Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 13
Introduction to bootstrap
5 May 2017

Plan of today's lecture

1. Example: Korsbetningen
2. The frequentist approach to inference (Ch. 6)
   - Statistics and sufficiency—small overview
   - Designing estimators
   - Uncertainty of estimators
3. Introduction to bootstrap (Ch. 7)
   - Empirical distribution functions
   - The bootstrap in a nutshell

Outline

Example: Korsbetningen

I Herrens år 1361, tredje dagen efter S:t Jacob, föll utanför Visbys portar gutarna i danskarnas händer. Här är de begravda. Bed för dem.

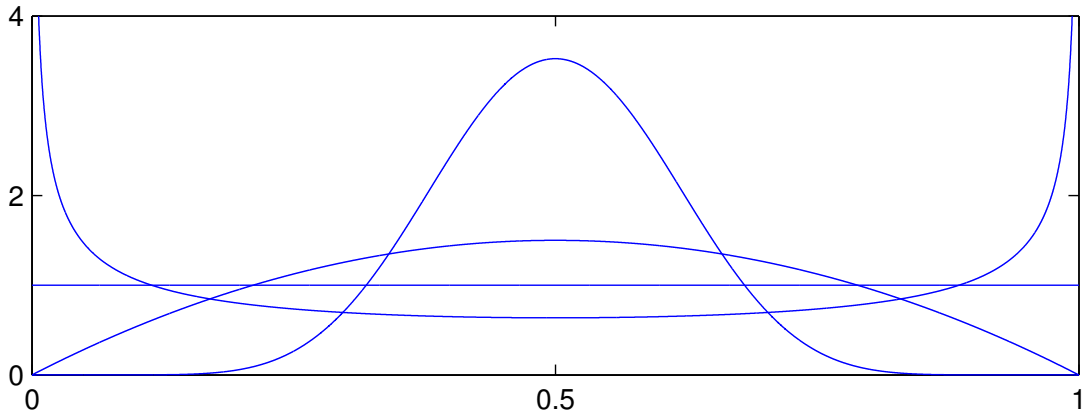
On the third day after saint Jacob, in the year of our lord 1361, the Goths fell outside the gates of Visby at the hands of the Danish. They are buried here. Pray for them.

Example: Korsbetningen—background

- In 1361 the Danish king Valdemar Atterdag conquered Gotland and captured the rich Hanseatic town of Visby.
- The conquest was followed by a plunder of Visby (brandskattning).
- Most of the defenders were killed in the attack and are buried in a field, Korsbetningen, outside of the walls of Visby.
- In 1929–1930 the gravesite and was excavated. A total of 493 femurs, 237 right and 256 left, were found.
- We want to estimate the number of buried bodies.

Example: Korsbetningen—model

- Model the numbers y_1 and y_2 of left and right legs, respectively, as independent observations from a Bin(n, p) distribution.
- Here n is the total number of people buried and p is the probability of finding a leg, left or right, of a person.
- We put a conjugate Beta(a, b)-prior on p and a U$(256, 2500)$ prior on n.



