functional is a function that assigns a given Cumulative distribution function a real number. Probably the simplest example is the mean of the distribution: for a given CDF $F(x)$, a single value of the mean exists. Another obvious example is the variance. So far we have seen that an estimator is a quantity that is constructed to estimate such statistical functionals from a given sample. Although we have tried to decide how good a given estimator is, using quantities such as bias and consistency, we have not seen any method to construct such estimators in the first place. Here we give first such method.

A statistical functional T is *linear* if there exists a function $\eta(x)$ such that:

$$T(F) = \int \eta(x)p(x)dx$$

where $p(x)$ is the probability density function corresponding to the CDF $F(x)$ (As always I am assuming that we have continuous random variables. You can write a similar equation for the discrete case). The

reason for calling this a linear functional is that if we have two CDFs F and G, and constants A and B, then:

$$T(aF + bG) = \int \eta(x)(ap(x) + bq(x))dx = a\int \eta(x)p(x)dx + b\int \eta(x)q(x)dx = aT(F) + bT(G)$$

Here we have assumed that $q(x)$ is the probability density function corresponding to the CDF $G(x)$. Now suppose we don't know the true CDF but would like to estimate a functional $T(F) = \int \eta(x)p(x)dx$ (which is linear as shown above). We can achieve this by "plugging-in" $\widehat{F}_n$ in place of F. This principle is known as the **plug-in principle**, and the estimator $T(\widehat{F}_n)$ obtained in this fashion is called a **plug-in estimator**.

How do we actually replace F by $\widehat{F}_n$ in $T(F)$? Observe that to compute $T(F)$, we use the corresponding density $p(x)$. Now consider a totally different random variable Y whose CDF is **exactly** $\widehat{F}_n$. Obviously, Y is a discrete random variable independent of whether X_i are discrete or continuous. Moreover, we also know the PMF for Y exactly! This PMF is given below:

$$\mathbb{P}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = y)$$

Since we know the PMF exactly, we can calculate $T(\widehat{F})$ as:

$$T(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} \eta(X_i) \tag{5}$$

This gives the plug-in estimator for the functional T. For example, for the mean, $\eta(x) = x$. Then the plug-in estimator for mean is:

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

which is just the sample mean! Let's try to find out a plug-in estimator for the variance. We know that:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \int x^2 p(x)dx - \left(\int xp(x)dx\right)^2$$

Using plug-in estimators for individual terms, we get:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n}\left(X_i \frac{1}{n}\sum_{j=1}^{n} X_j\right)$$

$$\therefore \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + X_i\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}\right) + \overline{X}\frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\therefore \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}\right) + \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)$$

Hence finally,

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

Recall that we have already seen this estimator, and have also proved that it is biased. We have also seen that it can be made unbiased by replacing n by $n-1$. However, this is the first time we have seen a construction of this estimator using some fundamental argument (in this case, the *plug-in principle*).