Part III

Bootstrap and Resampling Methods

Resampling Approaches to Statistical Inference

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

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

Notation: independence case

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

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

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

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

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

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

Notation: independence case

Recall that, for each y, under mild regularity conditions

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

but also

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

as $n \longrightarrow \infty$. Therefore

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

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

Notation: independence case

The bias and variance of the estimator are

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

both of which depend on the true F.

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

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

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

The Bootstrap

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

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

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

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

Example (Sample mean standard error)

Consider data

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

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

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

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

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

Example (Sample mean standard error)

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

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

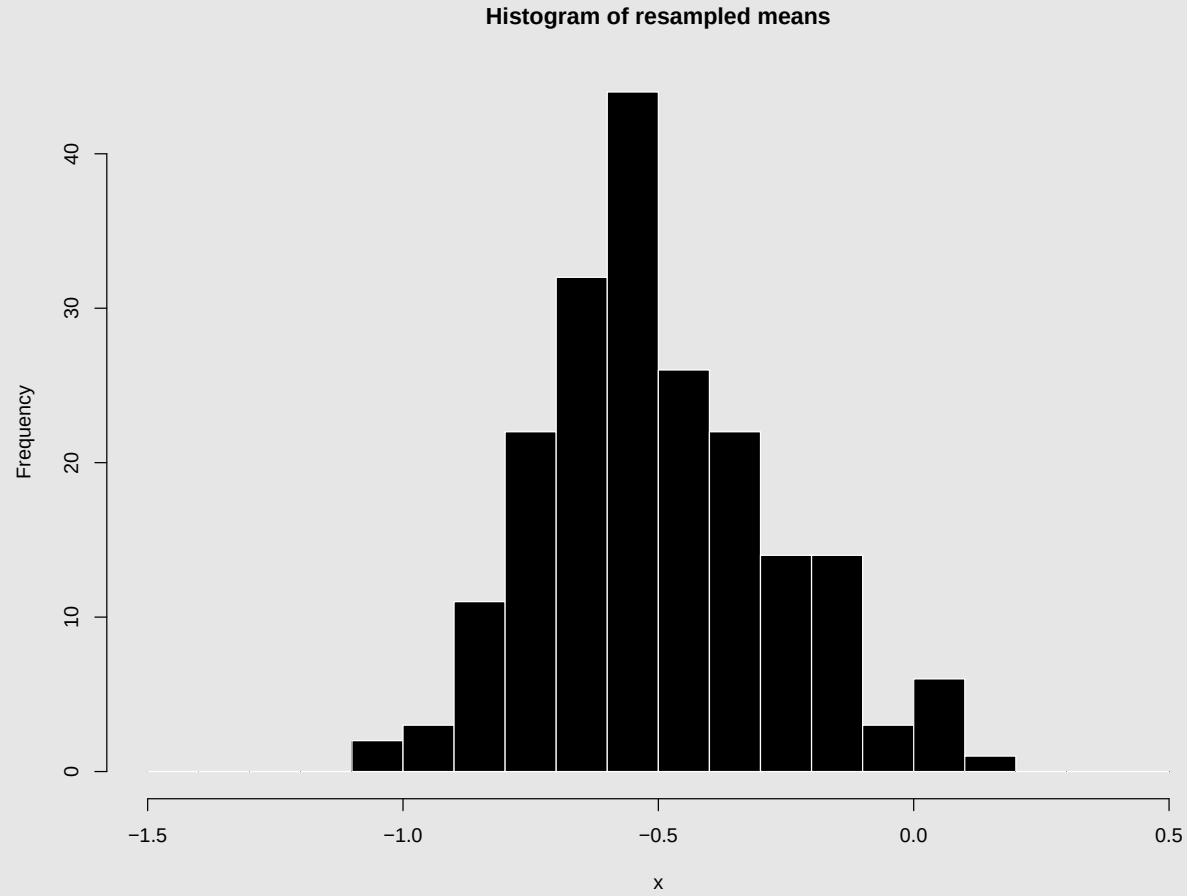
with sample mean -0.2221 and standard deviation 0.9589.

Example (Sample mean standard error)

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

Example (Sample mean standard error)

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



Example (Sample kurtosis standard error)

The sample kurtosis

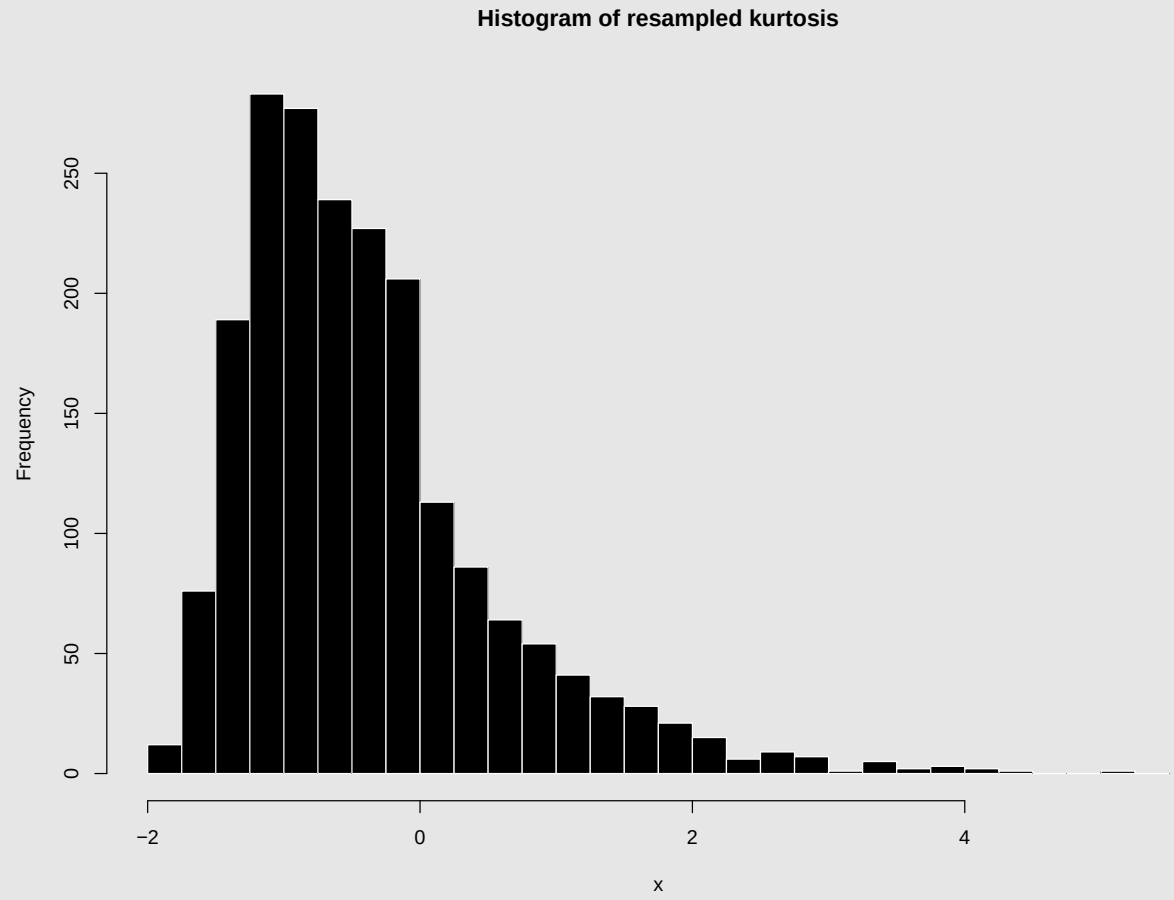
$$t_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^2} - 3$$

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

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

Example (Sample kurtosis standard error)

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



The Bootstrap: notation

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

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

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

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

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

and

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

with

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

The Bootstrap: choosing B

For a sample of size n, there are

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

possible distinct resampled data sets:

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

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

The Bootstrap: choosing B

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

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

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

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

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

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

The Bootstrap: choosing B

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

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

It can be shown that

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

and

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

The Bootstrap: choosing B

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

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

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

estimated by

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

But, by standard theory, under F

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

The Bootstrap: choosing B

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

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

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

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

The larger B gets, we expect to see

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

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

The Bootstrap: choosing B

Example (Simulation study)

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

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

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

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

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

The Bootstrap: choosing B

Example (Simulation study)

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

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

The Bootstrap: choosing B

Note: the estimator used here for σ^2 is

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

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

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

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

The Bootstrap: Further Examples

In the case of i.i.d. univariate sampling, the bootstrap extends easily to more general settings. For example, extension to the i.i.d. vector case is straightforward.

1. A two-sample problem.
2. The Pearson correlation coefficient.
3. Principal components analysis.

The Bootstrap: Two-sample Estimation

Suppose that two random samples are available from different distributions

$$\begin{aligned} F_1 &: y_{11}, \ldots, y_{1n_1} \\ F_2 &: y_{21}, \ldots, y_{2n_2} \end{aligned}$$

and the parameter of interest is the difference in means

$$\mu_1 - \mu_2.$$

In the normal case, under equal variances for the two populations, the standard error of the estimator

$$\overline{y}_1 - \overline{y}_2$$

can be studied analytically; more generally, outside of the normal case and for other estimands, it cannot.

The Bootstrap: Two-sample Estimation

The bootstrap operates in this case by

- resampling $y_{11}^{(b)}, \ldots, y_{1n_1}^{(b)}$ (size n_1) from the first sample,
- resampling $y_{21}^{(b)}, \ldots, y_{2n_2}^{(b)}$ (size n_2) from the second sample,
- forming the difference

$$\overline{y}_1^{(b)} - \overline{y}_2^{(b)}$$

for $b = 1, \ldots, B$. Other estimators can be studied in this way.

The Bootstrap: Pearson correlation

Suppose that pairs of observations $(x_i, y_i), i = 1, \ldots, n$ are available. The Pearson product-moment correlation is

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \overline{xy})}{\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right\}}}$$

which measures the linear correlation between the two variables in the sample.

r is the moment-based estimator of the population correlation ρ.

The Bootstrap: Pearson correlation

The bootstrap operates in this case by

- resampling **pairs** $((x_1^{(b)}, y_1^{(b)}), \ldots, (x_n^{(b)}, y_n^{(b)}))$ (size n)
- forming the resampled value of r

$$r^{(b)}$$

for $b = 1, \ldots, B$.

Note that again it is the i.i.d. units that are being resampled.

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

The relationship between GPA (x) and LSAT score (y) in $n = 15$ law school students is to be studied.

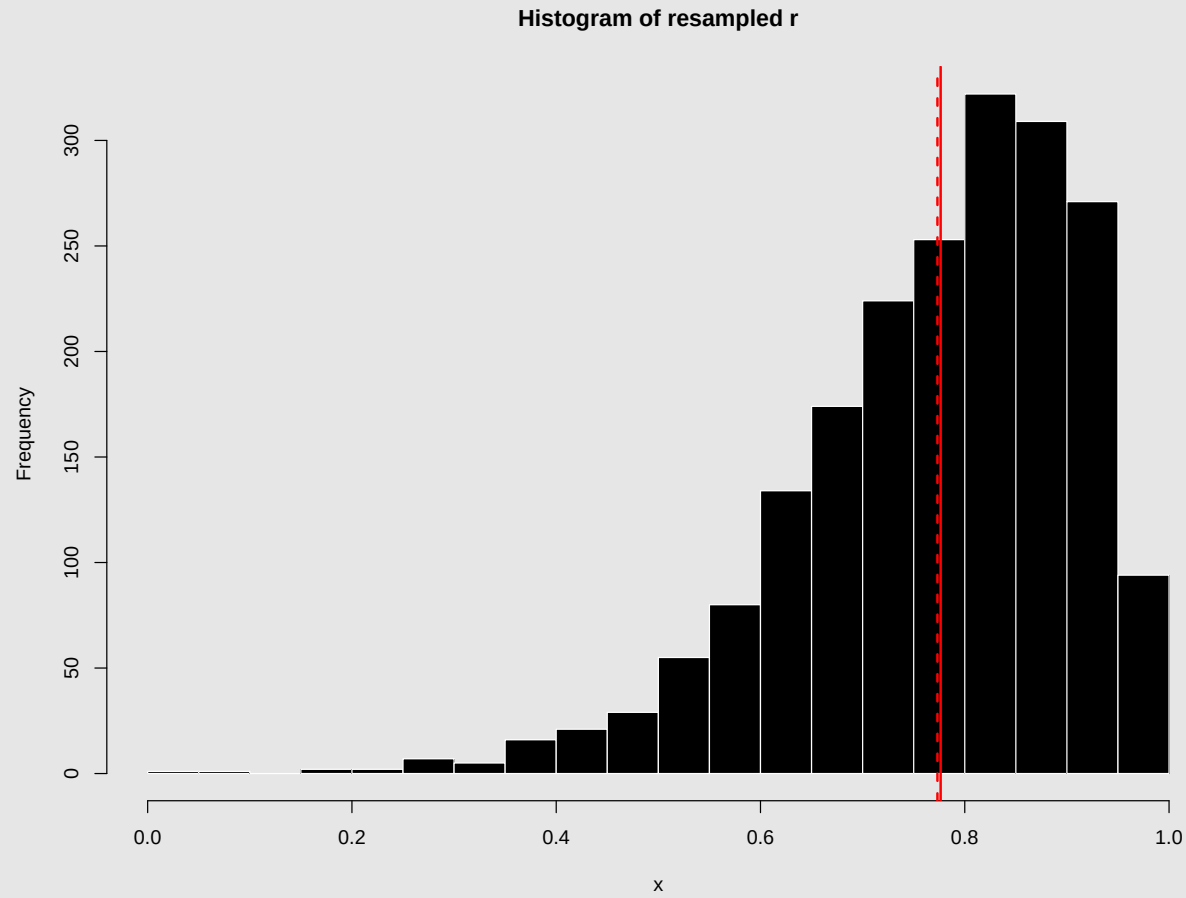
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GPA	339	330	281	303	344	307	300	343	336	313	312	274	276	288	296
LSAT	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594

For these data $r = 0.776$.

