Jackknife And Bootstrap

Adrija Bhar
Srijani Das
Yenisi Das

- Definitions
- Resampling Methods
- Resampling Techniques: *Jackknife*
- Resampling Technique: Bootstrap

Statistical Functional

Definitions

Statistical functional is a function $\Psi(.)$ that maps a distribution F to a real number (or vector).

Examples include:

$$\text{Mean:} \quad \Psi(F) = \int x dF(x)$$

$$\text{Variance:} \quad \Psi(F) = \int x^2 dF(x) - \left(\int x dF(x) \right)^2$$

$$\text{Median:} \quad \Psi(F) = F^{-1}(0.5)$$

where, $F^{-1}(0.5) = inf\{x : F(x) \geq 0.5\}$

The Plug-in Principle

Definitions

The **plug-in principle** is a simple method of estimating parameters from sample.

Let $\hat{F}_n$ denote the empirical distribution corresponding to a distribution function F based on n samples. Then the plug-in estimate of a parameter

$$\theta = \psi(F)$$

is defined to be,

$$\hat{\theta} = \psi(\hat{F}_n)$$

.

In other words, we estimate the function $\theta = \psi(F)$ of the probability distribution F by the same function of the empirical distribution function $\hat{F}_n$, i.e, $\hat{\theta} = \psi(\hat{F}_n)$.

- Definitions
- **Resampling Methods**
- Resampling Techniques: *Jackknife*
- Resampling Technique: Bootstrap

Statistics and their Sampling Distributions

Resampling Methods

The basic objective of statistical analysis is ”extracting all the information from the data”(Rao, 1989) to deduce the properties of the population that generated the data. Statistical analyses are based on *statistics* which are functions of data. Most statistical procedures require some knowledge of the sampling distribution of the statistic being used for analysis.

The sampling distributions of a statistics and its characteristics depend on the underlying population and therefore are unknown. Hence, they have to be estimated/approximated from the data in most of the estimation and inference problem. In most situations the relative accuracy of the estimators depend of the underlying population and we have to use the data to estimate the relative accuracy for selecting an estimator.

Traditional Approach

Resampling Methods

In the traditional approach an accuracy measure is estimated by an empirical analogue of an explicit theoretical formula of the accuracy measure or its approximation, which is derived from a postulated model. Let us use the variance as an illustration.

Let $X_1, X_2, ..., X_n$ be *iid* observations from unknown distribution F and $T_n = T_n(X_1, X_2, ..., X_n)$ be a given statistic. Then the variance of T_n can be written as

$$Var(T_n) = \int \left[T_n(\mathbf{x}) - \int T_n(\mathbf{y}) d \prod_{i=1}^{n} F(y_i) \right]^2 d \prod_{i=1}^{n} F(x_i)$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$. When T_n is *simple* for eg. the *sample mean* we can obtain an explicit expression of $var(T_n)$ as a function of some unknown quantities. And then estimate $Var(T_n)$ by substituting the unknown quantities with their estimates.

However there are not many statistics as simple as the *sample mean*. For most statistics the expression of $Var(T_n)$ is too complicated and it is hard to obtain an exact and explicit form of $Var(T_n)$.

For eg, 5% *trimmed sample mean*,

$$T_n = \bar{X_n}^{0.05} = \frac{1}{n - 2[0.05n]} \sum_{i=[0.05n]+1}^{n-[0.05n]} X_{(i)}$$

Classical statisticians responded to these by

1. Restrict to tracktable special cases
2. Using Asymptotics

As computers became advanced and fast, in mid 1940 a new world of statistics emerged which gave the statisticians a new way to use complex statistical models. Instead of dealing with it mathematically, we make the computer to perform the random experiment following the model.

Now coming to our existing problem, we have a random sample from an unknown distribution F, we have a statistic T_n based on the data. Suppose we want to estimate its bias given by, $\mathbb{E}(T_n - \theta)$, it is not enough to know $\mathbb{E}(T_n)$ because we have to also know θ. In such case resampling techniques comes to play.