Example: Korsbetningen—a hybrid MCMC

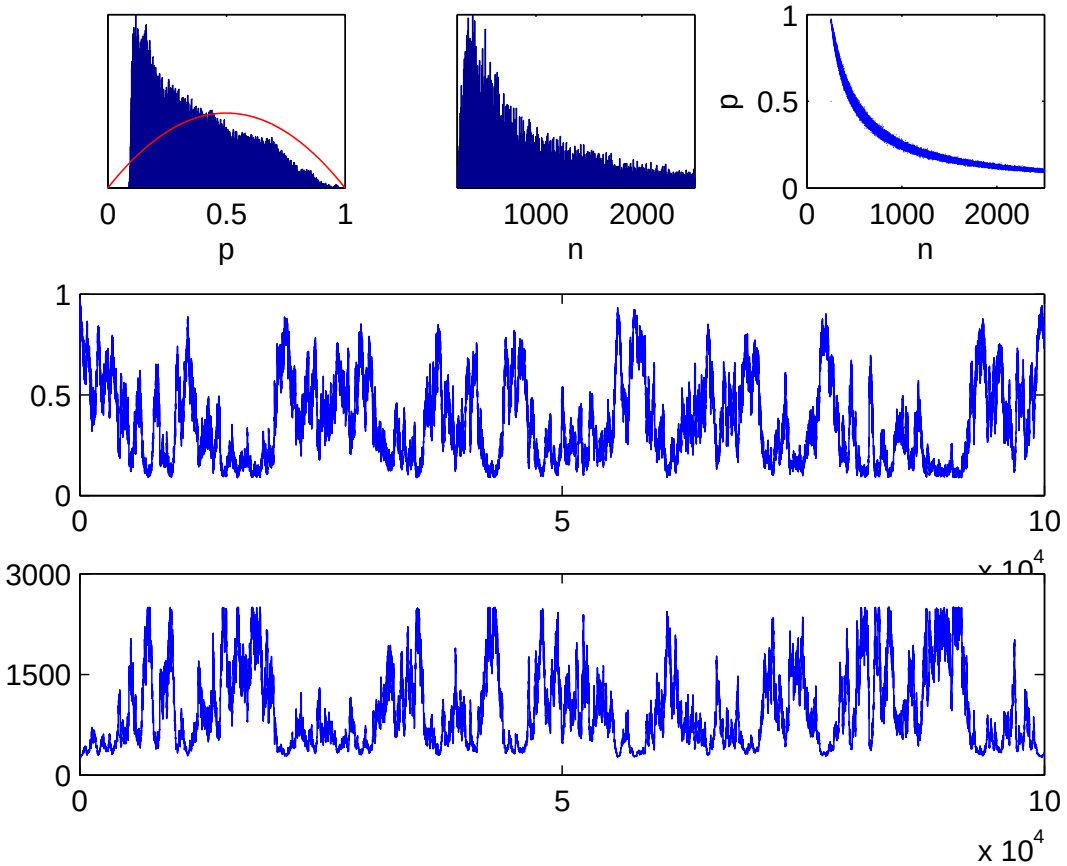
- We use a standard Gibbs step for

$$p \mid n, y_1, y_2 \sim \text{Beta}(a + y_1 + y_2, b + 2n - (y_1 + y_2)).$$

- MH for n, with a symmetric proposal obtained by drawing, given n, a new candidate n^* among the integers $\{n - R, \ldots, n, \ldots, n + R\}$.
- The acceptance probability becomes

$$\alpha(n, n^*) = 1 \wedge \frac{(1-p)^{2n^*}(n^*!)^2(n - y_1)!(n - y_2)!}{(1-p)^{2n}(n!)^2(n^* - y_1)!(n^* - y_2)!}.$$

Example: Korsbetningen—a hybrid MCMC



Example: Korsbetningen—improved sampler

- However, the previous algorithm mixes slowly.
- Thus, use instead the following scheme:
  1. First, draw a new n^* from the symmetric proposal as previously.
  2. Then draw, conditionally on n^*, also a candidate p^* from $f(p \mid n = n^*, y_1, y_2)$.
  3. Finally, accept or reject both n^* and p^*.
- This is a standard MH sampler!

Example: Korsbetningen—improved sampler (cont'd)

