

STAT 460/560 Class 6: Estimating the CDF and Functionals of the CDF

Daniel McDonald & Ben Bloem-Reddy

Reading: Chapter 7, [Was04]; Chapter 2.1-2.3, [Was06].

1. CDF, empirical CDF

Let $X_1, \dots, X_n$ be an IID random sample of variables taking values in $\mathbb{R}$, with unknown CDF F . Today's class focuses on estimating the CDF from the sample. Recall that the CDF is

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x dF(t) = \begin{cases} \sum_{x_j \leq t} f(x_j) & \text{if } X \text{ is discrete;} \\ \int_{-\infty}^x f(t) dt & \text{if } X \text{ is continuous.} \end{cases} \quad (6.1)$$

The **empirical CDF** (eCDF) is exactly what it sounds like:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(X_i \leq t). \quad (6.2)$$

This is just the CDF that puts mass $1/n$ at each data point in the sample. Figure 1 shows two examples.

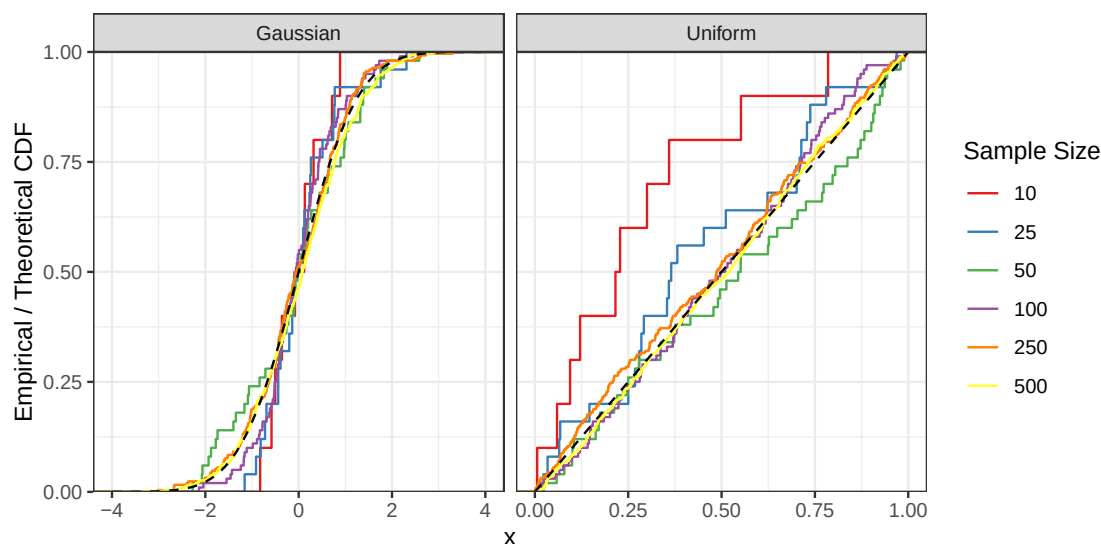


Figure 1: The empirical CDF of samples from the uniform distribution (right) and standard Gaussian distribution (left). Dashed lines are the CDF of the underlying distribution.

It suggests that as $n \rightarrow \infty$, the eCDF converges pointwise. Indeed, this is the case.

Theorem 6.1. *At any fixed value of $x \in \mathbb{R}$,*

$$\mathbb{E}(\hat{F}_n(x)) = F(x) \quad (6.3)$$

$$\text{Var}(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)) \quad (6.4)$$

$$\text{MSE}_F(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)) \rightarrow 0 \quad (6.5)$$

$$\hat{F}_n(x) \xrightarrow{p} F(x) . \quad (6.6)$$

Activity 6.1. Prove Theorem 6.1.

Solution: Since the X_i 's are assumed to be i.i.d., for fixed x the variables $Y_i = \mathbf{I}(X_i \leq x)$ are also i.i.d., with distribution Bernoulli(p_x), $p_x = P(X \leq x)$. The mean and variance follow from properties of the Bernoulli distribution. Since the MSE is equal to the bias squared (which is zero here) plus the variance, the expression for MSE follows. Finally, convergence in quadratic mean implies convergence in probability.

Exercise 6.1. Use the CLT to find the limiting distribution for $\hat{F}_n(x)$ for fixed x .

