Resampling Methods: Jackknife-Bootstrap

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester

Jackknife-General information

- Method developed around 1950 as a method to eliminate the bias of estimators. The method leads to standard errors for the estimators that can be easily computed.
- Let $X \sim F(\mu_F, \sigma_F^2)$ (F: cdf, $\mu_F = \mathbb{E}(X), \sigma_F^2 = \mathbb{V}(X)$). Then, if we want to estimate $\mu_F \to \bar{X} \underset{\text{CLT}}{\dot{\sim}} \mathcal{N}(\mu_F, \sigma_F^2/n) \Rightarrow se(\bar{X}) = \sigma_F/\sqrt{n}$ ($\sigma_F^2 \to S_F^2$ unbiased sample variance). But what happens with computing the standard errors of more complex estimators (e.g. median)?

  Thus Jackknife
  - → estimator with small bias
  - → easy computation of s.e.
- Let $X_1, X_2, \ldots X_n$ be a random sample from $f(x;\theta)$ and $\hat{\theta} = T(X_1, \ldots X_n)$. We denote by $\hat{\theta}_{(i)} = T(X_1, \ldots X_{i-1}, X_{i+1} \ldots X_n)$ (i.e. removing observation i).
- **Idea**: removing observations from the original sample and estimating again the parameter of interest, we can get information about the stability and variability of our estimator. That is, if we remove each time one observation and examine how much the values of our estimator change, we have an image of the variance of our estimator.

Jackknife estimator

The Jackknife estimator $\hat{\theta}_J$ of the parameter θ is:

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)}\,, \quad \text{where} \quad \bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$$

$\hat{\theta}_J \rightarrow$ we repeat the estimation of the parameter n times where each time our sample is the original by omitting one observation.

From the definition, we observe that we simply correct $\hat{\theta}$ using the estimators from the samples with one observation out each time.

Alternatively, $\hat{\theta}_J$ can be computed with the help of pseudovalues $p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$ as follows:

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i$$

Jackknife - Example 1

Observations $x_1, x_2, \ldots x_6$ corresponding to the waiting time of a bus (in mins)

	Estimator after removing i obs.		Pseudovalues	
Observations	Sample mean	Sample median	Sample mean	Sample median
4	6	6	4	3
3	6.2	6	3	3
7	5.4	5	7	8
6	5.6	5	6	8
5	5.8	6	5	3
9	5	5	9	8
34	34	33	34	33

where the last row depicts the sum.
The last column has only two values!

Jackknife - Example 1 (cont'd)

- First of all $\bar{x} = 5.666$, median $= 5.5$.
- Further, $\widehat{\bar{x}}_J = 6 \times 5.666 - 5 \times \frac{1}{6} \times 34 = 5.666$ and $\widehat{\text{median}}_J = 6 \times 5.5 - 5 \times \frac{1}{6} \times 33 = 5.5$
- pseudovalues
  e.g. $p_1 = 6 \times 5.666 - 5 \times 6 = 4$ (for the mean for example)
  Thus again we get

$$\widehat{\bar{x}}_J = \frac{1}{6} \times 34 = 5.666$$

$$\widehat{\text{median}}_J = \frac{33}{6} = 5.5$$

- We see that the Jackknife estimators coincide with the estimators we had. Is this always the case?

Different Jackknife estimators - Mean

A. Mean

Let $\hat{\theta} = \bar{X}$ and let us compute $\hat{\theta}_J$ with the help of pseudovalues:

$$
\begin{aligned}
p_i &= n\hat{\theta} - (n-1)\hat{\theta}_{(i)} = n\hat{\theta} - \cancel{(n-1)} \frac{\sum_{j=1}^{n} X_j - X_i}{\cancel{n-1}} \\
&= n\hat{\theta} - (n\hat{\theta} - X_i) = X_i
\end{aligned}
$$

Thus $\hat{\theta}_J = \frac{1}{n} \sum_{i=1}^{n} X_i = \hat{\theta} \rightarrow$ in other words for the mean value there is no reason to use the Jackknife method.

In a similar way, it can be shown that for every estimator of the form $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$ the Jackknife estimator coincides with the simple estimator.

Different Jackknife estimators - Median

B. Median

i) Sample size n (**even**)

Without loss of generality, let $X_1, X_2, \ldots X_n$ be the random sample sorted in increasing order. Then,

$$\hat{\theta}_{(i)} = \begin{cases} X_{n/2} & i \geq n/2+1 \quad (n/2 \text{ such observations}) \\ X_{n/2+1} & i < n/2+1 \quad (n/2 \text{ such observations}) \end{cases}$$

Thus,

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\frac{n}{2}\left(X_{n/2} + X_{n/2+1}\right) = \frac{X_{n/2} + X_{n/2+1}}{2} = \hat{\theta} \text{ (: sample median)}$$

Consequently, $\hat{\theta}_J = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)} = \hat{\theta}$, i.e. COINCIDE

ii) Sample size n (**odd**)

Let again the observations be in increasing order. Then,

$$2\hat{\theta}_{(i)} = \begin{cases} X_{(n+1)/2} + X_{(n+1)/2+1} & i < (n+1)/2 \quad ((n-1)/2 \text{ such observations}) \\ X_{(n+1)/2+1} + X_{(n+1)/2-1} & i = (n+1)/2 \quad (1 \text{ such observation}) \\ X_{(n+1)/2} + X_{(n+1)/2-1} & i > (n+1)/2 \quad ((n-1)/2 \text{ such observations}) \end{cases}$$

Different Jackknife estimators - Median (cont'd)

Thus,

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\Big[\frac{n-1}{2}\frac{1}{2}\left(X_{(n+1)/2}+X_{(n+1)/2+1}\right)+\frac{1}{2}\left(X_{(n+1)/2+1}+X_{(n+1)/2-1}\right) + \frac{n-1}{2}\frac{1}{2}\left(X_{(n+1)/2}+X_{(n+1)/2-1}\right)\Big]$$