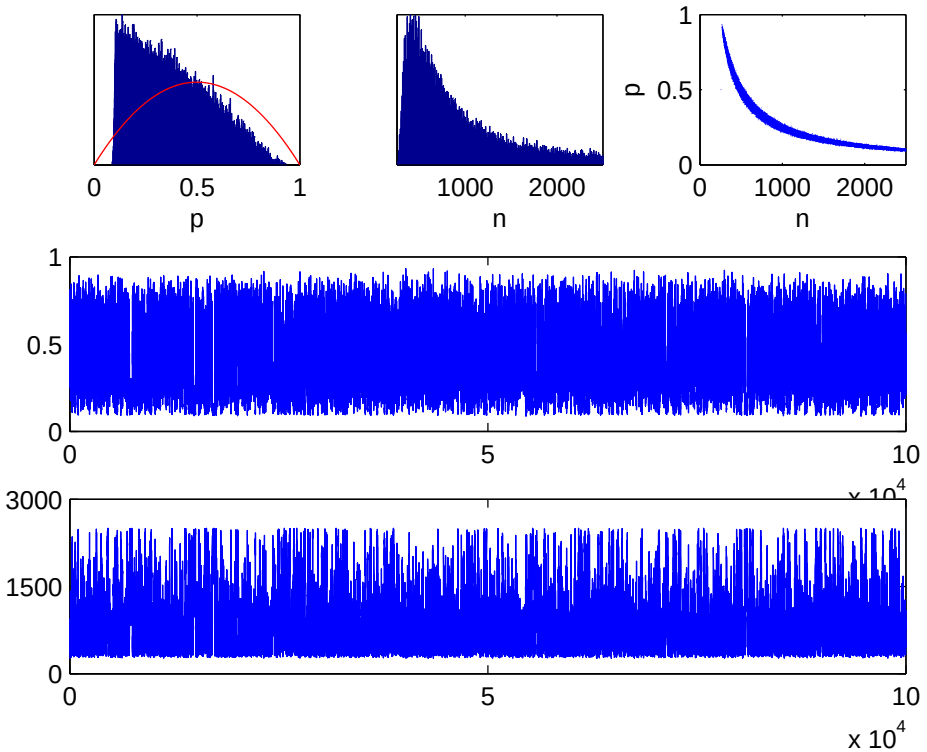
- For the new sampler, the proposal kernel becomes

$$q(n^*, p^* \mid n, p) \propto \frac{(2n^* + a + b - 1)!}{(a + y_1 + y_2 - 1)!(2n^* + b - y_1 - y_2 - 1)!} \times (p^*)^{a+y_1+y_2-1}(1-p^*)^{b+2n^*-(y_1+y_2)-1} \mathbb{1}_{|n-n^*| \leq R}.$$

- This gives the acceptance probability

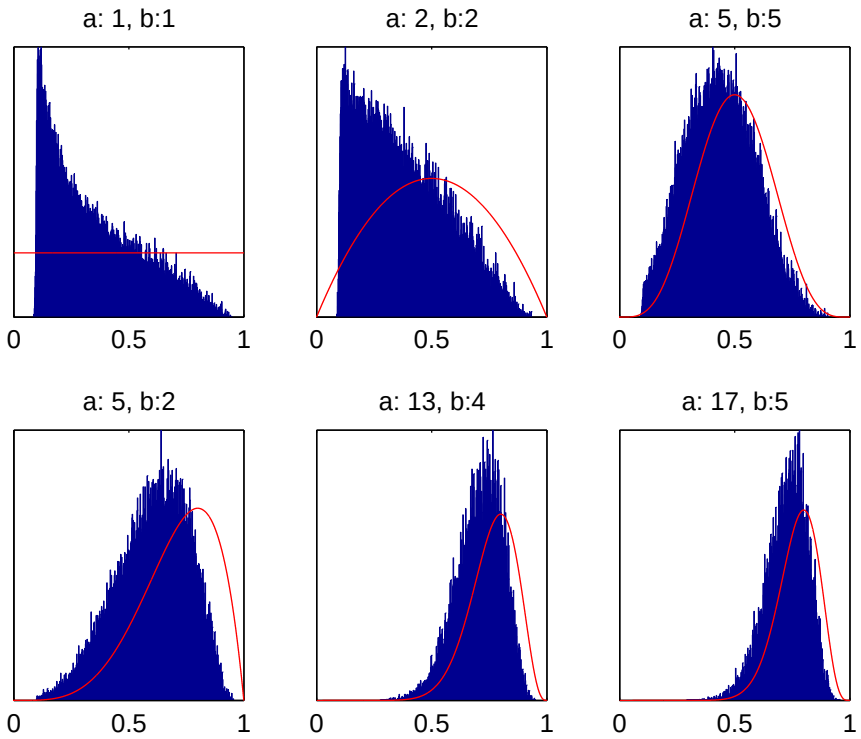
$$\begin{aligned}\alpha((n,p),(n^*,p^*)) &= 1 \wedge \frac{f(n^*,p^*)q(n,p \mid n^*,p^*)}{f(n,p)q(n^*,p^* \mid n,p)} \\ &= 1 \wedge \left(\frac{(n^*!)^2(n-y_1)!(n-y_2)!}{(n!)^2(n^*-y_1)!(n^*-y_2)!} \right. \\ &\quad \left. \times \frac{(2n+a+b-1)!(2n^*+b-y_1-y_2-1)!}{(2n^*+a+b-1)!(2n+b-y_1-y_2-1)!} \right).\end{aligned}$$