- Definitions
- Resampling Methods
- Resampling Techniques: *Jackknife*
- Resampling Technique: Bootstrap

Jackknife

Resampling Techniques: *Jackknife*

Quenouille(1949) introduced a method later named the *Jackknife* to estimate the *bias* of an estimator by deleting one datum each time from the original dataset and recalculating the estimator based on the rest of the data.

Let $T_n = T_n(X_1, ..., X_n)$ be an estimator for the unknown parameter θ. Bias of $T_n = \mathbb{E}(T_n - \theta)$. Let,

$$T_{n-1,i} = T_{n-1}(X_1, .., X_{i-1}, X_{i+1}, ..., X_n)$$

be the given statistics but based on $(n-1)$ observations $(X_1, .., X_{i-1}, X_{i+1}, ..., X_n)$, $i = 1, 2, ..., n$.

Quenouille's *Jackknife* bias estimator is

$$b_{JACK} = (n-1)(\bar{T}_n - T_n)$$

where $\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n-1,i}$.This leads to a bias reduced *Jackknife* estimator of θ,

$$T_{JACK} = T_n - b_{JACK}$$

The *Jackknife* estimators b_{JACK} and T_{JACK} can be heuristically justified as follows. Suppose that

$$bias(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

where a and b are unknown but do not depend on n. Since $T_{n-1,i}$, $i = 1, 2, .., n$, are identically distributed,

$$bias(T_{n-1}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right)$$

and $bias(\bar{T}_n)$ has the same expression.

Therefore, the expected value of b_{JACK} is given by,

$$\begin{aligned}\mathbb{E}(b_{JACk}) &= (n-1)[bias(\bar{T}_n) - bias(T_n)] \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right)\end{aligned}$$

Therefore it follows that,

$$bias(T_{JACk}) = bias(T_n) - \mathbb{E}(b_{JACk}) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right)$$

That is the $bias(T_{JACk})$ is of the order n^{-2}. the *Jackknife* produces a *bias* reduced estimator by removing the first order term in $bias(T_n)$.

The *Jackknife* technique gained more attention since Tukey(1958) found that the *Jackknife* can also be used to construct variance estimators.

The *jackknife* estimator of variance estimator is given by,

$$v_{JACK} = \frac{n-1}{n} \sum_{i=1}^{n} (T_{n-1,i} - \bar{T}_n)^2$$

Limitations of *Jackknife*

Jackknife requires computation of a statistic T repeatedly over many pseudo datasets. It computes T for the n *Jackknife* datasets. By looking only at the n *Jackknife* samples. Hence, the *Jackknife* uses only limited information about the statistic T.

During $1970 - 80$ the development in computer technology was very rapid which led to the development of new statistical methods that are computer intensive and are more reliable and have broader applications.

The ***bootstrap*** introduced by Efron(1979) is one of these methods.

- Definitions
- Resampling Methods
- Resampling Techniques: *Jackknife*
- Resampling Technique: Bootstrap

Bootstrap

Resampling Technique: Bootstrap

The idea of bootstrap is to approximate F by some distribution F^* that is,

- close to F
- completely known
- and can be easily simulated from

We shall use F^* as a proxy for F. The empirical distribution function F_n is a very obvious choice for F^*. This gives us **Nonparametric bootstrap technique**.

Non-Parametric Bootstrap

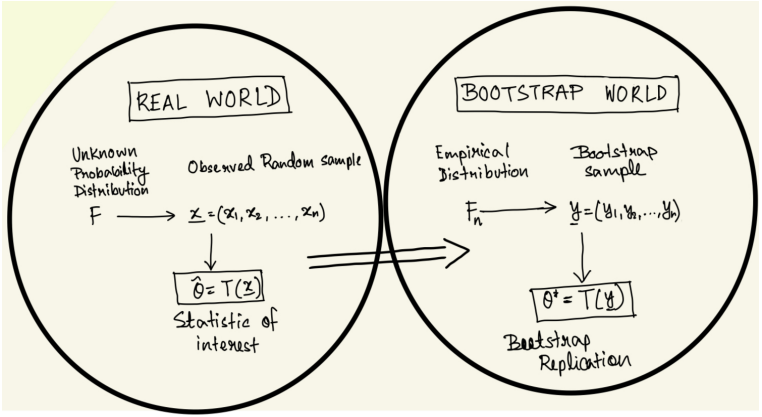
Resampling Technique: Bootstrap

Suppose we have a random sample of size n in our hand and we are interested in the standard error of a statistic $T := T(X_1, X_2, ..., X_n)$. The non-parametric bootstrap works in the following way:

