The Empirical Distribution Function and the Histogram

Rui Castro[*]

1 The Empirical Distribution Function

We begin with the definition of the empirical distribution function.

Definition 1. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, with distribution function $F(x) = \mathbb{P}(X_1 \leq x)$. The Empirical Cumulative Distribution Function (ECDF), also known simply as the empirical distribution function, is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} \ ,$$

where $\mathbf{1}$ is the indicator function, namely $\mathbf{1}\{X_i \leq x\}$ is one if $X_i \leq x$ and zero otherwise.

Note that this is simply the distribution function of a discrete random variable that places mass $1/n$ in the points $X_1, \ldots, X_n$ (provided all these are distinct). Note also that one can also define the empirical probability of any Borel-measurable set B as $\mathbb{P}_n(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in B\}$.

Note that the order of the samples is not important in the computation of F_n. For that reason it is useful to define the *order statistics.*

Definition 2. Let $X_1, \ldots, X_n$ be set of random variables. Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation operator such that $X_{\pi(i)} \leq X_{\pi(j)}$ if $i < j$. We define the order statistics as $X_{(i)} = X_{\pi(i)}$. Therefore

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \ .$$

Note that the order statistics are just a reordering of the data. Two particularly interesting order statistics are the minimum and the maximum, namely $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$. The ECDF can be obviously written as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_{(i)} \leq x\} \ .$$

It is easy to see that this stair-function, with jumps of height $1/n$ at the points $X_{(i)}$. It is therefore increasing, but also right-continuous and takes values on $[0, 1]$.

Clearly the ECDF seems to be a sensible estimator for the underlying distribution function. Next we try to understand a bit better the properties of this estimator.

[*] This set of notes was adapted from notes by Eduard Belitser.

Proposition 1. *For a fixed (but arbitrary) point $x \in \mathbb{R}$ we have that $nF_n(x)$ has a binomial distribution with parameters n and success probability $F(x)$. Therefore*

$$\mathbb{E}\left(F_n(x)\right) = F(x) \quad \textit{and} \quad \mathrm{var}(F_n(x)) = \frac{F(x)(1-F(x))}{n} \ .$$

This implies that $F_n(x)$ converges in probability to $F(x)$ as $n \to \infty$.

Proof. Clearly $\sum_{k=1}^{n} \mathbf{1}\{X_k \leq x\}$ is the sum of n independent Bernoulli random variables (with success probability $F(x)$), therefore $nF_n(x)$ is a binomial random variable. Therefore the mean and variance characterization in the proposition follows easily. The second statement of the proposition follows simply by Chebyshev's inequality: for any $\epsilon > 0$

$$\mathbb{P}(|F_n(x) - F(x)| \geq \epsilon) \leq \frac{F(x)(1-F(x))}{n\epsilon^2} \ .$$

□

Note that stronger statements about $F_n(x)$ can be made, namely the strong law of large numbers applies, meaning that $\hat{F}_n(x)$ converges to $F(x)$ almost surely, for every fixed $x \in \mathbb{R}$.

Chebyshev's inequality is rather loose, and although it can be used to get an idea on how fast $F_n(x)$ converges to x, it is far from tight. In particular in light of the central limit theorem $\mathbb{P}(|F_n(x)-F(x)| \geq \epsilon)$ should scale roughly like $e^{-n\epsilon^2}$ instead of $1/(n\epsilon^2)$. A stronger concentration of measure inequality that can be applied in this setting is Hoeffding's inequality, which implies that for any $\epsilon > 0$.

$$\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \ .$$

Note that the above results were all about *pointwise* convergence. That is, we examined what happens to $F_n(x)$ for a *fixed point x*. But do we have convergence simultaneously for *any* point $x \in \mathbb{R}$? The answer is affirmative, and formalized in the following important result.

Theorem 1 (Glivenko-Cantelli Lemma)**.** *The empirical distribution converges uniformly to $F(x)$, namely*

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \overset{a.s.}{\to} 0 \ ,$$

as $n \to \infty$, where the superscript a.s. denotes convergence almost surely.