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

2000 (resampled) sample correlation values, with standard deviation 0.1382.



The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

A common transformation to approximate normality (for large n) uses

$$Z_n = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

for which

$$\sqrt{n-3}Z_n \xrightarrow{d} Z \sim \mathcal{N}(\mu, 1)$$

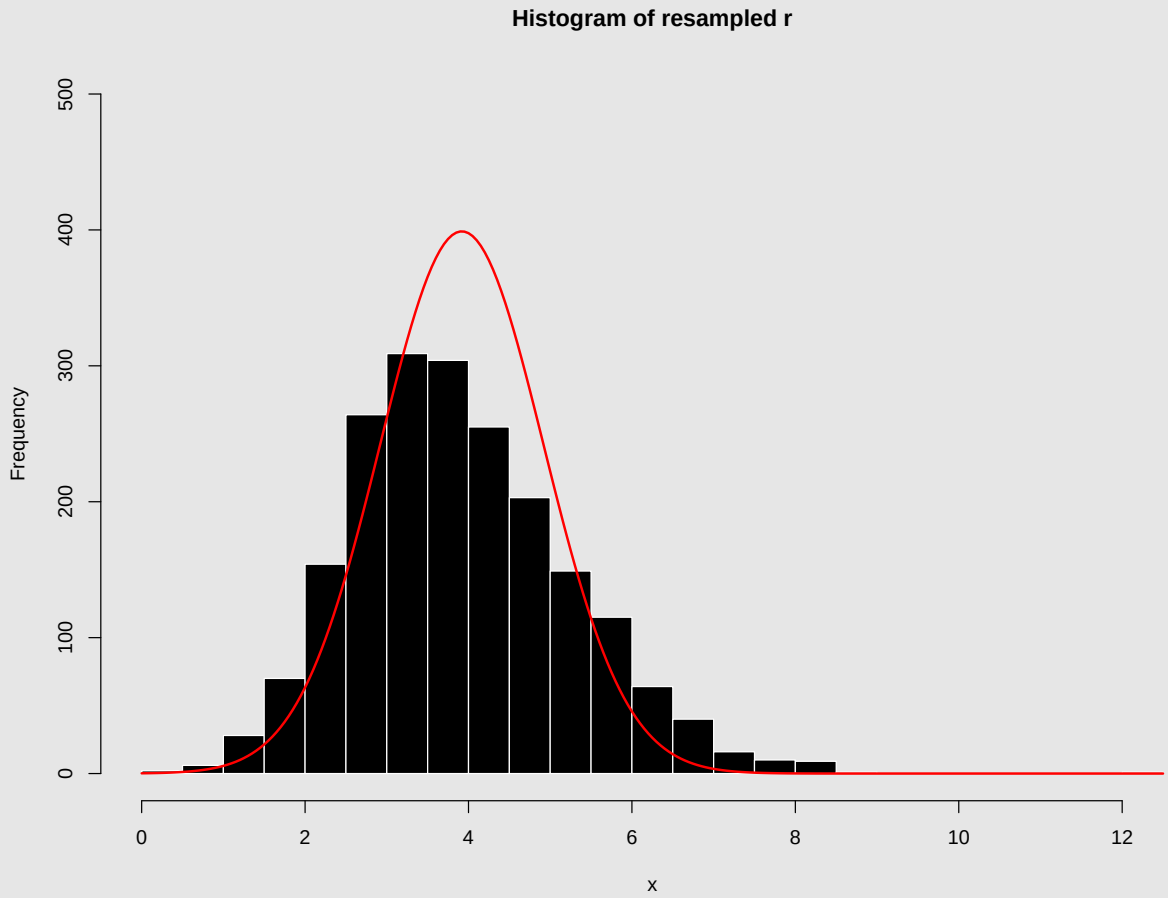
where

$$\mu = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

2000 Z_n values, and normal approximation: skewness is apparent, and the bootstrap variance is 1.793, rather than 1.



The Bootstrap: Principal components

Suppose that vectors of observations of length p, $\boldsymbol{x}_i, i = 1, \ldots, n$, are available. Let X denote the $n \times p$ matrix formed by row-binding the n i.i.d. vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, after *column standardization*

- each column of X has its mean subtracted.

Then the sample covariance matrix is

$$S = \frac{1}{n} X^\intercal X$$

This is a positive-definite matrix, that can be decomposed using the eigen-decomposition

$$S = VDV^\intercal$$

where D is the diagonal matrix of eigenvalues, and V is the corresponding eigenvector matrix with orthonormal columns.

The Bootstrap: Principal components

The diagonal elements of D are the sample eigenvalues

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$$

and the scaled quantities

$$\widehat{\theta}_j = \frac{\widehat{\lambda}_j}{\sum_{l=1}^{p} \widehat{\lambda}_l} \qquad j = 1, \ldots, p.$$

is an estimate of the proportion of variation explained by the first eigenvector.

The statistical properties of $\widehat{\theta}_j$ are hard to study analytically.

The Bootstrap: Pearson correlation

The bootstrap operates in this case by

- resampling **vectors** $(\boldsymbol{x}_1^{(b)}, \ldots, \boldsymbol{x}_n^{(b)})$ (size n)
- forming the resampled matrix $X^{(b)}$,
- forming the resampled covariance matrix $S^{(b)}$,
- computing eigenvalues $\widehat{\lambda}_1^{(b)}, \ldots, \widehat{\lambda}_p^{(b)}$ and scaling

for $b = 1, \ldots, B$.

Note that again it is the i.i.d. units that are being resampled.

The Bootstrap: Principal components

Example (Education test scores)

88 students each took five tests on mechanics, vectors, algebra, analysis and statistics. This yielded the S matrix

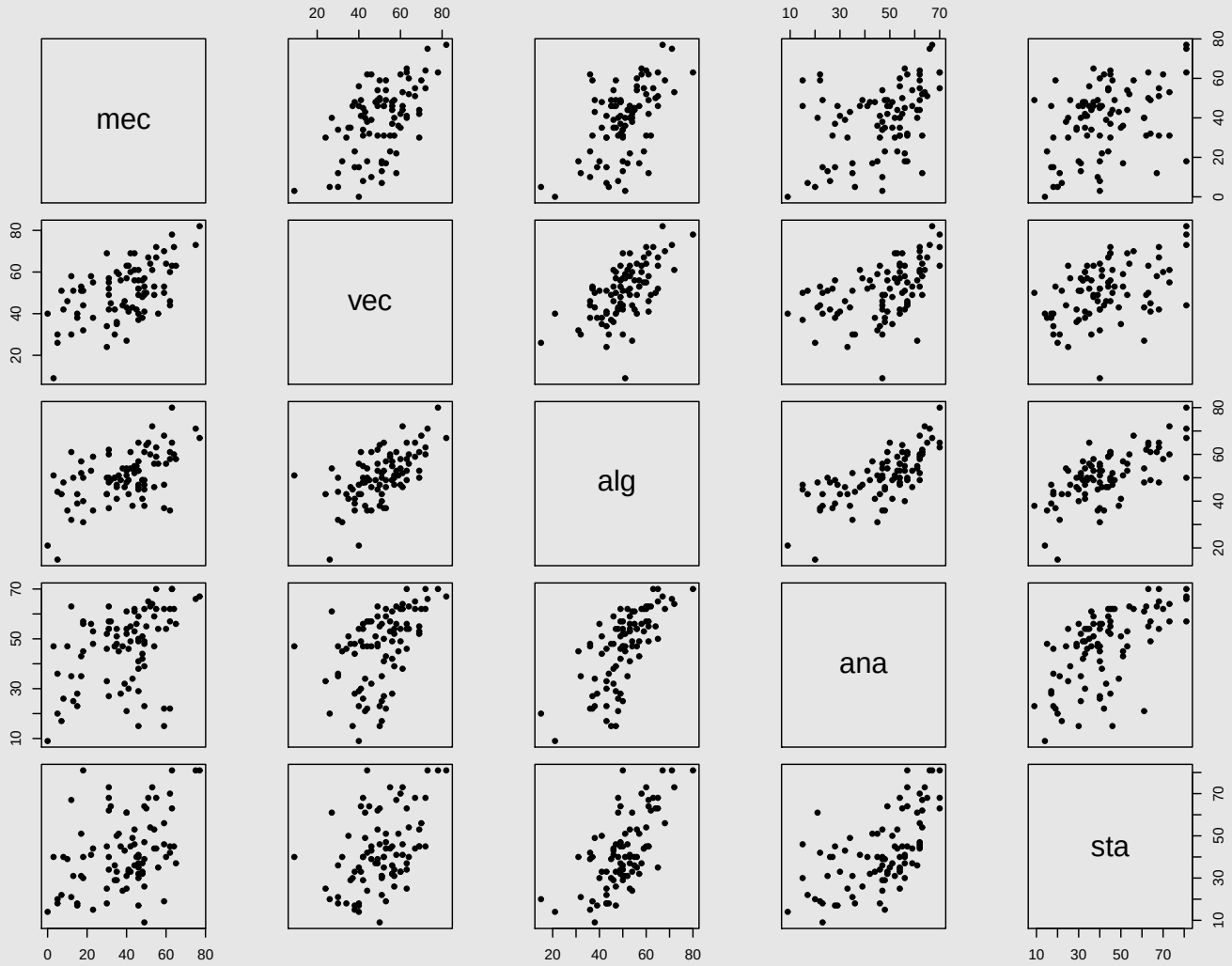
$$S = \begin{bmatrix} 302.29 & 125.78 & 100.43 & 105.07 & 116.07 \\ 125.78 & 170.88 & 84.19 & 93.60 & 97.89 \\ 100.43 & 84.19 & 111.60 & 110.84 & 120.49 \\ 105.07 & 93.60 & 110.84 & 217.88 & 153.77 \\ 116.07 & 97.89 & 120.49 & 153.77 & 294.37 \end{bmatrix}$$

and eigenvalues

$$\widehat{\boldsymbol{\lambda}} = (679.18, 199.81, 102.57, 83.67, 31.79)$$

The Bootstrap: Principal components

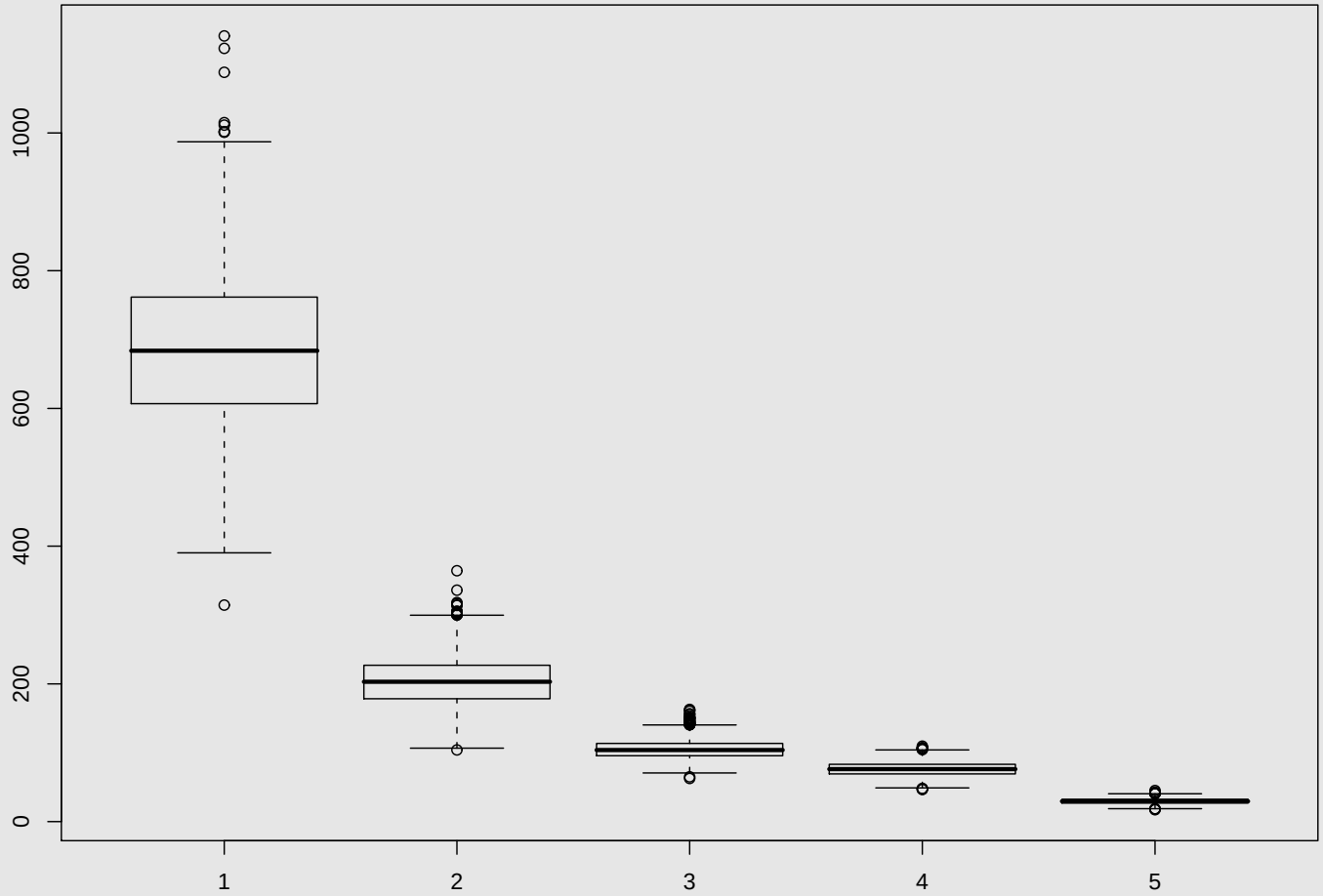
Example (Education test scores)



