Lecture #6

- To get a sense of non-parametric estimation, let us consider a classical problem: estimating a cdf.
- Let X be an unknown r.v. with cdf F, i.e. $\mathbb{P}(X \le x) = F(x)$. We would like to estimate F from data $x_1, x_2, \dots, x_n \overset{iid}{\sim} X$. The idea is simple: put some probability mass at each sample.

Defn: Let X be a r.v., and let $x_1, x_2, \dots, x_n \overset{iid}{\sim} X$. The empirical cdf of X is defined to be $\hat{F}_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{1}(x - x_i)$,

where $\mathbb{1}(y) = \begin{cases} 1, & y \ge 0 \\ 0, & y < 0. \end{cases}$

- So, each $\mathbb{1}(x - x_i)$ "switches on" when x exceeds x_i for the first time.
- How well does this work? Pretty well!

Theorem (Unbiasedness of $\hat{F}_n$): For any x, $\mathbb{E}(\hat{F}_n(x)) = F(x)$.

Proof: $\mathbb{E}(\hat{F}_n(x))$

$$= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x - x_i)\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}(x - x_i)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X \leq x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} F(x)$$

$$= F(x). \blacksquare$$

- For homework, you will prove consistency, namely that $\hat{F}_n(x) \xrightarrow{P} F(x)$.
- Our goal is to say "how big of a problem is it to use $\hat{F}_n$ in place of F? That is, can the data-driven estimate be used in place of the genuine article?

Theorem (Dvoretsky-Kiefer-Wolfowitz): Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} X$. Let F be the cdf of X. Then $\forall \varepsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \hat{F}_n(x)| > \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right).$$

- We will not prove this, but it gives us concentration. This can in turn be used to construct a confidence interval: Let $L(x)$, $U(x)$ be lower and upper bounds defined as

$$L(x) = \max\left\{0, \hat{F}_n(x) - \sqrt{\frac{1}{2n}\cdot\log\left(\frac{2}{\alpha}\right)}\right\}$$

$$U(x) = \min\left\{1, \hat{F}_n(x) + \sqrt{\frac{1}{2n}\cdot\log\left(\frac{2}{\alpha}\right)}\right\}$$

Then DKW gives

$$\mathbb{P}\Big(L(x) \le F(x) \le U(x) \quad \text{for all } x \Big) \ge 1-\alpha.$$

- Our first confidence interval! What's going on?! A few comments are in order:

(1) Expanding out and assuming n is large enough, we can ignore the min/max and just say

$$\mathbb{P}\left(\hat{F}_n(x) - \sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)} \le F(x) \le \hat{F}_n(x) + \sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)} \right) \ge 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\underbrace{-\sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)}}_{\text{lower bound}} \le \underbrace{\hat{F}_n(x) - F(x)}_{\text{error at } x} \le \underbrace{\sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)}}_{\text{upper bound}} \right) \ge 1-\alpha$$

(2) As $\alpha \to 0$, $\log\left(\frac{2}{\alpha}\right)$ blows up and the interval

$$\left[-\sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)}, \sqrt{\tfrac{1}{2n} \cdot \log\left(\tfrac{2}{\alpha}\right)} \right]$$

widens. So, more certainty requires a wider interval.

(3.) As $n \to \infty$, $\sqrt{\frac{1}{2n} \cdot \log\left(\frac{2}{\alpha}\right)} \to 0$, so the interval tightens. More samples means better estimation!

· This is a basic theme in statistics, parametric and non-parametric: more samples improves matters, and there is a tradeoff between confidence and precision.

· The empirical cdf leads to a general approach for estimating things, known as "plug in" estimation: just "plug in" $\hat{F}_n$ in place of F!

· Let T be any function of the cdf F, e.g.

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \cdot F'(x)\,dx, \qquad \overbrace{\text{median of } X = F^{-1}\left(\tfrac{1}{2}\right)}^{\text{think about this!}}$$

plug in $\hat{F}_n$!

For any T, the plug-in estimator of $\theta = T(F)$ is $\hat{\theta}_n = T(\hat{F}_n)$

ex: Consider the expected value, i.e. $T(F) = \int_{\mathbb{R}} x \cdot F'(x)\,dx$. The plug-in estimator is

$$T(\hat{F}_n) = \int_{\mathbb{R}} x \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x - x_i)\right]' dx$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} x \cdot \frac{d}{dx}\left[\mathbb{1}(x - x_i)\right]$$

· One needs to be a bit careful here. What is $\frac{d}{dx}\mathbb{1}(x - x_i)$? It's a distribution, beyond our techniques. But, we can use IBP to justify

$$\int_{\mathbb{R}} x \cdot \frac{d}{dx}\left[\mathbb{1}(x - x_i) \right] dx = x_i .$$ This is the idea of a "weak" derivative.

So, the plug-in estimator for the expected value is just the empirical average:

$$T(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

ex: Suppose now $T(F) = F^{-1}(\frac{1}{2})$ is the median? There are some issues about handling $\hat{F}_n$ here. Why? Well, $\hat{F}_n$ is constant between observed data points. In particular, it is not invertible. But, if we set

$$\hat{F}_n^{-1}(\alpha) = \inf \{ x \mid \hat{F}_n(x) \geq \alpha \},$$ then for n odd, $\hat{F}_n^{-1}(\frac{1}{2})$

agrees with the gradeschool notion of median: sort and pick the middle value!

· Plug-in estimators are nice: simple and for 1-dimension, effective. There are fundamental problems when X takes values in $\mathbb{R}^d$, $d \geq 2$. Indeed, how to estimate F? Big issues for d large... "Curse of dimensionality".

· Non-parametrics are beautiful but struggle in high dimensions, without extra care.