The proof is given in the appendix. Note that this result tells us about the convergence, but nothing about the speed of convergence (unlike Hoeffding's inequality). However, a similar result exists.

Theorem 2 (Dvoretzky-Kiefer-Wolfowith (DKW) inequality)**.** *For any $\epsilon > 0$ and any $n > 0$*

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \ .$$

A similar result was first proved in the fifties, but with a leading constant that was way bigger than 2. It was only in the nineties that the result as stated here was shown (which is also the best one can hope for). The proof of this result is rather involved, but you can get a slightly weaker result by using the ideas in the proof of the Glivenko-Cantelli lemma and Hoeffding's inequality (see the homework exercises).

2 Using the ECDF

Most of the times, we are not necessarily interested in the distribution function F, but rather on some functions of the distribution, for instance the mean, variance or median. In most cases we can come up with estimators for these quantities by simply replacing F by F_n, in what is known as a *plug-in* estimate.

Example 1 (the mean)**.** Let $\mu = \mathbb{E}(X_1) = \int x dF(x)$ be the mean of the the distribution F (assume $\mathbb{E}(X_1)$ exists). A natural estimator for μ is simply $\hat{\mu}_n = \int x dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$, which is known as the sample mean.

More generally,

$$\hat{\theta} = \int g(x) dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

is an estimator of $\theta = \int g(x) dF(x)$. It is unbiased and strongly consistent estimator, with variance

$$\frac{1}{n}(\int g(x)^2 dF(x)) - (\int g(x) dF(x))^2 \ .$$

Not all quantities we might want to estimate can be written exactly as in the above example.

Example 2 (the variance)**.** Let $\sigma^2 = \text{var}(X_1) = \int x^2 dF(x) - (\int x dF(x))^2$ denote the variance of X_1. A natural estimator for σ^2 is

$$\hat{\sigma}_n^2 = \int x^2 dF_n(x) - \left(\int x dF_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \ .$$

This is not an unbiased estimator, but it is asymptotically unbiased and strongly consistent.

Example 3 (the median)**.** Define $F^{-1}(y) = \inf\{x : \ F(x) \geq y\}$ and $F^{-1}(y+) = \inf\{x : \ F(x) > y\}$. Let $\theta = F^{-1}(1/2)$. If F is continuous and strictly increasing in the neighborhood of θ then θ is precisely the median of F. A natural estimator for θ is simply $F_n^{-1}(1/2)$. Since F_n is not continuous this is not necessarily the median of F_n. A slightly more convenient (from a practical point of view) way to estimate the mean is given by

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

This is called the sample median, and it is again a consistent estimator of the median.

3 Histogram

Clearly the empirical distribution function is a very powerful object, but it has limitations. Suppose the random sample $X_1, \ldots, X_n$ comes from a distribution with density $f(\cdot)$ (with respect to the Lebesgue measure). We would like to estimate this density. Recall that the density of a continuous random variable with distribution $F(x)$ is given simply by $f(x) = \frac{\partial}{\partial x}F(x) = F'(x)$. However, if we try to derive F_n we ran into serious trouble, as F_n is not continuous (it is a staircase function, with flat plateaus).

To overcome this difficulty there are a couple of strategies. Since F_n will approach F as n increases, and F is assumed to be continuous, we might try to approximate the derivative of F_n instead. Here we will take a slightly different approach but, as we will see, this is essentially implementing such an idea.

Let $\ldots < r_{-1} < r_0 < r_1 < \ldots$ be such that $r_i \in \mathbb{R} \cup \{-\infty, \infty\}$ and that these points define a partition of the real line. Define $\nu_k = \nu(r_k, r_{k+1})$ the number of samples that fall inside the interval $(r_k, r_{k+1}]$, formally $\nu_k = \sum_{i=1}^n \mathbf{1}\{X_i \in (r_k, r_{k+1}]\}$.

Definition 3. The histogram is defined as

$$\hat{f}_n(x) = \frac{\nu_k}{n(r_{k+1} - r_k)} = \frac{1}{n(r_{k+1} - r_k)} \sum_{i=1}^n \mathbf{1}\{X_i \in (r_k, r_{k+1}]\} \ .$$

for $x \in (r_k, r_{k+1}]$.