Korsbetningen—an improved MCMC sampler

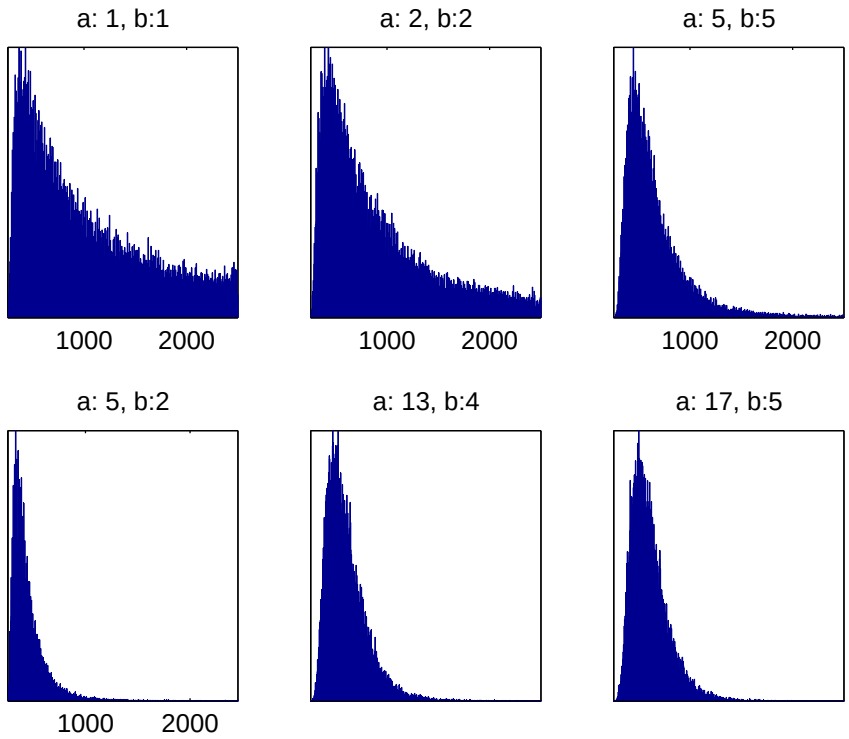


A one side 95% credible interval for n is $[344, \infty)$.

Korsbetningen—effect of the prior



Korsbetningen—effect of the prior



The lower side of a one sided 95% credible interval for n is $\{341, 345, 360, 297, 290, 289\}$.

Outline

The frequentist approach (again)

- Data y is viewed as an observation of a random variable Y with distribution $\mathbb{P}_0$, which most often is assumed to be a member of a parametric family

$$\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}.$$

  Thus, $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$ for some $\theta_0 \in \Theta$.
- Estimates $\widehat{\theta}(y)$ are realizations of random variables.
- A 95% confidence interval is calculated to cover the true value in 95% of the cases.
- Hypothesis testing is made by rejecting a hypothesis $\mathcal{H}_0$ if $\mathbb{P}(\text{data}||\mathcal{H}_0)$ is small.

The frequentist approach (again) (cont.)

- Let us extend the previous framework somewhat: given
  - observations y
  - and a model $\mathcal{P}$ for the data,

  we want to make inference about some property (estimand) $\tau = \tau(\mathbb{P}_0)$ of the distribution $\mathbb{P}_0$ that generated the data.
- For instance, denoting by f_0 the density of $\mathbb{P}_0$,

$$\tau(\mathbb{P}_0) = \int x f_0(x)\, dx \quad \text{(expectation value)}.$$

- The inference problem can split into two subproblems:
  1. How do we construct a data-based estimator of τ?
  2. How do we assess the uncertainty of the estimate?

