STA 532: Theory of Statistical Inference

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

2 Estimating CDFs and Statistical Functionals

Empirical CDFs

Let $\{X_i:\ i \le n\}$ be a "simple random sample", *i.e.*, let the $\{X_i\}$ be n iid replicates from the same probability distribution. We can't know that distribution exactly from only a sample, but we can estimate it by the "empirical distribution" that puts mass $1/n$ at each of the locations X_i (if the same value is taken more than once, its mass will be the sum of its $1/n$'s so everything still adds up to one). The CDF

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i,\infty)}(x)$$

of the empirical distribution will be piecewise-constant, with jumps of size $1/n$ at each observation point (or k/n in the event of k-way ties).

Since $\#\{i \le n:\ X_i \le x\}$ is just a Binomial random variable with $p = F(x)$ for the real PDF for the $\{X_i\}$, with mean np and variance $np(1-p)$, it is clear that for each $x \in \mathbb{R}$

- $\mathsf{E}\hat{F}_n(x) = F(x)$ and
- $\mathsf{V}\hat{F}_n(x) = F(x)[1-F(x)]/n$, so
- $\hat{F}_n(x)$ is an unbiased and MS consistent estimator of $F(x)$.

In fact something stronger is true— not only does $\hat{F}_n(x)$ converge to $F(x)$ pointwise in x, but also the *supremum* $\sup_x |\hat{F}_n(x) - F(x)|$ converges to zero. There are many ways a sequence of random variables might converge (studying those is the main topic of a measure-theoretic probability course like Duke's STA 711); the "Glivenko-Cantelli theorem" asserts that this maximum converges with probability one. Either Hoeffding's inequality (Wassily Hoeffding was a UNC statistics professor) or the DKW inequality of Dvoetzsky, Kiefer, and Wolfowitz give the strong bound

$$\mathsf{P}\big[\sup_x |\hat{F}_n(x) - F(x)| > \epsilon\big] \le 2e^{-2n\epsilon^2}$$

for every $\epsilon > 0$. It follows that, for any $0 < \gamma < 1$,

$$\mathsf{P}\big[L(x) \le F(x) \le U(x) \quad \text{for all } x \in \mathbb{R}\big] \ge \gamma$$

is a non-parametric confidence set for F, for $L(x) := 0 \vee \big(\hat{F}_n(x) - \epsilon_n\big)$, $U(x) := 1 \wedge \big(\hat{F}_n(x) + \epsilon_n\big)$, and $\epsilon_n := \sqrt{\log(2/(1-\gamma))/2n}$. Here $a \vee b$ denotes the maximum of $a, b \in \mathbb{R}$, $a \wedge b$ the minimum.

Statistical Functionals

Usually we don't want to estimate all of the CDF F for X, but rather some feature of it like its mean $\mathsf{E}X = \int xF(dx)$ or variance $\mathsf{V}X := \mathsf{E}\big(X - (\mathsf{E}X)\big)^2 = \int x^2F(dx) - (\mathsf{E}X)^2$ or the probability $[F(B) - F(A)]$ that X lies in some interval $(A, B]$.

Examples of Statistical Functionals

Commonly-studied or quoted functionals of a univariate distribution $F(\cdot)$ include:

- The **mean** $\mathsf{E}[X] = \mu := \int_{\mathbb{R}} x\,F(dx) = \int_0^\infty [1 - F(x)]\,dx - \int_{-\infty}^0 F(x)\,dx$, quantifying location;
- The qth **quantile** $z_q := \inf\{x < \infty : \ F(x) \ge q\}$, especially
- The **median** $z_{1/2}$, another way to quantify location;
- The **variance** $\mathsf{V}[X] = \sigma^2 := \int_{\mathbb{R}} (x - \mu)^2\,F(dx) = \mathsf{E}[X^2] - \mathsf{E}[X]^2$, quantifying spread;
- The **skewness** $\gamma_1 := \int_{\mathbb{R}} (x - \mu)^3\,F(dx)/\sigma^3$, quantifying asymmetry;
- The (excess) **kurtosis** $\gamma_2 := \int_{\mathbb{R}} (x - \mu)^4\,F(dx)/\sigma^4 - 3$, quantifying peakedness. "Lepto" is Greek for skinny, "Platy" for fat, and "Meso" for middle; distributions are called leptokurtic (t, Poisson, exponential), platykurtic (uniform, Bernoulli), or mesokurtic (normal) as γ_2 is positive, negative, or zero, respectively.
- The **expectation** $\mathsf{E}[g(X)] = \int_{\mathbb{R}} g(x)\,F(dx)$ for any specified problem-specific function $g(\cdot)$.