So,

$$\begin{aligned}
\hat{\theta}_J &= n\hat{\theta}-(n-1)\bar{\hat{\theta}}_{(\cdot)} \\
&= nX_{(n+1)/2}-(n-1)\frac{1}{n}\Big[\frac{n-1}{2}\frac{1}{2}\left(X_{(n+1)/2}+X_{(n+1)/2+1}\right) \\
&+ \frac{1}{2}\left(X_{(n+1)/2+1}+X_{(n+1)/2-1}\right)+\frac{n-1}{2}\frac{1}{2}\left(X_{(n+1)/2}+X_{(n+1)/2-1}\right)\Big] \\
&= nX_{(n+1)/2}-(n-1)\frac{1}{n}\Big[\frac{n-1}{4}\left(X_{(n+1)/2}+X_{(n+1)/2+1}\right) \\
&+ \frac{1}{2}\left(X_{(n+1)/2+1}+X_{(n+1)/2-1}\right)+\frac{n-1}{4}\left(X_{(n+1)/2}+X_{(n+1)/2-1}\right)\Big] \\
&= nX_{(n+1)/2}-\frac{n-1}{n}\Big[\left(\frac{n-1}{4}+\frac{1}{2}\right)X_{(n+1)/2+1}+\left(\frac{n-1}{4}+\frac{n-1}{4}\right)X_{(n+1)/2} \\
&+ \left(\frac{n-1}{4}+\frac{1}{2}\right)X_{(n+1)/2-1}\Big]
\end{aligned}$$

Different Jackknife estimators - Median (cont'd) (cont'd)

$$
\begin{aligned}
&= nX_{(n+1)/2} - \frac{n-1}{n}\Big[\left(\frac{n-1}{4} + \frac{1}{2}\right)\left(X_{(n+1)/2+1} + X_{(n+1)/2-1}\right) \\
&+ \left(\frac{n-1}{4} + \frac{n-1}{4}\right) X_{(n+1)/2}\Big] \\
&= nX_{(n+1)/2} - \frac{n-1}{n}\Big[\left(\frac{n+1}{4}\right)\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right) + \left(\frac{n-1}{2}\right) X_{(n+1)/2}\Big] \\
&= nX_{(n+1)/2} - \frac{(n-1)(n+1)}{4n}\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right) - \frac{(n-1)^2}{2n} X_{(n+1)/2} \\
&= \left[n - \frac{(n-1)^2}{2n}\right] X_{(n+1)/2} - \frac{(n-1)(n+1)}{4n}\left(X_{\frac{n+1}{2}+1} + X_{\frac{n+1}{2}-1}\right) \neq X_{\frac{n+1}{2}}
\end{aligned}
$$

e.g. if $n = 15$ (ordered observations)

$$\hat{\theta}_J = 8.47x_8 - 3.73(x_9 + x_7), \quad \hat{\theta} = x_8$$

Different Jackknife estimators - Variance

C. Variance

Consider the biased sample variance

$$
\begin{aligned}
\hat{\theta} = S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right] \overset{\sum_{i=1}^{n} X_i = n\bar{X}}{=} \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - n\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right] \\
&= \frac{1}{n^2}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right] = \frac{(n-1)\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_i X_j}{n^2} \\
\Rightarrow \quad n\hat{\theta} &= \frac{(n-1)\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_i X_j}{n}
\end{aligned}
$$

Omitting observation i it can be shown in a similar way that

$$
(n-1)\hat{\theta}_{(i)} = \left[(n-2)\sum_{j=1, j\neq i}^{n} X_j^2 - \sum_{k=1, k\neq i}^{n}\sum_{j=1, j\neq k, i}^{n} X_k X_j\right] / (n-1)
$$

Different Jackknife estimators - Variance (cont'd)

Thus,

$$
\begin{aligned}
(n-1)\bar{\hat{\theta}}_{(\cdot)} &= \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(n-2)\sum_{j=1,j\neq i}^{n}X_j^2-\sum_{k=1,k\neq i}^{n}\sum_{j=1,j\neq k,i}^{n}X_kX_j}{n-1}\right]
\end{aligned}
$$

But in the sum of squares, each term appears $(n-1)$ times, while in the product $(n-2)$ times. Thus,

$$
(n-1)\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\left[(n-2)\sum_{i=1}^{n}X_i^2 - \frac{n-2}{n-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}X_iX_j\right]
$$

Different Jackknife estimators - Variance (cont'd)

Therefore,

$$
\begin{aligned}
\hat{\theta}_J &= n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)} = \frac{(n-1)}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_i X_j \\
&- \frac{1}{n}\left[(n-2)\sum_{i=1}^{n} X_i^2 - \frac{n-2}{n-1}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_i X_j\right] \\
&= \left(\frac{n-1}{n} - \frac{n-2}{n}\right)\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} - \frac{n-2}{n(n-1)}\right)\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_i X_j \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_i X_j \\
&= \frac{(n-1)\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_i X_j}{n(n-1)} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, \quad \text{i.e. the unbiased variance estimator}
\end{aligned}
$$

Jackknife - Example 2

$X_1, X_2, \ldots X_n \sim$ Bernoulli(p) and we would like to estimate $\theta = p^2$.
The MLE of p is R/n, where $R = \sum_{i=1}^n X_i \quad \rightarrow \quad \hat{\theta} = (R/n)^2$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[(R/n)^2\right] = \frac{1}{n^2}\mathbb{E}[R^2].$$

But $R \sim \text{Bin}(n, p)$ $\begin{cases} \mathbb{E}[R] = np \\ \mathbb{V}[R] = np(1-p) \end{cases}$

Thus, $\mathbb{E}[R^2] = \mathbb{V}[R] + (\mathbb{E}[R])^2 = np(1-p) + n^2p^2$.

Consequently, $\mathbb{E}[\hat{\theta}] = p(1-p)/n + p^2 \neq p^2$: i.e. biased

Let us now find the Jackknife estimator:

$$\hat{\theta}_{(i)} = \left(\frac{R - X_i}{n-1}\right)^2 = \frac{R^2 - 2RX_i + {X_i}^2}{(n-1)^2}, \quad \text{and}$$

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^n \frac{R^2 - 2RX_i + {X_i}^2}{(n-1)^2} = \frac{nR^2 - 2R\sum_{i=1}^n X_i + \sum_{i=1}^n {X_i}^2}{n(n-1)^2}$$

Since $X_i = 0$ or $1 \rightarrow R = \sum_{i=1}^n X_i = \sum_{i=1}^n {X_i}^2$, so

$$\bar{\hat{\theta}}_{(\cdot)} = \left(nR^2 - 2R^2 + R\right)/n(n-1)^2$$

Jackknife - Example 2 (cont'd)

