Introduction to the Bootstrap

1 Motivation

The traditional approach to statistical inference relies on idealized models and assumptions. Often expressions for measures of accuracy such as the standard error are based on asymptotic theory and are not available for small samples. A modern alternative to the traditional approach is the bootstrapping method, introduced by Efron (1979). The bootstrap is a computer-intensive resampling method, which is widely applicable and allows the treatment of more realistic models.

As a motivation, we first discuss four examples of situations in which the exact sampling distribution of the statistic of interest is intractable. We will use these examples later to illustrate the application of the bootstrapping method.

Example *The accuracy of the sample mean*

Data: Mouse data

- Survival times of 16 mice after a test surgery
- 7 mice in treatment group (new medical treatment)
- 9 mice in control group (no treatment)

Group	*Survival time (in days)*									Mean
Treatment	94	197	16	38	99	141	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

Question: Did treatment prolong survival?

This question can be addressed by comparing the means for the two groups:

- $\bar{X} - \bar{Y} = 30.63$ indicates a life prolonging effect of the new treatment.
- *Problem:* samples show high fluctuation $\rightsquigarrow$ need to assess accuracy of estimates

Statistical theory for sample means:

Suppose $X_1, \ldots, X_n$ is an iid random sample with mean μ and variance σ^2. Then the standard error is the sample mean is

$$\mathrm{se}(\bar{X}) = \big[\mathrm{var}(\bar{X})\big]^{\frac{1}{2}} = \frac{\sigma}{\sqrt{n}}.$$

This suggests to estimate the standard error of $\bar{X}$ by

$$\widehat{\mathrm{se}}(\bar{X}) = \frac{s}{\sqrt{n}}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. The interpretation of the standard error as a measure of statistical accuracy is based on the central limit theorem, which (under quite general conditions on the distribution of the X_i) states that for large sample sizes n the sample mean $\bar{X}$ is approximately normally distributed,

$$\bar{X} \approx \mathcal{N}\big(\mu, \sigma/\sqrt{n}\big).$$

Thus we expect the mean $\bar{X}$ to be within two standard errors of the mean μ about 95% of the time. Substituting the above estimate for the standard error, we obtain a $(1-\alpha)$ confidence interval for μ,

$$\bar{X} \pm t_{n-1,\frac{\alpha}{2}}\, \widehat{\mathrm{se}}(\bar{X}).$$

In the mouse data example, we are interested in the question whether the new treatment lead to an increase in survival time. For this, we might consider the studentized test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\widehat{\mathrm{se}}(\bar{X})^2 + \widehat{\mathrm{se}}(\bar{Y})^2}}.$$

The observed value of T is 1.05, which indicates that the effect of the new treatment on survival is not significant.

Problems:

- The exact distribution of the two-sample test statistic T is not known (there are a number of approximations like Satterthwaite's approximation).

Example *Accuracy of the sample median*

Suppose that we want to compare the treatment and the control group in the mouse data example by their medians rather than their means. From the table above we find

$$\mathrm{med}(X) = 94, \qquad \mathrm{med}(Y) = 46, \qquad \text{and} \qquad T' = \mathrm{med}(X) - \mathrm{med}(Y) = 48.$$

In order to decide whether this is a significant difference, we need to quantify the accuracy of the sample medians.

Statistical theory for sample medians

- Unlike in the case of the sample mean there is no small sample formula for the standard error of the sample median.
- Suppose that the distribution P of the X_i is continuous with density $p(x)$. Then for large n, the median is approximately normally distributed,

  $$\mathrm{med}(X) \approx \mathcal{N}\Big(m_P, \frac{\sigma^2}{4n\, p(m_P)^2}\Big),$$

  where m_P is the median of the distribution P (i.e. $\mathbb{P}(X_i \le m_P) \ge \frac{1}{2}$ and $\mathbb{P}(X_i \ge m_P) \ge \frac{1}{2}$).

Problems:

- Are 7 (or 9) observations enough for the asymptotic approximation to work well?
- Can we reliably estimate the density p at m_p?
- How does the estimation of the (asymptotic) standard error affect the width of the confidence interval based on the normal approximation?

Example *Maximum likelihood estimation using the EM algorithm*

Consider a missing data problem with observed data Y_{obs} and missing data Y_{mis}. If observations are missing at random, the maximum likelihood estimator is derived from the observed-data log-likelihood function

$$l_n(\theta|Y_{\text{obs}}) = \int p(Y_{\text{obs}}, y_{\text{mis}}|\theta)\, dy_{\text{mis}},$$

where $p(y_{\text{obs}}, y_{\text{mis}}|\theta)$ is the density of the complete data $Y = (Y_{\text{obs}}, Y_{\text{mis}})^{\mathsf{T}}$. For large samples, the maximum likelihood estimator $\hat{\theta}$ is approximately normally distributed,

$$\hat{\theta} - \theta_0 \approx \mathcal{N}(0, I_{\text{obs}}(\theta_0)^{-1})$$

where θ_0 is the true parameter (assuming the model is correct) and

$$I_{\text{obs}}(\theta_0) = -\mathbb{E}\left(\frac{\partial^2 l_n(\theta|Y_{\text{obs}})}{\partial \theta^2}\bigg|_{\theta=\theta_0} \right)$$

is the observed information.

Problems:

- For many missing-data problems, the observed-data log-likelihood is too difficult to evaluate, and inference is based on the iterative EM algorithm instead.
- The EM algorithm does not automatically provide standard errors associated with the parameter estimates. The asymptotic covariance matrix $I_{\text{obs}}(\theta_0)^{-1}$ is not readily available because the implementation of the EM algorithm is based on the complete-data problem and does not require the evaluation of the observed-data log-likelihood or its derivatives.
- The EM algorithm can be extended to estimate also the observed information (eg SEM algorithm). This can be cumbersome.