The Bootstrap: Principal components

Example (Education test scores)

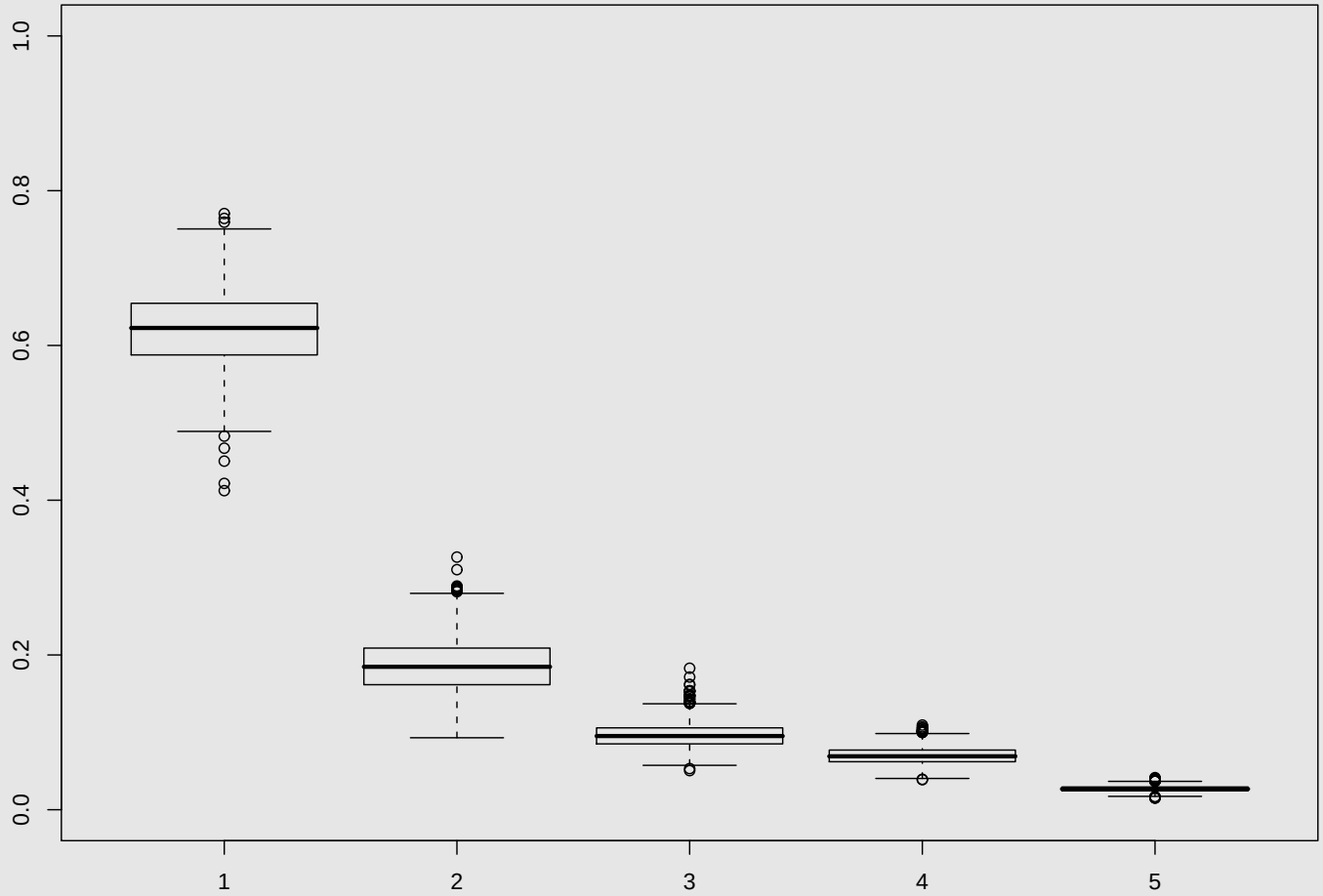
2000 resampled eigenvalues, $\lambda_j, j = 1, \ldots, 5$



The Bootstrap: Principal components

Example (Education test scores)

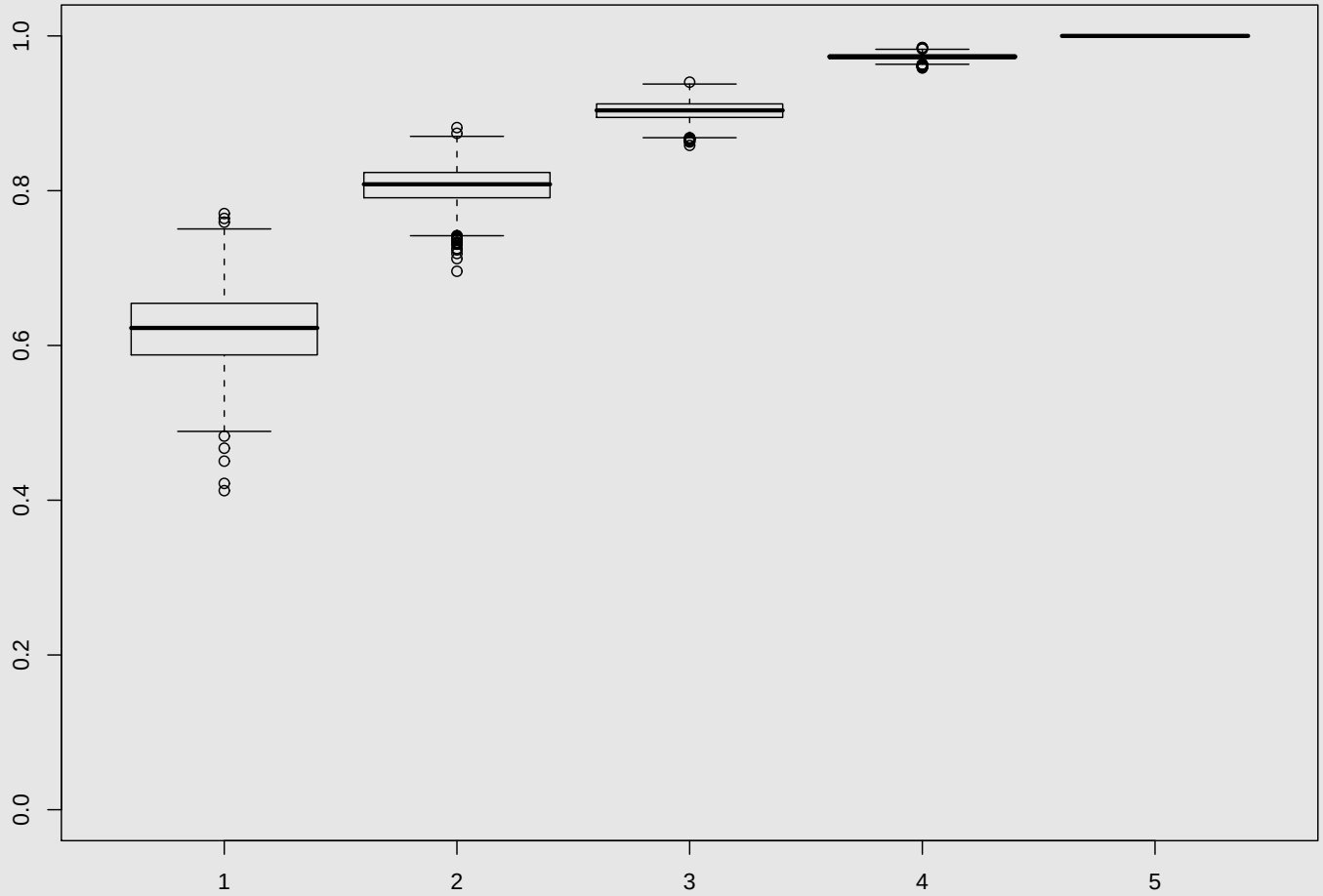
2000 resampled scaled eigenvalues, $\theta_j, j = 1, \ldots, 5$



The Bootstrap: Principal components

Example (Education test scores)

2000 resampled scaled eigenvalues, cumulative:



The Bootstrap: Principal components

Example (Education test scores)

We might conclude from this that really there are only two (orthogonal) dimensions of variation (rather than five) that explain the majority variation in the response.

The *principal components* are the vectors

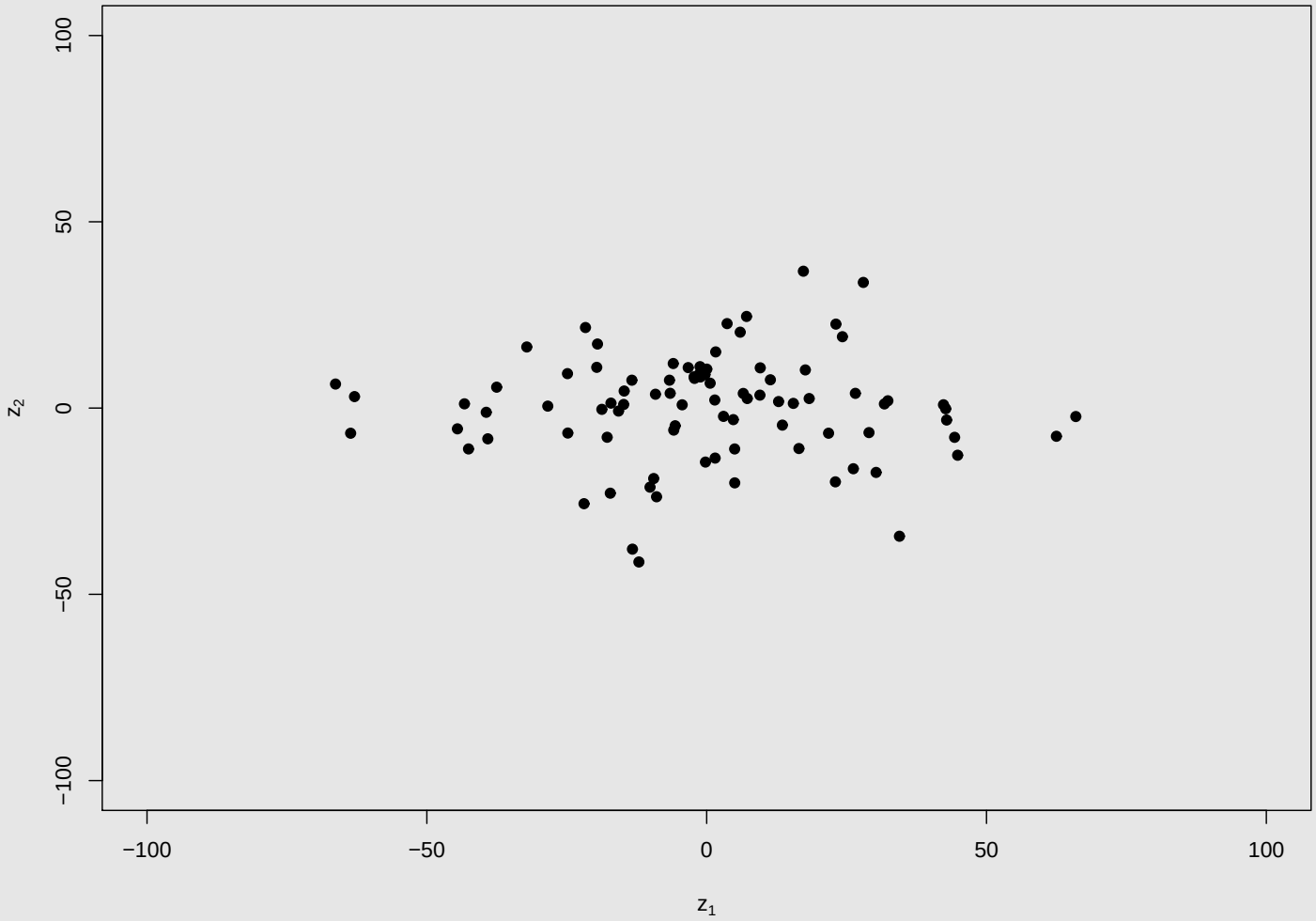
$$z_j = Xv_j \qquad j = 1, \ldots, p$$

where v_j is the jth column of V. The first two principal components carry greater than 80% of the variation observed.