Thus,

$$
\begin{aligned}
\hat{\theta}_J &= n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)} = n\left(\frac{R}{n}\right)^2 - (n-1)\frac{nR^2 - 2R^2 + R}{n(n-1)^2} \\
&= \frac{R^2}{n} - \frac{nR^2 - 2R^2 + R}{n(n-1)} = \frac{(n-1)R^2 - nR^2 + 2R^2 - R}{n(n-1)} \\
&= \frac{R^2 - R}{n(n-1)} = \frac{R(R-1)}{n(n-1)}.
\end{aligned}
$$

So,

$$
\begin{aligned}
\mathbb{E}[\hat{\theta}_J] &= \mathbb{E}\left[\frac{R(R-1)}{n(n-1)}\right] = \frac{1}{n(n-1)}\mathbb{E}[R^2 - R] = \frac{1}{n(n-1)}\left[\mathbb{E}[R^2] - \mathbb{E}[R]\right] \\
&= \frac{np(1-p) + n^2p^2 - np}{n(n-1)} = \frac{n^2p^2 - np^2}{n(n-1)} = \frac{n(n-1)p^2}{n(n-1)} = p^2
\end{aligned}
$$

i.e. unbiased

Standard errors and bias of Jackknife estimators

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} p_i \quad , \quad p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$$

Thus,

$$\mathbb{V}[\hat{\theta}_J] = \left(\sum_{i=1}^{n} \mathbb{V}[p_i]\right)/n^2 = \mathbb{V}[p_1]/n$$

since p_i are independent $\to$ uncorrelated (since X_i are independent) and all have the same variance (e.g. $\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$, $\hat{\theta}_{(i)} = \frac{\sum_{j=1,j\neq i}^{n} X_j}{n-1}$, $\mathbb{V}(\hat{\theta}_{(i)}) = \frac{\sigma_f^2}{n-1}$)

An unbiased estimator of the variance of the pseudovalues is $S_p^2 = \frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n-1}$, thus

$$\begin{aligned} S^2_{\hat{\theta}_J} &= \frac{\sum_{i=1}^{n}(p_i-\bar{p})^2}{n(n-1)} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{(i)} - \frac{1}{n}\sum_{j=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{(j)}\right)\right)^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\cancel{n\hat{\theta}} - (n-1)\hat{\theta}_{(i)} - \cancel{n\hat{\theta}} + \frac{n-1}{n}\sum_{j=1}^{n}\hat{\theta}_{(j)}\right)^2 \overset{\sum_{j=1}^{n}\hat{\theta}_{(j)}/n=\bar{\hat{\theta}}_{(\cdot)}}{=} \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left((n-1)\bar{\hat{\theta}}_{(\cdot)} - (n-1)\hat{\theta}_{(i)}\right)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\cdot)}\right)^2 \end{aligned}$$

Standard errors and bias of Jackknife estimators (cont'd)

$\rightarrow$ i.e. is the variance of the values of the estimator once removing one observation multiplied by $n-1$.
Thus,

$$se_{\hat{\theta}_J} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\cdot)}\right)^2}$$

Since as we saw in several examples $\hat{\theta}_J = \hat{\theta}$, we could estimate $se_{\hat{\theta}}$ from $se_{\hat{\theta}_J}$. Having computed the standard deviation, we could construct approximate confidence intervals

$$\left(\hat{\theta}_J \pm t_{n-1,\alpha/2} se_{\hat{\theta}_J}\right)$$

Finally, we could estimate the $Bias[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta$ of $\hat{\theta}$:

$$\widehat{Bias}[\hat{\theta}] = (n-1)(\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta})$$

Notice then that

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_{(\cdot)} = \hat{\theta} - \widehat{Bias}[\hat{\theta}]$$

Jackknife - Conclusions

- Jackknife estimators reduce bias compared to the simple estimators and we can easily compute their standard errors.
- If $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ then $\hat{\theta}_J = \hat{\theta}$, so we can get an estimate of the standard error simply by computing $se_{\hat{\theta}_J}$.
- However, if the form of $\hat{\theta}$ is not a linear function of the data (e.g. median or max) then the Jackknife method is not satisfactory. E.g. in the 1st example, observe that the pseudovalues take only two values and thus $se_{\hat{\theta}_J}$ is rather smaller than expected. It can be shown that as $n \to \infty$ $se_{\hat{\theta}_J}$ is not consistent, i.e. it does not converge to the real se. In those cases, it is better to use a generalization of the method called delete-d-Jackknife where we remove not 1 but more observations each time.
- Jackknife is a Non-parametric method $\rightarrow$ NO ASSUMPTION ABOUT THE POPULATION

Bootstrap method

- Resampling methods:
  - if we know the population distribution then Monte-Carlo methodology (parametric Bootstrap)
  - if not Jackknife (sample out of the sample) $\rightarrow$ issues because the different samples we get are very similar to each other so this does not work well for statistical functions which are not smooth like the median
- Bootstrap (Efron, 1979)

  IDEA: we do not know the distribution of our population so cannot work with Monte-Carlo $\rightarrow$ estimate it with the empirical distribution of our data $\rightarrow$ gives probability $1/n$ to each observation of the sample and 0 else:

$$\left.\begin{array}{llllll} \text{value} & x_1 & x_2 & \dots & x_n & \text{else} \\ \text{prob} & \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} & 0 \end{array}\right\} \rightarrow \hat{F}_n(x)$$

$$\hat{F}_n(x) = \frac{\#\text{observ} \leq x}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}}{n}$$

How do we sample from $\hat{F}_n(x)$?

Bootstrap method (cont'd)

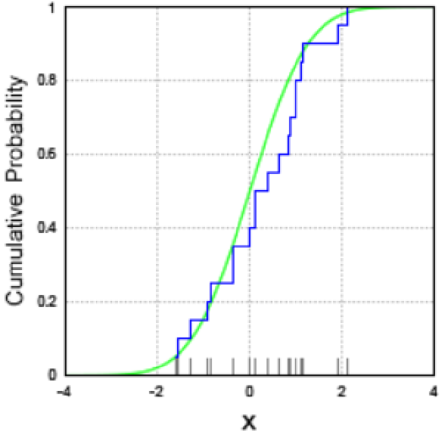
- To get a sample, of size n, from $\hat{F}_n(x)$, we perform sampling with replacement (if some value appears more than once, e.g. 2, then we have $n-1$ different values and one value with probability $2/n$. This is not a problem since it is equivalent with choosing each of the n observations with probability $1/n$).
- $\rightarrow$ Bootstrap sample:
  - it might contain some value(s) more than once (according to their appearance in the original sample)
  - some value(s) might not be present