Example *Number of modes of a density*

Suppose that $X_1, \ldots, X_n$ are an iid sample from a distribution P with continuous density $p(x)$. One important parameter of P is the number of modes of its density $p(x)$. Multimodality of the density indicates a heterogeneity in the data. As an illustration, we consider the following example.

Data: Galaxy data

- Velocities in km/sec of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region.
- Multimodality in the distribution of velocities is evidence for voids and superclusters in the far universe.

In this example, the structure in the distribution of velocities corresponds to the spatial distribution of galaxies in the far universe. Thus the question of existence of voids and superclusters can be addressed by testing

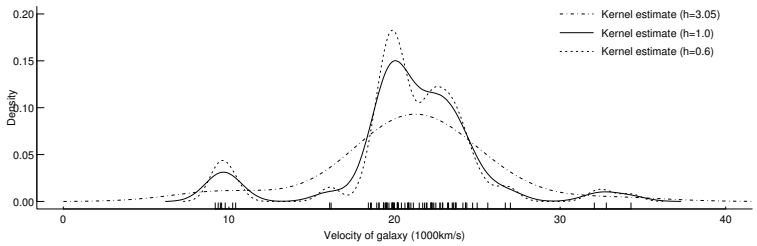
$$H_0 : n_{\text{mode}}(p) = 1 \qquad \text{vs} \qquad H_a : n_{\text{mode}}(p) > 1$$

where $n_{\text{mode}}(p)$ is the number of modes of the density.

The density of the velocities can be estimated nonparametrically by a kernel estimate

$$\hat{p}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big).$$

The bandwidth h determines the resolution of the density estimate. The following figure shows kernel estimates of the galaxy data for three different bandwidths.



As can be seen from the graphs, the number of modes exhibited by the density estimate $\hat{p}_{K,h}$ depends on the bandwidth h: For small h the estimate shows many modes some of which may be attributed to chance variation in the data, whereas for large h the estimate is much smoother but may exhibit too few modes due to oversmoothing. In particular, we can choose h large enough such that $\hat{p}_{K,h}$ only has one mode.

Let H_1 be the minimal bandwidth which leads to a unimodal density estimate, that is,

$$n_{\text{mode}}(\hat{p}_{K,H_1}) = 1 \qquad \text{and} \qquad n_{\text{mode}}(\hat{p}_{K,h}) > 1 \quad \text{for all } h < H_1.$$

Note that H_1 depends on X and is thus a random variable. Furthermore, let h_1 be the observed value for H_1. Large values of h_1 indicate oversmoothing and thus multimodality of the true density p. Thus H_1 can be used as a test statistic for the above test problem and the null hypothesis of unimodality is rejected at significance level α if

$$\mathbb{P}(H_1 > h_1 | H_0) \leq \alpha.$$

Problem:

- Distribution of H_1 under the null hypothesis H_0 is unknown.

2 The Bootstrap Principle

The basic idea of the bootstrapping method is that, in absence of any other information about the distribution, the observed sample contains all the available information about the underlying

distribution, and hence resampling the sample is the best guide to what can be expected from resampling from the distribution.

Suppose that a sample $X = (X_1, \ldots, X_n)^\mathsf{T}$ is used to estimate a parameter θ of the distribution and let $\hat{\theta} = s(X)$ be a statistic that estimates θ. For the purpose of statistical inference on θ, we are interested in the sampling distribution of $\hat{\theta}$ (or certain aspects of it) so as to assess the accuracy of our estimator or to set confidence intervals for our estimate of θ. In many applications, however, the sampling distribution of $\hat{\theta}$ is intractable.

If the true distribution P were known, we could draw samples $X^{(b)}$, $b = 1, \ldots, B$ from P and use Monte Carlo methods to estimate the sampling distribution of our estimate $\hat{\theta}$. Since P is unknown and we cannot sample from it, the bootstrapping idea suggests to resample the original sample instead. This distribution from which the bootstrap samples are drawn is the empirical distribution.

The empirical distribution For an sample $X_1, \ldots, X_n$ of independent real-valued random variables with distribution P, we define a probability distribution $\hat{P}$ by

$$\hat{P}(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i), \qquad \text{for (appropriate) } A \subseteq \mathbb{R}.$$

$\hat{P}$ is called the *empirical distribution* of the sample X. $\hat{P}$ can be thought as the distribution which puts mass $1/n$ on each observation X_i (for values that occurs more than once in the sample the mass will be a multiple of $1/n$). It follows that $\hat{P}$ is a discrete probability distribution with effective sample space $\{X_1, \ldots, X_n\}$.

It can be shown that $\hat{P}$ is a nonparametric maximum likelihood estimator of P which justifies to estimate P by $\hat{P}$ if no other information about P is available (such as e.g. P belongs to a parametric family).

Theoretical results: Let $A \subseteq \mathbb{R}$ (such that $P(A)$ is defined, i.e. A belongs to the Borel σ-algebra). Then we have

$$\hat{P}(A) \to P(A) \qquad \text{as } n \to \infty.$$

This result is a direct consequence of the law of large numbers since

$$n\,\hat{P}(A) = \sum_{i=1}^{n} 1_A(X_i) \sim \text{Bin}(n, P(A))$$

and thus $\hat{P}(A)$ tends to its expectation $P(A)$ as $n \to \infty$. This results can be strengthened to

$$\sup_{A \in I} \left| \hat{P}(A) - P(A) \right| \to 0 \qquad \text{as } n \to \infty,$$

where I is the set of all intervals of $\mathbb{R}$. In other words, the distribution $P(A)$ can be approximated by $\hat{P}(A)$ equally well for all $A \in I$.