This should be rather familiar to you from your first courses in statistics. This function is a stair function, with possibly discontinuities at the points $\{r_k\}$. Each cell in the histogram has width $b_k = r_{k+1} - r_k$ and height$\frac{\nu_k}{nb_k}$. It is easy to see that this function is always non negative, and the area between the function and the x-axis is exactly one. Therefore $\hat{f}_n(x)$ is a valid probability density function.

It seems believable that the histogram is, in some sense, and estimator for f, the density of X_i. Obviously the quality of this estimator is going to depend on the choice of partition $\{r_k\}$. The analysis below can be made more general, but to keep things simple lets consider the case where all the cells in the partition are the same size. Namely let assume $r_{k+1} - r_k = h$ for all $k \in \mathbb{N}$. Therefore the partition in completely defined by h and the point r_0. Define $\lfloor x \rfloor = \sup\{k \in \mathbb{Z} : x > k\}$, that is, the greatest integer strictly less than x. The histogram can be written as

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}\{X_i - r_0 \in (hk_x, h(k_x + 1)]\} \ ,$$

where $k_x = \lfloor (x - r_0)/h \rfloor$.

Remark 1. Note that we could have written the histogram as a function of the empirical distribution function, namely

$$\hat{f}_n(x) = \frac{F_n(r_{k+1}) - F_n(r_k)}{h}$$

for $x \in (r_k, r_{k+1}]$. This clearly looks like an approximation of the derivative of $F_n(x)$.

We would like to understand how good (or bad) is the histogram estimator, and what is the correct way of choosing h. Let us consider the following way of measuring the error.

Definition 4.

$$\text{MSE}(\hat{f}_n(x)) = \mathbb{E}((\hat{f}_n(x) - f(x))^2) \ .$$

As you know this way of measuring error has many advantages, in particular we have the convenient decomposition of the MSE in terms of a bias component and a variance component.

Lemma 1 (Bias-Variance Decomposition).

$$\mathit{MSE}(\hat{f}_n(x)) = \mathit{Bias}^2(\hat{f}_n(x)) \ + \ \text{var}(\hat{f}_n(x)) \ .$$

The MSE analysis of the histogram is rather simple, and illustrative of the techniques that frequently appear in non-parametric statistics. A key point to note is that $\nu_k \sim \text{Bin}(n, p_k)$ with $p_k = P(r_k < X_1 \leq r_k + h) = \int_{r_k}^{r_k+h} f(t)dt$. Therefore $\mathbb{E}(\hat{f}_n(x)) = p_k/h$ and $\text{var}(\hat{f}_n(x)) = p_k(1-p_k)/(nh^2)$.

In order to have a non-asymptotic analysis let's make some assumptions about the unknown density f. We begin with the following definition.

Definition 5. A function g is Lipschitz continuous in interval B if there is a constant $L > 0$ such that

$$|g(x) - g(y)| \leq L|x - y| \ , \qquad \text{for all } x, y \in B \ .$$

Lemma 2. *If f is Lipschitz continuous (with Lipschitz constant L) in $\mathbb{R}$ then*

$$\left|\mathbb{E}(\hat{f}_n(x)) - f(x)\right| \leq Lh \ .$$

Proof. Let k be such that $x \in (r_0 + hk, r_0 + h(k+1)]$.

$$\begin{aligned}
\left|\mathbb{E}(\hat{f}_n(x)) - f(x)\right| &= |p_k/h - f(x)| \\
&= \left|\frac{1}{h}\int_{r_0+hk}^{r_0+h(k+1)} f(t)dt - f(x)\right| \\
&= \left|\frac{1}{h}\int_{r_0+hk}^{r_0+h(k+1)} (f(t) - f(x))dt\right| \\
&\leq \frac{1}{h}\int_{r_0+hk}^{r_0+h(k+1)} |f(t) - f(x)|\, dt \\
&\leq \frac{1}{h}\int_{r_0+hk}^{r_0+h(k+1)} Lhdt \\
&= Lh \ .
\end{aligned}$$

□

This small lemma tell us that the bias of the histogram vanishes as $n \to \infty$ if we take $h \equiv h_n \to 0$. That is

$$\text{Bias}(\hat{f}_n(x)) = \mathbb{E}(\hat{f}_n(x)) - f(x) \to 0 \ ,$$

as $n \to \infty$. We can likewise study the variance of $\hat{f}_n(x)$.