  If we take B such samples then all the observations will appear with the frequency assumed by the empirical distribution and thus there is no problem.
- Thus, the basic idea of the method is that we perform sampling with replacement from the existing/original sample and thus make the assumption that the empirical distribution is a good approximation of the population distribution. When this is not the case (e.g. small n, multivariate problems, etc.) the bootstrap method does not work well.
- $$\text{Applications of the method} = \begin{cases} \text{s.e. estimation} \\ \text{bias} \\ \text{hypothesis testing} \\ \text{confidence intervals} \\ \text{approximating distributions} \end{cases}$$

Empirical distribution

Let $X_1, \ldots, X_n$ be a random sample of size n. Then the empirical distribution function is

$$\hat{F}_n(x) = \frac{\text{number of elements in the sample } \leq x}{n}$$



Parametric bootstrap illustration

Known Population with mean μ and SD σ. Create simple random samples (with simulation).

simple random sample $\nearrow$ $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow$ mean $\bar{x}$ and SD s

population $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$ imaginary samples $\rightarrow$ $\mathbf{x}_1^* = \begin{bmatrix} x_{11}^* \\ x_{12}^* \\ \vdots \\ x_{1n}^* \end{bmatrix} \rightarrow \bar{X}_1^*$ $\rightarrow \begin{bmatrix} \bar{X}_1^* \\ \bar{X}_2^* \\ \vdots \\ \bar{X}_B^* \end{bmatrix}$

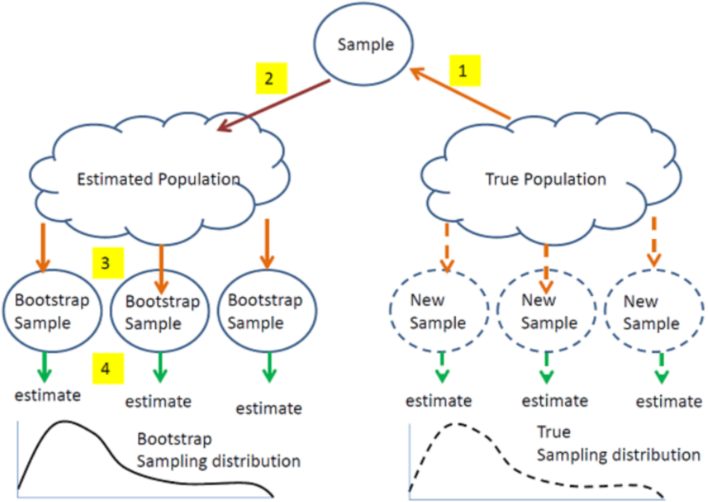
$\vdots$

imaginary samples $\searrow$ $\mathbf{x}_B^* = \begin{bmatrix} x_{B1}^* \\ x_{B2}^* \\ \vdots \\ x_{Bn}^* \end{bmatrix} \rightarrow \bar{X}_B^*$

$\mathbb{E}[\bar{X}] = \mu, se(\bar{X}) = s/\sqrt{n}$

In the parametric bootstrap, put the estimators (e.g. MLE) in the place of the unknown parameters.

Non-parametric bootstrap illustration



Non-parametric bootstrap illustration (cont'd)

$$\text{Sample} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \underset{\text{with replacement}}{\overset{\text{IID}}{=}} \begin{array}{l} \nearrow \mathbf{x}_1^* = \begin{bmatrix} x_{11}^* \\ x_{12}^* \\ \vdots \\ x_{1n}^* \end{bmatrix} \to \hat{\theta}_1^* \\ \to \qquad \vdots \\ \searrow \mathbf{x}_B^* = \begin{bmatrix} x_{B1}^* \\ x_{B2}^* \\ \vdots \\ x_{Bn}^* \end{bmatrix} \to \hat{\theta}_B^* \end{array} \to \begin{bmatrix} \hat{\theta}_1^* \\ \hat{\theta}_2^* \\ \vdots \\ \hat{\theta}_B^* \end{bmatrix} = \hat{\boldsymbol{\theta}}^*$$

$$\bar{\hat{\theta}}^* \equiv \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* \to \mathbb{E}[\hat{\theta}]$$

$$\sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \bar{\hat{\theta}}^* \right)^2} \to se(\hat{\theta})$$

Bootstrap algorithm

1) Simulate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$
2) Compute $\hat{\theta}^*$ for the above sample
3) Repeat the previous two steps B times

$$\hat{\boldsymbol{\theta}}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots \hat{\theta}_B^*)$$

Then $\hat{\boldsymbol{\theta}}^*$ is a sample from the unknown distribution of $\hat{\theta}$, so we learn any information we want with Monte-Carlo techniques.

Example: data $\mathbf{x} = (x_1, \dots x_n)$. We use the sample median $\hat{\theta}$ to estimate the true median θ. $se(\hat{\theta}) =?$

* Create $\mathbf{x}_1^*$: a sample of size n; sampling with replacement from $\mathbf{x}$.
* Compute the sample median $\hat{\theta}_1^*$ for $\mathbf{x}_1^*$
* Repeat the previous two steps B times $\rightarrow \hat{\boldsymbol{\theta}}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots \hat{\theta}_B^*)$
* The sample standard deviation of the $\hat{\theta}_i^*$ $(i = 1, \dots, B)$ values can be used to estimate $se(\hat{\theta})$.

Bootstrap in R: `bootstrap` or `boot` package

Bootstrap hypothesis testing

Example

Consider the following $n = 10$ observations

$$-0.89, -0.97, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96$$

We would like to test $\mathcal{H}_0: \mu = 1$ vs $\mathcal{H}_1: \mu \neq 1$.
(Note: we have no information about the population distribution. Further, the sample being small, we cannot apply CLT.)
We choose as test statistic $T = |\bar{X} - 1|$ which will be close to 0 if $\mathcal{H}_0$ is true, while large values suggest deviations from $\mathcal{H}_0$. We do not know the exact distribution of $T \rightarrow$ bootstrap

$$\bar{x} = 0.598 \rightarrow T = 0.402$$

IDEA: create samples under $\mathcal{H}_0$
Problem: $\mathcal{H}_0: \mu = 1$ while in our data $\bar{x} = 0.598$
Solution: Add 0.402 to each observation

- Simulate B bootstrap samples (e.g. 99) from the empirical distribution shifted by 0.402 so that $\mathcal{H}_0$ is satisfied
- Compute T for each obtained sample
- $\hat{p}_{\text{value}} = \frac{m+1}{B+1}$, m : # samples with $T > 0.402$

Equivalently, we could construct a CI for μ (using bootstrap, we will see how in the next slides) and check if it contains the value of 1.

