Part III

Bootstrap and Resampling Methods

Resampling Approaches to Statistical Inference

Frequentist properties of statistical estimators are often difficult to study in a finite sample setting: for example, the finite sample variance of the sample median can only be computed via numerical integration.

Resampling methods allow the study of frequentist properties of statistical quantities by producing pseudo-replicate data sets of the same size as the observed data, and examining the statistical variation across these replicate data sets.

Notation: independence case

Suppose $Y_1, \ldots, Y_n \sim F$ are a random sample, and let $\theta = \theta(F)$ be the focus of inferential interest. For example

$$\theta(F) = \int y \, dF(y) \qquad \text{or} \qquad \theta(F) = \inf_y \{F(y) \geq p\}$$

etc. Let $y_1, \ldots, y_n$ be the observed data.

If $\widehat{F}_n$ is the empirical cdf,

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[y_i, \infty)}(y) \qquad\qquad d\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i\}}(y)$$

then a natural estimator of $\theta(F)$ is $T_n = \theta(\widehat{F}_n)$,

Notation: independence case

Recall that, for each y, under mild regularity conditions

$$\widehat{F}_n(y) \xrightarrow{p} F(y)$$

but also

$$\sup_y |\widehat{F}_n(y) - F(y)| \xrightarrow{p} 0$$

as $n \longrightarrow \infty$. Therefore

$$\theta(\widehat{F}_n) \xrightarrow{p} \theta(F)$$

which justifies (asymptotically) the use of the plug-in estimator.

Notation: independence case

The bias and variance of the estimator are

$$\begin{aligned} b_{T_n}(F) &= \mathbb{E}_F[T_n] - \theta(F) \\ v_{T_n}(F) &= \mathbb{V}ar_F[T_n] \end{aligned}$$

both of which depend on the true F.

We wish to study these properties of the estimator. In some cases, it is possible to study these quantities analytically. Suppose, however, $\widehat{\theta}$ is the solution of

$$\sum_{i=1}^{n} m(X_i; \theta) = 0$$

for some m-estimating function. The estimator is not analytically available, so its properties are hard to study.

The Bootstrap

For a random sample, $y_1, \ldots, y_n$, the bootstrap proceeds as follows:

1. Set B (the number of bootstrap resamples)
2. For $b = 1, \ldots, B$,
   (a) generate a sample of size n $y_1^{(b)}, \ldots, y_n^{(b)}$ at random **with replacement** from $\widehat{F}_n$
   (b) form the statistic of interest $t_n^{(b)}$
3. Summarize the resampled estimates

$$t_n^{(1)}, \ldots, t_n^{(B)}$$

   using the desired statistical summary, $s(.)$.

Example (Sample mean standard error)

Consider data

-0.1568	0.4439	-0.7865	-1.6531	-0.6037
-0.6231	0.9061	-0.8215	-1.2829	-0.3538

with sample mean -0.4932 and standard deviation 0.7572, yielding the estimated standard error of the mean (s.e.m.)

$$\widehat{\text{se}}(\widehat{\theta}) = \frac{0.7572}{\sqrt{10}} = 0.2395.$$

If the presumption of normality is made, then this is the *plug-in* estimate of the standard error

- the **true** s.e.m. is $\text{se}(\widehat{\theta}) = \sigma/\sqrt{n}$
- we estimate σ by the standard deviation under $\widehat{F}_n$.

Example (Sample mean standard error)

Set $B = 200$. For $b = 1, \ldots, B$, resample, and compute the appropriate summary. For example, for $b = 1$,

