Bootstrap, Jackknife and other resampling methods

R. Dahyot

Web: https://roznn.github.io/
Twitter: @RDahyot

Introduction to resampling methods

- **Definitions and Problems**
- Non-Parametric Bootstrap
- Parametric Bootstrap
- Jackknife
- Permutation tests
- Cross-validation

References

Books

- An Introduction to Bootstrap, B. Efron and R. J. Tibshirani, Chapman & Hall, 1998.
- Bootstrap Methods and Their Application, A. Davidson and D. Hinkley, Cambridge University Press, 1997.
- Randomization, Bootstrapping, and Monte Carlo Methods in Biology, Manly, Chapman & Hall, 1997.

Special issues

- Silver Anniversary of the Bootstrap, Statistical Science, Vol. 18, nb. 2, May 2003.
- Signal Processing Applications Of The Bootstrap by S. Shamsunder; Computer-intensive methods in statistical analysis by D.N. Politis; The bootstrap and its application in signal processing, A.M. Zoubir and B. Boashash, in IEEE Signal Processing Magazine, January 1998.

The Empirical density function

Lets considerer a random sample (observations):

$$\mathrm{x} = (x_1, x_2, \cdots, x_n)$$

We wish to infer properties of the complete population $\mathcal{X}$ that yielded the sample. Lets define the population density function $f(.)$ such that

$$f \rightsquigarrow \mathrm{x} = (x_1, x_2, \cdots, x_n)$$

Definition

The empirical density function $\hat{f}(.)$ is defined as:

$$\hat{f}(x) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \delta(x - x_i)$$

where $\delta(\cdot)$ is the Dirac delta function.

Parameters

Definition

A parameter, θ, is a function of the probability density function (p.d.f.) f, e.g.:

$$\theta = t(f)$$

if θ is the mean

$$\theta = \mathbb{E}_f(x) = \int_{-\infty}^{+\infty} x \ f(x) dx = \mu_f$$

if θ is the variance

$$\theta = \mathbb{E}_f[(x - \mu_f)^2] = \int_{-\infty}^{+\infty} (x - \mu_f)^2 \ f(x) dx = \sigma_f^2$$

Statistics or estimates

Definition

A statistic (also called estimates, estimators) $\hat{\theta}$ is a function of $\hat{f}$ or the sample x, e.g.:

$$\hat{\theta} = t(\hat{f})$$

or also written $\hat{\theta} = s(\mathrm{x})$.

if $\hat{\theta}$ is the mean:

$$\begin{aligned}
\hat{\theta} \quad = t(\hat{f}) \quad &= \int_{-\infty}^{+\infty} x \ \hat{f}(x) dx \\
&= \int_{-\infty}^{+\infty} x \ \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \ dx \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= s(\mathrm{x}) = \overline{x}
\end{aligned}$$

Statistics or estimates

if $\hat{\theta}$ is the variance:

$$\begin{aligned}\hat{\theta} &= \int_{-\infty}^{+\infty} (x-\overline{x})^2 \ \hat{f}(x) dx \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ &= \hat{\sigma}^2\end{aligned}$$

The Plug-in principle

Definition

The Plug-in estimate of a parameter $\theta = t(f)$ is defined to be:

$$\hat{\theta} = t(\hat{f}).$$

The function $\theta = t(f)$ of the probability density function f is estimated by the same function $t(.)$ of the empirical density $\hat{f}$.

- $\overline{x}$ is the plug-in estimate of μ_f.
- $\hat{\sigma}$ is the plug-in estimate of σ_f

Computing the mean knowing f

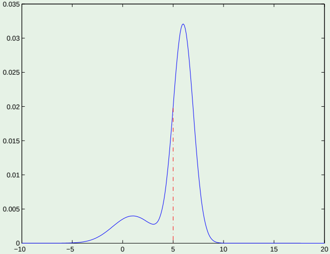
Example A

Lets assume we know the p.d.f. f:

$$f(x) = 0.2\ \mathcal{N}(\mu{=}1, \sigma{=}2) + 0.8\ \mathcal{N}(\mu{=}6, \sigma{=}1)$$

Then the mean is computed:

$$\begin{aligned}\mu_f = \mathbb{E}_f(x) &= \int_{-\infty}^{+\infty} x\ f(x)\ dx \\ &= 0.2 \cdot 1 + 0.8 \cdot 6 \\ &= 5\end{aligned}$$



Estimating the mean knowing the observations x

Example A

Observations $\mathrm{x} = (x_1, \cdots, x_{100})$:

$$\left\{\begin{array}{ccccc}
7.0411 & 4.8397 & 5.3156 & 6.7719 & 7.0616 \\
5.2546 & 7.3937 & 4.3376 & 4.4010 & 5.1724 \\
7.4199 & 5.3677 & 6.7028 & 6.2003 & 7.5707 \\
4.1230 & 3.8914 & 5.2323 & 5.5942 & 7.1479 \\
3.6790 & 0.3509 & 1.4197 & 1.7585 & 2.4476 \\
-3.8635 & 2.5731 & -0.7367 & 0.5627 & 1.6379 \\
-0.1864 & 2.7004 & 2.1487 & 2.3513 & 1.4833 \\
-1.0138 & 4.9794 & 0.1518 & 2.8683 & 1.6269 \\
6.9523 & 5.3073 & 4.7191 & 5.4374 & 4.6108 \\
6.5975 & 6.3495 & 7.2762 & 5.9453 & 4.6993 \\
6.1559 & 5.8950 & 5.7591 & 5.2173 & 4.9980 \\
4.5010 & 4.7860 & 5.4382 & 4.8893 & 7.2940 \\
5.5741 & 5.5139 & 5.8869 & 7.2756 & 5.8449 \\
6.6439 & 4.5224 & 5.5028 & 4.5672 & 5.8718 \\
6.0919 & 7.1912 & 6.4181 & 7.2248 & 8.4153 \\
7.3199 & 5.1305 & 6.8719 & 5.2686 & 5.8055 \\
5.3602 & 6.4120 & 6.0721 & 5.2740 & 7.2329 \\
7.0912 & 7.0766 & 5.9750 & 6.6091 & 7.2135 \\
4.9585 & 5.9042 & 5.9273 & 6.5762 & 5.3702 \\
4.7654 & 6.4668 & 6.1983 & 4.3450 & 5.3261
\end{array}\right\}$$

From the samples, the mean can be computed:

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{100} x_i}{100} \\
&= 4.9970
\end{aligned}$$

Accuracy of arbituary estimates $\hat{\theta}$

We can compute an estimate $\hat{\theta}$ of a parameter θ from an observation sample $\mathrm{x} = (x_1, x_2, \cdots, x_n)$. But

how accurate is $\hat{\theta}$ compared to the real value θ ?

Our attention is focused on questions concerning the probability distribution of $\hat{\theta}$. For instance we would like to know about:

- its standard error
- its confidence interval
- its bias
- etc.

Standard error of $\hat{\theta}$

Definition

The standard error is the standard deviation of a statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_f[\hat{\theta}]}$$

Standard error of $\overline{x}$

We have:

$$\mathbb{E}_f\left[(\overline{x} - \mu_f)^2\right] = \frac{\sum_{i=1}^{n} \mathbb{E}_f\left[(x_i - \mu_f)^2\right]}{n^2} = \frac{\sigma_f^2}{n}$$

Then

$$se_f(\overline{x}) = [\mathrm{var}_f(\overline{x})]^{1/2} = \frac{\sigma_f}{\sqrt{n}}$$

Plug in estimate of the standard error

Suppose now that f is unknown and that only the random sample $\mathrm{x} = (x_1, \cdots, x_n)$ is known. As μ_f and σ_f are unknown, we can use the previous formula to compute a plug-in estimate of the standard error.

Definition

The estimated standard error of the estimator $\hat{\theta}$ is defined as:

$$\hat{\mathrm{se}}(\hat{\theta}) = \mathrm{se}_{\hat{f}}(\hat{\theta}) = [\mathrm{var}_{\hat{f}}(\hat{\theta})]^{1/2}$$

Estimated standard error of $\overline{x}$

$$\hat{\mathrm{se}}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

Example on the mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery.

Did the treatment prolong survival ?

Example on the mouse data

Mean and Standard error for both groups

	$\overline{x}$	$\hat{\mathrm{se}}$
Treatment	86.86	25.24
Control	56.22	14.14

Conclusion at first glance

It seems that mice having the treatment survive $d = 86.86 - 56.22 = 30.63$ days more than the mice from the control group.

Example on the mouse data

Stantard error of the difference $d = \overline{x}_{Treat} - \overline{x}_{Cont}$

$\overline{x}_{Treat}$ and $\overline{x}_{Cont}$ are independent, so the standard error of their difference is $\hat{\mathrm{se}}(d) = \sqrt{\hat{\mathrm{se}}^2_{Treat} + \hat{\mathrm{se}}^2_{Cont}} = 28.93$. We see that:

$$\frac{d}{\hat{\mathrm{se}}(d)} = \frac{30.63}{28.93} = 1.05$$

This shows that this is an insignificant result as it could easily have arised by chance (i.e. if the test was reproduced, it is *likely possible* to measure datasets giving $d = 0$!).
Therefore, we can not conclude with certainty that the treatment improves the survival of the mice.

Confidence interval for $\hat{\theta}$

Definition

Assuming that the estimator $\hat{\theta}$ is normally distributed with unknown expectation θ and variance se^2, then :

$$\mathrm{Prob}\{\hat{\theta} - z^{(1-\alpha)}\mathrm{se} \leqslant \theta \leqslant \hat{\theta} - z^{(\alpha)}\mathrm{se}\} = 1 - 2\alpha$$

Therefore $1 - 2\alpha$ % confidence interval for θ is $[\hat{\theta} - z^{(1-\alpha)}\mathrm{se}; \hat{\theta} - z^{(\alpha)}\mathrm{se}]$ Confidence limits are the lower and upper boundaries values of a confidence interval. The confidence level is the probability value $100 \times (1 - 2\alpha)$ % associated with a confidence interval.