Bootstrap estimates of standard errors and bias

$X_1, \ldots X_n$ random sample and $\hat{\theta} = T(X_1, \ldots X_n)$. $se(\hat{\theta})$ =?, Bias($\hat{\theta}$)=?

- Create B bootstrap samples
- For each bootstrap sample i: $(X_1^*, X_2^* \ldots X_n^*)$ compute the value $\hat{\theta}_i^* = T(X_1^*, \ldots X_n^*)$. Then obtain the following estimates for bias and se of $\hat{\theta}$:
- $se(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \bar{\hat{\theta}}^*\right)^2}$, $\bar{\hat{\theta}}^* = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*$
- Bias($\hat{\theta}$) $= \bar{\hat{\theta}}^* - \hat{\theta}$ (the first term is an estimate of $\mathbb{E}[\hat{\theta}]$ and the second term is an estimate of θ)

Jackknife-Bootstrap comparison

n repetitions for Jackknife

B repetitions for Bootstrap $\rightarrow$ computational cost but works when Jackknife fails.

Bootstrap confidence intervals

- Suppose we have a random sample $X_1, \dots X_n$ from a population. Let $\hat{\theta} = T(X_1, \dots X_n)$ be a point estimate of an unknown population parameter θ.
- We wish to create a (symmetric) $100(1-\alpha)\%$ confidence interval (CI) for θ.
- In many cases, the general form of a confidence interval is

$$\left(\hat{\theta} - M \times se(\hat{\theta}), \hat{\theta} + M \times se(\hat{\theta})\right),$$

  where the multiplier M depends on our level of confidence and is coming from the sampling distribution g of the point estimate.
- $g = ?$, $M = ?$, $se(\hat{\theta}) = ?$
- If n is large, $\theta = \mu$ (i.e. the population mean) and $\hat{\theta} = T(X_1, \dots X_n) = \bar{X}$, we can apply the Central Limit Theorem (CLT) and work with percentiles of $N(0,1)$ or t_{n-1} in the place of M. Furthermore $\widehat{se}(\hat{\theta}) = S/\sqrt{n}$, with S being the sample s.d. of the data.
- In other cases?

Bootstrap confidence intervals (cont'd)

A) Classical bootstrap CIs

$$\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}) \quad \text{or} \quad \hat{\theta} \pm t_{n-1,\alpha/2} se(\hat{\theta}),$$

where $se(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$ using bootstrap.
Assumption: distribution of $\hat{\theta}$ normal. Issues when n is small or $\hat{\theta}$ not linear statistic.

B) Bootstrap t-CIs

Instead of using the percentiles of the normal distribution, we could use the percentiles of the distribution of $\hat{\theta} = T(X_1, \dots X_n)$ that we can estimate using bootstrap. Find the requested α-percentile $D(\alpha)$ from:

$$\{\#\text{values } z(\hat{\theta}_i^*) \leq D(\alpha)\}/B = \alpha$$

where $z(\hat{\theta}_i^*) = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$, i.e. we standardize the bootstrap values $\hat{\theta}_i^*$, $\forall i = 1, \dots = B$, and $se(\hat{\theta}_i^*)$ is the standard deviation of the i bootstrap sample. We might need bootstrap (or Jackknife) to find $se(\hat{\theta}_i^*)$ (unless we have $\theta = \mu \rightarrow se \rightarrow s/\sqrt{n}$, i.e. double bootstrap which increases the computational cost).
Note: B needs to be very large.

Bootstrap confidence intervals (cont'd)

C) Bootstrap CIs based on percentile points

$\hat{\boldsymbol{\theta}}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \dots \hat{\theta}_B^*)$

- We sort the values $\hat{\theta}_i^*$ in increasing order
- We find the $\alpha/2$, $1-\alpha/2$ percentile points of these values

The method is accurate only if the distribution of $\hat{\theta}$ is symmetric.

D) BCa CIs

Let $B = 2000$ and $\alpha = 0.1$ then C) gives us the following 90%-CI:

$$\left(\hat{\theta}^{*(0.05)}, \hat{\theta}^{*(0.95)}\right) \equiv (\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}})$$

$$\downarrow \qquad \downarrow$$

$$100^{\text{th}} \quad 1900^{\text{th}} \qquad \text{in order points}$$

As we said, these CIs work better when the distribution of $\hat{\theta}$ is symmetric.
Also what if $\hat{\theta}$ is a biased estimator of θ?
BCa $\rightarrow$ Bias Correction and acceleration
BCa: $(\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}}) = \left(\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}\right)$,

(BCa works again with percentiles of the bootstrap distribution)

Bootstrap confidence intervals (cont'd)

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(\alpha/2)})}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(1-\alpha/2)})}\right)$$

where $z^{(\beta)}$ is the β -percentile point of the standard normal distribution (e.g. $z^{(0.95)} = 1.645$) and $\Phi(\cdot)$ is the cdf of the standard normal (e.g. $\Phi(1.645) = 0.95$). In other words, it holds that $\Phi^{-1}(\alpha) = z^{(\alpha)}$.

$\hat{z}_0$ and $\hat{\alpha}$ are two quantities that correct for the bias and the deviation from the normal distribution respectively.

If $\hat{z}_0 = \hat{\alpha} = 0 \ \rightarrow$ case C) since $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - \alpha/2$.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\,\hat{\theta}_i^* < \hat{\theta}}{B}\right) \quad \text{bias correction}$$

$\downarrow$

proportion of $\hat{\theta}_i^* < \hat{\theta}$

(Note: if half of the $\hat{\theta}_i^*$ are $< \hat{\theta}$ and the rest half are $> \hat{\theta}$ then $\Phi^{-1}(0.5) = 0$, so we do not perform any bias correction.)

Bootstrap confidence intervals (cont'd)

and

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}\left(\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right]^{3/2}}$$

where $\hat{\theta}_{(i)}$ is the value of $\hat{\theta}$ after omitting observation i (Jackknife method) and $\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$.

$\hat{\alpha}$: It measures the asymmetry of the Jackknife pseudovalues.