The Bootstrap: Principal components

Example (Education test scores)

Principal components



The Bootstrap: Regression

The bootstrap can also be applied in a regression setting. Suppose pairs of data $(x_i, y_i), i = 1, \ldots, n$ are available, and the relationship is modelled as

$$y_i = g(x_i; \beta) + \epsilon_i \qquad \mathbb{E}[\epsilon_i | x] = 0$$

for some regression function g.

If g is taken to be linear in parameters β,

$$g(x; \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x)$$

and the residual errors are presumed Gaussian, then estimation can be carried out using least-squares in the usual way.

In this case, *pointwise* confidence intervals are easy to obtain. However, *simultaneous* confidence intervals are more complicated to compute.

The Bootstrap: Regression

In *harmonic regression*, the g_js are taken to be sin and cos functions. Assuming $p = 2q$ say, and $j = 1, \ldots, p$

$$g_j(x) = \begin{cases} \cos(2\pi x \phi_l) & j = 2l - 1 \\ \sin(2\pi x \phi_l) & j = 2l \end{cases}$$

for known constants $\phi_1, \ldots, \phi_q$.

For example, a typical choice is $q = 2$, with

$$\phi_1 = \frac{1}{2} \qquad \phi_2 = \frac{1}{4}$$

The Bootstrap: Regression

In such a model, the ML estimator is

$$\widehat{\beta} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{Y}$$

where $\boldsymbol{X}$ is the $n \times (p+1)$ matrix with ith row

$$[1, \cos(2\pi x_i \phi_1), \sin(2\pi x_i \phi_1), \cos(2\pi x_i \phi_2), \sin(2\pi x_i \phi_2)]$$

The variance of $\widehat{\beta}$ is

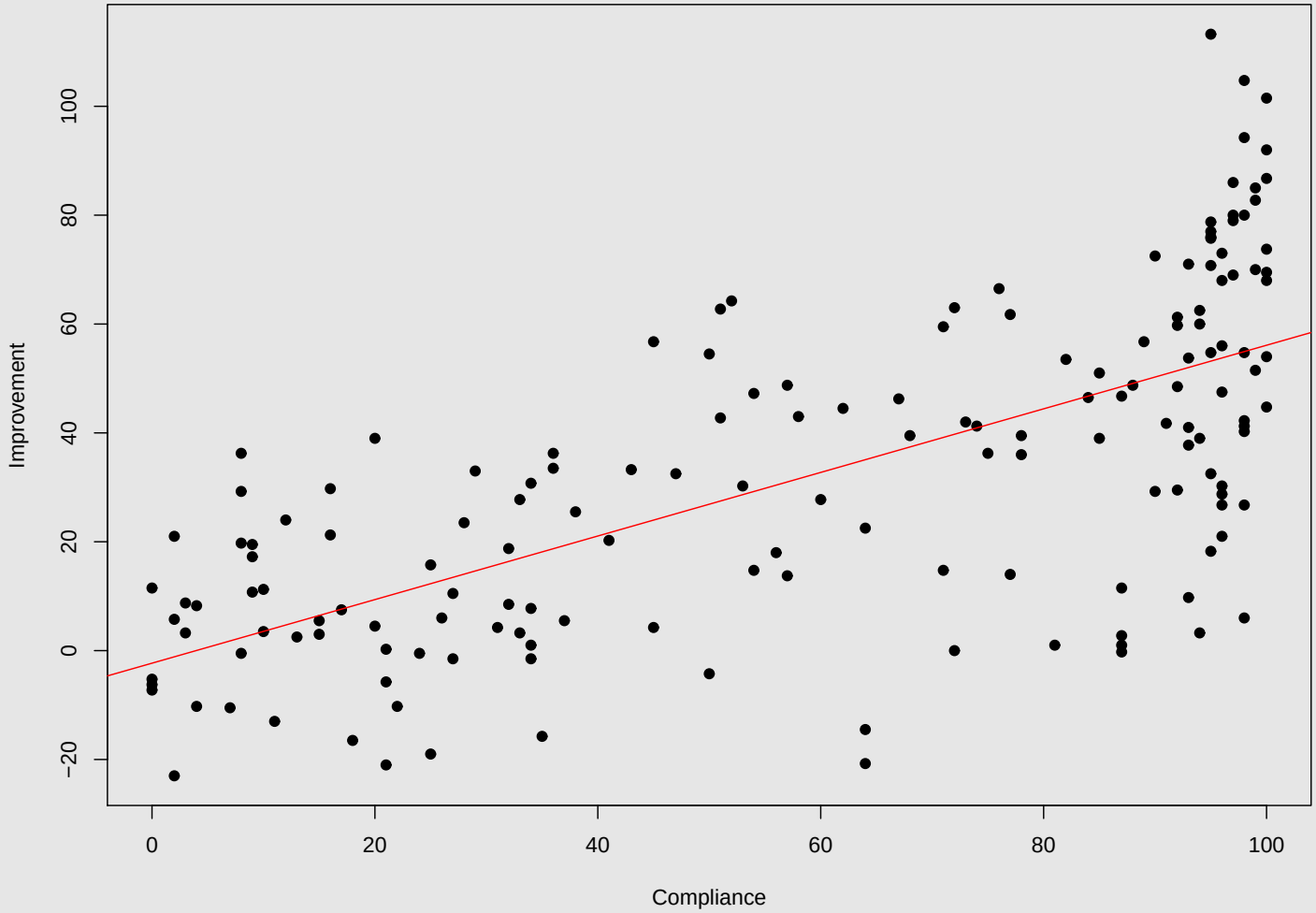
$$\widehat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

and pointwise confidence intervals can be obtained in the usual way.

The Bootstrap: Regression

Example (Cholostyramine Data)

$n = 164$ men took part in an experiment to see if the drug cholostyramine lowered blood cholesterol levels.



The Bootstrap: Regression

Example (Cholostyramine Data)

Consider the model

$$g(x;\beta) = \beta_{00} + \beta_{01}x + \sum_{j=1}^{q}\left[\beta_{j0}\cos(2\pi x\phi_l) + \beta_{j1}\sin(2\pi x\phi_l)\right]$$

on the range $[0, 100]$ with

$$\phi_1 = \frac{1}{50} \qquad \phi_2 = \frac{1}{66} \qquad \phi_3 = \frac{1}{75}$$

and the three fits obtained taking

- $q = 0$
- $q = 2$
- $q = 3$

The Bootstrap: Regression

The bootstrap operates in this case by

- resampling **pairs** $((x_1^{(b)}, y_1^{(b)}), \ldots, (x_n^{(b)}, y_n^{(b)}))$ (size n) and forming the boostrapped design matrix $\boldsymbol{X}_{(b)}$.
- fitting the regression to the resampled data obtaining

$$\widehat{\beta}^{(b)} = (\boldsymbol{X}_{(b)}^{\intercal}\boldsymbol{X}_{(b)})^{-1}\boldsymbol{X}_{(b)}^{\intercal}\boldsymbol{y}_{(b)}$$

- recording the fitted values

$$\hat{y}^{(b)} = \boldsymbol{X}_{(b)}\widehat{\beta}^{(b)}$$

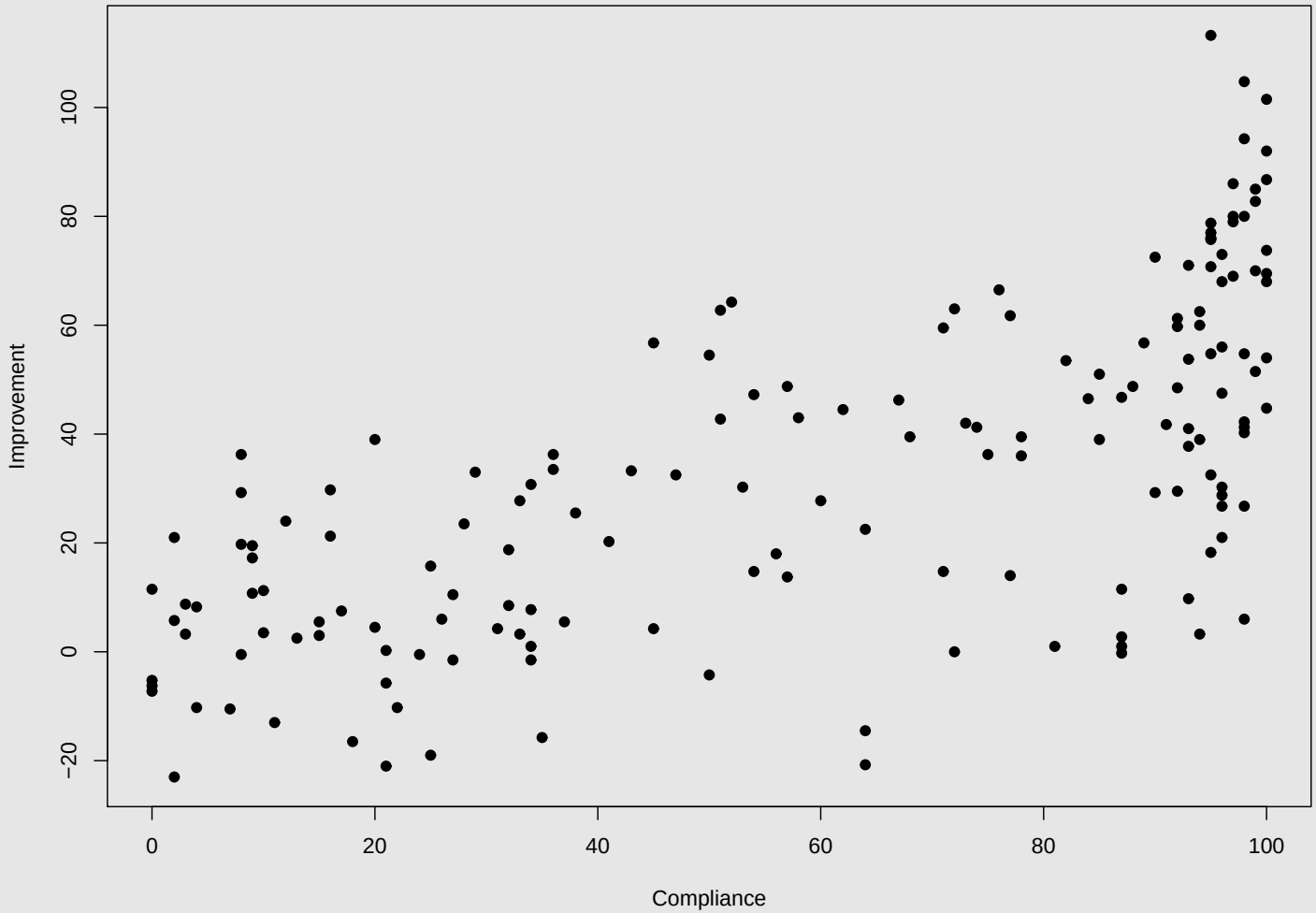
for $b = 1, \ldots, B$.

Again it is the i.i.d. units that are being resampled.