Sampling from the empirical distribution $\hat{P}$: Suppose we want to draw an iid sample $X^* = (X_1^*, \ldots, X_n^*)^\mathsf{T}$ from $\hat{P}$. As we have noted above, $\hat{P}$ puts mass $1/n$ on each observation X_i. Thus when sampling from $\hat{P}$, the ith observation X_i in the original sample is selected with probability $1/n$. This leads to the following two-step procedure:

- Draw $i_1, \ldots, i_n$ independently from the uniform distribution on $\{1, \ldots, n\}$.
- Set $X_j^* = X_{i_j}$ and $X^* = (X_1^*, \ldots, X_n^*)^\mathsf{T}$.

In other words, we sample with replacement from the original sample $X_1, \ldots, X_n$.

The bootstrap principle Suppose that

- $X = (X_1, \ldots, X_n)^\mathsf{T}$ is a random sample from a distribution P,
- $\theta = t(P)$ is some parameter of the distribution,
- $\hat{\theta} = s(X)$ is an estimator for θ.

For an evaluation of the statistical properties (such as bias or standard error) of the actual estimate $\hat{\theta}$, we wish to estimate the sampling distribution of $\hat{\theta}$.

The bootstrapping method mimics the data-generating process by sampling from an estimate $\hat{P}$ of the unknown distribution P. Thus the role of the above real quantities is taken by their analogous quantities in the "bootstrap world":

- $X^* = (X_1^*, \ldots, X_n^*)^\mathsf{T}$ is a bootstrap sample from $\hat{P}$,
- $\theta^* = t(\hat{P})$ is the parameter in the bootstrap world,
- $\hat{\theta}^* = s(X^*)$ is the bootstrap replication of θ.

The sampling distribution of $\hat{\theta}$ is then estimated by its bootstrap equivalent

$$\hat{\mathbb{P}}(\hat{\theta} \in A) = \mathbb{P}^*(\hat{\theta}^* \in A).$$

The bootstrap principle can be summarized by the following schematic diagram:

Real World			**Bootstrap World**	
Unknown probability distribution	Observed random sample		Empirical distribution	Bootstrap sample
$P \longrightarrow$	$X = (X_1, \ldots, X_n)$	$\Longrightarrow$	$\hat{P} \longrightarrow$	$X^* = (X_1^*, \ldots, X_n^*)$
	$\downarrow$			$\downarrow$
	$\hat{\theta} = s(X)$			$\hat{\theta}^* = s(X^*)$
	Statistic of interest			Bootstrap replication

In general, the estimate $\hat{P}$ will be determined by the available information about P. Only if the data comprise all available information about P, we estimate P by the empirical distribution.

Monte Carlo Approximation: Even though the distribution of the bootstrap sample X^* is known, the evaluation of the exact bootstrap sampling distribution of $\hat{\theta}^*$ can be still intractable. In fact, the sampling distribution has been derived only for special cases such as the median of an uneven number of observations.

In general, the bootstrap estimate of the sampling distribution of $\hat{\theta}$ is computed using Monte Carlo methods:

- Draw B independent bootstrap samples $X^{*(1)}, \ldots, X^{*(B)}$ from $\hat{P}$:

$$X_1^{*(b)}, \ldots, X_n^{*(b)} \overset{\text{iid}}{\sim} \hat{P} \qquad b = 1, \ldots, B$$

- Evaluate bootstrap replications

$$\hat{\theta}^{*(b)} = s(X^{*(b)}) \qquad b = 1, \ldots, B$$

- Estimate the sampling distribution of $\hat{\theta}$ by the empirical distribution of the bootstrap replications $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$:

$$\hat{\mathbb{P}}(\hat{\theta} \in A) = \frac{1}{B} \sum_{b=1}^{B} 1_A\big(\hat{\theta}^{*(b)}\big).$$

  for appropriate subsets $A \subseteq \mathbb{R}^p$ (if $\hat{\theta} \in \mathbb{R}^p$).

Often we are only interested in one characteristic of the sampling distribution of $\hat{\theta}$, for example the standard error or the bias. Estimates for these quantities can be straightforwardly obtained from the bootstrap replications.

The bootstrap algorithm for estimating standard errors Let $\hat{\theta} = s(X)$ be an estimator for θ and suppose we want to know the standard error of $\hat{\theta}$. A bootstrap estimate of standard error can be obtained by the following algorithm:

- Draw B independent bootstrap samples $X^{*(1)}, \ldots, X^{*(B)}$ from $\hat{P}$:

$$X_1^{*(b)}, \ldots, X_n^{*(b)} \overset{\text{iid}}{\sim} \hat{P} \qquad b = 1, \ldots, B.$$

- Evaluate the bootstrap replications

$$\hat{\theta}^{*(b)} = s(X^{*(b)}) \qquad b = 1, \ldots, B.$$

- Estimate the standard error $\text{se}(\hat{\theta})$ by the standard deviation of the B replications

$$\widehat{\text{se}}_{\text{boot}}(\hat{\theta}) = \left[\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*(b)} - \hat{\theta}^{*(\cdot)}\right)\right]^{\frac{1}{2}},$$

  where

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

Example *Mouse data*

As an example, consider the mouse data and suppose that we want to assess in the accuracy of the sample mean of the treatment group.