0.9061	0.4439	0.4439	-1.6531	-1.2829
-0.1568	-0.1568	-1.6531	0.4439	0.4439

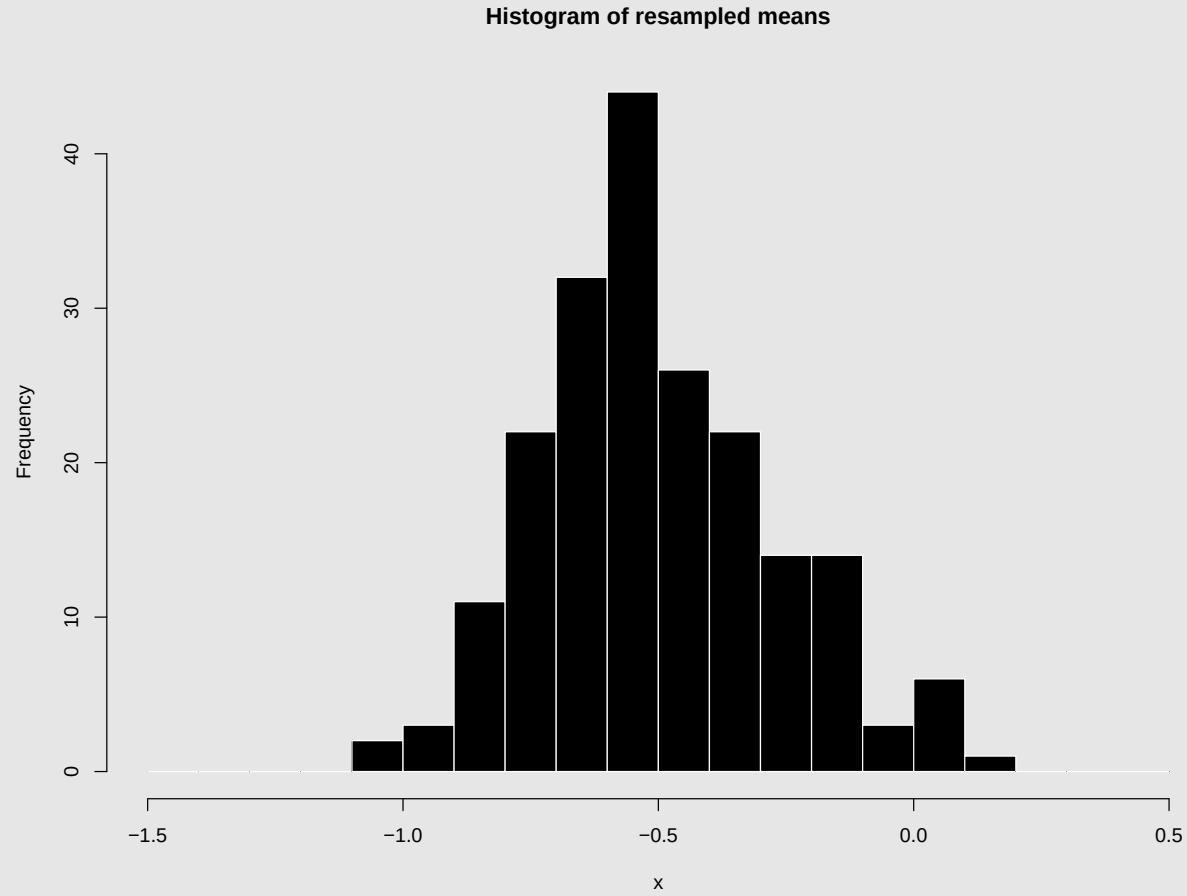
with sample mean -0.2221 and standard deviation 0.9589.

Example (Sample mean standard error)

b	Mean	Variance
1	-0.2221	0.9589
2	-0.7630	0.2068
3	-0.1794	0.7240
4	-0.5606	0.2927
5	-0.8929	0.4093
6	-0.5359	0.6291
7	-0.6969	0.6482
8	-0.4056	0.9639
⋮	⋮	⋮

Example (Sample mean standard error)

We now have 200 (resampled) sample means, with standard deviation $0.2327 \approx s/\sqrt{n}$, where s is the original sample s.d..



Example (Sample kurtosis standard error)

