36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 10: Estimate CDF and Functionals

Lecturer: Jing Lei

10.1 Review and Outline

Last lecture we discussed:

- Consistency of the MLE
- The Fisher Information
- Asymptotic Normality of MLE

In this lecture we will discuss some non-parametric estimation problems and discuss the plug-in method to estimate functionals. This is Chapter 7 of the Wasserman book.

10.2 Estimating the CDF

Assume we observe $X_1, \ldots, X_n \sim F$, and would like to estimate F. Perhaps worth noting that we impose absolutely no restrictions on F. Further, there is no notion of a (finite-dimensional) parameter that we can attempt to estimate in this context.

Some typical applications:

1. **Estimating (many) interval probabilities:** Suppose we observe a stochastic quantity many times, and are then interested in estimating the probability $\mathbb{P}(a \leq X \leq b)$ for some fixed $[a, b]$. In this case we would just use the empirical counts, and use the empirical variance to get some idea of the variability. We could even use the CLT/Hoeffding's inequality to obtain concentration bounds of the form:

$$\mathbb{P}\left(\left|\widehat{\mathbb{P}}(a \leq X \leq b) - \mathbb{P}(a \leq X \leq b)\right| \geq t\right) \leq 2e^{-cnt^2}.$$

Suppose now I wanted to estimate this probability for many intervals: $[a_1, b_1], [a_2, b_2], \ldots, [a_k, b_k]$ for some very large k, and I want *simultaneous* concentration, i.e.,

$$\mathbb{P}\left(\sup_k \left|\widehat{\mathbb{P}}(a_k \leq X \leq b_k) - \mathbb{P}(a_k \leq X \leq b_k)\right| \geq t\right) \leq \text{something small}.$$

The naive way to do this would be to estimate each probability and do a union bound. Wouldn't it be nice if we could instead estimate the entire CDF reliably?

2. **The Kolmogorov-Smirnov test:** This is only somewhat related to the estimation question we focus on today, but one other important use of the CDF is to test hypotheses about distributions, i.e. suppose I think my samples $X_1, \ldots, X_n$ have a $N(0,1)$ distribution. A natural way to test this hypothesis is by comparing the CDF of my samples to the CDF of a $N(0,1)$ distribution. In order to be more rigorous about the performance of such a test however, we need to understand a basic question: suppose the samples were truly from a $N(0,1)$ distribution, how far would we expect the sample CDF to be from the $N(0,1)$ CDF?

Our estimator for the CDF will just be the empirical CDF: the empirical CDF corresponds to the pmf that puts mass $1/n$ at each data point X_i, i.e.:

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

Let us try to investigate some basic properties of this estimator. Suppose we fix a value x:

1. **Bias:** The estimator we have proposed is unbiased, i.e.:

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

2. **Variance:** The variance of the estimator is:

$$\text{Var}(\widehat{F}_n(x)) = \frac{\mathbb{P}(X \leq x)(1 - \mathbb{P}(X \leq x))}{n}.$$

3. **MSE:** The MSE at x is just the squared bias + variance, i.e.,

$$\text{MSE} = \frac{\mathbb{P}(X \leq x)(1 - \mathbb{P}(X \leq x))}{n} \rightarrow 0,$$

as $n \rightarrow \infty$. From this we can conclude that for any fixed x our estimator converges in probability, i.e. that:

$$\mathbb{P}(|\widehat{F}_n(x) - F(x)| \geq \epsilon) \rightarrow 0,$$

as $n \rightarrow \infty$.

There are two additional important results that we will not prove but are worth knowing:

1. **Glivenko-Cantelli:** The Glivenko-Cantelli theorem is essentially a uniform LLN (we discussed these before briefly in the previous lecture). Precisely, it says that

   $$\sup_x |F(x) - \widehat{F}_n(x)| \to 0,$$

   almost surely. We have not seen almost sure convergence before but note that it implies convergence in probability. To emphasize, the previous result was a statement for a fixed x. The Glivenko-Cantelli theorem assures us that the empirical CDF converges to the true CDF *uniformly*, i.e. for every value x simultaneously.

2. **DKW (Dvoretzky-Kiefer-Wolfowitz):** The DKW inequality is a concentration inequality for CDFs. It implies the Glivenko-Cantelli theorem and is a more refined finite-sample bound:

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

It is again worth pondering why or how it is that ULLNs work, i.e. why is it possible that the empirical CDF is close to the true CDF for *every possible* x? ULLNs and more generally **empirical process theory** is at the heart of the more advanced statistical estimation results.

10.3 Estimating Statistical Functionals

We should first briefly remark on what exactly a functional is. We think of a function as a map from a point in some input space to the reals, i.e.,

$$f : x \mapsto f(x),$$

on the other hand a functional maps a function to a real number. A typical functional is the value of the function at some point x_0, i.e.

$$T(f) : f \mapsto f(x_0),$$

A *statistical functional* typically refers to a function of the CDF. Some canonical examples:

1. **Mean:** The mean can be thought of as a functional, i.e.:

   $$\mu(F) = \mathbb{E}_F X.$$

2. **Variance:** Similarly, the variance is a functional:

   $$\mathrm{Var}(F) = \mathbb{E}_F (X - \mu(F))^2 .$$

3. **Linear Functionals:** In general, we define linear functionals (like the mean) to be functionals of the form:

$$T(F) = \mathbb{E}_F r(X)\,,$$

for some function r. These are called linear because if we take $U = aF + bG$ then,

$$T(U) = aT(F) + bT(G).$$

The mean is a linear functional but the variance is not.

10.3.1 The plug-in estimator

A natural estimator for a linear functional is to plug-in the empirical CDF and use the resulting functional, i.e.:

$$\widehat{T}(F) := T(\widehat{F}_n) = \mathbb{E}_{\widehat{F}_n} r(X) = \frac{1}{n}\sum_{i=1}^{n} r(X_i).$$

Again the canonical example is estimating the mean of a distribution.

$$\hat{\mu} = \widehat{T}(F) := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This principle can also sometimes be used to estimate non-linear functionals like the variance.

$$\begin{aligned}\hat{\sigma}^2 &= \int x^2 \, d\widehat{F}_n(x) - \left(\int x \, d\widehat{F}_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.\end{aligned}$$

We will conclude this lecture with two more canonical examples:

Skewness: The skewness of a RV is:

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\mathbb{E}(X-\mu)^3}{(\mathbb{E}(X-\mu)^2)^{3/2}},$$

so we can see that we could use the plug-in principle separately on the numerator and denominator and then further use the plug-in principle to estimate μ. This leads to the estimator:

$$\widehat{\kappa} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right)^{3/2}}.$$

Correlation: The correlation between two RVs is a functional of the joint distribution of the pair (X, Y). The correlation is:

$$\rho = \frac{\mathbb{E}(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}.$$

As an exercise show that the plug-in estimator is the sample correlation:

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{\sqrt{\sum_{i=1}^n (X_i - \hat{\mu}_X)^2} \sqrt{\sum_{i=1}^n (Y_i - \hat{\mu}_Y)^2}}.$$

It is worth noting that in some sense all of these estimators are completely non-parametric, i.e. there are no parametric assumptions about the underlying distribution being made in order to derive estimators.