1. It draws B samples of size n from the empirical distribution function F_n. This is nothing but drawing simple random sample with replacement(SRSWR) of size n from the original data $(X_1, X_2, ..., X_n)$. So, we get B samples of size n denoted by: $\mathbf{X}_b^* = (X_{b_1}^*, X_{b_2}^*, ..., X_{b_n}^*)$, $b = 1, 2, ..., B$. These are called the **Empirical Bootstrap Samples**.
2. Compute statistic $T_b^* = T(X_{b_1}^*, X_{b_2}^*, ..., X_{b_n}^*)$ for all $b = 1, 2, ..., B$.
3. The empirical standard deviation of $T^* = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2}$ across the B bootstrap samples is our bootstrap estimator of standard deviation of T. Here, $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T_b^*$.

Visualization

Resampling Technique: Bootstrap



Notations

Resampling Technique: Bootstrap

Let $\{x_1, x_2, \ldots, x_n\}$ be a random sample of size n from a population with distribution $F(\cdot)$, and let $T(x_1, \ldots, x_n; F)$ be the random variable of interest - possibly depending on the unknown distribution $F(\cdot)$. Let $F_n(\cdot)$ denote the empirical distribution function *EDF* of $x_1, \ldots, x_n$. The Bootstrap method is to approximate the distribution of $T(x_1, \ldots, x_n; F)$ under F by $T_n(y_1, y_2, \ldots, y_n; F_n)$ under F_n, where $\{y_1 \ldots, y_n\}$ denotes a random sample of size n from $F_n(\cdot)$.

$$\cdot \bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\cdot s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

$$\cdot G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \leq x\}$$

Let, $\mathbb{P}$ and $\mathbb{P}^*$ denote probabilities under F and F_n, respectively and $\mathbb{E}$ and $\mathbb{E}^*$ denote expectations under F and F_n, respectively.

Theorem

If $\mathbb{E}(x^2) < \infty$, *then*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\bar{x}_n - \mu\right) \leq x\right) - \mathbb{P}^*\left(\sqrt{n}\left(\bar{y}_n - \bar{x}_n\right) \leq x\right) \right| \xrightarrow{a.s.} 0$$

Proof of Theorem

Resampling Technique: Bootstrap

Proof:

$$\mathrm{E}\ (x^2) < \infty \Rightarrow s_n^2 \xrightarrow{\text{a.s.}} \sigma^2 \text{ by SLLN}$$

Note that,

$$\left|\mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leqslant x\right) - \Phi(x) + \Phi(x) - \mathbb{P}^*\left(\frac{\sqrt{n}\,(\bar{y}_n - \bar{x}_n)}{s_n} \leqslant x\right)\right|$$

$$\leq \left|\mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leq x\right) - \Phi(x)\right| + \left|\mathbb{P}^*\left(\frac{\sqrt{n}\,(\bar{y}_n - \bar{x}_n)}{s_n} \leq x\right) - \Phi(x)\right| \quad , \forall x \in \mathbb{R}$$

$$\leq \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leq x\right) - \Phi(x)\right| + \sup_{x\in\mathbb{R}}\left|\mathbb{P}^*\left(\frac{\sqrt{n}\,(\bar{y}_n - \bar{x}_n)}{s_n} \leq x\right) - \Phi(x)\right| \quad , \forall x \in \mathbb{R}$$

Proof Contd.

Resampling Technique: Bootstrap

$$\Rightarrow \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leq x \right) - \mathbb{P}^*\left(\frac{\sqrt{n}\,(\bar{y}_n - \bar{x}_n)}{s_n} \leq x \right) \right|$$

$$\leq \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leq x \right) - \Phi(x) \right| + \sup_{x\in\mathbb{R}} \left| \mathbb{P}^*\left(\frac{\sqrt{n}\,(\bar{y}_n - \bar{x}_n)}{s_n} \leq x \right) - \Phi(x) \right|$$

By Polya's Theorem, $\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{s_n} \leq x \right) - \Phi(x) \right| \to 0.$

Thus, it's enough to show $\sup_{x\in\mathbb{R}} \left| \mathbb{P}^*\left(\frac{\sqrt{n}(\bar{y}_n - \bar{x}_n)}{s_n} \leq x \right) - \Phi(x) \right| \to 0$ a.s.