Confidence interval

The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter. A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.

percentile	confidence level	
$\alpha \times 100\ \%$	$(1 - 2\alpha) \times 100\ \%$	$z^{(1-\alpha)}$
10	80	1.28155
5	90	1.64485
2.5	95	1.95996
0.5	99	2.57583
0.25	99.5	2.80703
0.05	99.9	3.29053

Table: For a normal p.d.f $z^{(\alpha)} = -z^{(1-\alpha)}$

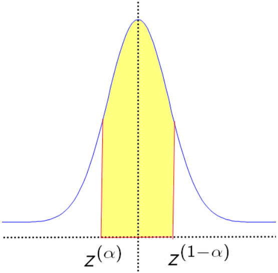


Figure: Density function $\mathcal{N}(0, 1)$.

Example Confidence interval

Confidence interval of the mean

Using the central limit theorem, the estimate $\overline{x}$ is following a normal density function $\mathcal{N}\left(\mu_f, \frac{\sigma_f^2}{n}\right)$. The 90% confidence interval is :

$$\overline{x} \pm 1.645 \frac{\sigma_f}{\sqrt{n}} \text{ estimated by} \pm 1.645 \frac{\hat{\sigma}}{\sqrt{n}}$$

confidence interval of the difference for the mouse data

The difference d in days of survival between the treatment group and the control group has a estimated 90% confidence interval defined as:

$$d = 30.63 \pm 1.645 \times 28.93 = 30.63 \pm 47.5898$$

Bias of $\hat{\theta}$

Definition

The Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$\mathrm{Bias}_f(\hat{\theta}, \theta) = \mathbb{E}_f(\hat{\theta}) - \theta$$

Bias of the mean $\overline{x}$

we have:

$$\mathbb{E}_f(\overline{x}) = \mathbb{E}_f\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{\sum_{i=1}^n \mathbb{E}_f(x_i)}{n} = \mu_f$$

then:

$$\mathrm{Bias}_f(\overline{x}, \mu_f) = \mathbb{E}_f(\overline{x}) - \mu_f = 0$$

Bias of $\hat{\theta}$

- A large bias is usually an undesirable aspect of an estimator's performance. Unbiased estimates (such $\mathbb{E}_f(\hat{\theta}) = \theta$) are interesting in practice as they promote a nice feeling of scientific objectivity in the estimation process.

Bias of $\hat{\sigma}^2$

$$\begin{aligned}\hat{\sigma}^2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2 = \tfrac{1}{n}\sum_{i=1}^{n}((x_i - \mu_f) + (\mu_f - \overline{x}))^2 \\ &= \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \mu_f)^2\right) - (\overline{x} - \mu_f)^2\end{aligned}$$

The first term has an expected value of σ_f^2 and the second term has expected value σ_f^2/n. So the bias of $\hat{\sigma}^2$ is:

$$\mathrm{Bias}_f(\hat{\sigma}^2, \sigma_f^2) = \sigma_f^2 - \frac{\sigma_f^2}{n} - \sigma_f^2 = -\frac{\sigma_f^2}{n}$$

Bias of $\hat{\theta}$

Instead of using $\hat{\sigma}^2$ as an estimate of the variance, you should try to choose an unbiased estimate.

Bias of $\overline{\sigma}^2$

Let define:

$$\overline{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

then by computing its bias:

$$\begin{aligned} \mathrm{Bias}_f(\overline{\sigma}^2, \sigma_f^2) &= \mathbb{E}_f(\overline{\sigma}^2) - \sigma_f^2 \\ &= 0 \end{aligned}$$

$\overline{\sigma}$ is an unbiased estimator of the standard deviation.

Summary

- Population density function $f(\cdot)$ and empirical density function $\hat{f}(\cdot)$,
- Plug-in principle: relation between θ and its estimate $\hat{\theta}$
- Standard error and confidence interval as a measure of accuracy of the estimate $\hat{\theta}$
- Accuracy of estimate is important to draw conclusions (e.g. mouse example).
- se has an explicit expression for the mean $\overline{x}$

Open problems

- f is generally unknown!

- Using the Plug-in principle, with more samples $\{\mathrm{x}\}$ drawn from f, we could estimate $\hat{se}$ or the bias.

- But the only information available is one sample $\mathrm{x} = (x_1, \cdot, x_n)$ drawn from f!

- Most of all, explicit expression of se of the estimate is not easy to get in most cases!

Introduction to resampling methods

- Definitions and Problems
- **Non-Parametric Bootstrap**
- Parametric Bootstrap
- Jackknife
- Permutation tests
- Cross-validation

Introduction

We want to assess the accuracy (bias, standard error, etc.) of an arbitrary estimate $\hat{\theta}$ knowing only one sample $\mathrm{x} = (x_1, \cdots, x_n)$ drawn from an unknown population density function f.

- We propose here one way, called *Bootstrap*, to do it using computer intensive techniques for resampling.

- Bootstrap is a data based simulation method for statistical inference. The basic idea of bootstrap is to use the sample data to compute a statistic and to estimate its sampling distribution, without any model assumption.

- No theoretical calculations of standard errors needed so we don't care how mathematically complex the estimator $\hat{\theta}$ can be!

Introduction

- The (non-parametric) bootstrap method is an application of the plug-in principle. By *non-parametric*, we mean that only x is known (observed) and no prior knowledge on the population density function f is available.

- Originally, the Bootstrap was introduced to compute standard error of an arbitrary estimator by Efron (1979) and to-date the basic idea remains the same.
- The term bootstrap derives from the phrase *to pull oneself up by one's bootstrap* (Adventures of Baron Munchausen, by Rudolph Erich Raspe). The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

Bootstrap samples and replications

Definition

A bootstrap sample $\mathrm{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $\mathrm{x} = (x_1, x_2, \cdots, x_n)$.

Considering a sample $\mathrm{x} = (x_1, x_2, x_3, x_4, x_5)$, some bootstrap samples can be:

$$\begin{aligned} \mathrm{x}^{*(1)} &= (x_2, x_3, x_5, x_4, x_5) \\ \mathrm{x}^{*(2)} &= (x_1, x_3, x_1, x_4, x_5) \\ &\text{etc.} \end{aligned}$$

Definition

With each bootstrap sample $\mathrm{x}^{*(1)}$ to $\mathrm{x}^{*(B)}$, we can compute a bootstrap replication $\hat{\theta}^*(b) = s(\mathrm{x}^{*(b)})$ using the plug-in principle.

How to compute Bootstrap samples

Repeat B times:

1. A random number device selects integers $i_1, \cdots, i_n$ each of which equals any value between 1 and n with probability $\frac{1}{n}$.
2. Then compute $\mathrm{x}^* = (x_{i_1}, \cdots, x_{i_n})$.

Some matlab code available on the web

See BOOTSTRAP MATLAB TOOLBOX, by Abdelhak M. Zoubir and D. Robert Iskander,
http://www.csp.curtin.edu.au/downloads/bootstrap_toolbox.html

How many values are left out of a bootstrap resample ?

Given a sample $\mathrm{x} = (x_1, x_2, \cdots, x_n)$ and assuming that all x_i are different, the probability that a particular value x_i is left out of a resample $\mathrm{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is:

$$\mathcal{P}(x_j^* \neq x_i, 1 \leqslant j \leqslant n) = \left(1 - \frac{1}{n}\right)^n$$

since $\mathcal{P}(x_j^* = x_i) = \frac{1}{n}$. When n is large, the probability $\left(1 - \frac{1}{n}\right)^n$ converges to $e^{-1} \approx 0.37$.

The Bootstrap algorithm for Estimating standard errors

1. Select B independent bootstrap samples $\mathrm{x}^{*(1)}, \mathrm{x}^{*(2)}, \cdots, \mathrm{x}^{*(B)}$ drawn from x
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathrm{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Estimate the standard error $\mathrm{se}_f(\hat{\theta})$ by the standard deviation of the B replications:

$$\hat{\mathrm{se}}_B = \left[\frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right]^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$

Bootstrap estimate of the standard Error

Example A

From the distribution f: $f(x) = 0.2\ \mathcal{N}(\mu{=}1, \sigma{=}2) + 0.8\ \mathcal{N}(\mu{=}6, \sigma{=}1)$. We draw the sample $\mathrm{x} = (x_1, \cdots, x_{100})$:

$$
\mathrm{x} = \left\{ \begin{array}{ccccc}
7.0411 & 4.8397 & 5.3156 & 6.7719 & 7.0616 \\
5.2546 & 7.3937 & 4.3376 & 4.4010 & 5.1724 \\
7.4199 & 5.3677 & 6.7028 & 6.2003 & 7.5707 \\
4.1230 & 3.8914 & 5.2323 & 5.5942 & 7.1479 \\
3.6790 & 0.3509 & 1.4197 & 1.7585 & 2.4476 \\
-3.8635 & 2.5731 & -0.7367 & 0.5627 & 1.6379 \\
-0.1864 & 2.7004 & 2.1487 & 2.3513 & 1.4833 \\
-1.0138 & 4.9794 & 0.1518 & 2.8683 & 1.6269 \\
6.9523 & 5.3073 & 4.7191 & 5.4374 & 4.6108 \\
6.5975 & 6.3495 & 7.2762 & 5.9453 & 4.6993 \\
6.1559 & 5.8950 & 5.7591 & 5.2173 & 4.9980 \\
4.5010 & 4.7860 & 5.4382 & 4.8893 & 7.2940 \\
5.5741 & 5.5139 & 5.8869 & 7.2756 & 5.8449 \\
6.6439 & 4.5224 & 5.5028 & 4.5672 & 5.8718 \\
6.0919 & 7.1912 & 6.4181 & 7.2248 & 8.4153 \\
7.3199 & 5.1305 & 6.8719 & 5.2686 & 5.8055 \\
5.3602 & 6.4120 & 6.0721 & 5.2740 & 7.2329 \\
7.0912 & 7.0766 & 5.9750 & 6.6091 & 7.2135 \\
4.9585 & 5.9042 & 5.9273 & 6.5762 & 5.3702 \\
4.7654 & 6.4668 & 6.1983 & 4.3450 & 5.3261
\end{array} \right\}
$$

We have $\mu_f = 5$ and $\overline{x} = 4.9970$.