Outline

Statistics

- A statistic t is simply a (possibly vector-valued) function of data.
- Some examples:
  1. The arithmetic mean: $t(y) = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
  2. The s^2-statistics: $t(y) = s^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.
  3. The ordered sample (order statistics):

$$t(y) = \{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\}.$$

  4. The maximum likelihood estimator (MLE):

$$t(y) = \text{argmax}_{\theta \in \Theta} f_\theta(y).$$

Sufficient statistics

- A statistic that summarizes completely the information contained in the data concerning the unknown parameter θ is called a sufficient statistic for θ.
- Mathematically, t is sufficient if the conditional distribution of Y given $t(Y)$ does not depend on the parameter θ.
- This means that given $t(Y)$ we may, by simulation, generate a sample Y' with exact the same distribution as Y without knowing the value of the unknown parameter θ_0.
- The factorization criterion (FC) says that $t(y)$ is sufficient if and only if the density of Y can be factorized as

$$f_\theta(y) = h(y)g_\theta(t(y)).$$

Example: a simple sufficient statistic

- For a simple example, let $y = (y_1, \ldots, y_n)$ be observations of n independent variables with $\mathrm{N}(\theta, 1)$-distribution. Then

$$f_\theta(y \mid \theta) = \prod_{i=1}^{n} f_\theta(y_i \mid \theta) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \theta)^2}{2} \right) \right\}$$
$$= \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} y_i^2 \right) \exp\left(\theta n \bar{y} - \frac{1}{2} n \theta^2 \right).$$

- Now, sufficiency of $t(y) = \bar{y}$ follows from the FC with

$$\begin{cases} t(y) \leftarrow \bar{y}, \\ g_\theta(t(y)) \leftarrow \exp\left(\theta n \bar{y} + \frac{1}{2} n \theta^2 \right), \\ h(y) \leftarrow \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} y_i^2 \right). \end{cases}$$

Completeness

- A data dependent statistics V is called ancillary if its distribution does not depend on θ and first order ancillary if $\mathbb{E}_\theta(V) = c$ for all θ (note that the latter is weaker than the former).
- Since a good sufficient statistics $T = t(Y)$ provides lots of information concerning θ it should not—if T is good enough—be possible to form even a first order ancillary statistics based on T, i.e.

$$\mathbb{E}_\theta(V(T)) = c \; \forall\theta \Rightarrow V(t) \equiv c \text{ (a.s)}.$$

- Subtracting c leads to the following definition: a sufficient statistics T is called complete if

$$\mathbb{E}_\theta(f(T)) = 0 \; \forall\theta \Rightarrow f(t) \equiv 0 \text{ (a.s)}.$$

Completeness (cont.)

- The following is a famous result in statistics:

Theorem (Lehmann-Scheffé)

Let T be an unbiased complete sufficient statistics for θ, i.e. $\mathbb{E}_\theta(T) = \theta$. Then T is the (uniquely) best unbiased estimator of θ in terms of variance.

- In example above, where $y = (y_1, \ldots, y_n)$ were observations of n independent variables with $\mathrm{N}(\theta, 1)$-distribution, one may show that the sufficient statistics $t(y) = \bar{y}$ is complete. Thus, t is the uniquely best unbiased estimator of θ!

Outline

Maximum likelihood estimators

- Our first task is to find a statistic that is a good estimate of the estimand $\tau = \tau(\mathbb{P}_0)$ of interest.
- Two common choices are
  - the MLE and
  - the least square estimator.
- As mentioned, the MLE is defined as the parameter value maximizing the likelihood function

$$\theta \mapsto f_\theta(y)$$

or, equivalently, the log-likelihood function

$$\theta \mapsto \log f_\theta(y).$$

Least square estimators

- When applying least squares we first find the expectation as a function of the unknown parameter:

$$\mu(\theta) = \int x f_\theta(x)\, dx.$$

- After this, we minimize the squared deviation

$$t(y) = \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} (\mu(\theta) - y_i)^2$$

between our observations and the expected value.

Outline

Uncertainty of estimators

- It is important to always keep in mind that the estimate $t(y)$ is an observation of a random variable $t(Y)$. If the experiment was repeated, resulting in a new vector y of random observations, the estimator would take another value.
- In the same way, the error $\Delta(y) = t(y) - \tau$ is a realization of the random variable $\Delta(Y) = t(Y) - \tau$.
- To assess the uncertainty of the estimator we thus need to analyze the distribution function F_Δ of the error $\Delta(Y)$ (error distribution) under $\mathbb{P}_0$.