The sample kurtosis

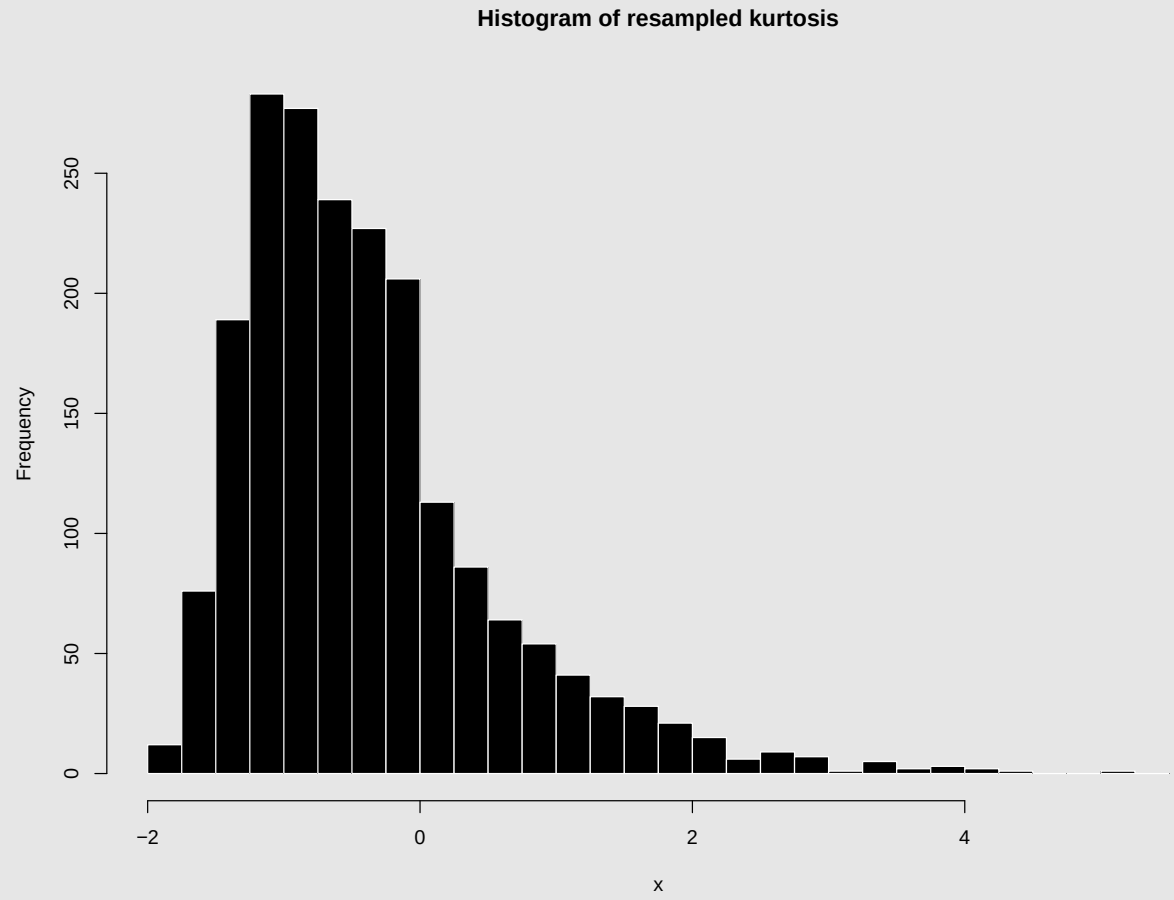
$$t_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^2} - 3$$

is a statistic that records the heavy-tailed nature of a sample.

There is no simple formula for the standard error for this statistic (although an asymptotic expression can be derived).

Example (Sample kurtosis standard error)

2000 (resampled) sample kurtosis values, with standard deviation 0.9689.



The Bootstrap: notation

Let $T_n = \widehat{\theta}$ be the statistic of interest, and let

$$s(\widehat{\theta}) \qquad \widehat{s}_B(\widehat{\theta}^{(*)})$$

respectively denote the true and bootstrap estimated value of summary quantity s of the sampling distribution of $\widehat{\theta}$.

For example, for the normal mean standard error, $\widehat{\theta} = \overline{y}$,

$$s(\widehat{\theta}) = \mathrm{se}(\widehat{\theta}) = \frac{\sigma}{\sqrt{n}} \qquad\qquad \widehat{\mathrm{se}}(\widehat{\theta}) = \frac{s}{\sqrt{n}}$$

and

$$\widehat{s}_B(\widehat{\theta}^{(*)}) = \widehat{\mathrm{se}}_B(\widehat{\theta}^{(*)}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\overline{y}^{(b)} - \overline{y}^{(*)})^2}$$

with

$$\overline{y}^{(*)} = \frac{1}{B}\sum_{b=1}^{B}\overline{y}^{(b)} \qquad\qquad \overline{y}^{(b)} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(b)}.$$

The Bootstrap: choosing B

For a sample of size n, there are

$$M = \binom{2n-1}{n}$$

possible distinct resampled data sets:

n	Number of distinct samples
10	9.238×10^4
20	6.892×10^{10}
50	5.045×10^{28}
100	4.527×10^{58}

How the value of B is chosen depends on the type of calculation being done, and the nature of the underlying data generating mechanism.

The Bootstrap: choosing B

Note that $B = \infty$ returns the plug-in estimate.