Applying the above algorithm for estimating the standard error, we first have to resample from the original seven observations and calculate for each bootstrap sample the sample mean. The following table shows the first 20 (out of $B = 1000$) bootstrap samples and their sample mean:

b	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	X_6^*	X_7^*	$\bar{X}^*$
1	38	141	94	16	99	197	23	86.9
2	94	23	197	16	141	38	94	86.1
3	16	141	94	23	94	38	99	72.1
4	16	94	94	23	99	197	16	77.0
5	38	141	16	99	16	141	141	84.6
6	197	16	197	94	16	16	16	78.9
7	99	23	94	23	38	197	99	81.9
8	38	38	38	23	16	99	38	41.4
9	23	38	141	94	23	94	23	62.3
10	38	23	141	94	38	141	197	96.0
11	38	38	38	99	197	141	141	98.9
12	38	23	38	99	23	38	99	51.1
13	23	94	197	99	99	16	99	89.6
14	38	16	16	38	141	38	141	61.1
15	94	38	16	94	23	38	141	63.4
16	23	197	94	16	38	99	99	80.9
17	38	99	16	38	16	197	38	63.1
18	197	16	141	16	16	94	197	96.7
19	141	38	94	197	38	23	16	78.1
20	23	99	23	16	197	99	23	68.6

From the bootstrap replications $\bar{X}^{*(b)}$, $b = 1, \ldots, B$, we obtain a bootstrap estimate for the standard error of the sample mean

$$\widehat{\mathrm{se}}_{\mathrm{boot}}(\bar{X}) = \frac{1}{B-1} \sum_{i=1}^{B} \left(\bar{X}^{*(b)} - \bar{X}^{*(\cdot)}\right)^2 = 23.53.$$

Note that this is a Monte Carlo approximation to the ideal bootstrap estimate, which in the special case of the sample mean is given by

$$\widehat{\mathrm{se}}_{\mathrm{boot}}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = 23.36.$$

The two estimates agree quite well.

The ideal bootstrap estimate of standard error for linear statistics In the special case of linear statistics, we can evaluate the ideal bootstrap estimate of standard error. For this, suppose that our statistic of interest is of the form

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i).$$

Then the bootstrap statistic is given by

$$\hat{\theta}^* = \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i^*) = \frac{1}{n} \sum_{j=1}^{n} N_j \alpha(X_j).$$

where N_j is the observed frequency of X_j in the bootstrap sample $X_1^*, \ldots, X_n^*$. Since X_i is resampled with probability $\frac{1}{n}$, the frequencies $N = (N_1, \ldots, N_n)$ are multinomially distributed

with parameter $(\frac{1}{n}, \ldots, \frac{1}{n})$ and we have

$$\mathbb{E}(N_i) = 1,$$
$$\operatorname{cov}(N_1, N_j) = \begin{cases} 1 - \frac{1}{n} & \text{if } i = j \\ -\frac{1}{n} & \text{if } i \neq j \end{cases}.$$

It follows that

$$\begin{aligned}
\mathbb{E}^*(\hat{\theta}^*) &= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}(N_j)\alpha(X_j) = \frac{1}{n} \sum_{j=1}^{n} \alpha(X_j) = \hat{\theta}, \\
\operatorname{var}^*(\hat{\theta}) &= \frac{1}{n^2} \sum_{i,j=1}^{n} \alpha(X_i)\, \alpha(X_j) \operatorname{cov}(N_i, N_j) \\
&= \frac{1}{n^2} \sum_{i}^{n} \alpha(X_i)^2 - \frac{1}{n^3} \Big(\sum_{i}^{n} \alpha(X_i) \Big)^2 \\
&= \frac{1}{n^2} \sum_{i}^{n} \big(\alpha(X_i) - \alpha(\cdot) \big)^2
\end{aligned}$$

where

$$\alpha(\cdot) = \frac{1}{n} \sum_{j=1}^{n} \alpha(X_j) = \hat{\theta}.$$

Thus the ideal bootstrap estimate of standard error is basically the same as the usual estimate $\widehat{\operatorname{se}}(\hat{\theta}) = s/\sqrt{n}$ (up to a factor $\big(\frac{n-1}{n}\big)^{\frac{1}{2}}$):

$$\widehat{\operatorname{se}}_{\text{boot}}(\hat{\theta}) = \Big(\frac{n-1}{n} \Big)^{\frac{1}{2}} \widehat{\operatorname{se}}(\hat{\theta}).$$

The bootstrap estimate of bias Suppose that we estimate the parameter $\theta = t(P)$ by the statistic

$$\hat{\theta} = s(X).$$

The bias of the estimator $\hat{\theta}$ is defined as

$$\operatorname{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Substituting the empirical distribution $\hat{P}$ for P, we obtain the bootstrap estimate of bias

$$\widehat{\operatorname{bias}}(\hat{\theta}) = \operatorname{bias}^*(\hat{\theta}^*) = \mathbb{E}^*(\hat{\theta}^*) - \theta^*,$$

where $\theta^* = t(\hat{P})$. Note that $\hat{\theta}$ and θ^* can be different. As an example, consider the trimmed mean as an estimate of the mean $\theta = \mathbb{E}(X_1)$. The trimmed mean is given by

$$\hat{\theta}_p = \frac{1}{n_p} \sum_{i=1}^{n} X_{(i)}\, 1\{\tfrac{np}{2} \leq i \leq \tfrac{n(1-p)}{2}\} \qquad \text{with } n_p = \sum_{i=1}^{n} 1\{\tfrac{np}{2} \leq i \leq \tfrac{n(1-p)}{2}\}.$$

On the other hand, the parameter θ^* in the bootstrap world is the expectation of an observation drawn from $\hat{P}$ und thus is equal to the sample mean of X,