Lemma 3. *Under the assumptions of Lemma 2*

$$\text{var}(\hat{f}_n(x)) \leq \frac{|f(x)|}{nh} + \frac{L}{n} \ .$$

Proof. Simply note that

$$\text{var}(\hat{f}_n(x)) = \frac{p_k(1-p_k)}{nh^2} \leq \frac{p_k}{nh^2} \leq \frac{h(|f(x)| + Lh)}{nh^2} \ ,$$

where the last inequality follows from the proof of Lemma 2 □

With these two results in hand we can evaluate the Mean Square Error (MSE) of the histogram estimator,

$$\text{MSE}(\hat{f}_n(x)) \leq (Lh)^2 + \frac{|f(x)|}{nh} + \frac{L}{n} \ .$$

So it is clear that we want to take h small to make the bias small, but also large enough to ensure the variance is also small. Clearly, if we want to minimize the MSE with respect to h we should take

$$h_n = \left(\frac{|f(x)|}{2L^2 n}\right)^{1/3} \ .$$

This will ensure that

$$\text{MSE}(\hat{f}_n(x)) = O(n^{-2/3})$$

as $n \to \infty$. Actually, the previous statement can be obtained if we take simply $h_n = cn^{-1/3}$, where $c > 0$. It turns out that this is (apart from constants) the best we can ever hope for if we only assume the density if Lipschitz! We'll see why this is the case later in the course.

It is worth mentioning that you can also get results without the Lipschitz assumption, but these are more asymptotic in nature

Proposition 2. *Suppose f is finite and continuous at point x. Then MSE$(\hat{f}_n(x)) \to 0$ as $n \to \infty$ provided both the histogram cell length h_n converges to zero and nh_n converges to infinity. This implies that*

$$\hat{f}_n(x) \xrightarrow{P} f(x) \ ,$$

as $n \to \infty$.

Proof sketch. The continuity assumption, together with the mean-value theorem for integrals allows you to conclude that $\mathbb{E}(\hat{f}_n(x)) \to f(x)$ as $n \to \infty$, provided $h_n \to 0$. On the other hand you can easily show that $\text{var}(\hat{f}_n(x)) \leq K/(nh_n)$ for some $K > 0$ and n large enough. Putting these facts together you conclude the result of the proposition. □

4 Homework Exercises

1. Prove Proposition 2, by formalizing all the steps in the sketch above.

2. In the analysis of the histogram above we considered only the behavior of the estimator at a point. However, we can consider a more global metric of performance, namely the Mean Integrated Squared Error (MISE) defined as

$$\text{MISE}(\hat{f}_n(x)) = \mathbb{E}\left(\int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx\right) \ .$$

Let $X_1, \ldots, X_n$ be samples from a density that is supported in $[0,1]$. Formulate conditions on f and h_n to ensure that the MISE converges to zero as $n \to \infty$.
Hint: The following fact from real analysis will be useful. Let $\|f\|_2^2 = \int_{\mathbb{R}} f^2(x)dx$ be the usual L_2 norm, and assume f is a density such that $\|f\|_2 < \infty$. Then, for any $\epsilon > 0$ there is a Lipschitz density g (for some $L_\epsilon > 0$) for which $\|f - g\|_2 \leq \epsilon$.

3. Using the ideas from the proof of Theorem 1 and Hoeffding's inequality show the following result, in the flavor of the DKW inequality. Namely for any $n \in \mathbb{N}$ and any $\epsilon > 0$

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \epsilon\right) \leq \left(\frac{2}{\epsilon} + 1\right) \ e^{-\frac{n\epsilon^2}{2}} \ .$$

   This result shows that, for any fixed $\epsilon > 0$, $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ converges exponentially fast to zero as $n \to \infty$. However, the dependence on ϵ is far from optimal.