By the fact that Lindeberg-Feller CLT holds provided,

$$\frac{1}{ns_n^2}\sum_{i=1}^{n}\mathbb{E}^*\left(y_i-\bar{x}_n\right)^2\mathbf{1}\left\{|y_i-\bar{x}_n|\geqslant\epsilon n^{1/2}s_n\right\}\to 0 \quad \text{a.s.}$$

Now note that,

$$\begin{aligned}&\frac{1}{ns_n^2}\sum_{i=1}^{n}\mathbb{E}^*\left(y_i-\bar{x}_n\right)^2\mathbf{1}\left\{|y_i-\bar{x}_n|\geqslant\epsilon n^{1/2}s_n\right\}\\ =\quad & s_n^{-2}\mathbb{E}^*\left(x-\bar{x}_n\right)^2\mathbf{1}\left\{|x-\bar{x}_n|\geqslant\epsilon n^{1/2}s_n\right\}\end{aligned}$$

Proof Contd.

Resampling Technique: Bootstrap

$$= \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{ |x_i - \bar{x}_n| \geqslant \epsilon n^{1/2} s_n \right\}$$

$$= \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{ |x_i - \mu + \mu - \bar{x}_n| \geqslant \epsilon n^{1/2} s_n \right\}$$

$$\leqslant \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{ |x_i - \mu| + |\mu - \bar{x}_n| \geqslant \epsilon n^{1/2} s_n \right\}$$

$$= \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{ |x_i - \mu| \geqslant \epsilon n^{1/2} s_n - |\mu - \bar{x}_n| \right\}$$

$$\leq \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{ |x_i| \geqslant \epsilon s_n n^{1/2} - |\mu - \bar{x}_n| - |\mu| \right\}$$

$$= \frac{1}{ns_n^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{x_i^2 \geqslant \left(\epsilon s_n n^{1/2} - |\mu - \bar{x}_n| + \mu\right)^2\right\}$$

Since $\bar{x}_n \xrightarrow{\text{a.s.}} \mu$ and $s_n^2 \xrightarrow{\text{a.s.}} \sigma^2$, to show

$$\frac{1}{ns_n^2} \sum_{i=1}^{n} \mathbb{E}^* (y_i - \bar{x}_n)^2 \mathbf{1}\left\{|y_i - \bar{x}_n| \geqslant \epsilon n^{1/2}\right\}$$

It is enough to show that

$$\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \mathbf{1}\left\{x_i^2 \geq \epsilon' n\right\} = o(n) \quad \forall \epsilon' > 0 - (*)$$

Consider an integer N and a function f defined on the unbounded interval $[N, \infty)$, on which it is monotone decreasing. Then the infinite series $\sum_{n=N}^{\infty} f(n)$ converges to a real number if and only if the improper integral $\int_N^{\infty} f(x)\, dx$ is finite.

$$\mathbb{E}\left(x^2\right) < \infty \Rightarrow \int_0^{\infty} \mathbb{P}\left(x^2 \geqslant t\right)\, dt < \infty$$

$$\Rightarrow \int_0^{\infty} \mathbb{P}\left(x^2 \geqslant t\epsilon\right)\, d(t\epsilon) < \infty$$

Therefore, $\sum_{i=1}^{\infty} P\left(x_i^2 \geqslant \epsilon i\right) < \infty$, hence $\{x_i^2 \geqslant \epsilon i\}$ happens only for finitely many i's almost surely. Thus the left-hand side of $(*)$ is in fact bounded almost surely:

$$\therefore \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2 \mathbf{1}\left\{x_i^2 \geqslant \epsilon n\right\} = o(n)$$

References

Resampling Technique: Bootstrap

- Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Statist 9 1187–1195
- The jackknife and bootstrap / Jun Shao, Dongsheng Tu. (Springer series in statistics)
- Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman Hall/CRC.

Thank You!

Any Questions?