Bootstrap estimate of the standard Error

Example A

1. $B = 1000$ bootstrap samples $\{\mathrm{x}^{*(b)}\}$
2. $B = 1000$ replications $\{\overline{x}^*(b)\}$
3. Bootstrap estimate of the standard error:

$$\widehat{\mathrm{se}}_{B=1000} = \left[\frac{\sum_{b=1}^{1000}[\overline{x}^*(b) - \overline{x}^*(\cdot)]^2}{1000-1}\right]^{\frac{1}{2}} = 0.2212$$

where $\overline{x}^*(\cdot) = 5.0007$. This is to compare with $\hat{se}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = 0.22.$

Distribution of $\hat{\theta}$

When enough bootstrap resamples have been generated, not only the standard error but any aspect of the distribution of the estimator $\hat{\theta} = t(\hat{f})$ could be estimated. One can draw a histogram of the distribution of $\hat{\theta}$ by using the observed $\hat{\theta}^*(b),\ b = 1, \cdots, B$.

Example A

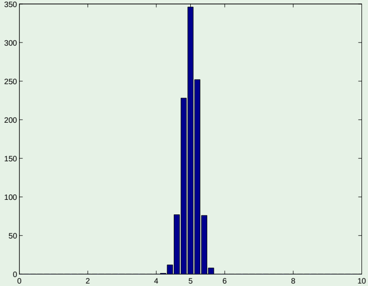


Figure: Histogram of the replications $\{\overline{x}^*(b)\}_{b=1\cdots B}$.

Bootstrap estimate of the standard error

Definition

The ideal bootstrap estimate $\mathrm{se}_{\hat{f}}(\hat{\theta}^*)$ is defined as:

$$\lim_{B\to\infty} \hat{\mathrm{se}}_B = \mathrm{se}_{\hat{f}}(\hat{\theta}^*)$$

$\mathrm{se}_{\hat{f}}(\hat{\theta}^*)$ is called a non-parametric bootstrap estimate of the standard error.

Bootstrap estimate of the standard Error

How many B in practice ?

you may want to limit the computation time. In practice, you get a good estimation of the standard error for B in between 50 and 200.

Example A

B	10	20	50	100	500	1000	10000
$\widehat{\mathrm{se}}_B$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187

Table: Bootstrap standard error w.r.t. the number B of bootstrap samples.

Bootstrap estimate of bias

Definition

The bootstrap estimate of bias is defined to be the estimate:

$$
\begin{aligned}
\mathrm{Bias}_{\hat{f}}(\hat{\theta}) &= \mathbb{E}_{\hat{f}}[s(\mathrm{x}^*)] - t(\hat{f}) \\
&= \hat{\theta}^*(\cdot) - \hat{\theta}
\end{aligned}
$$

Example A

B	10	20	50	100	500	1000	10000
$\mathbb{E}_{\hat{f}}(\overline{x}^*)$	5.0587	4.9551	5.0244	4.9883	4.9945	5.0035	4.9996
$\widehat{\mathrm{Bias}}$	0.0617	-0.0419	0.0274	-0.0087	-0.0025	0.0064	0.0025

Table: $\widehat{\mathrm{Bias}}$ of $\overline{x}^*$ ($\overline{x} = 4.997$ and $\mu_f = 5$).

Bootstrap estimate of bias

1. B independent bootstrap samples $\mathrm{x}^{*(1)}, \mathrm{x}^{*(2)}, \cdots, \mathrm{x}^{*(B)}$ drawn from x
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathrm{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Approximate the bootstrap expectation :

$$\hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b) = \frac{1}{B} \sum_{b=1}^{B} s(\mathrm{x}^{*(b)})$$

4. the bootstrap estimate of bias based on B replications is:

$$\widehat{\mathrm{Bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$$

Confidence interval

Definition

Using the bootstrap estimation of the standard error, the $100(1-2\alpha)\%$ confidence interval is:

$$\theta = \hat{\theta} \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

Definition

If the bias in not null, the bias corrected confidence interval is defined by:

$$\theta = (\hat{\theta} - \widehat{\mathrm{Bias}}_B) \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

Can the bootstrap answer other questions?

The mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice divided assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery. Did the treatment prolong survival ?

Can the bootstrap answer other questions?

The mouse data

- Remember in the first lecture, we compute $d = \overline{x}_{Treat} - \overline{x}_{Cont} = 30.63$ with a standard error $\hat{se}(d) = 28.93$. The ratio was $d/\hat{se}(d) = 1.05$ (an insignificant result as measuring $d = 0$ is likely possible).
- Using bootstrap method
  1. B bootstrap samples $\mathrm{x}^{*(b)}_{Treat} = (x^{*(b)}_{Treat\ 1}, \cdots, x^{*(b)}_{Treat\ 7})$ and $\mathrm{x}^{*(b)}_{Cont} = (x^{*(b)}_{Cont\ 1}, \cdots, x^{*(b)}_{Cont\ 9})$, $\forall 1 \leqslant b \leqslant B$
  2. B bootstrap replications are computed: $d^*(b) = \overline{x}^{*(b)}_{Treat} - \overline{x}^{*(b)}_{Cont}$
  3. The bootstrap standard error is computed for $B = 1400$: $\hat{se}_{B=1400} = 26.85$.
  4. The ratio is $d/\hat{se}_{1400}(d) = 1.14$.
- This is still not a significant result.

The Law school example

School	1	2	3	4	5	6	7	8
LSAT (X)	576	635	558	578	666	580	555	661
GPA (Y)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43

School	9	10	11	12	13	14	15
LSAT (X)	651	605	653	575	545	572	594
GPA (Y)	3.36	3.13	3.12	2.74	2.76	2.88	2.96

Table: Results of law schools admission practice for the LSAT and GPA tests. It is believed that these scores are highly correlated. Compute the correlation and its standard error.

Correlation

The correlation is defined :

$$\mathrm{corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]}{\left(\mathbb{E}[(X - \mathbb{E}(X))^2] \cdot \mathbb{E}[(Y - \mathbb{E}(Y))^2]\right)^{1/2}}$$

Its typical estimator is:

$$\widehat{\mathrm{corr}}(\mathrm{x}, \mathrm{y}) = \frac{\sum_{i=1}^{n} x_i \; y_i - n \; \overline{x} \; \overline{y}}{[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2]^{1/2} \cdot [\sum_{i=1}^{n} y_i^2 - n\overline{y}^2]^{1/2}}$$

The Law school example

- The estimated correlation is $\widehat{\text{corr}}(\mathrm{x}, \mathrm{y}) = .7764$ between LSAT and GPA.

Non-parametric Bootstrap estimate of the standard error

B	25	50	100	200	400	800	1600	3200
$\hat{\text{se}}_B$	.140	.142	.151	.143	.141	.137	.133	.132

Table: Bootstrap estimate of standard error for $\widehat{\text{corr}}(\mathrm{x}, \mathrm{y}) = .776$.

The standard error stabilizes to $\text{se}_{\hat{f}}(\widehat{\text{corr}}) \approx .132$.

The Law school example: Conclusion

- The textbook formula for the correlation coefficient is:

$$\hat{\mathrm{se}}(\widehat{\mathrm{corr}}) = (1 - \widehat{\mathrm{corr}}^2)/\sqrt{n-3}$$

- With $\widehat{\mathrm{corr}} = 0.7764$, the standard error is $\hat{se}(\widehat{\mathrm{corr}}) = 0.1147$.
- The estimated non-parametric bootstrap standard error $\mathrm{se}_{B=3200}$ is 0.132.

Summary

- Re-sampling of x to compute bootstrap samples x^*
- Computation of bootstrap replication of the estimator $\hat{\theta}^*(b)$ for $b = 1, \cdots, B$
- From replications, standard error $\widehat{\mathrm{se}}_B$, the bias $\widehat{\mathrm{Bias}}_B$ and the confidence interval.
- Non-parametric bootstrap estimations (no prior on f).

Introduction to resampling methods

- Definitions and Problems
- Non-Parametric Bootstrap
- **Parametric Bootstrap**
- Jackknife
- Permutation tests
- Cross-validation

Introduction

1. Nonparametric bootstrap estimates
2. Example of failure of the nonparametric bootstrap estimate
3. Parametric Bootstrap
4. Resampling and Monte Carlo Sampling
5. The law school example

Non-Parametric Bootstrap

Real World *Bootstrap World*

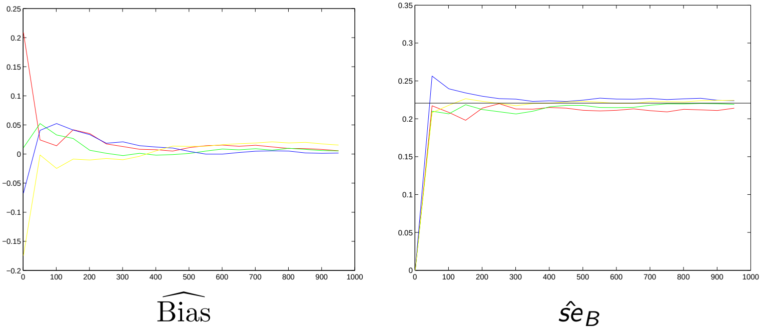
$$\begin{tikzcd}
f \arrow[r] & \mathrm{x} \arrow[r, Rightarrow] \arrow[d] & \hat{f} \arrow[r] & \mathrm{x}^* \arrow[d] \\
 & \hat{\theta} & & \hat{\theta}^*
\end{tikzcd}$$

Figure: Unknown probability model f gives observed data x and we wish to know the accuracy of the statistic $\hat{\theta} = s(\mathrm{x})$ for estimating the parameter of interest $\theta = t(f)$. No prior information is available on f, therefore $\hat{f}$ is estimated from x as the empirical distribution function. Accuracy is inferred from observed variability of bootstrap replication $\hat{\theta}^* = s(\mathrm{x}^*)$.