Bootstrap vs Jackknife

- $B \in [50, 200]$ gives good results for se, even for non-smooth estimators like the median. Larger B values are required for CIs. Thus, for $n < 50$ Jackknife which performs n repetitions is faster.
- Jackknife uses limited information about $\hat{\theta}$ (by looking only at the n Jackknife samples) which makes it less effective.
- Jackknife is proven to be an approximation of bootstrap.

E.g. Let $\hat{\theta} = \alpha + \frac{1}{n}\sum_{i=1}^{n} a(X_i)$ be a linear statistic (for $\alpha = 0$ and $a(X_i) = X_i \;\rightarrow\; \hat{\theta} = \bar{X}$). For such statistics, the standard deviations of the two methods agree except for a factor of $\left(\frac{n-1}{n}\right)^{1/2}$ used by Jackknife.

Take for instance $\hat{\theta} = \bar{X}$. Then,

$$se_J(\hat{\theta}) = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n(n-1)}\right)^{1/2},$$

$$se_B(\hat{\theta}) = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n^2}\right)^{1/2} = \left(\frac{n-1}{n}\right)^{1/2} se_J(\hat{\theta})$$

(For large n the 2 se's are almost the same.)

Bootstrap vs Jackknife (cont'd)

Will derive analytically here the two standard errors.

Let $X \sim F(\text{cdf})$ $\begin{cases} \mathbb{E}[X] = \mu_F \\ \mathbb{V}[X] = \sigma_F^2 \end{cases}$

Let $X_1, X_2, \ldots X_n$ be a random sample from F, then CLT yields

$$\bar{X} \dot{\sim} \mathcal{N}(\mu_F, \sigma_F^2/n) \;\Rightarrow\; se(\bar{X}) = \sqrt{\sigma_F^2/n}$$

If we do not know F, we can use $\hat{F}$ (plug-in principle)

$$\mu_{\hat{F}} = \bar{X} \;\text{ and }\; \sigma_{\hat{F}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Thus, $se_{\hat{F}}(\bar{X}) = \frac{\sigma_{\hat{F}}}{\sqrt{n}} = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n^2}\right)^{1/2}$. As $B \to \infty$, $\hat{se}_B(\bar{X}) = se_{\hat{F}}(\bar{X})$.
(most of the times we do not use the plug-in estimator $\sigma_{\hat{F}}^2$ for σ_F^2 but $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ which is unbiased).

Bootstrap vs Jackknife (cont'd)

For the Jackknife we have for $\hat{\theta} = \bar{X}$:

$$
\begin{aligned}
se_J(\hat{\theta}) &= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_{(\cdot)}\right)^2} \\
&= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j - \sum_{k=1}^{n}\frac{\frac{1}{n-1}\sum_{j=1, j\neq k}^{n} X_j}{n}\right)^2} \\
&= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j - \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{j=1, j\neq k}^{n} X_j\right)^2}
\end{aligned}
$$

Bootstrap vs Jackknife (cont'd)

But

$$\sum_{k=1}^{n} \sum_{j=1, j \neq k}^{n} X_j = (n-1) \sum_{i=1}^{n} X_i$$

and therefore

$$\frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{j=1, j \neq k}^{n} X_j = \bar{X}$$

Thus:

$$\begin{aligned} se_J(\hat{\theta}) &= \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} X_j - \bar{X} \right)^2} \\ &= \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \bar{X} \right)^2} \end{aligned}$$

Bootstrap vs Jackknife (cont'd)

But,

$$
\begin{aligned}
\left(\hat{\theta}_{(i)} - \bar{X}\right) &= \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j - \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \sum_{j=1, j\neq i}^{n}\left(\frac{X_j}{n-1} - \frac{X_j}{n}\right) - \frac{X_i}{n} \\
&= \sum_{j=1, j\neq i}^{n}\frac{X_j}{n(n-1)} - \frac{X_i}{n} \\
&= \left[\sum_{j=1, j\neq i}^{n}\frac{X_j}{n(n-1)} + \frac{X_i}{n(n-1)}\right] - \left[\frac{X_i}{n(n-1)} + \frac{X_i}{n}\right] \\
&= \frac{\bar{X}}{n-1} - \frac{X_i}{n-1} = \frac{1}{n-1}(\bar{X} - X_i)
\end{aligned}
$$

Therefore

$$
se_J(\hat{\theta}) = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n(n-1)}\right)^{1/2}
$$

Bootstrap vs Jackknife (cont'd)

- Thus, for linear statistical functions there is no loss of information when using Jackknife. On the contrary, for non-linear statistical functions there is a difference. Jackknife is a linear approximation of bootstrap, i.e. it agrees with bootstrap (except for the factor $\left(\frac{n-1}{n}\right)^{1/2}$) for a specific linear statistical function approximating $\hat{\theta}$. Thus, the efficiency of the Jackknife estimator of se depends on how close to linear is $\hat{\theta}$.
- Similarly for the bias, Jackknife is an approximation of bootstrap with the approximation now based on a quadratic and not a linear function (e.g. variance).
- Finally, we could say that Jackknife is like bootstrap but the sampling is without replacement and the samples are of size $n-1$ instead of n.

Modifications for Bootstrap

- Smoothed Bootstrap

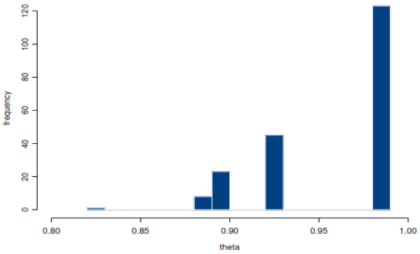
  Instead of $\hat{F}_n$ use kernel

- Iterated Bootstrap

  We take the values $\hat{\theta}_i^*$ and perform Bootstrap on them as well.

Example where Bootstrap fails

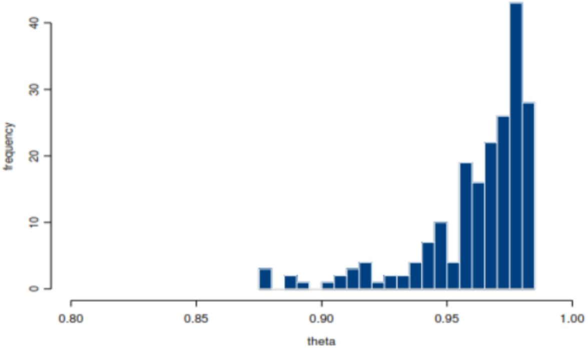
Let $X_1, \dots X_n \sim U(0, \theta)$. We know that the MLE $\hat{\theta} = \max(X_i) = X_{(n)}$.
Take $\theta = 1$ and $n = 50$.
Generating a sample from $U(0, 1)$ (runif(50,0,1)), we found that the largest value is 0.9832. Next we took 200 bootstrap samples and the values $\hat{\theta}_i^*$, $i = 1, \dots 200$ are shown below:



The graph we take is rather bad. The reason is that we attempt to approximate a continuous distribution (uniform) with a discrete (empirical). Given that $\hat{\theta} = \max(X_i)$ is a function using information from one observation alone, the largest, the result is not satisfactory.