The Bootstrap: Regression

Example (Cholostyramine Data)

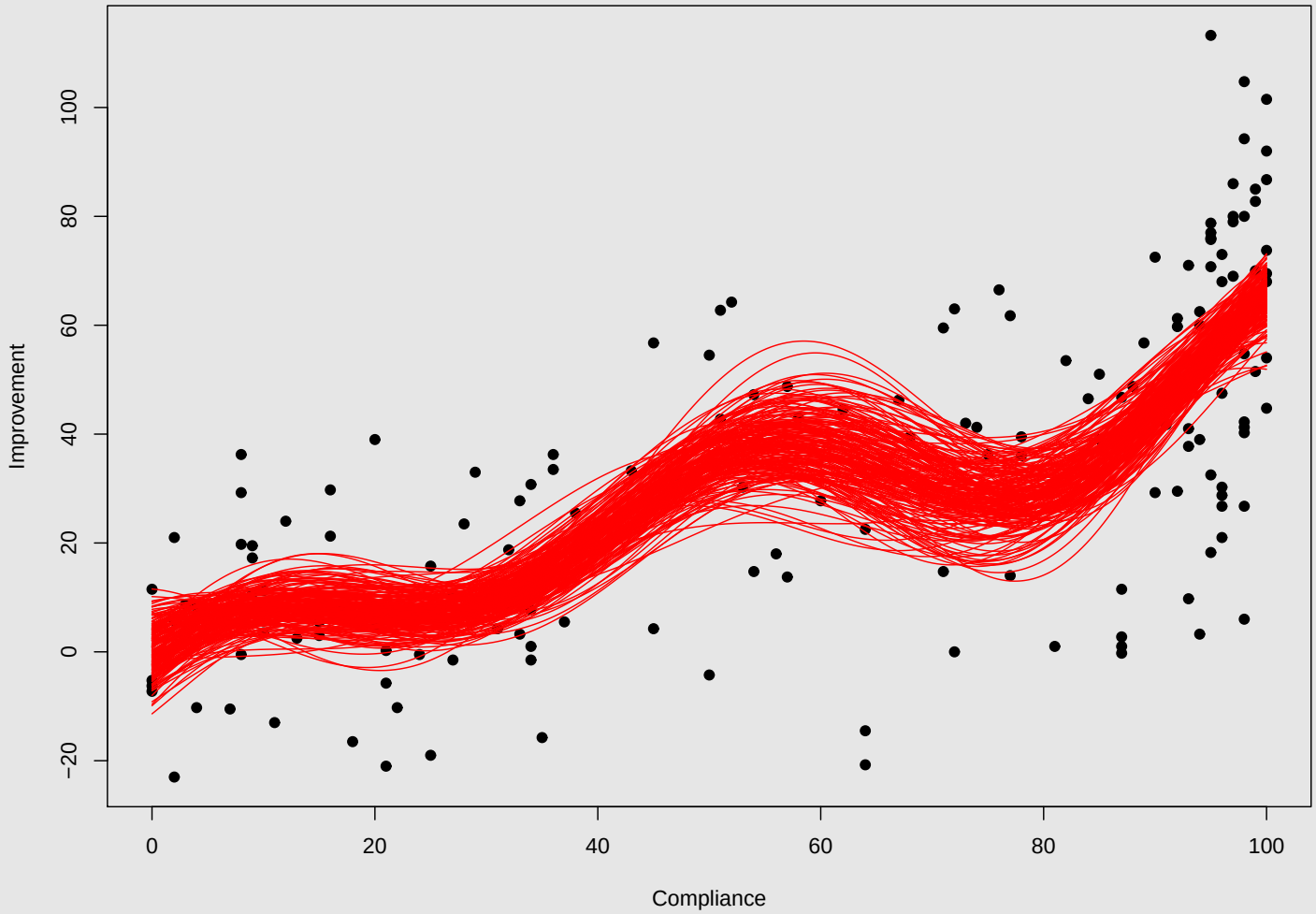
$B = 200$ bootstrapped curves with $q = 2$.



The Bootstrap: Regression

Example (Cholostyramine Data)

$B = 200$ bootstrapped curves with $q = 2$.



The Bootstrap: Loess

A more flexible model is based on *local smoothing* using *loess*.

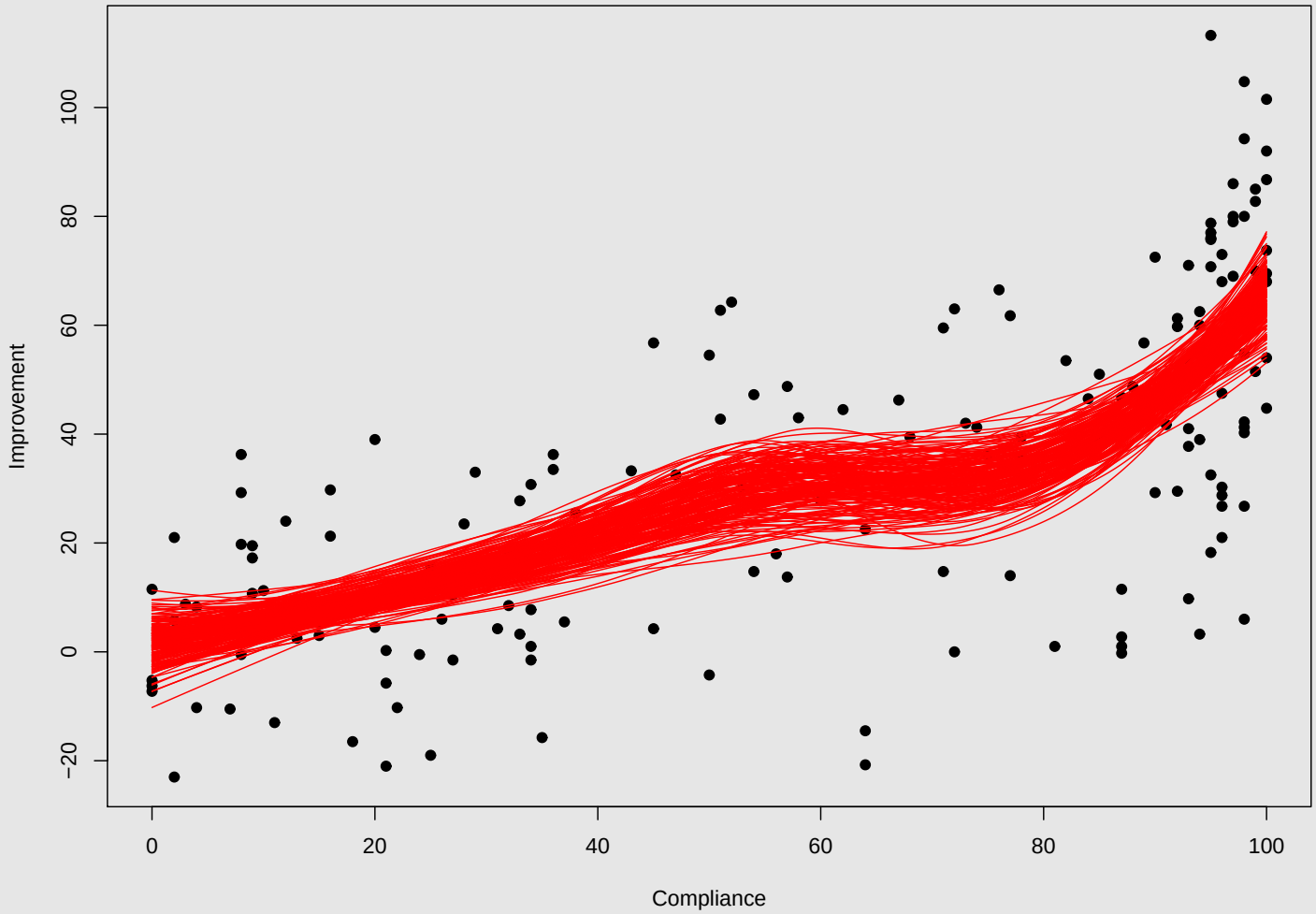
The loess approach fits a series of weighted local linear regression models restricted to a rolling neighbourhood of data points. The size of the neighbourhood (the *span*) determines the smoothness of the resulting fitted curve.

The statistical properties of the resulting fitted curve are hard to study analytically, but can be studied by the bootstrap.

The Bootstrap: Loess

Example (Cholostyramine Data)

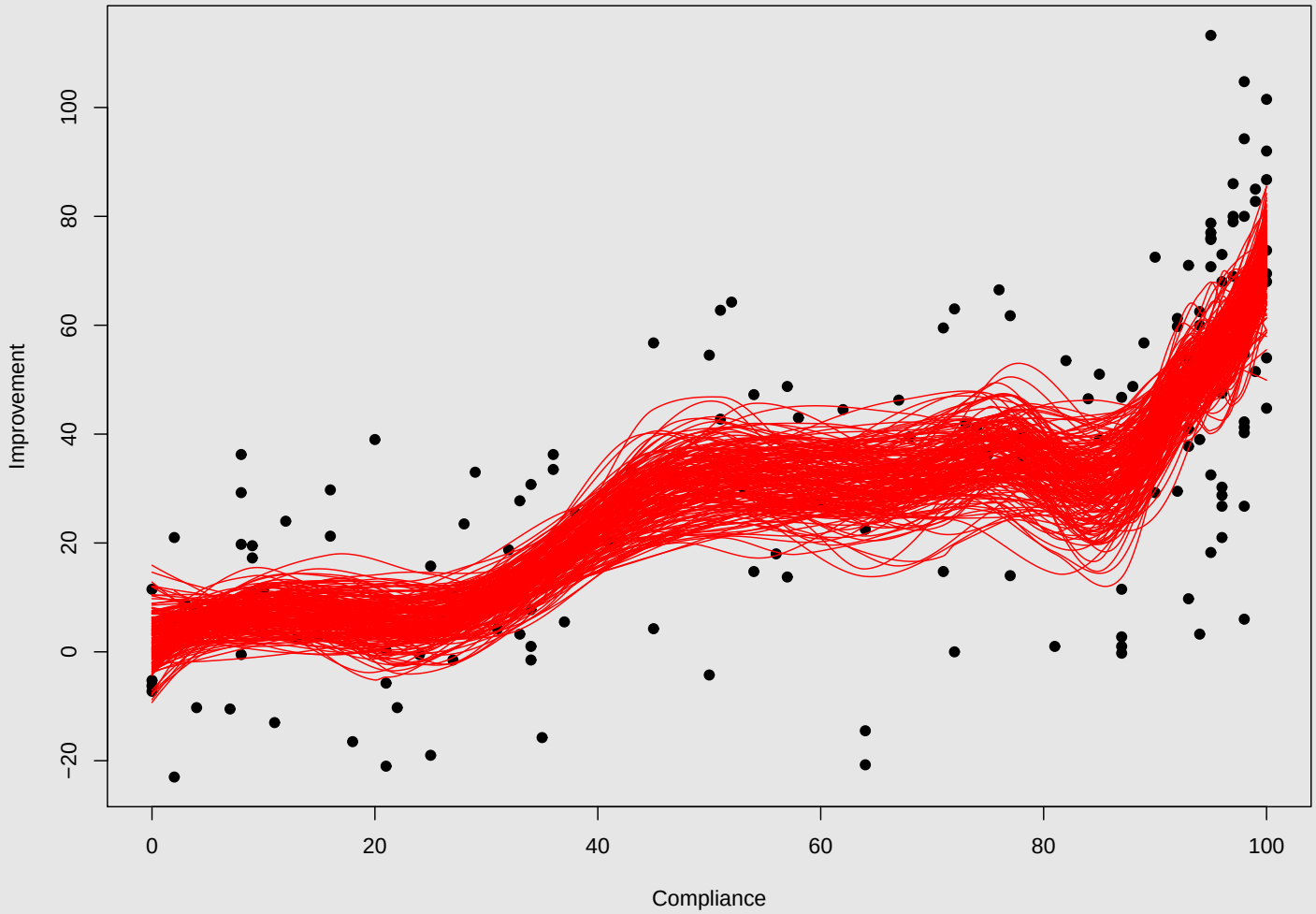
$B = 200$ bootstrapped loess curves: span is 0.75.



The Bootstrap: Loess

Example (Cholostyramine Data)

$B = 200$ bootstrapped loess curves: span is 0.375.