Convergence of the bootstrap estimates

Example A

$f(x) = 0.2\ \mathcal{N}(\mu{=}1, \sigma{=}2) + 0.8\ \mathcal{N}(\mu{=}6, \sigma{=}1) \rightsquigarrow x = (x_1, \cdots, x_{100}).$



$\widehat{\text{Bias}}$

$\hat{se}_B$

Figure: Bias and standard error bootstrap estimates w.r.t. B (4 experiments have been run).

Example of non-parametric bootstrap failure

Example B

Considering a sample x drawn from a uniform distribution $f = \mathcal{U}(0, \theta = 1)$, the statistics of interest is $\hat{\theta} = \max\{x_1, \cdots, x_n\}$, and x=(0.5729,0.1873,0.5984,0.2883,0.8722, 0.4320,0.4896,0.7106,0.2754,0.7637).

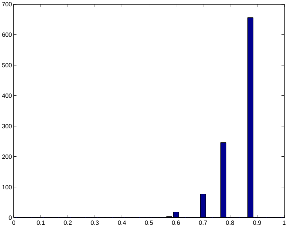


Figure: Histogram of the nonparametric bootstrap replications $\hat{\theta}^*$ with $n = 10$, $B = 1000$, $\hat{\theta} = 0.8722$. The maximum peak is at $\hat{\theta} = 0.8722$ with a probability of $\mathcal{P}(\hat{\theta} \in \mathrm{x}^*) = 0.6560 \approx 1 - (1 - 1/n)^n = 0.6513$.

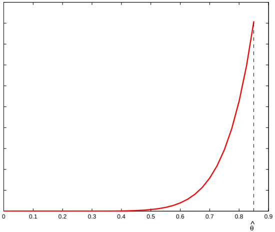


Figure: Theoretical results (extreme values) says that $\mathcal{P}(\hat{\theta}^*) = n\frac{(\hat{\theta}^*)^{n-1}}{\hat{\theta}^n}$.

Example of non-parametric bootstrap failure

Example B

What went wrong in this example ?

- The empirical density function $\hat{f}$ is not a good approximation of the true distribution $f = \mathcal{U}(0, \theta)$.
- Either parametric knowledge of f or some smoothing of $\hat{f}$ is needed to rectify matters.

Convergence of the bootstrap estimates

With $\mathrm{x} = (x_1, \cdots, x_n)$, n i.i.d. values, it is required:

1. Convergence of $\hat{f}$ to f for $n \to \infty$ (Glivenko-Cantelli lemma)
2. Estimate $\hat{\theta} = t(\hat{f})$ is the plug-in estimate of $\theta = t(f)$
3. Smoothness condition on the functional. E.g
   - Smooth functionals: means, variance, etc.
   - Not smooth: extreme order statistics (minimum, maximum)

Parametric Bootstrap

$$\begin{tikzcd}
 & \textit{Real World} & \textit{Estimation} & & \textit{Bootstrap World} \\
\textit{Prior} : f \simeq \mathcal{N}(\mu, \sigma) \arrow[r] & \mathrm{x} \arrow[r, Rightarrow] \arrow[d] & (\overline{x}, \hat{\sigma}) & \hat{f} \simeq \mathcal{N}(\overline{x}, \hat{\sigma}) \arrow[r] & \mathrm{x}^* \arrow[d] \\
 & \hat{\theta} & & & \hat{\theta}^*
\end{tikzcd}$$

Figure: Example of parametric Bootstrap. f is a normal distribution of unknown parameters (μ, σ). From the observed data x drawn from f, an estimation of the parameters is performed giving $(\overline{x}, \hat{\sigma})$. $\hat{f}$ is then modelled by a normal distribution $\mathcal{N}(\overline{x}, \hat{\sigma})$, from which bootstrap replications can be drawn x^*. Accuracy is inferred from observed variability of bootstrap replication $\hat{\theta}^* = s(\mathrm{x}^*)$.

Example with extreme value

Example B

We draw $B = 1000$ bootstrap replication of $\hat{\theta}^* = \max\{x^*\}$ using the parametric assumption $\mathcal{U}(0, \hat{\theta})$.

The extreme value distribution is $\mathcal{P}(\hat{\theta}^*) = n\frac{(\hat{\theta}^*)^{n-1}}{\hat{\theta}^n}$.

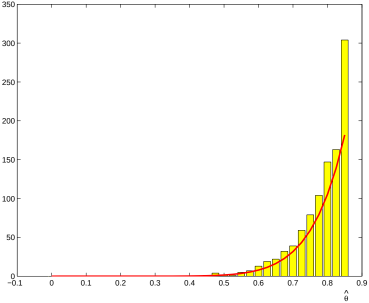


Figure: Histogram of the parametric bootstrap replications $\hat{\theta}^*$ with $n = 10,\ B = 1000,\ \hat{\theta} = 0.8722$.

The Law school example

School	1	2	3	4	5	6	7	8
LSAT (X)	576	635	558	578	666	580	555	661
GPA (Y)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43

School	9	10	11	12	13	14	15
LSAT (X)	651	605	653	575	545	572	594
GPA (Y)	3.36	3.13	3.12	2.74	2.76	2.88	2.96

Table: Results of law schools admission practice for the LSAT and GPA tests. It is believed that these scores are highly correlated. Compute the correlation and its standard error.

Correlation

The correlation is defined :

$$\mathrm{corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]}{\left(\mathbb{E}[(X - \mathbb{E}(X))^2] \cdot \mathbb{E}[(Y - \mathbb{E}(Y))^2]\right)^{1/2}}$$

Its typical estimator is:

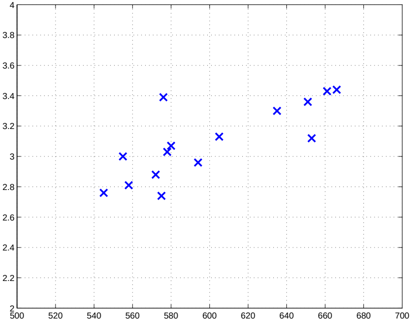
$$\widehat{\mathrm{corr}}(\mathrm{x}, \mathrm{y}) = \frac{\sum_{i=1}^{n} x_i \ y_i - n \ \overline{x} \ \overline{y}}{[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2]^{1/2} \cdot [\sum_{i=1}^{n} y_i^2 - n\overline{y}^2]^{1/2}}$$

The Law school example

Parametric Bootstrap approach

Assuming that f is a bivariate normale distribution, $\hat{f}_{norm}$ is estimated by computing the mean $\bar{z} = (\bar{x}, \bar{y})$ and the covariance matrix $\widehat{\Sigma}$ from the data.

Then B samples $(\mathrm{x}, \mathrm{y})^*$ can be drawn from $\hat{f}_{par}$ and the bootstrap estimate of the correlation coefficient can be performed.



The Law school example: Parametric Approach

Prior model

- **Assumption.** f is a bivariate normal density function of the form:

$$f(x,y) = \frac{\exp\left[-\frac{(\mathrm{z}-\boldsymbol{\mu}_{zf})^T\Sigma^{-1}(\mathrm{z}-\boldsymbol{\mu}_{zf})}{2}\right]}{(2\pi)|\det(\Sigma)|^{1/2}}$$

with $\mathrm{z} = \begin{pmatrix} x \\ y \end{pmatrix}$

- **Problem.** The parameters, mean $\boldsymbol{\mu}_{zf} = (\mu_{xf}, \mu_{yf})$ and the covariance matrix Σ, are unknown.

The Law school example: Parametric Approach

Estimate of the parametric p.d.f.

The parametric p.d.f. is estimated by:

$$\hat{f}_{par}(x,y) = \frac{\exp\left[-\frac{(\mathrm{z}-\overline{\mathrm{z}})^T\overline{\Sigma}^{-1}(\mathrm{z}-\overline{\mathrm{z}})}{2}\right]}{(2\pi)|\det(\overline{\Sigma})|^{1/2}}$$

Means are $\overline{\mathrm{z}} = (\overline{x} = \frac{\sum_{i=1}^n x_i}{n}, \overline{y} = \frac{\sum_{i=1}^n y_i}{n})$. The covariance matrix is defined as:

$$\overline{\Sigma} = \frac{1}{n-1}\begin{bmatrix} \sum_{i=1}^n (x_i - \overline{x})^2 & \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \\ \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) & \sum_{i=1}^n (y_i - \overline{y})^2 \end{bmatrix}$$

The mean is $\overline{\mathrm{z}} = (\overline{x} = 600.3, \overline{y} = 3.1)$ and $\overline{\Sigma} = \begin{bmatrix} 1747 & 0.0079 \\ 0.0079 & 0.0001 \end{bmatrix}$

Parametric Bootstrap estimate of standard error

1. Using the prior assumption and the available observation $\mathrm{x} = (x_1, x_2, \cdots, x_n)$, estimate $\hat{f}_{par}$
2. Instead of sampling with replacement from the data x, draw B samples $\mathrm{x}^{*(b)}$ of size n from $\hat{f}_{par}$
3. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathrm{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

4. Estimate the standard error $\mathrm{se}_f(\hat{\theta})$ by the standard deviation of the B replications:

$$\hat{\mathrm{se}}_B = \left[\frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right]^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$.

