Principle of Statistics (continued)

November 27, 2017

Contents

In the previous lecture, we saw that for a bootstrap sample $(X_1^b, ..., X_n^b)$ drawn from $\mathbb{P}_n(\mid X_1, ..., X_n)$, we have $\sup_{t\in\mathbb{R}} |\mathbb{P}_n(\sqrt{n}(\bar{X}_n^b - \bar{X}_n) \leq t(X_1, ..., X_n) - \Phi(t)| \xrightarrow{a.s.} 0$ as $n \to \infty$, where $\Phi(t) = \mathbb{P}(Z \leq t)$, $Z \sim N(0, \sigma^2)$.

Remark. As in the B.vM theorem, this theorem can be used to show that $\mathbb{P}(\mu \in \mathcal{C}_n) \to 1 - \alpha$ when $n \to \infty$.

Idea: For fiexd $(X_i)_{i\geq 1}$, $\mathbb{P}_n(\mid X_1, ..., X_N)$ is a sequence of distributions. Considering the X_i as independent random variables drawn from P allows us to make statements "with randomness". The idea is to fix a sequence X_i (equivalent to fix ω in the original probability space) if we can show that $\sup_{t\in\mathbb{R}} |\mathbb{P}_n(... \mid X_1, ..., X_n) - F(t)| \to 0$ "for almost all ω", then we have almost sure convergence.

Lemma. If $A_n \sim f_n \xrightarrow{d} A \sim f$, and F is continuous (c.d.f of f), then $\sup |F_n(t) - F(t)| \to 0$ as $n \to \infty$.

Proof. By continuity of F, there exists points $-\alpha_0 = x_0 < x_1 < ... < x_k = +\infty$ such that $F(x_i) = \frac{i}{k}$. Then for every $x \in [x_{i-1}, x_i]$, $F_n(x) - F(x) \leq F_n(x_i) - F(x_{i-1}) = F_n(xi) - F(x_i) + \frac{1}{k}$, and $F_n(x) - F(x) \geq F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k}$. For k large enough, $\frac{1}{k} < \frac{\varepsilon}{2}$, for n large enough (dependes on k), we have $\max_{0\leq i\leq k} |F_n(x_i) - F(x_i)| < \varepsilon/2$ (pointwise convergence) (dependes on k), we have

$$\max_{0\leq i\leq k} |F_n(x_i) - F(x_i)| < \varepsilon/2$$

(pointwise convergence). As a consequence,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \max_{0\leq i\leq k} |F_n(x_i) - F(x_i)| + \frac{1}{k} < \varepsilon$$

□

Definition. The sequence $(Z_{n,i}, i = 1, ..., n)_{n\geq 1}$ is a triangular array of i.i.d. random variables if:
• For all $n \geq 1$, $(Z_{n,1}, ..., Z_{n,i}, ..., Z_{n,n})$ is a sequence of i.i.d. random variables;
For example, $Z_{11} = (Z_{1,i})$,
$Z_{21}, Z_{22} = (Z_{2,i})$,
...,
$Z_{n,1}, ..., Z_{n,n} = (Z_{n,i})$.
We need independence on each line, but not across the lines. We don't need even need the same distribution at each line.

Proposition. (CLT for triangular arrays)
Let $(Z_{n,i}; i = 1, ..., n)$ be a triangular array of iid random variables, each with finite variance. We have $Var_{Q_n}(Z_{n,i}) = \sigma_n^2 \to \sigma_2$ as $n \to \infty$, each line consists of n independent draws from Q_n. Then, under the following hypotheses (1-3), we have

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i} - \mathbb{E}_{Q_n}[Z_{n,i}]) \xrightarrow{d} N(0, \sigma^2)$$

(1) $\forall \delta > 0$, $nQ_n(|Z_{n,1}| > \sqrt{n}\delta) \to 0$ as $n \to \infty$;
(2) $Var(Z_{n,1}1\{|Z_{n,1}| \leq \sqrt{n}\}) \to \sigma^2$ as $n \to \infty$;

(3) $\sqrt{n}\mathbb{E}[Z_{n,1}1\{|Z_{n,1}| > \sqrt{n}\}] \to 0$ as $n \to \infty$.
(The statement of these assumptions is not examinable.)

