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

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

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

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

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

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

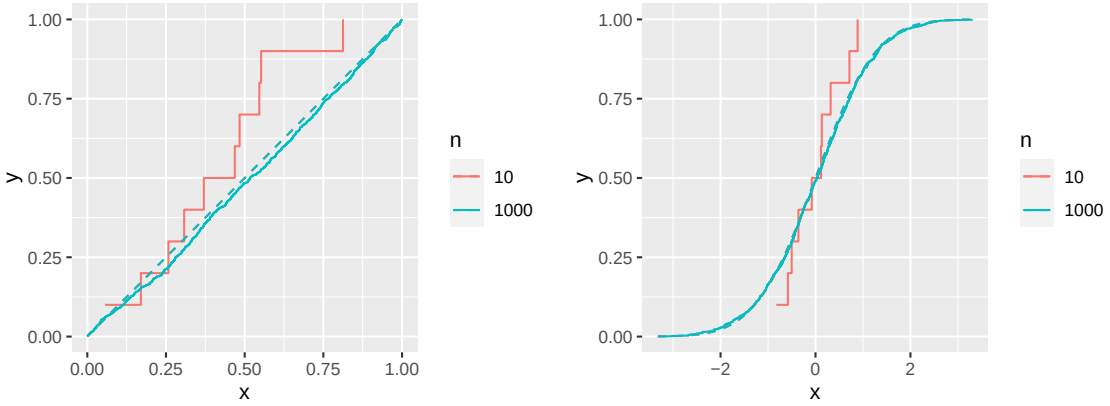

Figure 1: The empirical CDF of $n \in \{10, 1000\}$ samples from the uniform distribution (left) and standard normal distribution (right). Dashed lines are the CDF of the underlying distribution.

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

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

$$\mathbb{E}(\hat{F}_n(x)) = F(x) \qquad \text{Var}(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)) \tag{6.3}$$

$$\text{MSE}_F(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)) \to 0 \qquad \hat{F}_n(x) \xrightarrow{p} F(x) \; . \tag{6.4}$$

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 $n > N$,

$$\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \epsilon \, . \tag{6.5}$$

Pointwise convergence is nice, but may not be enough in some cases. For different x's, N may be very different. 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 $n > N$,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon\right) \leq \epsilon \, . \tag{6.6}$$

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

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

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0 \, . \tag{6.7}$$

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* $X_1, \ldots, X_n \sim_{\text{IID}} F$*. Then for any* $\epsilon > 0$*,*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|\right) \leq 2e^{-2n\epsilon^2} \, . \tag{6.8}$$

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

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)} \, , \tag{6.9}$$

so that with

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\} \qquad U(x) = \min\{\hat{F}_n(x) + \epsilon_n, 1\} \, , \tag{6.10}$$

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

Note that this is barely scratching the surface; the mathematical techniques of *empirical process theory* can be used to say much more.

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

$$\hat{\theta}_n = T(\hat{F}_n) \; . \tag{6.11}$$

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

Note that

$$\|F - G\|_\infty := \sup_{x \in \mathbb{R}} |F(x) - G(x)| \tag{6.12}$$

is a norm on the set of CDFs (called the sup-norm), and one way to show that an estimator of a functional is consistent is to show that the functional 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.

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

$$T(\hat{F}_n(x)) = \int r(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i) \; . \tag{6.13}$$

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

In other cases, for example, the quantile function, it's not clear how to do even basic things like estimate $\mathsf{se}(\hat{F}_n(x))$. 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 derviative** of a functional T at the CDF F, in the direction of the CDF G,

$$L_F(G) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon} \; . \tag{6.14}$$

This represents an "infinitesimal step" from $T(F)$ towards $T(G)$. If $G = \delta_x$, i.e., the point mass at x, then $L_F(x) := L_F(\delta_x)$ is called the **influence function** (IF). (The name comes from the field of robust statistics.) It 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),

$$\hat{L}(x) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\hat{F}_n + \epsilon \delta_x) - T(\hat{F}_n)}{\epsilon} \; . \tag{6.15}$$

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

1. $L_F(x) = a(x) - T(F)$ and $\hat{L}(x) = a(x) - T(\hat{F}_n)$.
2. For any G,

$$L_F(G) = T(G) - T(F) \; , \tag{6.16}$$

and

$$T(G) = T(F) + \int L_F(x)dG(x) \; . \tag{6.17}$$

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

Activity 6.2. Prove the three identities above.

Solution: We'll prove (6.16) first. By linearity of T,

$$\begin{aligned} L_F(G) &= \lim_{\epsilon \to 0} \frac{T(F) - \epsilon T(F) + \epsilon T(G) - T(F)}{\epsilon} \\ &= T(G) - T(F) \ . \end{aligned}$$

Plugging in δ_x for G yields the first identity,

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

For (6.17), we can integrate both sides of the first identity 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 (6.17). Finally, replacing G with F in (6.17) gives us $\int L_F(x)dF(x) = T(F) - T(F) = 0$.

Now, using (6.17) with $G = \hat{F}_n$ yields

$$T(\hat{F}_n) - T(F) = \frac{1}{n}\sum_{i=1} L_F(X_i) \ . \tag{6.18}$$

From $\int L_F(x)dF(x) = 0$, this converges in probability to 0. Moreover, we can apply the CLT to the right-hand side to get

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \rightsquigarrow \mathcal{N}(0, \tau^2) \ , \tag{6.19}$$

as long as

$$\tau^2 = \int L_F^2(x)dF(x) = \int (a(x) - T(F))^2 dF(x) < \infty \ . \tag{6.20}$$

We can estimate this quantity by

$$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^n \hat{L}^2(X_i) = \frac{1}{n}\sum_{i=1}^n (a(X_i) - T(\hat{F}_n))^2 \ . \tag{6.21}$$

It can be shown that $\hat{\tau}^2 \xrightarrow{\text{P}} \tau^2$. Moreover, if $\mathsf{se} = \sqrt{\text{Var}(T(\hat{F}_n))}$ and $\hat{\mathsf{se}} = \hat{\tau}/\sqrt{n}$ then $\hat{\mathsf{se}}/\mathsf{se} \xrightarrow{\text{P}} 1$. So we also have

$$\sqrt{n}\frac{(T(\hat{F}_n) - T(F))}{\hat{\tau}} \rightsquigarrow \mathcal{N}(0, 1) \ . \tag{6.22}$$

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.