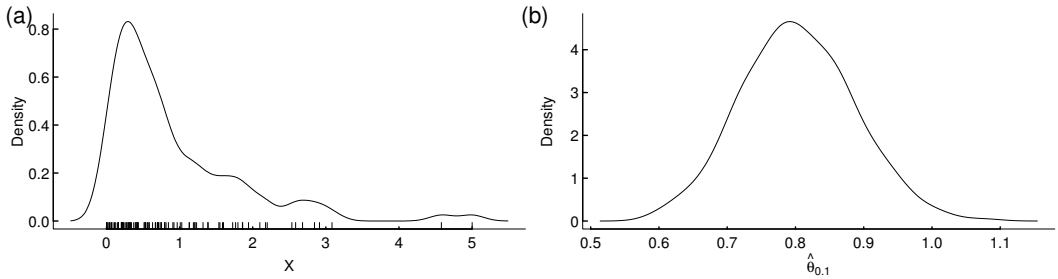
$$\theta^* = \mathbb{E}^*(X_j) = \bar{X}.$$

Example *Trimmed mean of exponentially distributed observations*

Consider observations $X_1, \ldots, X_{100}$ independently sampled from an unknown distribution P (Figure (a) below). The mean of P can be estimated by the trimmed mean

$$\hat{\theta}_{0.1} = \frac{1}{90} \sum_{i=6}^{95} X_{(i)}.$$

To estimate the bias of $\hat{\theta}_{0.1}$ we have created $B = 1000$ bootstrap samples X^* and for each computed the corresponding trimmed mean $\hat{\theta}^*_{0.1}$. The estimated sampling distribution of $\hat{\theta}_{0.1}$ is shown in Figure (b) below.



(a) Estimated density of the observations $X_1, \ldots, X_{100}$ and (b) estimated sampling distribution of the trimmed mean $\hat{\theta}_{0.1}$ (based on $B = 1000$ bootstrap replications).

The bootstrap estimate of the bias is given by

$$\widehat{\text{bias}}(\hat{\theta}_{0.1}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*\,(b)}_{0.1} - \bar{X} = -0.1058.$$

The estimate of bias can be used to correct the original estimate $\hat{\theta} = 0.8042$ such that it becomes less biased. The obvious bias-corrected estimator is

$$\tilde{\theta}_{0.1} = \hat{\theta}_{0.1} - \widehat{\text{bias}}(\hat{\theta}_{0.1}).$$

For the above data, we obtain with $\hat{\theta}_{0.1} = 0.8042$ the corrected estimate $\tilde{\theta}_{0.1} = 0.9101$.

Example *Allele frequency estimation*
Consider again the ABO blood type data $N_{\text{obs}} = (N_A, N_B, N_{AB}, N_O)$. By application of the EM algorithm, we were able to compute the maximum likelihood estimates for the allele frequencies p_A, p_B, p_O:

$$\hat{p}_A = 0.2136 \qquad \hat{p}_B = 0.0501 \qquad \hat{p}_O = 0.7363.$$

Asymptotic theory for maximum likelihood estimation implies that these estimators are approximately normally distributed and thus suggests to take

$$\hat{p}_A \pm z_{\alpha/2} \cdot \widehat{\text{se}}(\hat{p}_A),$$

as an approximate $(1 - \alpha)$ confidence interval for p_A with similar confidence intervals for p_B and p_O. Since the EM algorithm does not provide an estimate for the standard error, we can apply the bootstrapping method to obtain an estimate for $\text{se}(\hat{p}_A)$.

- Draw bootstrap sample $N^*_{\text{obs}} = (N^*_A, N^*_B, N^*_{AB}, N^*_O)^\mathsf{T}$:

$$N^*_{\text{obs}} \sim M\big(n, \tfrac{N_A}{n}, \tfrac{N_B}{n}, \tfrac{N_{AB}}{n}, \tfrac{N_O}{n}\big).$$

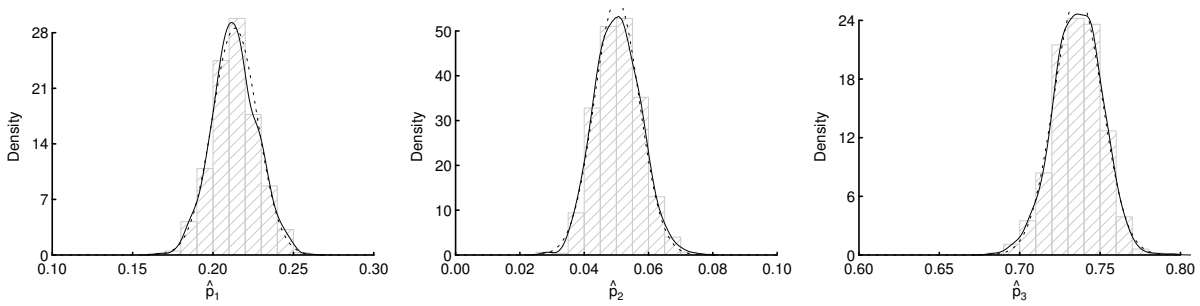
- Compute estimates $\hat{p}^*_A, \hat{p}^*_B, \hat{p}^*_O$ using the EM algorithm.
- Iterate previous two steps B times.
- Estimate standard error of $\hat{p}_A$ by

$$\widehat{\text{se}}(\hat{p}_A) = \left[\frac{1}{B-1}\sum_{b=1}^{B}\big(\hat{p}_A^{*(b)} - \hat{p}_A^{*(\cdot)}\big)^2\right]^{\frac{1}{2}}$$

where

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

The following figure shows the bootstrap estimates of the sampling distributions of the parameter estimators for p_A, p_B, and p_O.



Estimates of the sampling distribution of the allele frequency estimators $\hat{p}_A$, $\hat{p}_B$, and $\hat{p}_O$: Kernel estimate (solid), normal approximation (dashed), and histogram estimate (grey).