Proof. (of the main theorem)
Fix $(X_i)_{i\geq 1}$ (equivalent to fix ω in the original probability space). Under $Q_n = \mathbb{P}_n(\cdot \mid X_1, ..., X_n)$, $Z_{n,i} = X_i^{b(n)}$ (bootstrap on n observations), $\mathbb{E}_n[Z_{n,i}] = \mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)}] = \bar{X}_n$. Then, the $(X_i^{b(n)}; i = 1, ..., n)_{n\geq 1}$ are a triangular array of i.i.d. variables. We have that

$$Var_{\mathbb{P}_n}(X_i^{b(n)}) = \mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)^2}] - (\mathbb{E}_{\mathbb{P}_n}[X_i^{b(n)}])^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2 = \sigma_n^2$$

by definition of $\mathbb{P}_n(| X_1, ..., X_n)$. For almost all the ω, or almost all infinite sequences, $\sigma_n^2 \to \sigma^2$, and hypotheses (1-3).
• By the CLT for triangular arrays, we have that

$$\sqrt{n}(\bar{X}_n^{b(n)} - \bar{X}_n) \xrightarrow{d} N(0, \sigma^2)$$

as $n \to \infty$ 'for almost all ω'.
• By lemma,

$$\sup_{t\in\mathbb{R}} |\mathbb{P}_n(\sqrt{n}(\bar{X}_n^{b(n)} - \bar{X}_n) \leq t) - \Phi(t)| \geq 0$$

as $n \to \infty$ 'for almost all ω', meaning that it $\xrightarrow{a.s.} 0$. □

Remark. • This shows the validity of the bootstrap confidence interval for the mean.
• In genral, this can be extended to estimation of θ: Sampling from $\mathbb{P}_n$ as for the mean, and compute the bootstrap MLE $\hat{\theta}_n^b = \hat{\theta}(X_1^b, ..., X_n^b)$ then using $\sqrt{n}(\hat{\theta}_n^b - \hat{\theta}_n)$ as a proxy for $\sqrt{n}(\hat{\theta}_n - \theta_0)$, we will have similar results that show that taking R_n such that

$$\mathbb{P}_n(||\hat{\theta}_n^b - \hat{\theta}_n|| \leq \frac{R_n}{\sqrt{n}} \mid X_1, ..., X_n) = 1 - \alpha$$

can be used to construct a valid confidence region.
This approach is known as the non-parametric bootstrap.
Another approach is to do the same thing with $X_1^b, ..., X_n^b \sim \mathbb{P}_{\hat{\theta}_n}$, same types of results will hold.

Last lecutre on Monday. *No* Lecture on Wednesday!

0.1 Monte-Carlo methods

In statistics, we often cannot explicityl compute expectation/integrals, which can be problematic when:
• We want to obtain a posterior distribution;
• We want to obtain a posterior mean;
• We want to compute quantiles of a distribution.

One of the ideas of Monte-Carlo methods is to replace explicitly computation by *simulations*Last lecutre on Monday. *No* Lecture on Wednesday!

0.2 Monte-Carlo methods

In statistics, we often cannot explicityl compute expectation/integrals, which can be problematic when:
• We want to obtain a posterior distribution;
• We want to obtain a posterior mean;
• We want to compute quantiles of a distribution.

One of the ideas of Monte-Carlo methods is to replace explicitly computation by *simulations*. One of the first challenges is to draw from a given, general distribution.

Definition. A pseudorandom generator provides independent $U_i^* \sim Unif(0,1)$.

Remark. They are generated such that for all practical uses, $P(U_1^* \leq u_1, ..., U_N^* \leq u_n) = \prod_{i=1}^N u_i$ (up to 'machine precision'). Here it can be thought of as a black-box that outputs i.i.d. uniform numbers – this can be used as a starting point to generate other variables.

Proposition. The random variables $K_i = \sum_{k=1}^n k 1_{(\frac{k-1}{n}, \frac{k}{n}]}(U_i^*)$ are i.i.d. uniform on $\{1, ..., n\}$.

Proof. K_i is clearly uniform on $\{1, ..., n\}$, as each segment has length $\frac{1}{n}$. They are independent as functions of independent random variables U_i^*. □

Remark. Assigining other values to each of the intervals with uniform distribution on any set of size n. In particular, we can simulate bootstrap samples by writing

$$X_i^b = \sum_{k=1}^n X_k 1_{(\frac{k-1}{n}, \frac{k}{n}]}(U_i^*)$$

If the intervals are chosen with different lengths, we can generate any discrete distribution. For a general distribution with cdf F, we can generalize this idea.