4. Choose an arbitrary continuous distribution function F, and generate a random sample from this distribution (the integral probability transform might be handy in doing this). Compute estimates of the mean, variance and median and see how these behave when you vary the sample size. Compute the empirical distribution function and compare it to F for various random samples of different sizes. Repeat the previous experiment with the histogram, and compare it to the density $f = F'$ (note that, even if F is not differentiable everywhere, it will not be differentiable only is a small set of points (which has measure zero), so the density is well defined almost everywhere).

A Appendix

Proof of Theorem 1. Convergence almost-surely means that

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| = 0\right) = 1 \ .$$

Recall that the strong law of large numbers tells us that, for an arbitrary (but fixed) $x \in \mathbb{R}$

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{F}_n(x) = F(x)\right) = 1 \ .$$

The proof proceeds by reducing the supremum in the statement of the theorem to a maximum over a finite set. We will prove the theorem for the case when $F(x)$ is a continuous function. The proof can be easily extended for general distribution functions. Let $\epsilon > 0$ be fixed. Since F is continuous we can find m points such that $-\infty = x_0 < x_1 < \cdots < x_m = \infty$ and $F(x_j) - F(x_{j-1}) \leq \epsilon$ for $j \in \{1, \ldots, m\}$, where m is finite. Now take any point $x \in \mathbb{R}$. There is a $j \in \{1, \ldots, m\}$ such that $x_{j-1} \leq x \leq x_j$. As distribution functions are non-decreasing we conclude that

$$\begin{aligned}
\hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_j) - F(x_{j-1}) \\
&= (F_n(x_j) - F(x_j)) + (F(x_j) - F(x_{j-1})) \\
&\leq (F_n(x_j) - F(x_j)) + \epsilon \ . \\
&\leq \max_{j\in\{0,\ldots,m\}} \left|\hat{F}_n(x_j) - F(x_j)\right| + \epsilon \ .
\end{aligned}$$

In the same fashion

$$\begin{aligned}
\hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_{j-1}) - F(x_j) \\
&= (F_n(x_{j-1}) - F(x_{j-1})) + (F(x_{j-1}) - F(x_j)) \\
&\geq (F_n(x_{j-1}) - F(x_{j-1})) - \epsilon \ . \\
&\geq -\max_{j\in\{0,\ldots,m\}} \left|\hat{F}_n(x_{j-1}) - F(x_{j-1})\right| - \epsilon \ .
\end{aligned}$$

Therefore

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \leq \epsilon + \max_{j\in\{0,\ldots,m\}} \left|\hat{F}_n(x_j) - F(x_j)\right| \ . \tag{1}$$

Now let's make use of the strong law of large numbers. Define the events

$$A_j = \left\{\lim_{n\to\infty} \left|\hat{F}_n(x_j) - F(x_j)\right| \neq 0\right\} \ .$$

We know that $\mathbb{P}(A_j) = 0$. Define also the event

$$A = \left\{\lim_{n\to\infty} \max_{j\in\{0,\ldots,m\}} \left|\hat{F}_n(x_j) - F(x_j)\right| \neq 0\right\} \ .$$

Clearly $A = \cup_{j=0}^m A_j$ and so $\mathbb{P}(A) = \mathbb{P}(\cup_{j=0}^m A_j) \leq \sum_{j=0}^m \mathbb{P}(A_j) = 0$. So the second term in the right-hand-side of (1) converges almost-surely to zero. Therefore

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \leq \epsilon$$

almost surely. Or in other words, the event

$$B_\epsilon = \left\{\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \leq \epsilon\right\}$$

has probability one for any $\epsilon > 0$. Taking $\epsilon \to 0$ concludes the proof. To see this, and being overly formal, note that

$$\left\{\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| = 0\right\} = \cap_{\epsilon>0} B_\epsilon = \cap_{k\in\mathbb{N}} B_{1/k} \ .$$

Therefore, clearly

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| = 0\right) = \mathbb{P}\left(\cap_{k\in\mathbb{N}} B_{1/k}\right) = 1 \ .$$

□