Example where Bootstrap fails (cont'd)

The probability that the value 0.9832 is not included in a bootstrap sample is $(1 - 1/n)^n$, we thus expect it to be present in only $1 - (1 - 1/n)^n \to 1 - e^{-1} \approx 0.632$ of the samples. In fact, we had it in 124 out of the 200 samples. One solution in these cases is to use parametric bootstrap (Monte Carlo). That is, instead of generating data from the empirical distribution we generate from $U(0, \hat{\theta})$:



Bootstrap: Problems

Problems might appear in Bootstrap when

i) we try to estimate extreme proportion values (to get satisfactory results, B needs to be very large)

ii) n small $\Rightarrow \hat{F}_n$ is not a good approximation of F. Further, in that case we cannot take many Bootstrap samples because there is a chance to produce exactly the same samples.

iii) we try to estimate quantities that do not exist
e.g. $X \sim$ Cauchy $\rightarrow \mathbb{E}[X] \rightarrow \nexists$
Thus, if we try to estimate $se_{\bar{X}}$ this will fail.

Bootstrap in linear regression

Consider two r.v.'s Y and X and assume we would like to fit a linear regression model $Y = a + bX + \epsilon,\ \epsilon \sim N(0, \sigma^2)$ (what will say can be easily extended to more than one explanatory variables). Let (y_i, x_i), $i = 1, \ldots, n$, denote the original observed data.

1st idea: we draw n pairs, at random with replacement, from the original pairs $(y_1, x_1), \ldots, (y_n, x_n)$, B times in total. For each bootstrap sample j $(j = 1, \ldots, B)$: (y^*_{ji}, x^*_{ji}), $i = 1, \ldots, n$, we fit the linear regression model and we calculate the quantities of interest, e.g. $(\hat{a}^*_j, \hat{b}^*_j, R^2_j)$. Then we can compute their mean values, their standard error, we can use them to produce CI's, or to test hypotheses.

Problem: By using this method, we treat the values of the explanatory variable as random rather than fixed. We might want to treat them as fixed (e.g. data derived from an experimental design).

Bootstrap in linear regression (cont'd)

2nd idea:

- we fit the model using the original data (y_i, x_i), $i = 1, \ldots, n$, and find $\hat{a}, \hat{b}$.
- we compute the residuals $e_i = y_i - \hat{a} - \hat{b}x_i$ and the fitted values $\hat{y}_i = \hat{a} + \hat{b}x_i$, for $i = 1, \ldots, n$.
- we do bootstrap to the residuals, i.e. we draw n values, at random with replacement, from the residuals $e_1, \ldots, e_n$, B times in total. For each bootstrap sample j $(j = 1, \ldots, B)$: $(e^*_{j1}, \ldots e^*_{jn})$, we calculate the bootstrap response data $(y^*_{j1} = \hat{y}_1 + e^*_{j1}, \ldots, y^*_{jn} = \hat{y}_n + e^*_{jn})$.
- For each bootstrap sample j $(j = 1, \ldots, B)$: (y^*_{ji}, x_i), $i = 1, \ldots, n$, we fit the linear regression model and we calculate the quantities of interest, e.g. $(\hat{a}^*_j, \hat{b}^*_j, R^2_j)$. Then we can compute their mean values, their standard error, we can use them to produce CI's, or to test hypotheses.

Both approaches can be used to make inferences even if normality does not hold. The two methods give similar results. More statistically correct is the second (assumption of linear regression: the design matrix is already known).

Bootstrap in linear regression - Example

Ornithologist: 12 sparrows
- age in days : X
- length of wings in cm : Y

$$Y = 0.779\,(=\hat{a}) + 0.266\,(=\hat{b})X \quad \text{(least squares method)}$$

$$e_i = Y_i - 0.779 - 0.266X_i \quad \text{(normality not required)}$$

Table 1

Wings' length	Age	Residuals
1.40	3	-0.176623
1.50	3	-0.076623
2.20	5	0.091415
2.40	6	0.025435
3.10	8	0.193473
3.20	9	0.027492
3.20	10	-0.238489
3.90	11	0.195530
4.10	12	0.129549
4.70	14	0.197588
4.50	15	-0.268393
5.20	17	-0.100355

Bootstrap in linear regression - Example (cont'd)

Table 2

	Mean	SD	95%CI		95%CI	
$\hat{s}^2$	0.026	0.00776	0.011	0.041	0.012	0.042
$\hat{a}$	0.777	0.10961	0.563	0.991	0.554	0.986
$\hat{\beta}$	0.266	0.01052	0.245	0.286	0.246	0.286
$\hat{F}$	717.536	288.291	406.493	1566.891	417.432	1499.78
$\hat{R}^2$	0.984	0.0044	0.975	0.993	0.972	0.991

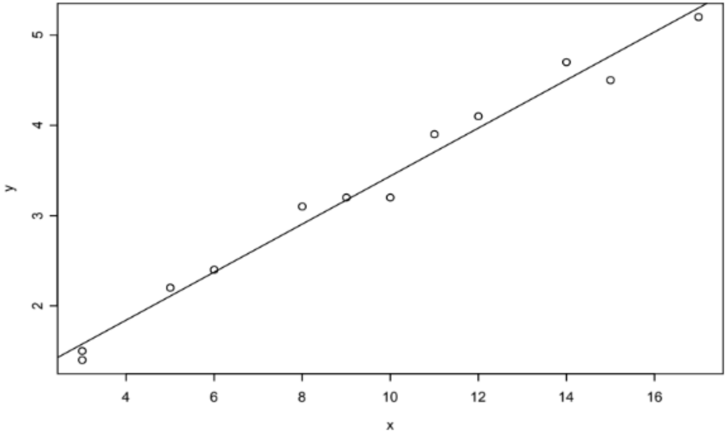
(columns 3-4: CI with classical method, columns 5-6: CI based on percentiles)