Parametric Bootstrap estimate of the bias

1. Using the prior assumption and the available observation $\mathrm{x} = (x_1, x_2, \cdots, x_n)$, estimate $\hat{f}_{par}$
2. Instead of sampling with replacement from the data x, draw B samples $\mathrm{x}^{*(b)}$ of size n from $\hat{f}_{par}$
3. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathrm{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

4. Estimate the bias:

$$\widehat{\mathrm{Bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$.

Resampling and Monte Carlo Simulation

- In resampling, one could do all possible combinations, but it would be too time-consuming and computing-intensive.

- The alternative is Monte Carlo sampling, which restricts the resampling to a certain number. It is used in the computation of the bootstrap samples in the nonparametric case when a random index $(i_1, \cdots, i_n)$ is simulated from the uniform distribution $[1; n]$ $(\mathrm{x}^* = (x_{i_1}, \cdots, x_{i_1}))$.

- The data could be totally hypothetical in Monte Carlo simulation, while in the resampling, the simulation is based upon some real observation $\mathrm{x} = (x_1, \cdots, x_n)$.

- In the parametric case, the bootstrap samples from $\hat{f}_{par}$ are computed using Monte Carlo methods and are not anymore resamples from x.

The Law school example: Parametric Bootstrap

For $B = 3200$ bootstrap replications, we compute $\widehat{\mathrm{corr}}^*(\cdot) = 0.7661$ and the parametric bootstrap standard error 0.1169.

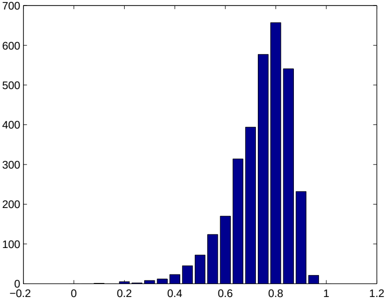


Figure: Histogram of 3200 parametric bootstrap replication of corr $\widehat{\mathrm{corr}}(x^*, y^*)$.

The Law school example: Conclusion

- The textbook formula for the correlation coefficient is:

$$\mathrm{se}_{\hat{f}} = (1 - \widehat{\mathrm{corr}}^2)/\sqrt{n-3}$$

  With $\widehat{\mathrm{corr}}(\mathrm{x}, \mathrm{y}) = 0.7764$, the standard error is $se_f = 0.1147$.
- The non-parametric bootstrap standard error for $B = 3200$ is 0.132.
- The parametric bootstrap standard error for $B = 3200$ is 0.1169.

Parametric and nonparametric bootstrap estimates

- The nonparametric approach leads to a finite number of possible replications $\hat{\theta}^*(b)$. In fact considering n distinct values in $\mathrm{x} = (x_1, \cdots, x_n)$, the maximum number of *different* bootstrap samples (and replications) is [1]:

$$B_{max} = \left(\begin{array}{c} 2n-1 \\ n-1 \end{array} \right) = \frac{(2n-1)!}{n!(n-1)!}$$

- The parametric approach has an unlimited number of different bootstrap samples and replications.

[1] $n = 11$, $B_{max} = 652716$: big enough to minimize the effect of the discreteness in the nonparametric approach.

When might the (parametric and non parametric) bootstrap fail?

Bootstrap might fail when:

- incomplete data (missing data) : incomplete observation x
- dependent data (e.g. correlated time series) $\mathrm{x} = (x_1, \cdots, x_n)$ dependent
- dirty data (outliers) : noisy observation x

For a critical view on bootstrap, see the publication *Exploring the limits of bootstrap* edited by Le Page and Billard 1990 (ISBN: 0-471-53631-8).

Conclusion

- In parametric bootstrap, $\hat{f}_{par}$ is not anymore the empirical density function.

- If the prior information on f is accurate, then $\hat{f}_{par}$ estimates better f than the empirical p.d.f.. In this case the parametric bootstrap gives better estimation for the standard errors.

- Most of the time, the point of making assumptions is to derive the textbook formulas.
  All models are wrong, but some are useful- G.E.P. Box, 1979.

- On the other hand, non-parametric bootstrap allows the computation of accurate standard errors (in many cases) without making any prior assumption.

Conclusion

- In non-parametric mode, the bootstrap method relieves the analyst from choosing a parametric assumption about the form of the underlying density function f.

- In both case, bootstrap can provide answers for problem for which no textbook formulae exists.

Introduction to resampling methods

- Definitions and Problems
- Non-Parametric Bootstrap
- Parametric Bootstrap
- **Jackknife**
- Permutation tests
- Cross-validation

Introduction

The bootstrap method is not always the best one. One main reason is that the bootstrap samples are generated from $\hat{f}$ and not from f. Can we find samples/resamples exactly generated from f?

- If we look for samples of size n, then the answer is no!
- If we look for samples of size m ($m < n$), then we can indeed find (re)samples of size m exactly generated from f simply by looking at different subsets of our original sample x!

Looking at different subsets of our original sample amounts to sampling without replacement from observations $x_1, \cdots, x_n$ to get (re)samples (now called subsamples) of size m. This leads us to subsampling and the jackknife.

Jackknife

- The jackknife has been proposed by Quenouille in mid 1950's.
- In fact, the jackknife predates the bootstrap.
- The jackknife (with $m = n - 1$) is less computer-intensive than the bootstrap.
- *Jackknife* describes a swiss penknife, easy to carry around. By analogy, Tukey (1958) coined the term in statistics as a general approach for testing hypotheses and calculating confidence intervals.

Jackknife samples

Definition

The Jackknife samples are computed by leaving out one observation x_i from $\mathrm{x} = (x_1, x_2, \cdots, x_n)$ at a time:

$$\mathrm{x}_{(i)} = (x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$$

- The dimension of the jackknife sample $\mathrm{x}_{(i)}$ is $m = n - 1$
- n different Jackknife samples : $\{\mathrm{x}_{(i)}\}_{i=1 \cdots n}$.
- No sampling method needed to compute the n jackknife samples.

Available BOOTSTRAP MATLAB TOOLBOX, by Abdelhak M. Zoubir and D. Robert Iskander,
http://www.csp.curtin.edu.au/downloads/bootstrap_toolbox.html

Jackknife replications

Definition

The ith jackknife replication $\hat{\theta}_{(i)}$ of the statistic $\hat{\theta} = s(\mathrm{x})$ is:

$$\hat{\theta}_{(i)} = s(\mathrm{x}_{(i)}), \quad \forall i = 1, \cdots, n$$

Jackknife replication of the mean

$$\begin{aligned} s(\mathrm{x}_{(i)}) &= \tfrac{1}{n-1} \textstyle\sum_{j \neq i} x_j \\ &= \tfrac{(n\overline{x} - x_i)}{n-1} \\ &= \overline{x}_{(i)} \end{aligned}$$

Jackknife estimation of the standard error

1. Compute the n jackknife subsamples $\mathrm{x}_{(1)}, \cdots, \mathrm{x}_{(n)}$ from x.

2. Evaluate the n jackknife replications $\hat{\theta}_{(i)} = s(\mathrm{x}_{(i)})$.

3. The jackknife estimate of the standard error is defined by:

$$\hat{\mathrm{se}}_{jack} = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2 \right]^{1/2}$$

where $\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$.

Jackknife estimation of the standard error of the mean

For $\hat{\theta} = \overline{x}$, it is easy to show that:

$$\begin{cases} \overline{x}_{(i)} = \frac{n\overline{x} - x_i}{n-1} \\ \\ \overline{x}(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \overline{x}_{(i)} = \overline{x} \end{cases}$$

Therefore:

$$\begin{aligned} \widehat{\mathrm{se}}_{jack} &= \left\{ \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(n-1)n} \right\}^{1/2} \\ &= \frac{\overline{\sigma}}{\sqrt{n}} \end{aligned}$$

where $\overline{\sigma}$ is the unbiased variance.

Jackknife estimation of the standard error

- The factor $\frac{n-1}{n}$ is much larger than $\frac{1}{B-1}$ used in bootstrap.

- Intuitively this inflation factor is needed because jackknife deviation $(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2$ tend to be smaller than the bootstrap $(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2$ (the jackknife sample is more similar to the original data x than the bootstrap).

- In fact, the factor $\frac{n-1}{n}$ is derived by considering the special case $\hat{\theta} = \overline{x}$ (somewhat arbitrary convention).

Comparison of Jackknife and Bootstrap on an example

Example A: $\hat{\theta} = \overline{x}$

$f(x) = 0.2\ \mathcal{N}(\mu=1,\sigma=2) + 0.8\ \mathcal{N}(\mu=6,\sigma=1) \rightsquigarrow \mathrm{x} = (x_1, \cdots, x_{100})$.

- Bootstrap standard error and bias w.r.t. the number B of bootstrap samples:

B	10	20	50	100	500	1000	10000
$\widehat{\mathrm{se}}_B$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187
$\widehat{\mathrm{Bias}}_B$	0.0617	-0.0419	0.0274	-0.0087	-0.0025	0.0064	0.0025

- Jackknife: $\widehat{\mathrm{se}}_{jack} = 0.2207$ and $\widehat{\mathrm{Bias}}_{jack} = 0$

- Using textbook formulas: $\mathrm{se}_{\hat{f}} = \frac{\hat{\sigma}}{\sqrt{n}} = 0.2196$ ($\frac{\overline{\sigma}}{\sqrt{n}} = 0.2207$).

Jackknife estimation of the bias

1. Compute the n jackknife subsamples $\mathrm{x}_{(1)}, \cdots, \mathrm{x}_{(n)}$ from x.

2. Evaluate the n jackknife replications $\hat{\theta}_{(i)} = s(\mathrm{x}_{(i)})$.

3. The jackknife estimation of the bias is defined as:

$$\widehat{\mathrm{Bias}}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

where $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$.

Jackknife estimation of the bias

- Note the inflation factor $(n-1)$ (compared to the bootstrap bias estimate).

- $\hat{\theta} = \overline{x}$ is unbiased so the correspondence is done considering the plug-in estimate of the variance $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$.