Definition. For a general cdf F, we define the generalized inverse of F as

$$F^-(u) = \inf\{x : u \leq F(x)\}$$

Remark. For a fixed value $t \in \mathbb{R}$, the function F gives $F(t) \in [0,1]$ a probability $\mathbb{P}(X \leq t)$. For a fixed value $u \in [0,1]$, the function F gives $t = F^-(u)$ such that approximately $\mathbb{P}(X \leq t) = u$.

Proposition. $X = F^-(u)$ for $U \sim Unif(0,1)$ has a distribution cdf F.

Proof. (Example sheet)
$\mathbb{P}(X \leq t) = \mathbb{P}(F^-(U) \leq t) = ... = F(t)$, and use that $\mathbb{P}(U \leq z) = z$ for $z \in (0, 10.$ □

Conclusion (of the first part): If F is known, explicit, we can generated $(X_1^*, ..., X_N^*) = (F^-(U_1^*), ..., F^-(U_N^*))$ that are i.i.d., each with cdf F. If we

want to compute $\mathbb{E}_{X\sim f}[g(X)]$, we can approximate it by $\frac{1}{N}\sum_{i=1}^{N} g(X_i^*)$, and use the fact that $\frac{1}{N}\sum_{i=1}^{N} g(X_i^*) \xrightarrow{a.s.} \mathbb{E}[g(X)]$ by the LLN.

In certain situations, the distributino might be complex, and F, F^- are not explicit. For example, $N(\mu, \sigma^2)$ can be solved by looking up in a table, but $\Pi(\cdot \mid X)$ can involve complicated integrals (the density) which make computation of F_Π impossible.

There are several ways to tackle this and approximately sample from distributions.

(1) Importance sampling: Let F have density f, and random variables i.i.d. $X_i^* \sim h$.

Proposition. $\mathbb{E}_h\left[\frac{g(x)}{h(x)}f(x)\right] = \mathbb{E}_f[g(x)]$.

Proof. The above is equal to

$$\int_\chi \frac{g(x)}{h(x)} f(x) \cdot h(x) dx = \int_\chi g(x) f(x) dx$$

As a consequence,

$$\frac{1}{N}\sum_{i=1}^{n} \frac{g(X_i^*)}{h(X_i^*)} f(X_i^*) \xrightarrow{a.s.} \mathbb{E}_{X\sim f}[g(X)]$$

. □

(2) Accept/Reject algorithm. As in (1), but $f \leq M \cdot h$ for some constant M.
Step 1: generate $X \sim h$ and $U \sim U(0,1)$.
Step 2: $Y = X$ if $U \leq \frac{f(X)}{M \cdot h(X)}$, otherwise return to step 1. Then $Y \sim f$ (example sheet).

For multivariate problems where conditional distributions are aesy to compute but not joint distributions, we can then use the *Gribbs samples*: In the bivariate case (X, Y), start at the same $X = x_0$, and $Y_1 \sim f_{Y|X}(\cdot|x_0)$, and $X_1 \sim f_{X|Y}(\cdot|y_1)$, etc., $Y_t \sim f_{Y|X}(\cdot|X_{t-1})$, $X_t \sim f_{X|Y}(\cdot|Y_t)$. The sequences (X_t, Y_t),(X_t),(Y_t) are all Markov chains, with invariant distribution f, $f_{X|Y}$, $f_{Y|X}$. And we can use the ergodic theorem to approximate expectations

$$\frac{1}{N} g(X_t, Y_t) \to \mathbb{E}_{(X,Y)\sim f}[g(X, Y)]$$

This can be used in particular in the case $Q(x, \theta) = f(x, \theta)\pi(\theta)$.

Important: Last lecutre today, no lecutre on Wednesday!!

0.3 Nonparametric statistics

Consider observing $X_1, ..., X_n \sim P$ i.i.d. with the distribution P having cdf on $\mathbb{R}$: $F(t) = \mathbb{P}(X \leq t)$ for all $t \in \mathbb{R}$. Here we want to estimate directly the function F, without a parametric assumption: we cannot "estimate θ to estimate F_θ".

Remark. We note that

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{E}_p[1_{[-\infty,t]}(x)]$$

(or $\int_{\mathbb{R}} 1_{[-\infty,t]}(x) d\mathbb{P}(x)$) if the distribution is continuous.