The Bootstrap: Loess

The choice of the span can be determined by K-fold *cross-validation*:

1. split data at random into K equal portions of size m.
2. fit the loess model using $K-1$ *training* portions using a fixed value of the span.
3. use the fitted model to predict the values in the remaining *test* portion; record the prediction mean square error

$$\frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_i - y_i)^2$$

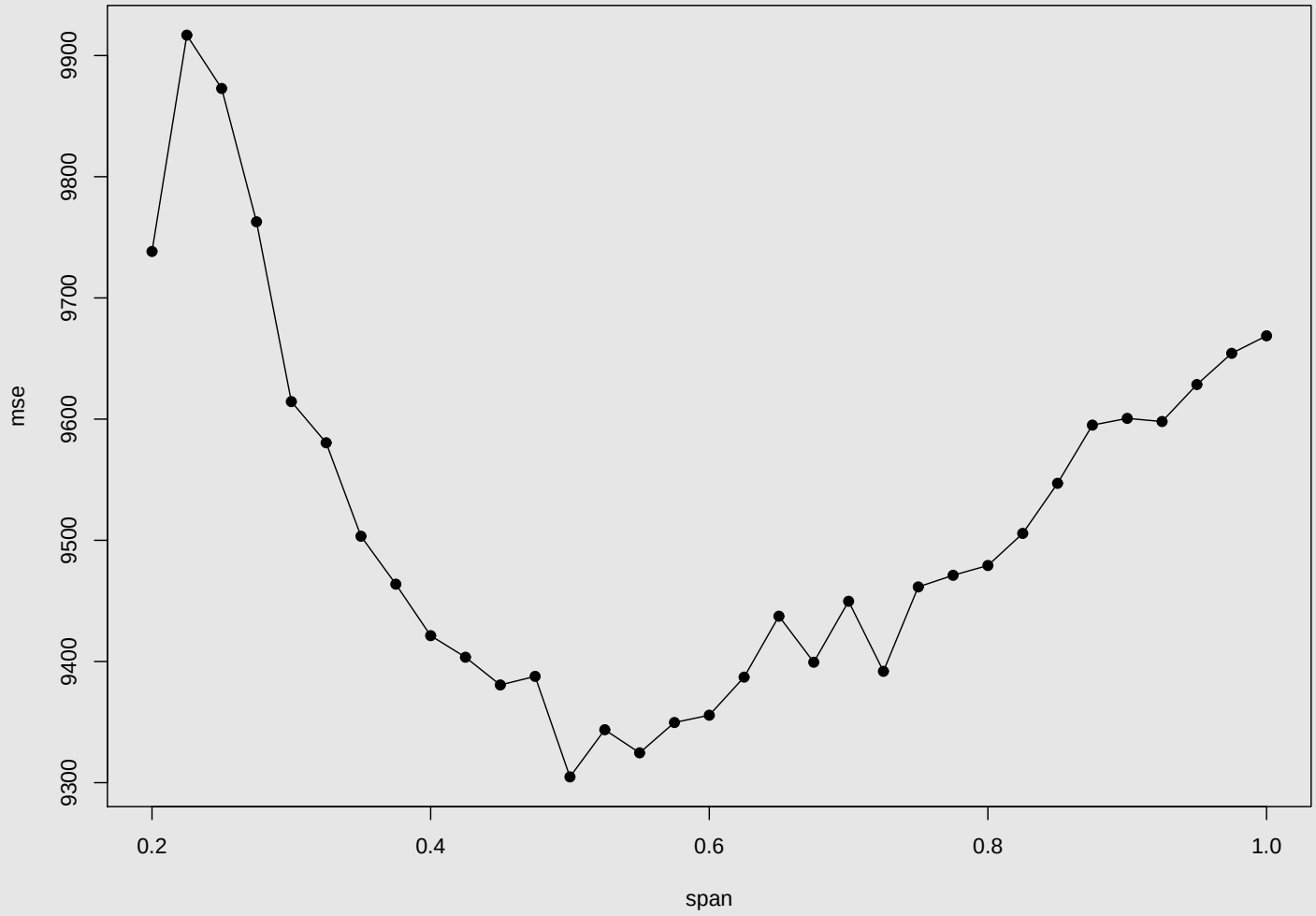
4. repeat 1. - 3. over different random splits, taking the average prediction MSE.

and repeat this whole process for different values of the span, recording the MSE every time. Choose the optimal span to minimize the prediction MSE.

The Bootstrap: Loess

Example (Cholostyramine Data)

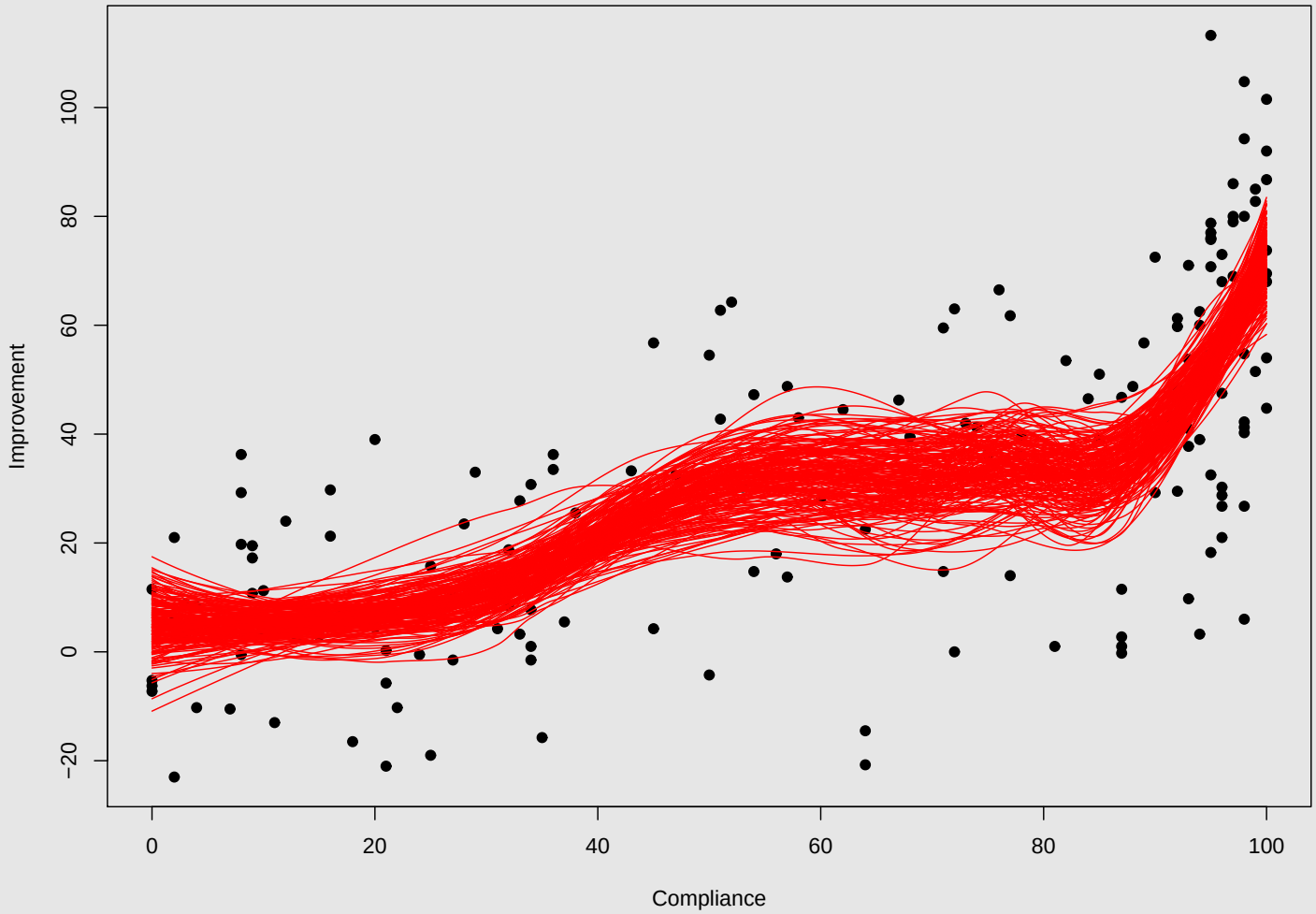
8-fold cross-validation: average prediction MSE from 20000 random splits.



The Bootstrap: Loess

Example (Cholostyramine Data)

$B = 200$ bootstrapped loess curves: span is 0.50.



The Bootstrap: Loess instability

Note: in this *semi-parametric* setting, the cross-validation can become unstable. A specific random split may disrupt the local structure of the data so that a local fit at any given x is not possible.

This instability can potentially occur in the parametric setting as well; if the x values contain ties (as for the cholostyramine data) then, by chance we may resample an $\boldsymbol{X}_{(b)}$ such that

$$\boldsymbol{X}_{(b)}^{\mathsf{T}}\boldsymbol{X}_{(b)}$$

is singular.

The Bootstrap: Problems

Suppose $Y_1, \ldots, Y_n \sim \mathcal{N}(0,1)$. Consider the statistic

$$T_n = \max\{Y_1, \ldots, Y_n\}$$

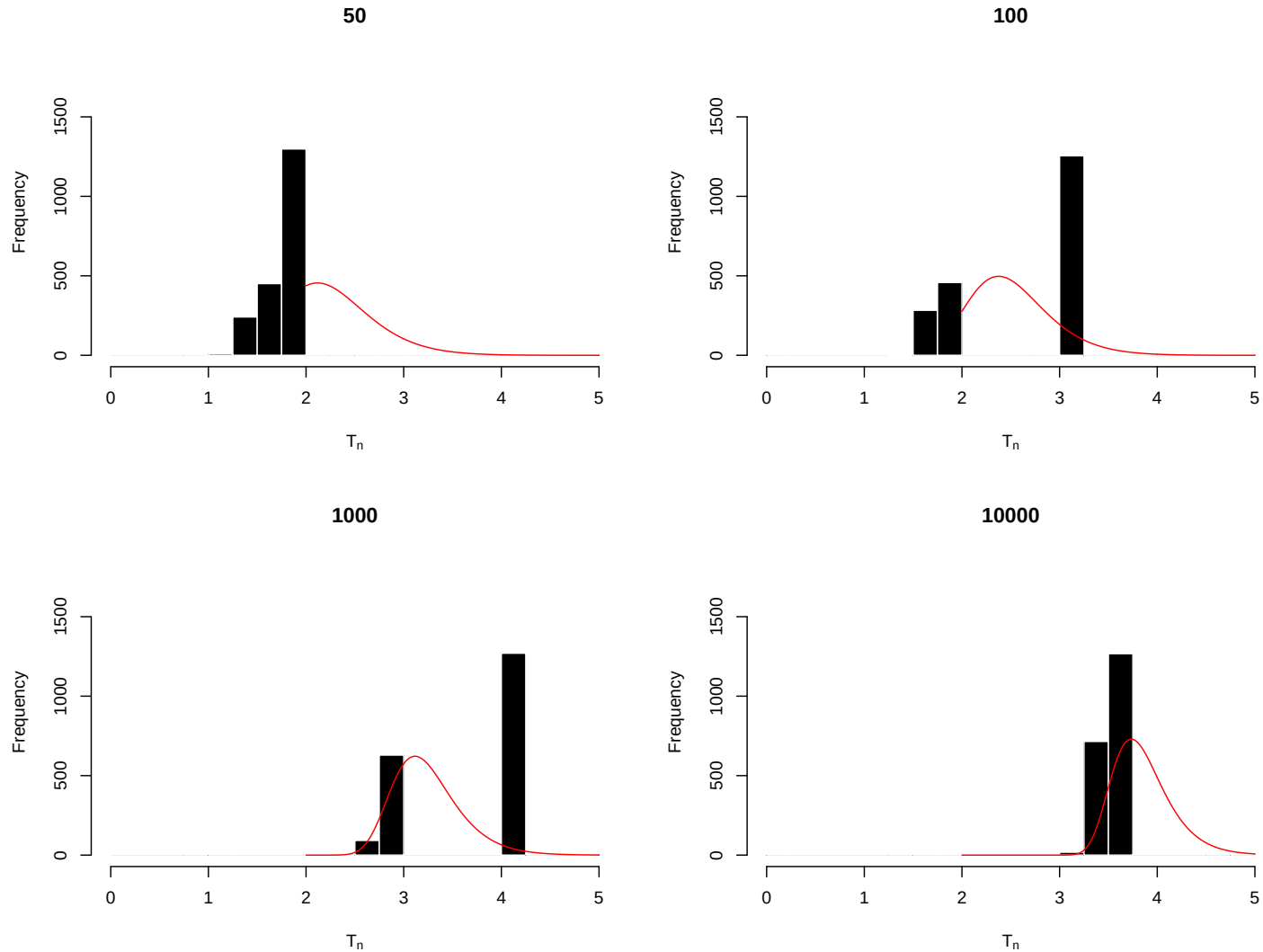
We know that for $t \in \mathbb{R}$

$$F_{T_n}(t) = \{\Phi(t)\}^n \qquad\qquad f_{T_n}(t) = n\phi(t)\{\Phi(t)\}^{n-1}$$

Can we study this statistic via the bootstrap ?

The Bootstrap: Problems

$B = 2000$ bootstrapped maxima for $n = 50, 100, 1000, 10000$.



The Bootstrap: Problems

The maximum order statistic cannot be studied in this setting using the bootstrap.

In this case, T_n has no limiting distribution, that is, as $n \longrightarrow \infty$

$$F_{T_n}(t) \longrightarrow 0$$

for all t. Is this the source of the problem ?

The Bootstrap: Problems

Suppose $Y_1, \ldots, Y_n \sim \mathit{Uniform}(0, \theta)$. Consider again the statistic T_n, which in this case is the ML estimator of θ:

$$\widehat{\theta}_n = T_n = \max\{Y_1, \ldots, Y_n\}.$$

We know that for $t \in \mathbb{R}$

$$F_{T_n}(t) = \left\{\frac{t}{\theta}\right\}^n \qquad\qquad f_{T_n}(t) = n\frac{1}{\theta}\left\{\frac{t}{\theta}\right\}^{n-1}$$

The same problem arises in this setting.

The Bootstrap: Problems

For the bth bootstrap resampled data set

$$\begin{aligned}\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] &= \mathbb{P}[\max\{Y_1^{(b)}, \ldots, Y_n^{(b)}\} = \max\{Y_1, \ldots, Y_n\}] \\ &= 1 - \left(1 - \frac{1}{n}\right)^n\end{aligned}$$

and as $n \longrightarrow \infty$

$$1 - \left(1 - \frac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \doteq 0.632$$

Therefore, even for large n, there is a probability mass at $\widehat{\theta}$ in the sampling distribution.

