LECTURE NOTES 13

Now we will discuss nonparametric inference (Chapter 7 in the book).

Notation: Recall that $\int g(x)dF(x)$ means $\int g(x)p(x)$ if F has denity p and means $\sum_j \int g(x_j)p(x_j)$ in the discrete case.

1 Estimating the CDF

We observe $X_1, \ldots, X_n \sim F$, and would like to estimate F. We impose absolutely no restrictions on F so this is completely nonparametric. usually, estimating $F$is actually just the first step in some other task.

Our estimator for the CDF is the **empirical CDF** that puts mass $1/n$ at each observation:

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

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:

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

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

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

   as $n \to \infty$. From this we can conclude that for any fixed x our estimator converges in probability, i.e. that: $\widehat{F}_n(x) \xrightarrow{P} F(x)$.

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

1. **Glivenko-Cantelli:** The Glivenko-Cantelli theorem says

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

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:

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

From DKW we get confidence bands. Let

$$L(x) = \max\left\{\widehat{F}_n(x) - \epsilon_n, 0\right\},$$
$$U(x) = \min\left\{\widehat{F}_n(x) + \epsilon_n, 1\right\},$$

where

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

Then,

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

2 Estimating Statistical Functionals

A **Statistical Functional** is a map T that assigns a number $T(F)$ to each distribution. Sometimes we write $\theta = T(F)$.

Note: People often call θ a parameter but this does not mean we are using a parametric model.

Some examples:

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

$$\mu(F) = \int x \, dF(x).$$

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

$$\text{Var}(F) = \int (x - \mu(F))^2 \, dF(x).$$

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

$$T(F) = \int r(x) \, dF(x),$$

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

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

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

2.1 The plug-in estimator

We define the **plug-in** estimator by

$$\widehat{\theta} = T(\widehat{F}_n).$$

For example, suppose that $\mu = T(F) = \int x dF(x)$. The plug-in estimator is

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

Similarly, if $T(F) = \int r(x) dF(x)$ is a linear functional, then

$$T(\widehat{F}_n) = \int r(x) \, d\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} r(X_i).$$

For the variance we have

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

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:

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

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:

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

3 Quantiles

The β-quantile is defined to be

$$\psi = T(F) = \inf\{x : \ F(x) > \beta\}.$$

When $\beta = 1/2$, this is the median. The plug-in estimator is $\widehat{\psi} = T(\widehat{F}_n)$. Define the order statistics, that is, the ordered data:

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

Then $\widehat{\psi} = T(\widehat{F}_n) = X_{(j)}$ where $j \approx n\beta$.

We can compute an exact confidence interval for ψ as follows. Recall that, if $X \sim F$ then $F(X) \sim \text{Unif}(0,1)$. Hence, let $U_1, \ldots, U_n \sim \text{Unif}(0,1)$ and let

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$$

be the uniform order statistic. Then $F(X_{(j)}) \stackrel{d}{=} U_{(j)}$. Therefore,

$$\begin{aligned} P(X_{(j)} \leq \psi \leq X_{(k)}) &= P(X_{(j)} \leq F^{-1}(\beta) \leq X_{(k)}) \\ &= P(F(X_{(j)}) \leq \beta \leq F(X_{(k)})) \\ &= P(U_{(j)} \leq \beta \leq U_{(k)}). \end{aligned}$$

If we can somehow choose j and k so that $P(U_{(j)} \leq \beta \leq U_{(k)}) \geq 1 - \alpha$, then $[X_{(j)}, X_{(k)}]$ is an exact (non-asymptotic) confidence interval for ψ.

Let N be the number of $U_i's$ less than β. Note that $U_{(j)} > \beta$ if and only if $N \leq j$. Also note that $N \sim \text{Binom}(n, \beta)$. Let G be the cdf of $\text{Binom}(n, \beta)$. So

$$P(U_{(j)} > \beta) = P(N \leq j) = G(j) = \alpha/2$$

if we choose $j = G^{-1}(\alpha/2)$. Similalrly,

$$P(U_{(k)} < \beta) = P(N \geq k) = 1 - P(N \leq k) = \alpha/2$$

if we choose $k = G^{-1}(1 - \alpha/2)$. Thus

$$\begin{aligned} P(\psi \notin [X_{(j)}, X_{(k)}]) &= P(\beta \notin [U_{(j)}, U_{(k)}]) \\ &\leq P(U_{(j)} > \beta) + P(U_{(k)} < \beta) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha. \end{aligned}$$

To summarize,

$$P(\psi \in [X_{(j)}, X_{(k)}]) = 1 - \alpha$$

where $j = G^{-1}(\alpha/2)$ and $k = G^{-1}(1 - \alpha/2)$.

Remark: Note that, by the CLT, $(N - n\beta)/\sqrt{n\beta(1-\beta)} \approx Z = N(0, 1)$. Hence,

$$j \approx n\beta - z_{\alpha/2}\sqrt{n\beta(1-\beta)}, \quad k \approx n\beta + z_{\alpha/2}\sqrt{n\beta(1-\beta)}.$$