For all $t \in \mathbb{R}$, the real number $F(t)$ is the expectation of the random variable $1_{[-\infty,t]}(X) \in \{0,1\}$ of which we observe n i.i.d. draws.

Definition. The *empirical distribution function* is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty,t]}(X_i)$$

Remark. If we are interested in the value of the c.d.f., for some fixed t, $F_n(t)$ is a consistent estimator of $F(t)$ by Law of large numbers, and we can control the rate of estimation by limiting distribution of $\sqrt{n}(F_n(t) - F(t))$ (CLT gives $N(0, F(t)(1 - F(t)))$). This is just a Bernoulli model.
Because we are interested in the overall (all of $\mathbb{R}$) behaviour of F_n, and see F_n as the estimator of a function, we have to understand it s dependency structure.

Theorem. (Glivenko-Cantelli theorem)
We have, as $n \to \infty$, that

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s.} 0$$

Proof. If f is continuous in t, writing $q(X_i, t) = 1_{[-\infty,t]]}(X_i)$, we have $\mathbb{E}_p[q(X,t)] = F(t)$ and the uniform law of large numbers applies directly:

$$\sup_{t \in \mathbb{R}} \left| \underbrace{\frac{1}{n} \sum_{i=1}^{n} q(X_i, t)}_{F_n(t)} - \underbrace{\mathbb{E}_p[q(X,t)]}_{F(t)} \right| \xrightarrow{a.s.} 0$$

The case where F is not continuous can be handled as well and the result holds by using that F is non-decreasing and cutting $[0,1]$ into smaller intervals of size $\leq \varepsilon$. □

Theorem. (Donskin-Kolmogorov-Doob theorem)
As $n \to \infty$, the random function $\sqrt{n}(F_n - F)$ converges $\sqrt{n}(F_n - F) \xrightarrow{"d"} \mathcal{G}_F$ "in distribution over the space of functions". Here $\mathcal{G}_F$ is a random function from $\mathbb{R}$ to $\mathbb{R}$ such that $\mathcal{G}_F(t)$ is normally distributed $N(0, F(t)(1 - F(t)))$, and $Cov(\mathcal{G}_F(s), \mathcal{G}_F(t)) = F(s)(1 - F(t))$ for $s \leq t$.

Construction of $\mathcal{G}_F$:
Informal definition: A Brownian motion, a Wiener process is defined as
• $W_j = 0$ a.s.;
• $t \to W_t$ is continuous a.s.;
• For $s \leq t$, $W_t - W_s$ is independent of $(W_{s'})_{s' \leq s}$ and has distribution $N(0, t - s)$.

The Brownian bridge is "tied to 0 at 0 and 1", and defined as $B_t = W_t - tW_1$.

The variance of $B_t = t(1-t)$ and $Cov(B_s, B_t) = s(1-t)$ for $s \leq t$. Taking $\mathcal{G}_F(t) = B_{F(t)}$ gives a construction of this process.

Remark. If $U_1, ..., U_n \overset{i.i.d.}{\sim} U[0,1]$, then $F(t) = t$ and $\sqrt{n}(F_n - F)$ will converge directly to a Brownian bridge.

Theorem. (Kolmogorov-Smirnov theorem)
$\sqrt{n}||F_n - F||_\infty \xrightarrow{d} ||B||_\infty$, where $||B||_\infty = \sup_{t\in(0,1)} |B_t|$.

Proof. $||\mathcal{G}_F||_\infty = \sup_{t\in\mathbb{R}} |B_{F(t)}| = \sup_{t\in(0,1)} |B_t|$. □

Remark. $||B||_\infty$ doesn't depend on F, there are tables so it can be used in many inference tasks.
(1) Non-parametric hypothesis testing: we have $H_0 : F = F_0$, or $H_1 : F \neq F_0$. Then $\sqrt{n}||F_n - F_0||_\infty \xrightarrow{d} ||B||_\infty$.
(2) Confidence bands for F: we can define $C_n(x)$ for all x in $\mathbb{R}$ around $F_n(x)$, and study $\mathbb{P}(F(x) \in C_n(x) \forall x \in \mathbb{R}) \to$
Other application of non-parametric statistics:
(1) Regression, wher $y_i = f(X_i) + \varepsilon_i$; unknown function f.
(2) Density estimation: $X_i \sim P$ have density f. In many applications $\mathbb{E}|\hat{f}_n - f| \gg \frac{1}{\sqrt{n}}$.

—end of lecture notes—