Not all of these exist for some distributions— for example, the mean, variance, skewness, and kurtosis are all undefined for heavy-tailed distributions like the Cauchy or α-Stable. There are quantile-based alternative ways to quantify location, spread, asymmetry, and peakedness, however— for example, the interquartile range IQR $:= [z_{3/4} - z_{1/4}]$ for spread, for example.

Any of these can be estimated by the same expression computed with the *empirical* CDF $\hat{F}_n(x)$ replacing $F(x)$, without specifying a parametric model for F. There are methods (one is the "jackknife"; another, the "bootstrap", is described below) for trying to estimate the mean and variance of any of these functionals from a sample $\{X_1, \cdots, X_n\}$.

Later we'll see ways of estimating the functionals that *do* require the assumption of particular parametric statistical models. There's something of a trade-off in deciding which approach to take. The parametric models typically give more precise estimates and more powerful tests, *if* their underlying assumptions are correct. BUT, the non-parametric approach will give sensible (if less precise) answers even if those assumptions fail. In this way they are said to be more "robust".

Simulation

The Bootstrap

One way to estimate the probability distribution of a functional $T_n(X) = T(X_1, \ldots, X_n)$ of n iid replicates of a random variable $X \sim F(dx)$, called the "bootstrap" (Efron, 1979; Efron and

Tibshirani, 1993), is to approximate it by the empirical distribution of $T_n(\hat{X})$ based on draws *with replacement* from a sample $\{X_1, \ldots, X_n\}$ of size n. The underlying idea is that these would be drawn from exactly the right distribution of $T(X)$ if we could possibly repeat draws of $X = (X_1, \ldots, X_n)$ from the *population*; if the sample is large enough, we can hope that the empirical distribution will be close to the population distribution, and so the bootstrap sample will be much like a true random sample from the population (but without the expense of drawing new data).

Bootstrap Variance

For example, the population median

$$M = T(F) := \inf\{x \in \mathbb{R} : \; F(x) \ge 1/2\}$$

might be estimated by the sample median $M_n = T(\hat{F}_n)$, but how precise is that estimate? One measure would be its *standard error*

$$\mathsf{se}(M_n) := \left\{\mathsf{E}|M_n - M|^2\right\}^{1/2}$$

but to calculate that would require knowing the distribution of X, while we only have a sample. The Bootstrap approach is to use some number B of repeated draws with replacement of size n from this sample as if they were draws from the population, and estimate

$$\hat{\mathsf{se}}(M_n) \approx \left\{\frac{1}{B}\sum_{b=1}^{B} |M_n^b - \hat{M}_n|^2\right\}^{1/2}$$

where $\hat{M}_n$ is the sample average of the B medians $\{M_n^b\}$.

Bootstrap Confidence

Interval estimates $[L, U]$ of a real-valued parameter θ, intended to cover θ with probability at least $100\gamma\%$ for any θ, can also be constructed using a bootstrap approach. One way to do that is to begin with an iid sample $X = \{X_1, \ldots, X_n\}$ from the uncertain distribution F; draw B independent size-n draws with replacement from the sample X; for each, compute the statistic $T_n(X^b)$; and set L and U to the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of $\{T_n(X^b)\}$, respectively, for $\alpha = (1-\gamma)$. Wasserman (2004, §8.3) argues why this should work and gives two alternatives.

Bayesian Simulation

Bayesian Bootstrap

Rubin (1981) introduced the "Bayesian bootstrap" (BB), a minor variation on the bootstrap that leads to a simulation of the posterior distribution of the parameter vector θ governing a distribution $F(\cdot \mid \theta)$ in a parametric family, from a particular (and, in Rubin's view, implausible) improper prior distribution. This five-page paper is a good read, and argues that neither the BB nor the original bootstrap is suitable as a "general inferential tool" because of its implicit use of this prior.