- The jackknife estimate of the bias for the plug-in estimate of the variance is then:

$$\widehat{\mathrm{Bias}}_{jack} = \frac{\overline{-\sigma}^2}{n}$$

Histogram of the replications

Example A

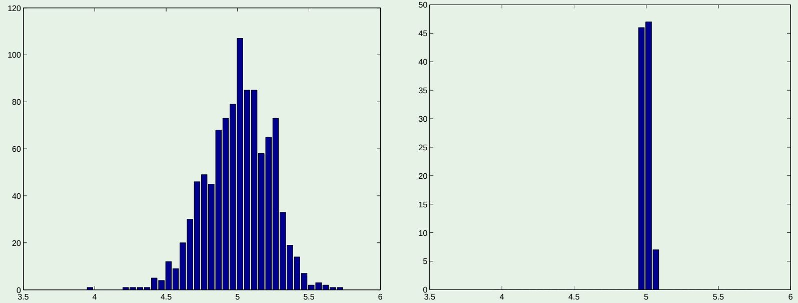


Figure: Histograms of the bootstrap replications $\{\hat{\theta}^*(b)\}_{b\in\{1,\cdots,B=1000\}}$ (left), and the jackknife replications $\{\hat{\theta}_{(}i)\}_{i\in\{1,\cdots,n=100\}}$ (right).

Histogram of the replications

Example A

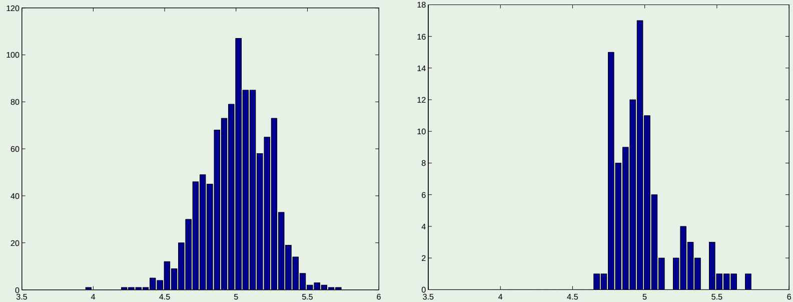


Figure: Histograms of the bootstrap replications $\{\hat{\theta}^*(b)\}_{b\in\{1,\cdots,B=1000\}}$ (left), and the inflated jackknife replications $\{\sqrt{n-1}(\hat{\theta}_{(i)}-\hat{\theta}_{(\cdot)})+\hat{\theta}_{(\cdot)}\}_{i\in\{1,\cdots,n=100\}}$ (right).

Relationship between jackknife and bootstrap

- When n is small, it is easier (faster) to compute the n jackknife replications.

- However the jackknife uses less information (less samples) than the bootstrap.

- In fact, the jackknife is an approximation to the bootstrap!

Relationship between jackknife and bootstrap

- Considering a linear statistic :

$$\begin{aligned}\hat{\theta} &= s(\mathrm{x}) = \mu + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\alpha(x_i)\\ &= \mu + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\alpha_i\end{aligned}$$

Mean $\hat{\theta} = \overline{x}$

The mean is linear $\mu = 0$ and $\alpha(x_i) = \alpha_i = x_i, \quad \forall i \in \{1, \cdot, n\}$.

- There is no loss of information in using the jackknife to compute the standard error (compared to the bootstrap) for a linear statistic. Indeed the knowledge of the n jackknife replications $\{\hat{\theta}_{(i)}\}$, gives the value of $\hat{\theta}$ for any bootstrap data set.
- For non-linear statistics, the jackknife makes a linear approximation to the bootstrap for the standard error.

Relationship between jackknife and bootstrap

- Considering a quadratic statistic

$$\hat{\theta} \quad = s(\mathrm{x}) = \mu + \tfrac{1}{n} \sum_{i=1}^{n} \alpha(x_i) + \tfrac{1}{n^2} \beta(x_i, x_j)$$

Variance $\hat{\theta} = \hat{\sigma}^2$

$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ is a quadratic statistic.

- Again the knowledge of the n jackknife replications $\{s(\hat{\theta}_{(i)})\}$, gives the value of $\hat{\theta}$ for any bootstrap data set. The jackknife and bootstrap estimates of the bias agree for quadratic statistics.

Relationship between jackknife and bootstrap

The Law school example: $\hat{\theta} = \widehat{\mathrm{corr}}(x, y)$.

The correlation is a non linear statistic.

- From B=3200 bootstrap replications, $\hat{\mathrm{se}}_{B=3200} = 0.132$.
- From $n = 15$ jackknife replications, $\hat{\mathrm{se}}_{jack} = 0.1425$.
- Textbook formula: $\mathrm{se}_{\hat{f}} = (1 - \widehat{\mathrm{corr}}^2)/\sqrt{n-3} = 0.1147$

Failure of the jackknife

The jackknife can fail if the estimate $\hat{\theta}$ is not smooth (i.e. a small change in the data can cause a large change in the statistic). A simple non-smooth statistic is the median.

On the mouse data

Compute the jackknife replications of the median $x_{Cont} = (10, 27, 31, 40, 46, 50, 52, 104, 146)$ (Control group data).

- You should find 48,48,48,48,45,43,43,43,43 [a].
- Three different values appears as a consequence of a lack of smoothness of the median[b].

[a] The median of an even number of data points is the average of the middle 2 values.

[b] the median is not a differentiable function of x.

Delete-d Jackknife samples

Definition

The delete-d Jackknife subsamples are computed by leaving out d observations from x at a time.

- The dimension of the subsample is $n - d$.
- The number of possible subsamples now rises $\left(\begin{array}{c} n \\ d \end{array} \right) = \frac{n!}{d!(n-d)!}$.
- Choice: $\sqrt{n} < d < n$

Delete-d jackknife

1. Compute all $\begin{pmatrix} n \\ d \end{pmatrix}$ d-jackknife subsamples $\mathrm{x}_{(1)}, \cdots, \mathrm{x}_{(n)}$ from x.
2. Evaluate the jackknife replications $\hat{\theta}_{(i)} = s(\mathrm{x}_{(i)})$.
3. Estimation of the standard error (when $n = r \cdot d$):

$$\widehat{\mathrm{se}}_{d-jack} = \left\{ \frac{r}{\begin{pmatrix} n \\ d \end{pmatrix}} \sum_i (\hat{\theta}_{(i)} - \hat{\theta}(\cdot))^2 \right\}^{1/2}$$

where $\hat{\theta}(\cdot) = \frac{\sum_i \hat{\theta}_{(i)}}{\begin{pmatrix} n \\ d \end{pmatrix}}$.

Concluding remarks

- The inconsistency of the jackknife subsamples with non-smooth statistics can be fixed using delete-d jackknife subsamples.

- The subsamples (jackknife or delete-d jackknife) are actually samples (of smaller size) from the true distribution f whereas resamples (bootstrap) are samples from $\hat{f}$.

Summary

- Bias and standard error estimates have been introduced using jackknife replications.
- The Jackknife standard error estimate is a linear approximation of the bootstrap standard error.
- The Jackknife bias estimate is a quadratic approximation of the bootstrap bias.
- Using smaller subsamples (delete-d jackknife) can improve for non-smooth statistics such as the median.

Introduction to resampling methods

- Definitions and Problems
- Non-Parametric Bootstrap
- Parametric Bootstrap
- Jackknife
- **Permutation tests**
- Cross-validation

So far

The resampling methods are:

- Bootstrap **re**sampling: generate samples with the same size n as x with replacement.

- Jackknife **sub**sampling : generate samples with a smaller size than x without replacement.

Used for:

- Compute accuracy measures (standard error, bias, etc.) of a statistic $\hat{\theta}$ from one set $\mathrm{x} = (x_1, \cdots, x_n)$.
- Compare two sets of observations: the example of the mouse data

Example on the mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery.

Did the treatment prolong survival ?

Example on the mouse data

1. Compute B bootstrap samples for each group
   - $\mathrm{x}^{*(b)}_{Treat} = (x^{*(b)}_{Treat\ 1}, \cdots, x^{*(b)}_{Treat\ 7})$
   - $\mathrm{x}^{*(b)}_{Cont} = (x^{*(b)}_{Cont\ 1}, \cdots, x^{*(b)}_{Cont\ 9})$
2. B bootstrap replications are computed: $\hat{\theta}^*(b) = \overline{x}^{*(b)}_{Treat} - \overline{x}^{*(b)}_{Cont}$
3. you can approximate the p.d.f. of the replications by a histogram.

Example on the mouse data

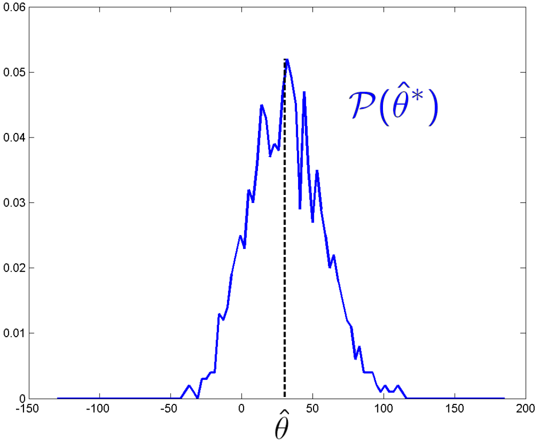


Figure: P.d.f. $\mathcal{P}(\hat{\theta}^*)$ (histogram) of the replication $\hat{\theta}^*$ ($\hat{\theta} = 30.63$ and $\hat{se}_B = 26.85$).