The Bootstrap: Problems

Other problems can occur for the bootstrap when resampling can fundamentally change the nature of the data set:

- **Mixture distributions:** if the Y_i follow a mixture distribution with K components, then resampling may only yield observations from $L < K$ of the components. This changes the statistical behaviour of some estimators.
- **Logistic Regression (or classification):** For binary data, resampling may eliminate **all** of the data with response value 0 (or 1), in which case the logistic regression coefficients are not defined.

The Parametric Bootstrap

The *parametric* bootstrap replaces **resampling** from $\widehat{F}_n$ with **sampling** from a parametric approximation $F(.; \widehat{\vartheta}_n)$, where $\widehat{\vartheta}_n$ is the parameter estimate derived from the original sample.

Essentially, this is a *Monte Carlo* technique, where $\widehat{s}(\widehat{\theta}^{(*)})$ is approximated using B random samples from $F(.; \widehat{\vartheta}_n)$.

This can be useful if a good (and tractable) parametric approximation exists.

The Parametric Bootstrap

Suppose $Y_1, \ldots, Y_n \sim \mathit{Uniform}(0, \theta)$.

- **Nonparametric bootstrap:** resampling from $\{y_1, \ldots, y_n\}$ leads to problems as

$$\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] > 0$$

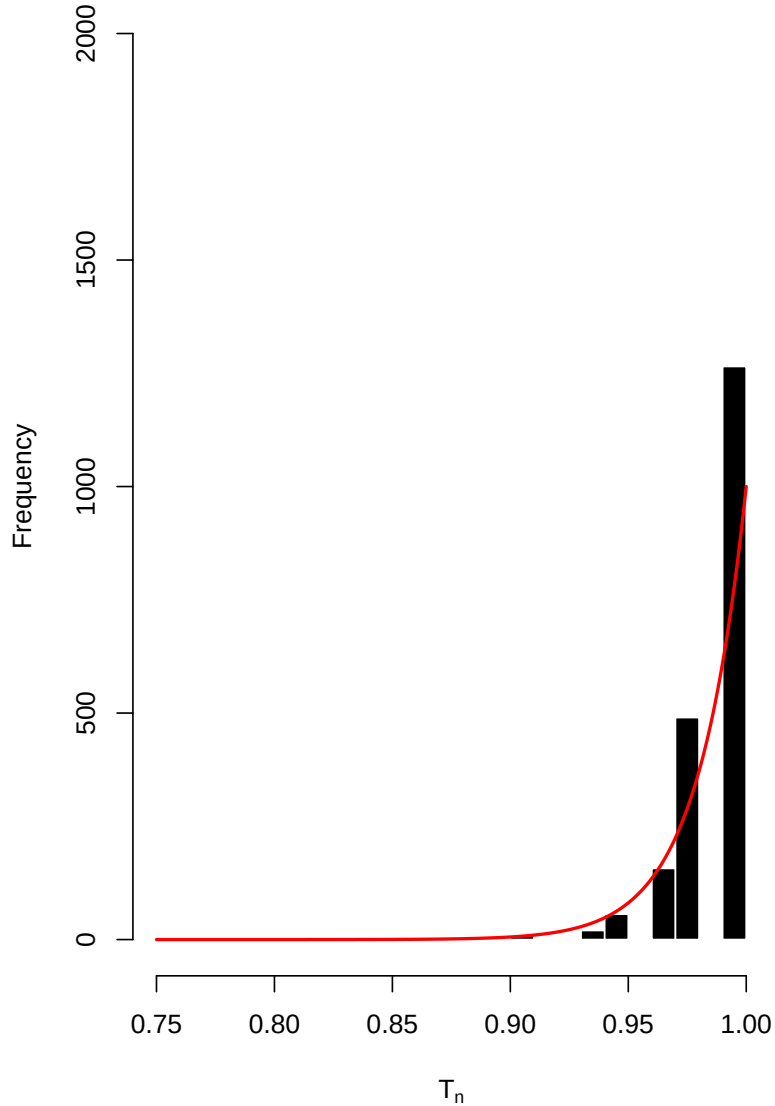
- **Parametric bootstrap:** sampling from $\mathit{Uniform}(0, \widehat{\theta})$ yields

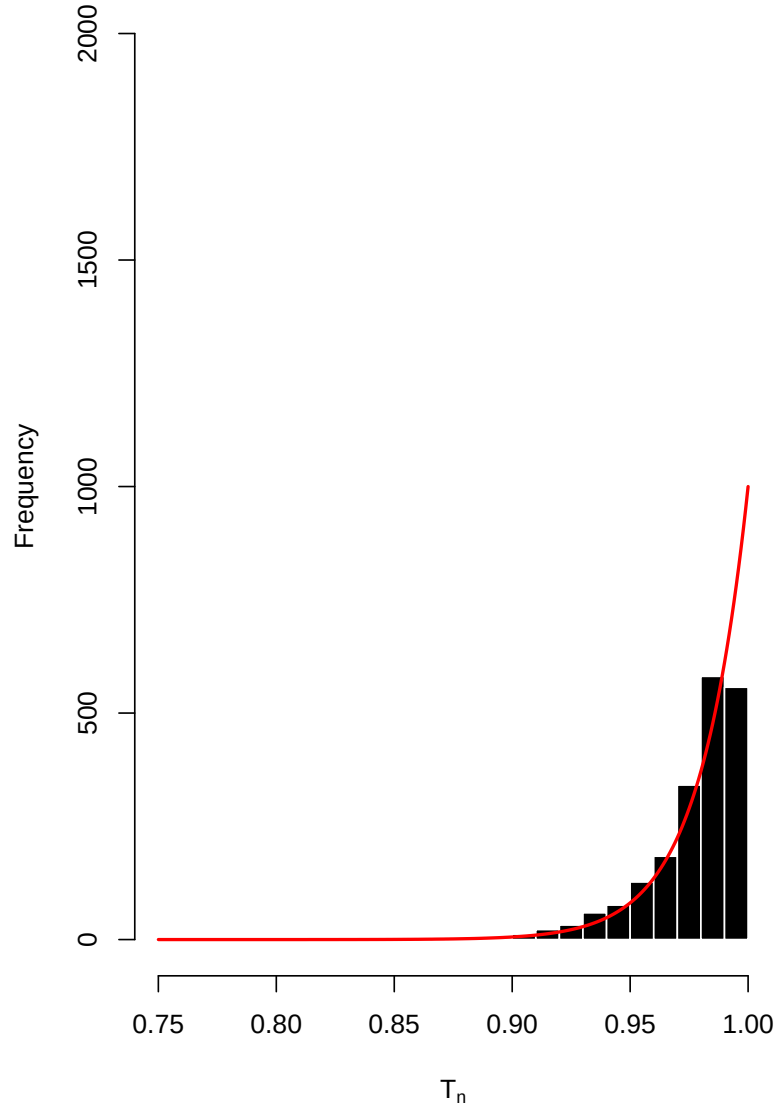
$$\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] = 0$$

  and a reasonable approximation to the distribution of $\widehat{\theta}$.

The Parametric Bootstrap

Non-parametric (left) vs Parametric (right)





Bootstrap regression: alternative approach

In an alternative formulation of the bootstrap for linear regression, consider

$$y_i = \boldsymbol{x}_i^\mathsf{T} \beta + \epsilon_i$$

where $\epsilon_i \sim F(.)$, where F is constrained to have mean zero.

If the distribution F is regarded as unknown, but the form of the linear regression is satisfactory (up to unknown parameters), attention can focus on bootstrapping the residual errors $\epsilon_1, \ldots, \epsilon_n$.

Bootstrap regression: alternative approach

In this β can be estimated by least-squares

$$\widehat{\beta} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

and then the residual cdf F estimated by the empirical cdf of the residuals

$$e_i = y_i - \boldsymbol{x}_i^\intercal \widehat{\beta} \qquad i = 1, \ldots, n$$

denoted $\widehat{F}_n$.

The bootstrap data $y_1^{(b)}, \ldots, y_n^{(b)}$ are then computed by producing a bootstrap resample $e_1^{(b)}, \ldots, e_n^{(b)}$, and setting

$$y_i^{(b)} = \boldsymbol{x}_i^\intercal \beta + e_i^{(b)}$$

Bootstrap regression: alternative approach

This then yields the bootstrap estimate

$$\widehat{\beta}^{(b)} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}^{(b)}.$$

In this case, the idealized bootstrap estimate ($B = \infty$) of the standard error is available analytically via

$$\mathbb{V}ar[\widehat{\beta}^{(b)}] = \widehat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

where

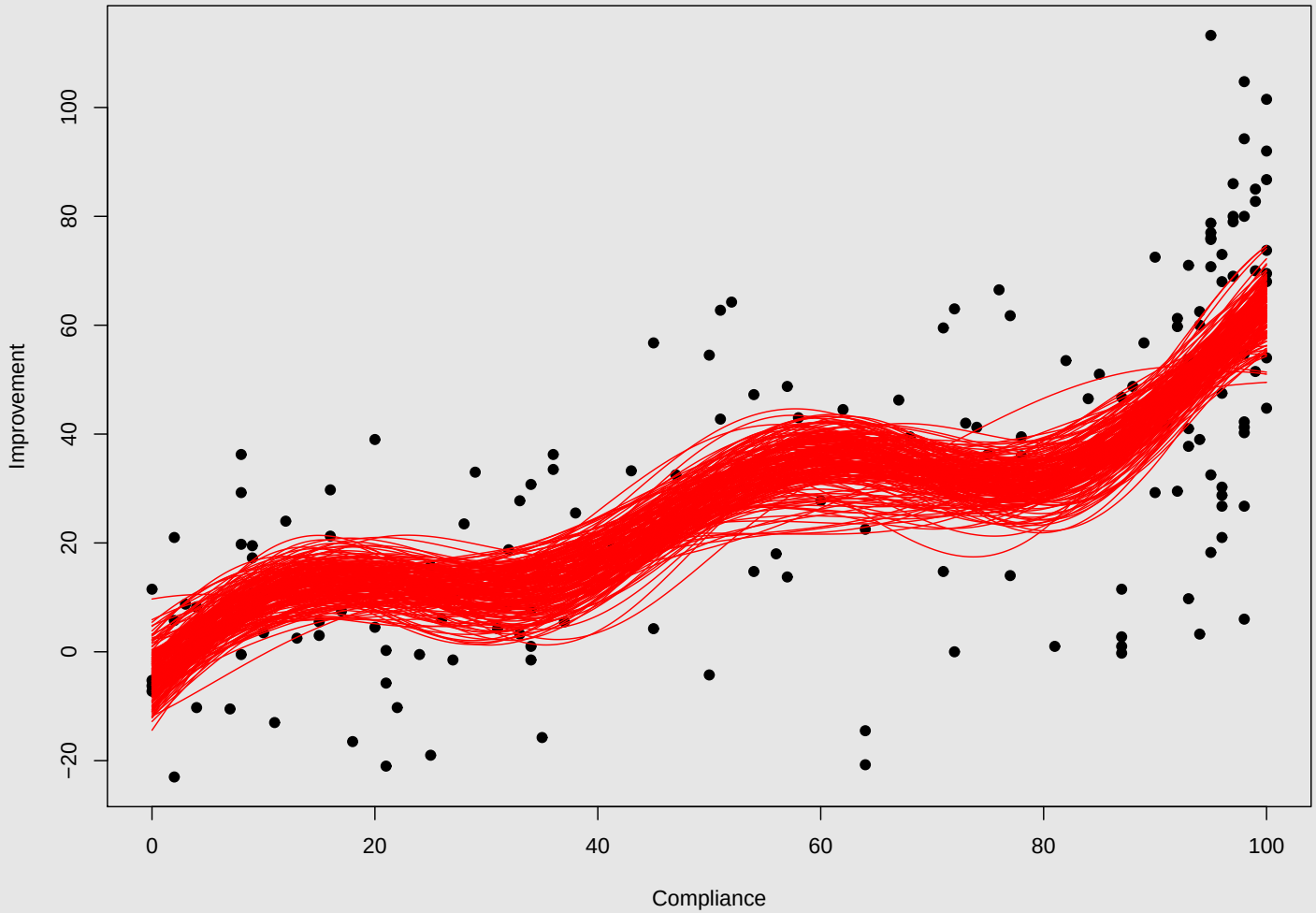
$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2$$

that is, identical to the usual variance in least-squares.

Bootstrap regression: alternative approach

Example (Cholostyramine Data)

$B = 200$ harmonic regressions with bootstrapped residuals.