Solution: Using the mean and variance from above, we have for fixed x ,

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \rightsquigarrow \mathcal{N}(0, F(x)(1 - F(x))) .$$

Pointwise convergence says that the eCDF will converge at each point x , in the sense that for each $\epsilon > 0$, there is some N such that for all $n > N$,

$$\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \epsilon . \quad (6.7)$$

Pointwise convergence is nice, but may not be enough in some cases. For different x 's, N may be very different—i.e., the rate of convergence may vary a lot over $x \in \mathbb{R}$. Since we're estimating a *function* we'd like some guarantee that the entire function is converging more or less together. **Uniform convergence** expresses that: for each $\epsilon > 0$, there is some N such that for all $n > N$,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon\right) \leq \epsilon . \quad (6.8)$$

In short, $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{p} 0$. The Glivenko–Cantelli theorem establishes something even stronger.

Theorem 6.2 (Glivenko–Cantelli). *Let $X_1, \dots, X_n \sim_{\text{iid}} F$. Then*

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0 . \quad (6.9)$$

The proof is beyond the scope of this course. See, for example, Wainwright [Wai19, Ch. 4].

A related result can be used to obtain an approximate **confidence band**.

Theorem 6.3 (Dvoretzky–Kiefer–Wolfowitz (DKW)). *Let $X_1, \dots, X_n \sim_{\text{iid}} F$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} . \quad (6.10)$$

Since this is uniform over $x \in \mathbb{R}$, we can invert it for a confidence band: for $\alpha \in (0, 1)$, let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)} , \quad (6.11)$$

so that with

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\} \quad U(x) = \min\{\hat{F}_n(x) + \epsilon_n, 1\}, \quad (6.12)$$

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

Note that this is barely scratching the surface; the mathematical techniques of *empirical process theory* can be used to say much more. Chapters 18+ in [van98] are a good starting point if you're curious.

2. Functionals of the CDF

It turns out that many common estimands can be expressed as a **functional** of the CDF, i.e., some function of the CDF, denoted $T(F)$.

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

$$\hat{\theta}_n = T(\hat{F}_n). \quad (6.13)$$

Note that

$$\|F - G\|_\infty := \sup_{x \in \mathbb{R}} |F(x) - G(x)| \quad (6.14)$$

is a norm on the set of CDFs (called the sup-norm), and one way to show that an estimator of a functional T is consistent is to show that T is continuous with respect to the sup-norm and then appeal to Glivenko–Cantelli. We won't pursue that here, but see [Wai19, Ch. 4] for more if you're interested. Instead, we'll focus on a special (simple) class of functionals called **linear functionals**.

A **linear functional** of F is any functional that can be written as $T(F) = \int a(x)dF(x)$, for some function a .

Plug-in estimators of linear functionals have especially nice properties because if T is linear then

$$T(\hat{F}_n(x)) = \int a(x)d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n a(X_i). \quad (6.15)$$

In this case, the statistical properties of the estimator don't require any sophisticated techniques.

Example 6.1.

1. Let $T(F) = \int x dF(x)$ be the mean of the distribution F . Then T is a linear functional with $a(x) = x$. In this way, the sample mean is the plug-in estimator of the population mean: $T(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n X_i$.
2. Let $T(F)(t) = \int \mathbf{I}(x \leq t) dF(x)$ be the CDF itself. Then T is a linear functional with $a(x) = \mathbf{I}(x \leq t)$. In this way, the eCDF is the plug-in estimator of the CDF: $T(\hat{F}_n)(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(X_i \leq t)$.

In other cases, for example, the quantile function, it's not clear how to do even basic things like estimate $\text{se}(\hat{F}_n(x))$. Wasserman [Was06, Ch. 2.3] has some techniques for approximating this with a nonparametric version of the delta method. Often, those methods are not practically very useful, but they are the starting point of theoretical analysis of nonparametric methods. We'll just touch the surface to get an idea of the big picture.

3. Functional delta method at a high level

The main difficulty here is that we're working in a set of functions, which is infinite-dimensional and doesn't necessarily have the familiar structure of $\mathbb{R}^d$, so basic things like taking derivatives (for, e.g., a delta method) aren't immediately clear. The way to generalize the derivative (precisely, the directional derivative) is via the **Gâteaux derivative** of a functional T at the CDF F , in the direction of the CDF G ,

$$L_F(G) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)F + \epsilon G) - T(F)}{\epsilon}. \quad (6.16)$$