Introduction

- Two sample problem : definitions
- Parametric solution
- Non parametric solution:
  - permutation test
  - randomization test
  - bootstrap test

The two sample problem

Two independent random sample are observed x_a and x_b drawn from possibly different probability density functions:

$$f_a \rightsquigarrow \mathrm{x}_a = \{x_{a,1}, \cdots, x_{a,n}\}$$

$$f_b \rightsquigarrow \mathrm{x}_b = \{x_{b,1}, \cdots, x_{b,m}\}$$

Definition

The null hypothesis $\mathcal{H}_0$ assumes that there is no difference in between the density function $f_a = f_b$.

Hypothesis test and Achieved significance level (ASL)

Definition

A hypothesis test is a way of deciding whether or not the data decisively reject the hypothesis $\mathcal{H}_0$.

Definition

The achieved significance level of the test (ASL) is defined as:

$$\begin{aligned} \mathrm{ASL} &= \boldsymbol{\mathcal{P}}(\hat{\theta}^* \geqslant \hat{\theta} | \mathcal{H}_0) \\ &= \int_{\hat{\theta}}^{+\infty} \mathcal{P}(\hat{\theta}^* | \mathcal{H}_0) \ d\hat{\theta}^* \end{aligned}$$

The smaller ASL, the stronger is the evidence of $\mathcal{H}_0$ false. The notation star differentiates between an hypothetical value $\hat{\theta}^*$ generated according to $\mathcal{H}_0$, and the actual observation $\hat{\theta}$.

Parametric test

- A tradionnal way is to consider some hypotheses: $f_a \sim \mathcal{N}(\mu_a, \sigma^2)$ and $f_b \sim \mathcal{N}(\mu_b, \sigma^2)$, and the null hypothesis becomes $\mu_a = \mu_b$.

- Under $\mathcal{H}_0$, the statistic $\hat{\theta} = \overline{x}_a - \overline{x}_b$ can be modelled as a normal distribution with mean 0 and variance $\sigma^2_{\hat{\theta}} = \sigma^2(\frac{1}{m} + \frac{1}{n})$.

- The ASL is then computed:

$$\mathrm{ASL} = \int_{\hat{\theta}}^{+\infty} \frac{e^{\frac{-(\hat{\theta}^* - \hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}}}{\sqrt{2\pi}\sigma_{\hat{\theta}}} \, d\hat{\theta}^*$$

Parametric test

- σ is unknown and has to be estimated from the data:

$$\overline{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_{ai} - \overline{x}_a)^2 + \sum_{i=1}^{m}(x_{bi} - \overline{x}_b)^2}{m + n - 2}$$

- For the mouse data $\mathrm{ASL} = .131$: the null hypothesis cannot be rejected.

- However, this (parametric) method relies on the hypotheses made while calculating the ASL.

Permutation tests

- *Permutation tests* are a computer-intensive statistical technique that predates computers.
- This idea was introduced by R.A. Fisher in the 1930's.
- The main application of permutation tests is the two-sample problem.

Computation of the two sample permutation test statistic

Notation m number of values in observation x_{Treat}, n number of values in observation x_{Cont}.

If $\mathcal{H}_0$ is true, then:

1. We can combine the values from both observations in one of size $m + n = N$: $\mathrm{x} = \{\mathrm{x}_{Treat}, \mathrm{x}_{Cont}\}$.
2. Take a subsample x^*_{Treat} from x of size m. The remaining n values constitute the subsample x^*_{Cont}.
3. Compute the replication $\overline{x}^*_{Treat}$ and $\overline{x}^*_{Cont}$ on x^*_{Treat} and x^*_{Cont} respectively.
4. Compute the replication of the difference $\hat{\theta}^* = \overline{x}^*_{Treat} - \overline{x}^*_{Cont}$.

Example on the mouse data

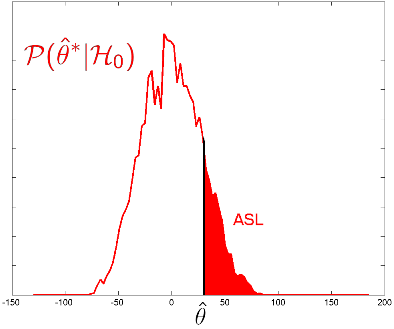


Figure: Histogram of the permutation replications $\mathcal{P}(\hat{\theta}^*|\mathcal{H}_0)$. ASL is the red surface ($\mathrm{ASL}_{perm} = 0.14$).

If the original difference $\hat{\theta} = d = \overline{x}_{Treat} - \overline{x}_{Cont}$ falls outside the 95% of the distribution of the permutation replication (i.e. $\mathrm{ASL}_{perm} < 0.05$), then the null hypothesis is rejected.

Computation of the two sample permutation test statistic

1. $\mathrm{x} = \{\mathrm{x}_a; \mathrm{x}_b\}$ of size $n + m = N$.

2. Compute all :
   - $\begin{pmatrix} N \\ n \end{pmatrix}$ permutation samples x^*. Select the n first values to define x_a^* and the last m ones to define x_b^*
   - $\begin{pmatrix} N \\ n \end{pmatrix}$ replications $\hat{\theta}^*(b) = \overline{x}_a^* - \overline{x}_b^*$

3. Approximate ASL_{perm} by:

$$\widehat{\mathrm{ASL}}_{perm} = \frac{\#\{\hat{\theta}^* \geqslant \hat{\theta}\}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Remark on the permutation test

- The histogram of the permutation replications $\hat{\theta}^*$ approximates $\mathcal{P}(\hat{\theta}^*|\mathcal{H}_0)$.

- The resamples are not really permutations but more combinations.

- $\left(\begin{array}{c} N \\ n \end{array}\right)$ can be huge so in practice, ASL_{perm} is approximated by Monte Carlo methods.

Computation of the two sample randomization test statistic

1. $\mathrm{x} = \{\mathrm{x}_a; \mathrm{x}_b\}$ of size $n + m = N$.

2. Compute B times:
   - Randomly selected permutation samples x^*. Select the n first values to define x_a^* and the last m ones to define x_b^*
   - Compute the replications $\hat{\theta}^*(b) = \overline{x}_a^* - \overline{x}_b^*$

3. Approximate ASL_{perm} by:

$$\widehat{\mathrm{ASL}}_{perm} = \frac{\#\{\hat{\theta}^* \geqslant \hat{\theta}\}}{B}$$

Remarks

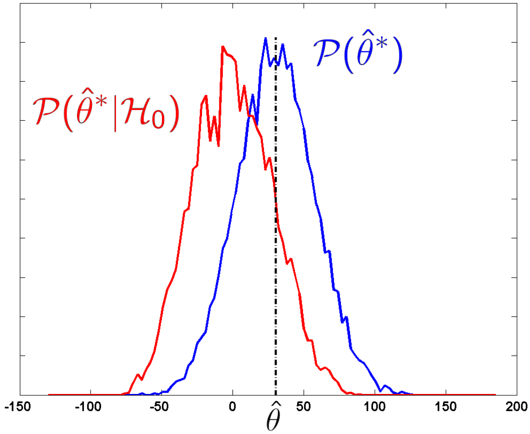


Figure: Histograms of the bootstrap replications $\mathcal{P}(\hat{\theta}^*)$ (blue), and the permutation replications $\mathcal{P}(\hat{\theta}^*|\mathcal{H}_0)$ (red).

Remarks

- Permutation replications are computed without replacement.
- The distribution of permutation replications approximates $\mathcal{P}(\theta^*|\mathcal{H}_0)$.
- The bootstrap replications presented in the introduction are computed on resamples with replacements. The distribution of those bootstrap replications defines $\mathcal{P}(\theta^*)$.
- Is there a way to get $\mathcal{P}(\theta^*|\mathcal{H}_0)$ using a bootstrap method ?

Computation of the two sample bootstrap test statistics

1. $\mathrm{x} = \{\mathrm{x}_a; \mathrm{x}_b\}$ of size $n + m = N$.

2. Compute B times:
   - Bootstrap samples from x. Select the n first values to define x_a^* and the last m ones to define x_b^*.
   - Compute the replications $\hat{\theta}^*(b) = \overline{x}_a^* - \overline{x}_b^*$

3. Approximate ASL_{boot} by:

$$\widehat{\mathrm{ASL}}_{boot} = \frac{\#\{\hat{\theta}^*(b) \geqslant \hat{\theta}\}}{B}$$

Example on the mouse data

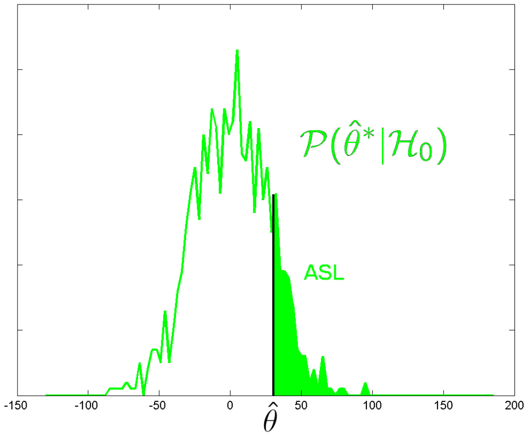


Figure: Histogram of the bootstrap replications in the two sample test $\mathcal{P}(\hat{\theta}^*|\mathcal{H}_0)$. ASL is the green surface ($\mathrm{ASL}_{boot} = 0.13$).

Relationship between the permutation test and the bootstrap test

- Very similar results in between the permutation test and the bootstrap test.

- ASL_{perm} is the exact probability.

- ASL_{boot} is not an exact probability but is guaranteed to be accurate as an estimate of the ASL, as the sample size goes to infinity.

- In the two-sample problem, the permutation test can only test the null hypothesis $f_a = f_b$ while the bootstrap can perform other hypothesis testing.

Summary

- Hypothesis testing has been introduced, involving the computation of a probability ASL
- Permutation, Randomization and bootstrap tests have been introduced as alternative to parametric tests.
- Again the main difference in between those nonparametric tests, is the way the resamples are computed (with or without replacements).