Estimating Bias

The bias of an estimator T_n of $\theta(F)$ is

$$b_{T_n}(F) = \mathbb{E}_F[T_n] - \theta(F)$$

which can be estimated by

$$\widehat{b}_{T_n}(\widehat{F}_n) = \mathbb{E}_{\widehat{F}_n}[T_n^{(*)}] - \theta(\widehat{F}_n)$$

where $T_n^{(*)}$ is the bootstrap estimator, derived from bootstrap resamples

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

and $\theta(\widehat{F}_n)$ is the plug-in estimate. Typically $\mathbb{E}_{\widehat{F}_n}[T_n^{(*)}]$ is approximated by

$$\frac{1}{B}\sum_{b=1}^{B} t_n^{(b)}$$

Estimating Bias

If $T_n = \theta(\widehat{F}_n)$, that is, the estimator is the plug-in estimator, then an alternative estimate of bias can be obtained for $B = \infty$.

Denote by $P_j^{(b)}, j = 1, \ldots, n$ denote the proportion of a bootstrap resampled data set $y_1^{(b)}, \ldots, y_n^{(b)}$ that are equal to y_j.

$$P_j^{(b)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_j\}}(y_i^{(b)})$$

The (random) probability distribution $P^{(b)} = (P_1^{(b)}, \ldots, P_n^{(b)})$ (the *resampling vector*) can be used to construct the bootstrap estimate in terms of the original data; for example,

$$\overline{y}^{(b)} = \sum_{j=1}^{n} P_j^{(b)} y_j.$$

Estimating Bias

The bootstrap resampling vectors are thus

$$P^{(1)}, \ldots, P^{(B)}.$$

In general, denote the bootstrap estimator in terms of the

$$T_n^{(b)} = T_n(P^{(b)})$$

Let $P_0 = \frac{1}{n}\mathbf{1}_n$ denote the vector of length n containing identical elements $1/n$. Then, in this notation

$$T_n(P_0) = \theta(\widehat{F}_n)$$

Estimating Bias

Let

$$\overline{P}^{(*)} = \frac{1}{B}\sum_{b=1}^{B} P^{(b)}$$

This allows a new bias estimate

$$\widetilde{b}_{T_n}(\widehat{F}_n) = \frac{1}{B}\sum_{b=1}^{B} t_n^{(b)} - T_n(\overline{P}^{(*)})$$

to be compared with the plug-in estimate

$$\widehat{b}_{T_n}(\widehat{F}_n) = \frac{1}{B}\sum_{b=1}^{B} t_n^{(b)} - T_n(P_0)$$

Estimating Bias

Example (Hormone patch data)

Eight subjects were treated with two different hormone patches, and the subsequent level of hormone in the bloodstream measured.

ID	Placebo	Old Patch	New Patch	old-plac. z	new - old y
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

Estimating Bias

Example (Hormone patch data)

The parameter of interest here is

$$\theta = \frac{\mathbb{E}[Y]}{\mathbb{E}[Z]}$$

which is a measure of bio-equivalence. The plug-in estimate is

$$t_n = \theta(\widehat{F}_n) = \frac{\overline{y}}{\overline{z}} = -0.0713.$$

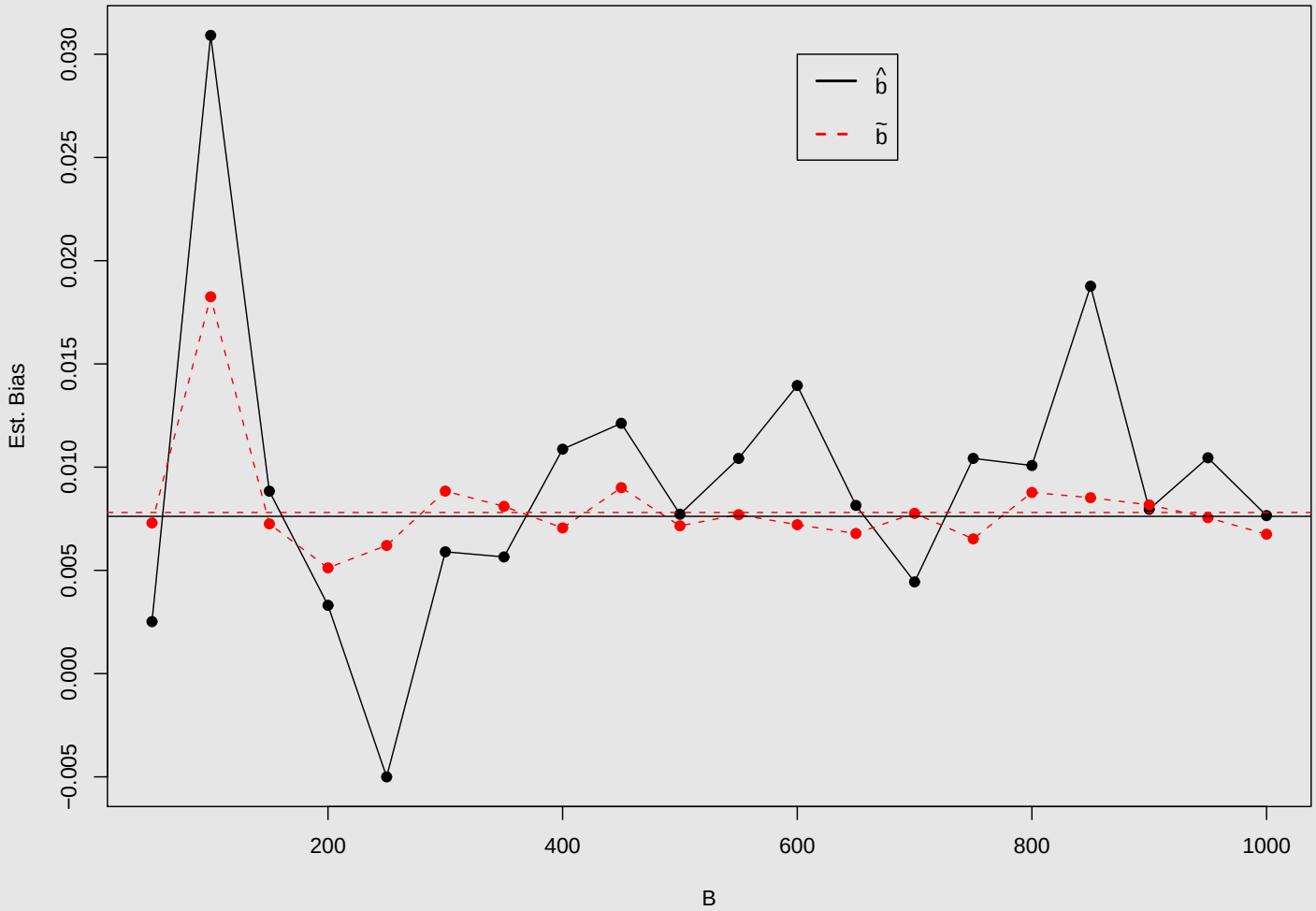
This parameter can be studied using the bootstrap, resampling individual subjects as the i.i.d. units.

For different values of B, we study the bias estimates given by the previous formulae.

Estimating Bias

Example (Hormone patch data)

Bias estimates $\widehat{b}$ and $\widetilde{b}$ for B up to 1000.



Estimating Bias

Example (Hormone patch data)

The bias estimate $\widetilde{b}$ is much less variable than $\widehat{b}$.

The two bias estimates agree approximately when $B = 100000$

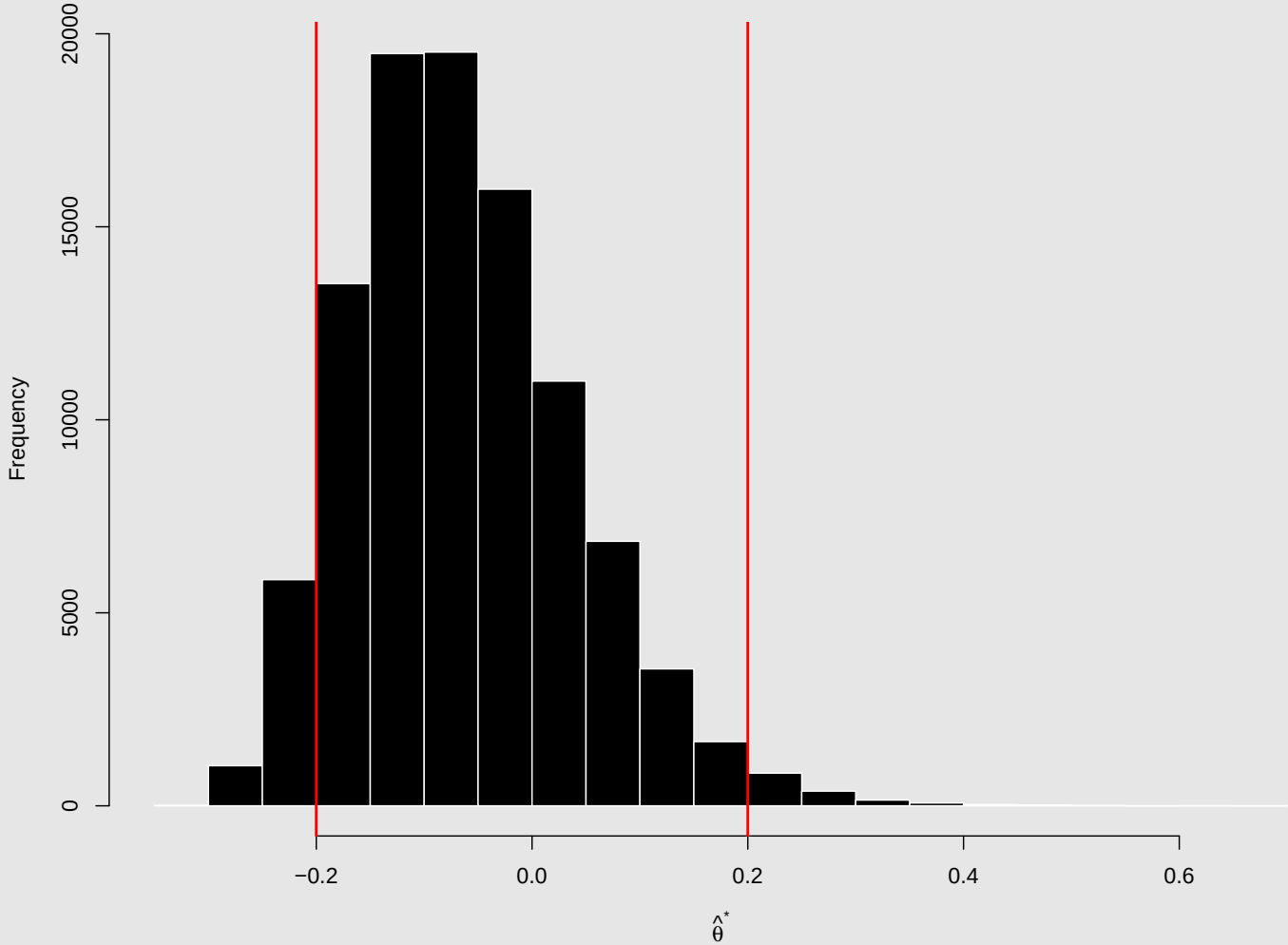
$$\widehat{b}_{T_n}(\widehat{F}_n) = 0.00762 \qquad \widetilde{b}_{T_n}(\widehat{F}_n) = 0.00781$$

but for B up to 1000, $\widetilde{b}$ is very stable, whereas $\widehat{b}$ fluctuates considerably.

Estimating Bias

Example (Hormone patch data)

Sampling distribution of $\widehat{\theta}$: $\mathbb{P}[|\widehat{\theta}| < 0.2] \simeq 0.9158$.



The jackknife

The *jackknife* was introduced as another bias estimation procedure. For $i = 1, \ldots, n$, let

$$\boldsymbol{y}_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$$

be the *i-deleted* data set (with the ith datum deleted). Let

$$\widehat{\theta}_{(i)} = t_{n-1}(\boldsymbol{y}_{(i)})$$

be the estimate derived from the i-deleted data, and let

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

A bias estimate for a plug-in statistic $\widehat{\theta}$ is then

$$\widehat{b}_{\text{JACK}} = (n-1)(\widehat{\theta}_{(.)} - \widehat{\theta})$$

The jackknife

In this calculation

$$\widehat{\theta} = \theta(\widehat{F}_n) \qquad\qquad \widehat{\theta}_{(i)} = \theta(\widehat{F}_{n(i)})$$

are the plug-in estimators derived from the full and i-deleted data respectively.

A jackknife estimate for the standard error of a statistic is

$$\widehat{\text{se}}_{\text{JACK}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)})^2}$$

This formula applies for *smooth* functions t of the data.

Note: There is no resampling here, so precisely n computations are needed.

The jackknife

Formally, the jackknife is an approximation to the bootstrap that arises from a linear approximation to the statistic of interest, derived from a Taylor expansion.

The above bias/s.e. formulae are derived under the assumption that the identity

$$\widehat{\theta} = t_n(\boldsymbol{y}) = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(y_i)$$

holds exactly from some function $\alpha(.)$. The jackknife arises from regarding a first-order Taylor **approximation** for arbitrary t_n.

The jackknife procedure fails when such an expansion fails; for example, for studying the **median**, the jackknife provides an *inconsistent* estimator.