Lecture 01/22/20

Dr. Syring

Statistical Inference

Collect data $X \sim P$. Want to infer something about P . Cases:

a) Know nothing about P and wish to infer the precise distribution. (This is very challenging).

b) Know very little about P and wish to infer some "small" thing about P , such as its mean. (This is not so hard as a)).

c) Know a parametric family to which P belongs and wish to infer the parameter, thus identifying P (This is also not so hard as a)).

We will mostly be concerned with c) but today we talk about a).

Histogram Estimates

Section 4.1.2

Histograms are not just statistical graphics; they are also estimators of probability mass/density functions.

Sometimes you actually want to estimate the whole distribution of a r.v. But, even if you want to estimate something simpler like a mean or variance, you can estimate the whole distribution and then derive an estimate of a simpler quantity by using its relationship with the distribution. For example, suppose we estimate a pdf of a continuous r.v. by $\hat{f}(x)$. Then, we can estimate the mean of X by $\hat{\mu} = \int x\hat{f}(x)dx$. This is called “plug-in” estimation.

Histogram Estimates, Finite Discrete Case

Suppose X has, essentially, a multinomial distribution, meaning that X takes one of a finite number of values, e.g. $X \in \{1, 2, 3, \ldots, k\}$.

By the LLN,

$$\hat{p}_j := \frac{1}{n} \sum_i 1\{X_i = j\} \overset{i.p.}{\to} p_j.$$

(And, since k is finite, this convergence can be made uniform.)

So, the histogram with bars at $1, 2, 3, \ldots, k$ of heights $\hat{p}_j$ converges to the graph of the pmf $p(x) = p_j$ for $x = j$.

Histogram Estimates, Continuous Case

Supose our observed data is $x_1, x_2, \ldots, x_n$ and denote $a := \min x_i$ and $b := \max x_i$. Then, for some “small” $h > 0$ and $m = (b - a + 2h)/2h$

$$(a - h, b + h) = \bigcup_{j=1}^{m} (a + (2j - 3)h, a + (2j - 1)h).$$

For $x \in A_j := (a + (2j - 3)h, a + (2j - 1)h)$ approximate the pdf by

$$\hat{f}(x) = \frac{1}{2hn} \sum_i \#(x_i \in A_j).$$

So, there is a different histogram/density estimate for each h used! Much more difficult to show that this “works” than in the discrete case (Glivenko-Cantelli Thm).

But, these histogram estimates are all valid densities since

$$\int \hat{f}(x)dx = \sum_{j=1}^{m} \int_{a+(2j-3)h}^{a+(2j-1)h)} \frac{\#(x \in A_j)}{2hn} dx$$

$$= \sum_{j=1}^{m} \#(x \in A_j) \frac{2h}{2hn} = \frac{2hn}{2hn} = 1.$$

Glivenko-Cantelli Theorem

Slightly different than histogram, we actually estimate the CDF $F(x)$. Suppose $X_1, \ldots, X_n$ iid with CDF $F(x)$ defined on the real line. Let $\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$.

Clear that for a fixed x $\hat{F}(x)$ converges to $F(x)$ i.p. and $n\hat{F}(x)$ is binomial so the CLT says that for every fixed x $\sqrt{n}(\hat{F}(x) - F(x)) \sim N(0, F(x)(1 - F(x)))$.

Glivenko-Cantelli makes the LLN “uniform”.

Theorem: $\lim \sup_x |\hat{F}(x) - F(x)| = 0$ in probability. The difficulty (why this isn’t immediate) is because we are taking the maximum over an uncountable set.

Proof: Let $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ such that

$$F(t_i-) - F(t_{i-1}) < \epsilon.$$

As a result, for any $t \in (t_{i-1}, t_i)$

$$\hat{F}(t) - F(t) \leq \hat{F}(t_i-) - F(t_i-) + \epsilon$$

and

$$\hat{F}(t) - F(t) \geq \hat{F}(t_{i-1}) - F(t_{i-1}) - \epsilon.$$

For any fixed $p > 0$ and $\epsilon > 0$ there exists an $n(t_i-)$ and $n(t_{i-1})$ s.t. for all n larger than both these we have both $P(|\hat{F}(t_i-) - F(t_i-)| > \epsilon) < p$ and $P(|\hat{F}(t_{i-1}) - F(t_{i-1})| > \epsilon) < p$. Take $N > \max\{\max_i n(t_i-), \max_i n(t_{i-1})\}$ and these probability statements hold uniformly (that is for every i). Therefore,

$$\lim_{n\to\infty} \sup_x |\hat{F}(x) - F(x)| < \epsilon$$

for every ϵ and hence the limit is zero.