This represents an “infinitesimal step” from $T(F)$ towards $T(G)$. Suppose that $G = G(t) = \mathbf{I}(x \leq t)$, i.e., the CDF of a point mass at x (the PMF that puts all mass at x). For convenience, in this case we’ll write $G = \delta_x$, even though δ_x is not actually a CDF. In this case, $L_F(x) := L_F(\delta_x)$ is called the **influence function** (IF).¹ The IF represents the change in $T(F)$ when an infinitesimal amount ϵ of mass is subtracted from F , and replaced with a point mass at x . We can estimate the IF with the empirical IF (dropping the subscript F),

$$\hat{L}(x) = \lim_{\epsilon \rightarrow 0} \frac{T((1-\epsilon)\hat{F}_n + \epsilon\delta_x) - T(\hat{F}_n)}{\epsilon} . \quad (6.17)$$

If T is a *linear* functional, so that $T(F) = \int a(x)dF(x)$, then the following identities follow from the definitions:

1. For any G ,

$$L_F(G) = T(G) - T(F) , \quad (6.18)$$

2. $L_F(x) = a(x) - T(F)$ and therefore $\hat{L}(x) = a(x) - T(\hat{F}_n)$.

3. For any G ,

$$T(G) = T(F) + \int L_F(x)dG(x) . \quad (6.19)$$

4. $\int L_F(x)dF(x) = 0$.

Activity 6.2. Prove the three identities above.

Solution: We’ll prove (6.18) first. Using linearity of T and the definition of $L_F(G)$, we have

$$\begin{aligned} L_F(G) &= \lim_{\epsilon \rightarrow 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{\int a(x)d((1-\epsilon)F + \epsilon G)(x) - \int a(x)dF(x)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \frac{T(F) - \epsilon T(F) + \epsilon T(G) - T(F)}{\epsilon} \\ &= T(G) - T(F) . \end{aligned}$$

For 2., we plug in δ_x for G in (6.18) to get

$$L_F(x) = \int a(x')d\delta_x(x') - T(F) = a(x) - T(F) ,$$

and similarly for $\hat{L}(x)$ (plug in $\hat{F}_n$ for F).

For (6.19), we can integrate both sides of (6.18) with respect to G , so that

$$\int L_F(x)dG(x) = \int a(x)dG(x) - \int T(F)dG(x) = T(G) - T(F) .$$

Rearranging terms yields (6.19).

Finally, replacing G with F in (6.19) gives us $\int L_F(x)dF(x) = T(F) - T(F) = 0$.

¹The name comes from the field of robust statistics, where the IF measures the “influence” of a contamination of the underlying distribution F with an observation at x .

Now, using (6.19) with $G = \hat{F}_n$ yields

$$T(\hat{F}_n) - T(F) = \int L_F(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n L_F(X_i) . \quad (6.20)$$

Because $\mathbb{E}[L_F] = \int L_F(x) dF(x) = 0$ (by point 4), $\frac{1}{n} \sum_{i=1}^n L_F(X_i) = o_p(1)$. Therefore, $T(\hat{F}_n) - T(F) = o_p(1)$. Moreover, we can apply the CLT to $\frac{1}{n} \sum_{i=1}^n L_F(X_i)$ to get

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \rightsquigarrow \mathcal{N}(0, \tau^2) , \quad (6.21)$$

as long as

$$\tau^2 = \int L_F^2(x) dF(x) = \int (a(x) - T(F))^2 dF(x) < \infty . \quad (6.22)$$

By similarly using the plug-in estimator (with the empirical distribution), we can estimate this quantity by

$$\hat{\tau}^2 = \frac{1}{n} \sum_{i=1}^n \hat{L}^2(X_i) = \frac{1}{n} \sum_{i=1}^n (a(X_i) - T(\hat{F}_n))^2 . \quad (6.23)$$

Similar arguments can be used to show that $\hat{\tau}^2 \xrightarrow{p} \tau^2$. Moreover, if $\text{se} = \sqrt{\text{Var}(T(\hat{F}_n))}$ and $\hat{\text{se}} = \hat{\tau}/\sqrt{n}$ then $\hat{\text{se}}/\text{se} \xrightarrow{p} 1$. So we also have

$$\sqrt{n} \frac{(T(\hat{F}_n) - T(F))}{\hat{\tau}} \rightsquigarrow \mathcal{N}(0, 1) . \quad (6.24)$$

These arguments extend beyond linear functionals, to so-called Hadamard differentiable functionals, which are “approximately linear” in a small neighborhood around each F . See van der Vaart [van98, Ch. 20] for all the details, and the Appendix of Wasserman [Was06, Ch. 2] for a few of the details.

References

- [van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- [Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- [Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.
- [Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer New York, 2006.