Introduction to resampling methods

- Definitions and Problems
- Non-Parametric Bootstrap
- Parametric Bootstrap
- Jackknife
- Permutation tests
- **Cross-validation**

Type of resampling

- **Randomization exact test** (or permutation test) developed by R. A. Fisher (1935/1960), the founder of classical statistical testing.
- **Jackknife** invented by Maurice Quenouille (1949) and later developed by John W. Tukey (1958).
- **Bootstrap** invented by Bradley Efron (1979, 1981, 1982) and further developed by Efron and Tibshirani (1993).
- **Cross-validation**

Resampling

Definition

Resampling means that the inference is based upon repeated sampling within the same sample. Resampling is tied to the Monte Carlo simulation. Some may distinguish both in that resampling consider all possible replications (permutation test, jackknife) and the Monte-Carlo sampling restrict the resampling to a certain number of replications (bootstrap, randomisation test).

Another difference in between resampling methods is the nature of the resamples: computed with or without replacement.

Introduction to cross-validation

- **Cross-validation** was proposed by Kurtz (1948-simple cross-validation) , extended by Mosier (1951-double cross-validation) and by Krus and Fuller (1982- multicross-validation).

- The original objective of cross-validation is to verify replicability of results.

- Similarly with hypothesis testing, the goal is to find out if the result is replicable or just a matter of random.

prediction error

Definition

A prediction error measures how well a model predicts the response value of a future observation. It is often used for model selection.
It is defined as:

- the expected square difference between a future response and the prediction from the model in regression models:

$$\mathbb{E}(y - \hat{y})$$

- the probability of incorrect classification in classification problem:

$$\mathfrak{P}(y \neq \hat{y})$$

The linear regression model

- We have a set of (2-dimensional) points $z = \{(x_i, y_i)\}_{i=1,\cdots,n}$.
- An unknow relation is linking y_i to x_i such as:

$$\begin{aligned} y_i &= \beta(x_i) + \epsilon_i \\ &= \sum_{q=0}^{p} \beta_q \, x_i^q + \epsilon_i \end{aligned}$$

- The error terms ϵ_i are assumed to be random sample from a random distribution having expectation 0 and variance σ^2.

The Model Selection Problem

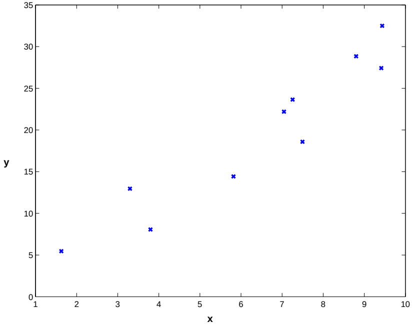
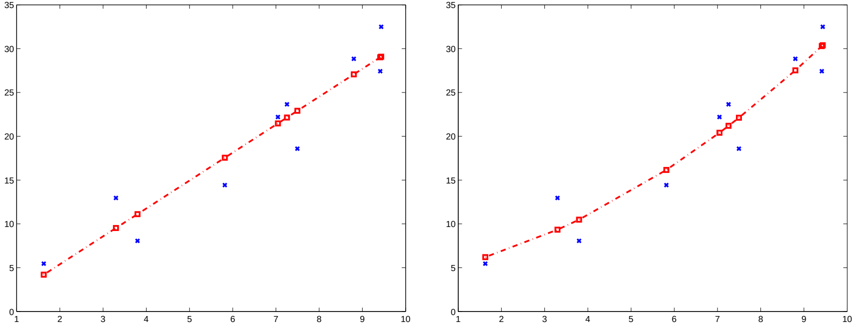


Figure: Exemple of data set (In this example $(\beta_0, \beta_1, \beta_2) = (1, 3, 0)$, $\sigma = 3$ and $n = 10$).

The Model Selection Problem



Model A: $\hat{y} = \hat{\beta}_1 \; x + \hat{\beta}_0$

Model B: $\hat{y} = \hat{\beta}_2 \; x^2 + \hat{\beta}_1 \; x + \hat{\beta}_0$

Figure: The data points (blue cross) fitted assuming a linear model (left), and a quadratic model (right). Regression estimates are $(\hat{\beta}_0, \hat{\beta}_1) = (-0.9802, 3.1866)$ for model A, and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (4.2380, 0.8875, 0.1997)$ for model 2. Which model is the best ?

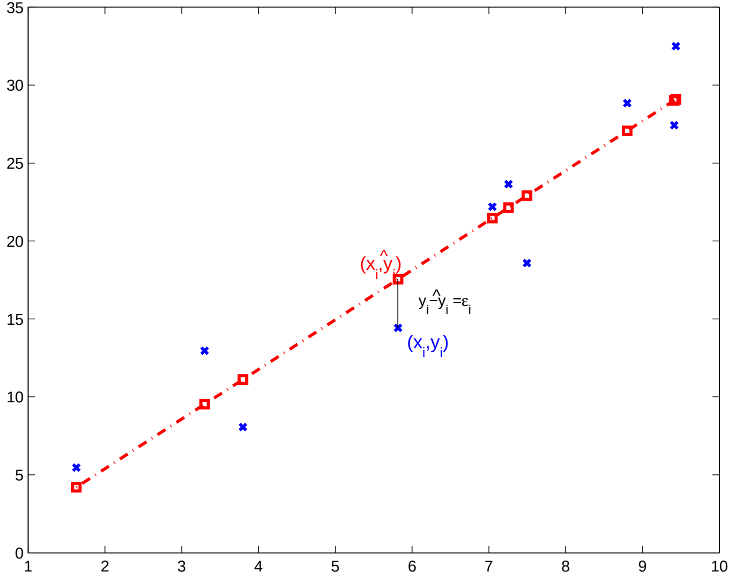
The Model Selection Problem

- Which model is the best ? $=$ How well are you going to predict future data drawn from the same distribution?
- As a prediction error, we can compute the average Residual Squared Error (RSE):

$$\mathrm{PE} = \quad \frac{\mathrm{RSE}}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

  where $\hat{y}_i$ is the prediction by the model $\hat{y}_i = \hat{\beta}(x_i)$.
- To start, y_i is taken in the sample z (instead of a future response). In this case, the prediction error is then called the apparent prediction error (APE)

The Model Selection Problem



The Model Selection Problem

For our two models:

- For model A, the apparent prediction error is 7.1240.
- For model B, the apparent prediction error is 5.8636

The apparent prediction error computed with model B is better (smaller) than with the model A. Note that the family of functions in B contains the ones defined in A. So B can capture more variabilities in the data set z, and can fit (in the sense of the APE) better the observation z.

- But this is an *apparent* prediction error. What is the prediction error for a new observation ?
- How to get a new sample to compute this prediction?

Cross-validation

- For a more realistic estimate of the prediction error, we need a test sample different from the training sample.
- One way is to split the observation z into two sets, one for training and the other one for testing.
- Using a part of the available observation to fit the model, and a different part to test in the computation of predication error is known as the cross-validation.

Simple and double cross-validation

- **Simple cross-validation**. Divide the set z into two groups, one for *training* and the other one for *testing*. The parameters β are estimated on the training data set. The cross-validation is the prediction error computed using the test sample.

- **Double cross-validation**. Models are generated for both sub-samples, and then both equations are used to generate cross-validation.

K-Fold cross-validation

1. Split z into K equal subsets z_k

2. Do K times:
   - Estimate the model $\beta_{(k)}$ on $z_{(k)} = \{z_1, \cdot, z_{k-1}, z_{k+1}, \cdots, z_K\}$
   - Compute the prediction error $\mathrm{PE}(k)$ between the test sample z_k and the predicted model by $\beta_{(k)}$.

3. Compute the average of those K prediction errors as the overall estimate of the prediction error

$$\mathrm{CV} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{PE}(k)$$

K-Fold cross-validation

- The number K of subsets depend on the size n of the observation z.
- For large datasets, even 3-Fold Cross Validation will be quite accurate.
- For small datasets, we may have to use **leave-one-out** cross validation where $K = n$.
- **Multicross-validation** is an extension of double cross-validation. Double cross-validation procedures are repeated many times by randomly selecting sub-samples from the data set.

Cross-validation and parameter selection

- We just show how to perform model selection (i.e. choice of the number of parameters p in the regression model).
- Cross-validation can also be used for parameter estimation by choosing the parameter value which minimises the prediction error.
- The cross-validation is computed using subsamples of z. An alternative consists in considering bootstrap samples.

Bootstrap estimates of the prediction error

1. Compute B boostrap resamples z^* from the observation z

2. Compute the model parameter $\beta^{*(b)}$ from z^*, and the corresponding prediction error $\mathrm{PE}(b)$ with the testing observation being the original sample z.

3. Compute the average of those B prediction errors as the overall estimate of the prediction error

$$\mathrm{CV}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\mathrm{PE}(b)$$

Bootstrap estimates of the prediction error

- This bootstrap extension to cross-validation turns out to not work very well. But it can be improved.
- Keep the record of the results from the previous procedure. Run a second experiment but choosing the bootstrap sample z^* itself as a test sample. Compute the difference of the two previous results (this difference is called optimism). This optimism is then added to the APE (as a bias correction).
- Which is best between CV and bootstrap alternatives is not really clear.

Other estimates of the prediction error

- So far, we have define the PE using the average RSE. This is the prediction error measure used in cross-validation.
- One can think of using other measures such as the Bayesian Information Criterion (BIC)

$$\frac{RSE}{n} + \log n \cdot p\hat{\sigma}^2/n$$

- BIC penalises the model as the number of parameter p increases. It is a consistent criterion i.e. it chooses the good model as $n \to \infty$.
- However, two drawbacks of the BIC compared to CV:
  - you need an estimate $\hat{\sigma}$.
  - you need the knowledge of p.

Summary

- Cross validation method applied to selection model (regression).
- Other applications such as classification use the cross-validation. In this case, y_i is a label indicating the class. The prediction error is defined as a misclassification rate.
- Those methods are very much used in machine learning.