Confidence intervals and bias

- Assume that we have found the error distribution F_Δ. For a confidence interval $(L(y), U(y))$ for τ on level α,

$$\begin{aligned} 1-\alpha &= \mathbb{P}_0\left(L(Y) \leq \tau \leq U(Y)\right) \\ &= \mathbb{P}_0\left(t(Y) - L(Y) \geq t(Y) - \tau \geq t(Y) - U(Y)\right) \\ &= \mathbb{P}_0\left(t(Y) - L(Y) \geq \Delta(Y) \geq t(Y) - U(Y)\right). \end{aligned}$$

- Thus,

$$\begin{cases} t(Y) - L(Y) = F_\Delta^{-1}(1-\alpha/2) \\ t(Y) - U(Y) = F_\Delta^{-1}(\alpha/2) \end{cases} \Leftrightarrow \begin{cases} L(Y) = t(Y) - F_\Delta^{-1}(1-\alpha/2) \\ U(Y) = t(Y) - F_\Delta^{-1}(\alpha/2) \end{cases}$$

and the confidence interval becomes

$$\mathrm{I}_\alpha = \left(t(y) - F_\Delta^{-1}(1-\alpha/2), t(y) - F_\Delta^{-1}(\alpha/2)\right).$$

Confidence intervals and bias

- The bias of the estimator is

$$B_t = \mathbb{E}_0\left(t(Y) - \tau\right) = \mathbb{E}_0\left(\Delta(Y)\right) = \int z f_\Delta(z)\, dz,$$

  where $f_\Delta(z) = \frac{d}{dz} F_\Delta(z)$ denotes the density of $\Delta(Y)$.
- Consequently, finding the error distribution F_Δ is essential for making qualitative statements about the estimator.
- In the previous normal distribution example,

$$\Delta(Y) = \bar{Y} - \theta_0 \sim \mathrm{N}(0, 1/n),$$

  yielding $\mathbb{E}_0(\Delta(Y)) = 0$ and

$$\begin{cases} F_\Delta^{-1}(1-\alpha/2) = \lambda_{\alpha/2}\frac{1}{\sqrt{n}} \\ F_\Delta^{-1}(\alpha/2) = -\lambda_{\alpha/2}\frac{1}{\sqrt{n}} \end{cases} \Rightarrow \mathrm{I}_\alpha = \left(\bar{y} \pm \lambda_{\alpha/2}\frac{1}{\sqrt{n}}\right).$$

Outline

Overview

- So, we need $F_\Delta(z)$ (or $f_\Delta(z)$) to evaluate the uncertainty of t. However, here we generally face two obstacles:
  1. We do not know $F_\Delta(z)$ (or $f_\Delta(z)$); these distributions may for instance depend on the quantity τ that we want to estimate. Sometimes we do not even have at hand a well-specified model, just data!
  2. Even if we knew $F_\Delta(z)$, finding the quantiles $F_\Delta^{-1}(p)$ is typically complicated.
- The bootstrap algorithm deals with these problems by
  1. replacing $\mathbb{P}_0$ by an data-based approximation and
  2. analyzing the variation of $\Delta(Y)$ using MC simulation from the approximation of $\mathbb{P}_0$, respectively.

Outline

The empirical distribution function (EDF)

- The empirical distribution (ED) $\widehat{\mathbb{P}}_0$ associated with the data $y = (y_1, y_2, \ldots, y_n)$ gives equal weight $(1/n)$ to each of the y_is (assuming that all values of y are distinct).
- Consequently, if $Z \sim \widehat{\mathbb{P}}_0$ is a random variable, then Z takes the value y_i with probability $1/n$.
- The empirical distribution function (EDF) associated with the data y is defined by

$$\widehat{F}_n(z) = \widehat{\mathbb{P}}_0(Z \leq z)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \leq z\}} = \text{fraction of } y_i\text{s that are less than } z.$$

Properties of the EDF

- It holds that

$$\lim_{z \to -\infty} \widehat{F}_n(z) = \lim_{z \to -\infty} F(z) = 0,$$
$$\lim_{z \to \infty} \widehat{F}_n(z) = \lim_{z \to \infty} F(z) = 1.$$