Bootstrap in linear regression - Example (cont'd)

1) From the scatterplot on the original data (see next slide), there seems to be a linear relationship between the two variables.
2) For the results to be correct, the residuals need to be normal $\rightarrow$ Diagnostic checking
   Further homoscedasticity, ... For Bootstrap no assumption.
3) With classical statistics, even if the assumptions hold true we can estimate standard errors, and thus generate CIs only for a, b and not for quantities like $F, R^2, \sigma^2, Corr(\hat{a}, \hat{b})$.
4) Bootstrap CIs $\nearrow$ classical method, $\searrow$ percentile points
   For a, b, σ^2 and R^2 they are almost identical because of normality as shown in the diagrams (slide #47). For F this is not true though.
5) CI for a does not contain 0 $\rightarrow$ statistically significant
   CI for b does not contain 0 $\rightarrow$ statistically significant

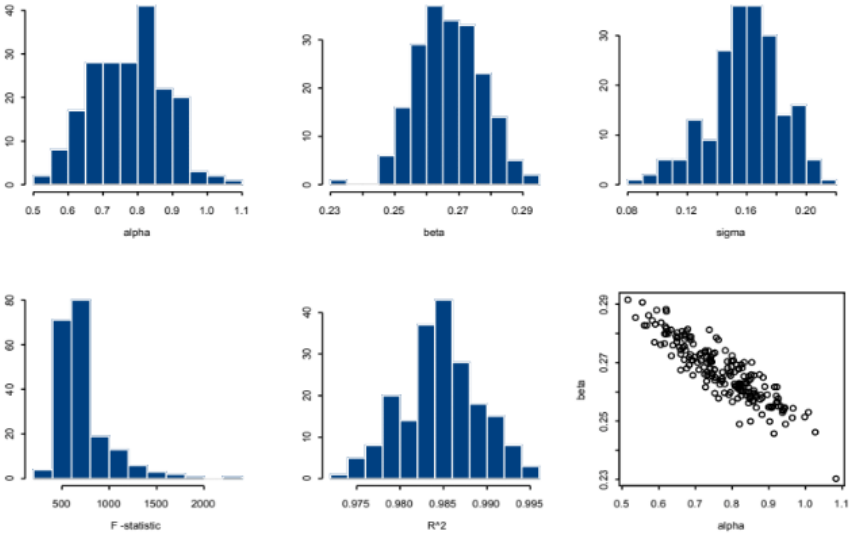
Bootstrap in linear regression - Example (cont'd)

Scatterplot between x and y:



Bootstrap in linear regression - Example (cont'd)

Histograms of bootstrap values of the different quantities:



Bootstrap in linear regression - Example (cont'd)

- We have used $B = 1000$ (large value in order to get "accurate" CIs)
- The covariance between $\hat{a}$ and $\hat{b}$ can be estimated from the covariance of their bootstrap values and is found to be -0.001 while $Corr(\hat{a}, \hat{b}) = -0.902$.
- From Table 2, we see that the 2 CI's for all quantities except for F are almost identical and this is due to normality. Further, the bootstrap averages are very close to $\hat{a}, \hat{b}$ from the linear regression with the original data, which means small bias.
- Scatterplot for $\hat{a}, \hat{b}$ bootstrap values reveals a high dependence between the estimators.
- CI for b does not contain 0 and thus there is a linear relationship between X and Y.

Parametric Bootstrap in regression

Let the errors follow any distribution with mean 0, e.g. t_ν, ν unknown. We cannot apply linear regression but we can apply parametric Bootstrap (assume that the homoscedasticity assumption holds). Then:

- Fit the model using least squares method (this step is independent of the errors' distribution, it only asks for homoscedasticity)
- Estimate ν from e_i and denote it by $\hat{\nu}$ (using for example a qqplot)
- Simulate $e_1^*, \ldots e_n^*$ from $t_{\hat{\nu}}$
- Compute $Y_i^* = \hat{a} + \hat{b}x_i + e_i^*$
- Fit the linear model on the (Y_i^*, x_i) and compute the quantities of interest
- Repeat the last three steps B times

Bootstrap for dependent data

Basic principle for Bootstrap is that $X_1, \dots X_n$ are independent. Additionally, $\hat{F}$ is a good estimator of F. None of these is true when X_i are dependent (e.g. time-series).

Block Bootstrap

To keep in our data some of the dependency they have, instead of performing sampling with replacement from the observations we perform it on blocks of observations. If the blocks are well chosen, they keep a lot of the information which is of interest to us.
Of the n observations we generate b blocks of length l. If $bl \neq n$ then one of the blocks might have fewer than l, observations. In this way, we keep the information for dependency up to order l. We lose information though at the points where the blocks are split.

Bootstrap for dependent data (cont'd)

Example: $x_1, \dots x_{12}$ observed dependent data
$y_1 = (x_1, x_2, x_3), \;\; y_2 = (x_4, x_5, x_6), \;\; y_3 = (x_7, x_8, x_9), \;\; y_4 = (x_{10}, x_{11}, x_{12})$
i.e. $b = 4, l = 3$
Next, we draw at random with replacement from the $y_i, \; i = 1, \dots 4$.
Since $l = 3$ we are not able to represent autocorrelation of order larger than 2 in our data.

Moving Blocks

We create overlapping blocks, i.e. different blocks contain the same observation. In the previous example,
$y_1 = (x_1, x_2, x_3), \;\; y_2 = (x_2, x_3, x_4), \;\; y_3 = (x_3, x_4, x_5), \;\; \dots$
$\dots, \;\; y_{11} = (x_{11}, x_{12}, x_1), \;\; y_{12} = (x_{12}, x_1, x_2)$
and perform sampling with replacement from the $y_i, \; i = 1, \dots 12$.
In this way we keep more information.

Subsampling

Let us denote the observed data by $x_1, \ldots x_n$

- Jackknife removes one observation. This idea, as already mentioned, can be generalized by removing d observations (delete-d-Jackknife). In that case, we need to generate all possible samples of size $n - d$ which are $\binom{n}{d}$. This number might be huge though. Thus, it suffices to take random samples out of the $\binom{n}{d}$.
- $\rightarrow$ Subsampling: there is a difference between this approach and Bootstrap. We take the samples from F and not $\hat{F}$ which is an advantage. However, the samples are of size $n - d$ instead of n leading to loss of information. Also, in subsampling the samples are taken without replacement.