Importance Sampling

Most Bayesian analyses require the evaluation of one or more integrals, often in moderately high-dimensional spaces. For example: if $\pi(\theta)$ is a prior density function on $\Theta \subset \mathbb{R}^d$, and if $\mathcal{L}(\theta \mid X)$ is the likelihood function for some observed quantity $X \in \mathcal{X}$, then the posterior expectation of any function $g : \Theta \to \mathbb{R}$ is given by the ratio

$$\mathsf{E}\left[g(\theta) \mid X\right] = \frac{\int_\Theta g(\theta)\, \mathcal{L}(\theta \mid X)\, \pi(\theta)\, d\theta}{\int_\Theta \quad \mathcal{L}(\theta \mid X)\, \pi(\theta)\, d\theta}. \tag{1a}$$

Often the integrals in both numerator and denominator are intractable analytically, so we must resort to numerical approximation. Let $f(\theta)$ be any pdf such that the ratio $w(\theta) := \mathcal{L}(\theta \mid X)\, \pi(\theta)/f(\theta)$ is bounded (for this, $f(\theta)$ must have "fatter tails" than $\mathcal{L}(\theta \mid X)\pi(\theta)$), and let $\{\theta_m\}$ be iid replicates from the distribution with pdf $f(\theta)$. Then

$$= \frac{\int_\Theta g(\theta)\, w(\theta)\, f(\theta)\, d\theta}{\int_\Theta \quad w(\theta)\, f(\theta)\, d\theta} = \lim_{M\to\infty} \frac{\sum_{m=1}^M g(\theta_m)\, w(\theta_m)}{\sum_{m=1}^M \quad w(\theta_m)} \tag{1b}$$

so $\mathsf{E}[g(\theta) \mid X]$ can be evaluate as the limit of weighted averages of $g(\cdot)$ at the simulated points $\{\theta_m\}$. Provided that $\int_\Theta g(\theta)^2 f(\theta)\, d\theta < \infty$, the mean-square error of the sequence of approximations in (1b) will be bounded by σ^2/M for a number σ^2 that can also be estimated from the same Monte Carlo sample $\{\theta_m\}$, giving a simple measure of precision for this estimate.

This simulation-based approach to estimating integrals, called "Monte Carlo Importance sampling", works well in dimensions up to six or seven or so. A number of ways have been discovered and exploited to reduce the stochastic error bound $\sigma/\sqrt{M}$. These include "antithetic variables", in which the iid sequence $\{\theta_m\}$ is replaced by a sequence of negatively-correlated pairs; "control variates", in which one tries to estimate $[g(\theta) - h(\theta)]$ for some quantity h whose posterior mean is known; and "sequential MC", in which the sampling function $f(\theta)$ is periodically replaced by a "better" one.

MCMC

A similar approach to (1) that succeeds in many higher-dimensional problems is Markov Chain Monte Carlo, based on sample averages of $\{g(\theta_m) :\ 1 \le m < \infty\}$ for an ergodic sequence $\{\theta_m\}$ constructed so that it has stationary distribution $\pi(\theta \mid X)$. You'll see much more about that in other courses at Duke, so we won't focus on it here.

Particle Methods, Adaptive MCMC, Variational Bayes, ...

There are a number of variations on MCMC methods, as well. Some of these involve averaging $\{g(\theta_m^{(k)}) :\ 1 \le m < \infty\}$ for a number of streams $\theta_m^{(k)}$ (here the streams are indexed by k), possibly by a variable number of streams whose distributions may evolve through the computation. This is an area of active research; ask any Duke statistics faculty member if you're interested.

References

Efron, B. (1979), "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1–26, doi:10.1214/aos/1176344552.40.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Boca Ratan, FL: Chapman & Hall/CRC.

Rubin, D. B. (1981), "The Bayesian Bootstrap," *Annals of Statistics*, 9, 130–134.

Wasserman, L. (2004), *All of Statistics*, New York, NY: Springer-Verlag.