Sectioning and bootstrapping

Andreas Sorge

August 25, 2014

The following presentation is based on the excellent textbook by Asmussen and Glynn [1].

Given a random element X, its distribution F, and some real-valued functional ψ, we would like to estimate $\psi(F)$ and its $1-\alpha$ confidence interval without further assumptions. For example, the mean is the functional $\psi(F) = \int xF(dx)$, where dx is the probability to find X in dx.

Given R independent samples $X_1, \ldots, X_R$ from F, an estimate for $\psi(F)$ is $\psi(\hat{F}_R)$, where

$$\hat{F}_R(dx) := \frac{1}{R}\sum_{r=1}^{R} \delta_{X_r}(dx)$$

is the *empirical distribution* and $\delta_{X_r}(A) = 1 \Leftrightarrow X_r \in A$.

For real-valued random variables, the empirical cumulative distribution function is

$$\hat{F}_R(x) := \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{\{X_r \leq x\}}.$$

As $R \to \infty$, we have $\psi(\hat{F}_R) \to \psi(F)$ almost surely [1]. Furthermore, we have a central limit theorem such that $\psi(\hat{F}_R)$ is distributed as $\psi(F) + Y$, where $Y \sim \mathcal{N}(0, \sigma/\sqrt{R})$.

0.1 Sectioning

Sectioning means splitting the sample into N subsamples (sections) of size K. The empirical distribution of the n-th section is

$$\hat{F}_{n,K}(dx) := \frac{1}{K}\sum_{r=(n-1)K+1}^{nK} \delta_{X_r}(dx).$$

The $1-\alpha$ confidence interval for the estimator $\psi(\hat{F})$ is

$$\psi(\hat{F}) \pm t_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{N}},$$

where $t_{1-\alpha/2}$ is the critical value of the Student t distribution with $N-1$ degrees of freedom and the estimator for the variance

$$\hat{\sigma}^2 := \frac{1}{N-1}\sum_{n=1}^{N}\left(\psi(\hat{F}_{n,K}) - \psi(\hat{F}_R)\right)^2.$$

The number of sections needs to be sufficiently large in order for the central limit theorem to approximately hold.

0.2 Bootstrapping

When a model for the distribution F is lacking, or too complicated for statistical inference, bootstrapping methods provide alternatives. Bootstrapping takes the empirical distribution $\hat{F}_R$ as a surrogate for the true distribution F. Instead of drawing more samples from F, bootstrapping involves resampling from $\hat{F}_R$.

The true $1-\alpha$ confidence interval of the estimator $\psi(\hat{F}_R)$ is

$$(\psi(\hat{F}_R) - z_2, \psi(\hat{F}_R) - z_1)$$

with the $\alpha/2$ and $1-\alpha/2$ quantiles z_1, z_2

$$P(\psi(\hat{F}_R) - \psi(F) < z_1) = P(\psi(\hat{F}_R) - \psi(F) > z_2) = \frac{\alpha}{2}$$

such that

$$P(\psi(F) \in (\psi(\hat{F}_R) - z_2, \psi(\hat{F}_R) - z_1)) = 1 - \alpha.$$

Assuming that $\hat{F}_R \approx F$, the empirical quantiles z_1^*, z_2^* satisfy [1]

$$P_{\hat{F}_R}(\psi(\hat{F}_R) - \psi(F) < z_1^*) = P_{\hat{F}_R}(\psi(\hat{F}_R) - \psi(F) > z_2^*) = \frac{\alpha}{2}$$

We draw B bootstrap samples of size R from $\hat{F}_R$. The b-th bootstrap sample is $X^*_{1,b}, \ldots, X^*_{R,b}$ with each random variable $X^*_{r,b}$ drawn independently from $X_1, \ldots, X_R$ with equal probabilities $P(X^*_{r,b} = X_{r'}) = \frac{1}{R}$. The empirical distribution of the b-th bootstrap sample is

$$\hat{F}^*_{R,b}(dx) := \frac{1}{R} \sum_{r=1}^{R} \delta_{X^*_{r,b}}(dx).$$

Then the empirical quantiles z_1^*, z_2^* are the $\lfloor \frac{\alpha}{2}(B+1) \rfloor$-th and $\lfloor (1-\frac{\alpha}{2})(B+1) \rfloor$-th order statistic, respectively, of the B independent and identically distributed random variables

$$\left(\psi(\hat{F}^*_{R,b}) - \psi(\hat{F}_R) \right)_{b=1}^{B}.$$

These quantiles z_1^*, z_2^* approximate the quantiles z_1, z_2 of the true distribution F, and hence, yield the approximate $1-\alpha$ confidence interval [1]

$$\left(\psi(\hat{F}_R) - z_2^*, \psi(\hat{F}_R) - z_1^* \right).$$

References

[1] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, Stochastic Modelling and Applied Probability, Vol. 57 (Springer New York, 2007).