The normal approximation shows that the estimators are indeed approximately normally distributed. The bootstrap estimates of standard error thus lead to the following approximate 95% confidence interval for the parameters:

Parameter	*ML estimate*	*Std error*	*95% confidence interval*
p_1	0.214	0.014	[0.187, 0.240]
p_2	0.050	0.007	[0.037, 0.064]
p_3	0.736	0.015	[0.708, 0.765]

These confidence intervals agree quite well with the posterior intervals obtained by the data augmentation algorithm.

Remark: Note that application of maximum likelihood does not require that P belongs to the fitted parametric family. For general P, the MLE $\hat{\theta}$ estimates the parameter

$$\theta_0 = \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}\big(\log p(X|\theta)\big),$$

where X is distributed according to P. The parameter θ_0 characterizes the distribution P_θ which comes closest to P (where "distance" is measured by the Kullback-Leibler distance). For the empirical distribution $\hat{P}$ we obtain

$$\theta_0^* = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}^*\big(\log p(X^*|\theta)\big) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\theta),$$

that is, θ_0^* is equal to the maximum likelihood estimator $\hat{\theta}$.

The parametric bootstrap Suppose we know that the distribution P belongs to a parametric family of distributions P_θ with densities $p(x|\theta)$. If $\hat{\theta}$ is an estimate of the true parameter, θ_0 say, an obvious estimate of P is the distribution $\hat{P} = P_{\hat{\theta}}$ with density $p(x|\hat{\theta})$. In this case, we can still use the bootstrapping method to obtain an estimate of the sampling distribution of $\hat{\theta}$ (or any statistic $g(\hat{\theta})$). Our knowledge about P is incorporated into the bootstrap algorithm by substituting the parametric distribution $P_{\hat{\theta}}$ for the empirical distribution. This is called the *parametric bootstrap*:

- Draw B independent bootstrap samples $X^{*(1)}, \ldots, X^{*(B)}$ from $P_{\hat{\theta}}$:

  $$X_1^{*(b)}, \ldots, X_n^{*(b)} \overset{\text{iid}}{\sim} P_{\hat{\theta}} \qquad b = 1, \ldots, B$$

- Evaluate bootstrap replications

  $$\hat{\theta}^{*(b)} = s(X^{*(b)}) \qquad b = 1, \ldots, B$$

- Estimate the sampling distribution of $\hat{\theta}$ by the empirical distribution of the bootstrap replications $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$:

  $$\hat{\mathbb{P}}(\hat{\theta} \in A) = \frac{1}{B}\sum_{b=1}^{B} 1_A\big(\hat{\theta}^{*(b)}\big).$$

  for appropriate subsets $A \subseteq \mathbb{R}^p$ (if $\hat{\theta} \in \mathbb{R}^p$).

Example *Allele frequency estimation*
Consider for the last time the ABO blood type data. According to the Hardy-Weinberg law, the genotype counts $N = (N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}, N_O)$ are multinomially distributed with parameters p_A^2, $2p_Ap_O$, p_B^2, $2p_Bp_O$, $2p_Ap_B$, and p_O^2. Furthermore, $N_A = N_{AA} + N_{AO}$ and $N_B = N_{BB} + N_{BO}$. This defines a distribution for the observed data $N_{\text{obs}}^* = (N_A, N_B, N_{AB}, N_O)$, from which we can sample the bootstrap samples $N_{\text{obs}} = (N_A^*, N_B^*, N_{AB}^*, N_O^*)$.

The parametric bootstrap yields the following estimates of standard error and 95% confidence intervals for the allele frequencies:

Parameter	*ML estimate*	*Std error*	*95% confidence interval*
p_1	0.214	0.014	[0.186, 0.240]
p_2	0.050	0.012	[0.027, 0.074]
p_3	0.736	0.017	[0.703, 0.769]

3 Testing Hypotheses

Let X and Y be two samples from two possibly different, unknown distributions P and Q and suppose that we want to test whether the two distributions are equal. Thus we have

$$H_0 : P = Q \qquad \text{vs} \qquad H_a : P \neq Q.$$

Furthermore let us assume that T is an appropriate test statistic for this test problem. Then if we observe the value $T = t$ for the test statistic, the null hypothesis will be rejected at significance level α if

$$\mathbb{P}(T \geq t) \leq \alpha$$

under the null hypothesis. In many applications, the sampling distribution of the test statistic is not known (exactly) and the p-value cannot be calculated. This suggests to use the bootstrap instead and estimate the p-value by

$$\hat{\mathbb{P}}(T \geq t) = \mathbb{P}^*(T^* \geq t).$$

One complication which arises in bootstrapping test problems is that we need to sample under the null hypothesis.

For the above test problem this can be achieved by resampling $X^{*(b)}$ and $Y^{*(b)}$ from the joint sample (X, Y). From these bootstrap samples, we can then compute the bootstrap replications of the test statistic

$$T^{*(b)} = s(X^{*(b)}, Y^{*(b)})$$

and estimate the p-value by

$$\hat{\mathbb{P}}(T \geq t) = \frac{1}{B} \sum_{b=1}^{B} 1\{T^{*(b)} \geq t\}.$$

Example *Mouse data*

Consider again the mouse data, where we were interested whether the new treatment prolonged the survival time. One standard solution to this problem is to test for equality of the means of the two groups, that is, to consider the test problem

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_a : \mu_X \neq \mu_Y.$$

Unlike in the situation above, the null hypothesis requires only equality in the means but not e.g. in the variances. Obviously the means of the two groups are not equal, but this can be corrected by a small transformation of the original data. Let

$$\tilde{X}_i = X_i - \bar{X} + \bar{Z}$$
$$\tilde{Y}_i = Y_i - \bar{Y} + \bar{Z}$$

where

$$\bar{Z} = \frac{1}{n_X + n_Y} \Big[\sum_{i=1}^{n_X} X_i + \sum_{i=1}^{n_Y} Y_i \Big].$$