- For the s.e.m., at $B = \infty$

$$\widehat{\text{se}}_\infty(\widehat{\theta}^{(*)}) = \sqrt{\sum_{j=1}^{M} w_j(\overline{y}_j - \overline{y}_\cdot)^2}$$

  where $w_j, j = 1, \ldots, M$ represents the relative frequency of the resampled data sets yielding mean $\overline{y}_j$, and $\overline{y}_\cdot$ denotes the weighted mean

$$\overline{y}_\cdot = \frac{1}{M}\sum_{j=1}^{M} w_j \overline{y}_j.$$

- This quantity arises from the **exact** sampling distribution of the $\widehat{\theta}$, given $\widehat{F}_n$.
- Recall that there may not be a simple analytical formula for the required statistic.

The Bootstrap: choosing B

Let $\text{se}(\widehat{\theta})$ denote the standard error functional for statistic $\widehat{\theta}$, and, suppressing the dependence on $\widehat{\theta}$, let

- $\widehat{\text{se}}_B$ denote the estimate of the standard error from B bootstrap resamples
- $\widehat{\text{se}}_\infty$ denote the ideal estimated bootstrap standard error at $B = \infty$. Recall that $\widehat{\text{se}}_\infty \equiv \widehat{\text{se}}(\widehat{\theta})$ is the plug-in estimated standard error of $\widehat{\theta}$ (that is, the standard deviation of the sampling distribution of $\widehat{\theta}$ under $\widehat{F}_n$).

It can be shown that

$$\mathbb{E}_F[\widehat{\text{se}}_B^2] = \mathbb{E}_F[\widehat{\text{se}}_\infty^2]$$

and

$$\mathbb{V}ar_F[\widehat{\text{se}}_B^2] \geq \mathbb{V}ar_F[\widehat{\text{se}}_\infty^2]$$

The Bootstrap: choosing B

Example ($\mathcal{N}(\mu, \sigma^2)$)

The sample variance of $\widehat{\theta} = \overline{Y}$ is

$$\text{se}(\widehat{\theta})^2 = \frac{\sigma^2}{n}$$

estimated by

$$\widehat{\text{se}}(\widehat{\theta})^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}{n} = \frac{(n-1)s^2}{n^2}$$

But, by standard theory, under F

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The Bootstrap: choosing B

Example ($\mathcal{N}(\mu, \sigma^2)$)

Hence, as $\widehat{\text{se}}_\infty \equiv \widehat{\text{se}}$, we have

$$\mathbb{E}_F[\widehat{\text{se}}^2_\infty] = \frac{(n-1)\sigma^2}{n^2} \qquad \mathbb{V}ar_F[\widehat{\text{se}}^2_\infty] = \frac{2(n-1)\sigma^4}{n^4}$$

These results are **not** identical to those obtained taking expectations under $\widehat{F}_n$.

The larger B gets, we expect to see

$$\mathbb{V}ar_F[\widehat{\text{se}}^2_B] \longrightarrow \mathbb{V}ar_F[\widehat{\text{se}}^2_\infty]$$

Note that we cannot compute the limiting value as $B \longrightarrow \infty$ in many cases.

The Bootstrap: choosing B

Example (Simulation study)

Suppose F is $\mathcal{N}(0,1)$, so that

$$\mathbb{E}_F[\widehat{\text{se}}^2_\infty] = \frac{(n-1)}{n^2} \qquad \mathbb{V}ar_F[\widehat{\text{se}}^2_\infty] = \frac{2(n-1)}{n^4}.$$

For $n = 10$, these values are 0.090 and 0.0424^2 respectively.

Using 2000 replicate runs, for different values of B, the estimated following values were obtained:

B	50	100	200	500	1000	10000	20000	∞
Mean	0.0910	0.0910	0.0907	0.0906	0.0907	0.0907	0.0907	0.0900
s.d.	0.0469	0.0450	0.0437	0.0432	0.0428	0.0427	0.0427	0.0424

The Bootstrap: choosing B

Example (Simulation study)

For $n = 50$, these values are 0.0196 and 0.0040^2 respectively.

B	50	100	200	500	1000	10000	20000	∞
Mean	0.0195	0.0196	0.0197	0.0197	0.0197	0.0197	0.0197	0.0196
s.d.	0.0056	0.0049	0.0044	0.0042	0.0041	0.0040	0.0040	0.0040

The Bootstrap: choosing B

Note: the estimator used here for σ^2 is

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

that is, the **ML** estimator, not the **unbiased** estimator.

This estimator yields the "correct" variance for the bootstrap estimator at $B = \infty$, as S^2 is the correct plug-in estimator, that is, it is the variance under $\widehat{F}_n$.

The unbiased estimator could be used in practice; the difference is negligible for large n.