Obviously the empirical distributions of the two transformed samples $\tilde{X}$ and $\tilde{Y}$ have equal means and thus satisfy the condition of the null hypothesis. We obtain the following bootstrap algorithm:

- Sample $X_1^{*(b)}, \ldots, X_{n_X}^{*(b)}$ independently from $\tilde{X}$.
- Sample $Y_1^{*(b)}, \ldots, Y_{n_Y}^{*(b)}$ independently from $\tilde{Y}$.
- Evaluate bootstrap replications

$$T^{*(b)} = \frac{\bar{X}^{*(b)} - \bar{Y}^{*(b)}}{\sqrt{\frac{s^2_{X^{*(b)}}}{n_X} + \frac{s^2_{Y^{*(b)}}}{n_Y}}} \qquad b = 1, \ldots, B.$$

- Estimate the p-value

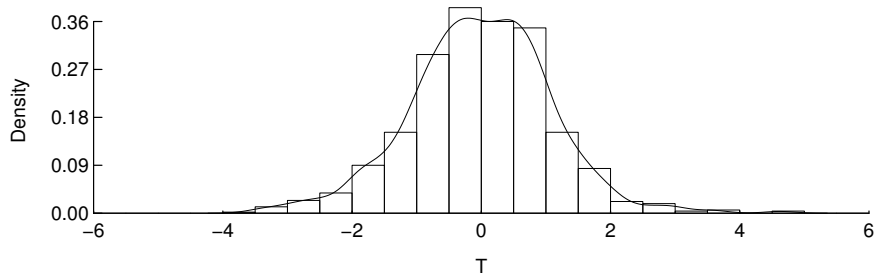
$$\hat{\mathbb{P}}(T \geq t) = \frac{1}{B} \sum_{b=1}^{B} 1\{T^{*(b)} \geq t\},$$

where t is the observed value of the two-sample t test statistic.

For the mouse data, the observed value of T was $t = 1.06$. From $B = 1000$ bootstrap replications of T 133 were greater than or equal to t. Thus the bootstrap estimate of the p-value is

$$\hat{\mathbb{P}}(T \geq t) = 0.133$$

and we do not reject the null hypothesis. Thus the observed difference in the means between the two groups is not significant.



Bootstrap estimate of the sampling distribution of the two-sample t test statistic for the mouse data.

Example *Number of modes of a density*

The Galaxy data consist of the velocities (in km/sec) of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. As pointed out earlier, the structure in the distribution of velocities corresponds to the spatial distribution of galaxies in the far universe. In particular, a multimodal distribution of velocities indicates a strong heterogeneity in the spatial distribution of the galaxies and thus is seen as evidence for the existence of voids and superclusters in the far universe.

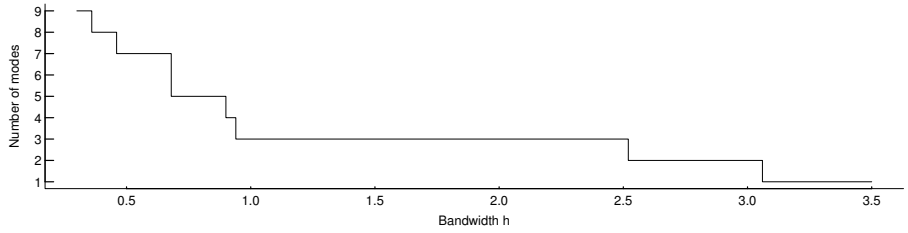
Statistically, the question of multimodality can be formulated as a test problem

$$H_0 : n_{\text{mode}}(p) = 1 \qquad \text{vs} \qquad H_a : n_{\text{mode}}(p) > 1$$

where $n_{\text{mode}}(p)$ is the number of modes of the density of the velocities. To develop an appropriate test statistic for this problem, we considered nonparametric kernel density estimates

$$\hat{p}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big).$$

It can be shown that the number of modes of $\hat{p}_{K,h}(x)$ decreases monotonically as h increases. For the galaxy data, this relationship between the number of modes of $\hat{p}_{K,h}(x)$ and the bandwidth h is shown in the following figure.



Let H_1 be the minimal bandwidth for which $\hat{p}_{K,H_1}$ is unimodal, that is,

$$n_{\text{mode}}(\hat{p}_{K,H_1}) = 1 \qquad \text{and} \qquad n_{\text{mode}}(\hat{p}_{K,h}) > 1 \quad \text{for all } h < H_1.$$

Furthermore, let h_1 be the actual observed value for the galaxy data ($h_1 = 3.05$). Since multimodal densities need more smoothing to become unimodal, the minimal bandwidth H_1 can be used as a test statistic for the above problem. The null hypothesis is rejected at significance level α if the corresponding p-value is smaller than α:

$$\mathbb{P}_0(H_1 > h_1) \leq \alpha.$$

Since the sampling distribution of H_1 is unknown, we use the bootstrapping method to estimate the p-value. For this, we need to sample from an estimate of P_0, the distribution under the null hypothesis. Restricting ourselves to distributions with densities $p_{K,h}$, we find that the distribution $\hat{P}_0$ with density $p_{K,h_1}(x)$ is in a sense closest to the empirical distribution $\hat{P}$ as it needs the least amount of smoothing. More precisely, it can be shown that

$$\sum_{i=1}^{n} \log p_{K,h_1}(X_i) \leq \sum_{i=1}^{n} \log p_{K,h}(X_i)$$

for all $h > h_1$, that is, $p_{K,h_1}(x)$ maximizes the log-likelihood under the restriction of unimodality (within the class of distributions with densities $p_{K,h}$).

Let Z^* be a sample from the empirical distribution $\hat{P}$. Then

$$X_i^{*\,(b)} = Z_i^* + h_1 \varepsilon_i, \qquad i = 1, \ldots, n,$$

is an iid sample from $\hat{P}_0$. Since the variance of the bootstrap sample has been increased by adding the normal error term, the data are usually rescaled to have the sample variance as the original observations. This leads to the following algorithm:

- Draw B independent bootstrap samples $X^{*\,(1)}, \ldots, X^{*\,(B)}$ from $\hat{P}_0$:

  $$X_i^{*\,(b)} = \bar{Z}^* + \left(1 + h_1^2/\hat{\sigma}^2\right)^{-\frac{1}{2}} \left(Z_i^* - \bar{Z}^* + h_1 \varepsilon_i\right), \quad i = 1, \ldots, n,$$

  where $Z_1^*, \ldots, Z_n^*$ are independently sampled from $\hat{P}$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- Evaluate the bootstrap replications $H_1^{*\,(b)}$ for $b = 1, \ldots, B$.
- Estimate the p-value (or achieved significance level) by

  $$\hat{\mathbb{P}}(H_1 \geq h_1) = \frac{1}{B} \sum_{b=1}^{B} 1\{H_1^{*\,(b)} \geq h_1\}.$$

- Reject the null hypothesis if $\hat{\mathbb{P}}(H_1 \geq h_1) \leq 0.05$.

4 Confidence Intervals

Having generated the bootstrap replications $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$, we have an estimate of the sampling distribution of $\hat{\theta}$. From this, we can construct confidence intervals for θ.

Standard confidence interval: Suppose that $\hat{\theta}$ is approximately normally distributed with mean θ and variance $\mathrm{se}(\hat{\theta})^2$. Then an approximate $(1-\alpha)$ confidence interval for θ is given by

$$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\,\widehat{\mathrm{se}}_{\mathrm{boot}}(\hat{\theta}) \qquad \text{and} \qquad \hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\,\widehat{\mathrm{se}}_{\mathrm{boot}}(\hat{\theta}),$$

where z_α is the α critical value of the standard normal distribution.

Bootstrap t interval: Again suppose that $\hat{\theta}$ is approximately normally distributed with mean θ and variance $\mathrm{se}(\hat{\theta})^2$. Furthermore let $\widehat{\mathrm{se}}_X(\hat{\theta})$ be an estimator of $\mathrm{se}(\hat{\theta})$ based on the sample X. From the bootstrap samples $X^{*(b)}$, we then calculate

$$T^{*(b)} = \frac{\hat{\theta}^{*(b)} - \hat{\theta}}{\widehat{\mathrm{se}}_{X^*}(\hat{\theta})}.$$

From these values $T^{*(b)}$, we can estimate the critical values $t_{1-\alpha/2}$ and $t_{\alpha/2}$ by $\hat{t}_{1-\alpha/2}$ and $\hat{t}_{\alpha/2}$, respectively, such that

$$\frac{1}{B}\sum_{b=1}^{B} 1\{T^{*(b)} \le \hat{t}_{1-\alpha/2}\} \approx \frac{\alpha}{2} \qquad \text{and} \qquad \frac{1}{B}\sum_{b=1}^{B} 1\{T^{*(b)} \ge \hat{t}_{\alpha/2}\} \approx \frac{\alpha}{2}.$$

Then

$$\hat{\theta}_L = \hat{\theta} + \hat{t}_{1-\alpha/2}\mathrm{se}(\hat{\theta}) \qquad \text{and} \qquad \hat{\theta}_U = \hat{\theta} + \hat{t}_{\alpha/2}\mathrm{se}(\hat{\theta})$$

defines an approximate $(1-\alpha)$ confidence interval for θ.

Percentile interval: The $(1-\alpha)$ confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$ is given by the empirical quantiles of the boostrap replications, that is,

$$\hat{\mathbb{P}}^*(\hat{\theta}^* \le \hat{\theta}_L) = \frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}^{*(b)} \le \hat{\theta}_L\} \approx \frac{1}{2}\,\alpha,$$
$$\hat{\mathbb{P}}^*(\hat{\theta}^* \ge \hat{\theta}_U) = \frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}^{*(b)} \ge \hat{\theta}_U\} \approx \frac{1}{2}\,\alpha.$$

Bias corrected percentile interval: The confidence interval should have equal probability to both sides of $\hat{\theta}$, that is,

$$\mathbb{P}(\hat{\theta} \le \theta \le \hat{\theta}_U) = \mathbb{P}(\hat{\theta}_L \le \theta \le \hat{\theta}).$$

If $\hat{\theta}$ is not the median of the boostrap distribution, this condition is not fulfilled. An appropriate correction is given by

$$\frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}^{*(b)} \le \hat{\theta}_L\} \approx \Phi(2\,z^* - z_{\alpha/2})$$
$$\frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}^{*(b)} \le \hat{\theta}_U\} \approx \Phi(2\,z^* + z_{\alpha/2}),$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution and

$$z'^* = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} 1\{\hat{\theta}^{*\,(b)} \leq \hat{\theta}\}\right) = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^{*\,(b)} \leq \hat{\theta}\}}{B}\right).$$

Roughly speaking, z_0 measures the discrepancy between the median of $\hat{\theta}^*$ and $\hat{\theta}$ in normal units.

There exists an extension of the bias corrected percentile interval, the BC_a interval (eg Efron and Tibshirani, 1993).

References

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.