- In addition, $n\widehat{F}_n(z) \sim \text{Bin}(n, F(z))$.
- This implies the LLN (as $n \to \infty$)

$$\widehat{F}_n(z) \to F(z) \quad \text{(a.s.)}$$

- as well as the CLT

$$\sqrt{n}(\widehat{F}_n(z) - F(z)) \xrightarrow{\text{d.}} \text{N}(0, \sigma^2(z)),$$

where

$$\sigma^2(z) = F(z)(1 - F(z)).$$

Outline

The bootstrap

- Having access to data y, we may now replace $\mathbb{P}_0$ by $\widehat{\mathbb{P}}_0$.
- Any quantity involving $\mathbb{P}_0$ can now be approximated by plugging $\widehat{\mathbb{P}}_0$ into the quantity instead. In particular,

$$\tau = \tau(\mathbb{P}_0) \approx \widehat{\tau} = \tau(\widehat{\mathbb{P}}_0).$$

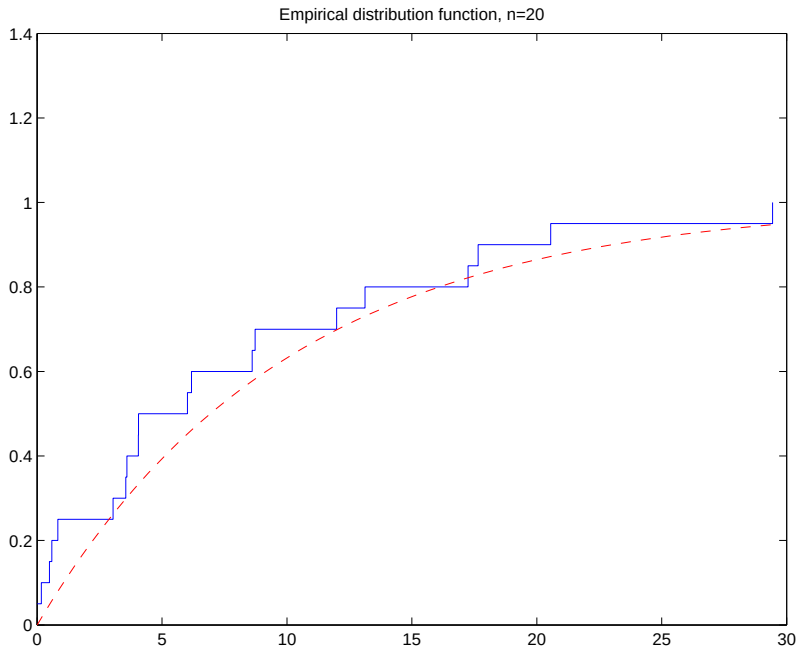
- Moreover, the uncertainty of $t(y)$ can be analyzed by drawing repeatedly $Y^* \sim \widehat{\mathbb{P}}_0$ and look at the variation (histogram) of $\Delta(Y^*) = t(Y^*) - \tau \approx \Delta(Y^*) = t(Y^*) - \widehat{\tau}$.
- Recall that the ED gives equal weight $1/n$ to all the y_is in y. Thus, simulation from $\widehat{\mathbb{P}}_0$ is carried through by simply drawing, with replacement, among the values $y_1, \ldots, y_n$.

The bootstrap: algorithm

- Construct the ED $\widehat{\mathbb{P}}_0$ from the data y.
- Simulate B new data sets Y_b^*, $b \in \{1, 2, \ldots, B\}$, where each Y_b^* has the size of y, from $\widehat{\mathbb{P}}_0$. Each Y_b^* is obtained by drawing, with replacement, n times among the y_is.
- Compute the values $t(Y_b^*)$, $b \in \{1, 2, \ldots, B\}$, of the estimator.
- By setting in turn $\Delta_b^* = t(Y_b^*) - \widehat{\tau}$, $b \in \{1, 2, \ldots, B\}$, we obtain values being approximately distributed according to the error distribution. These can be used for uncertainty analysis.

A toy example

- We let $y = (y_1, \ldots, y_{20})$ be i.i.d. with unknown mean θ (and unknown distribution). As estimator we take, as usual, $t(y) = \bar{y}$.



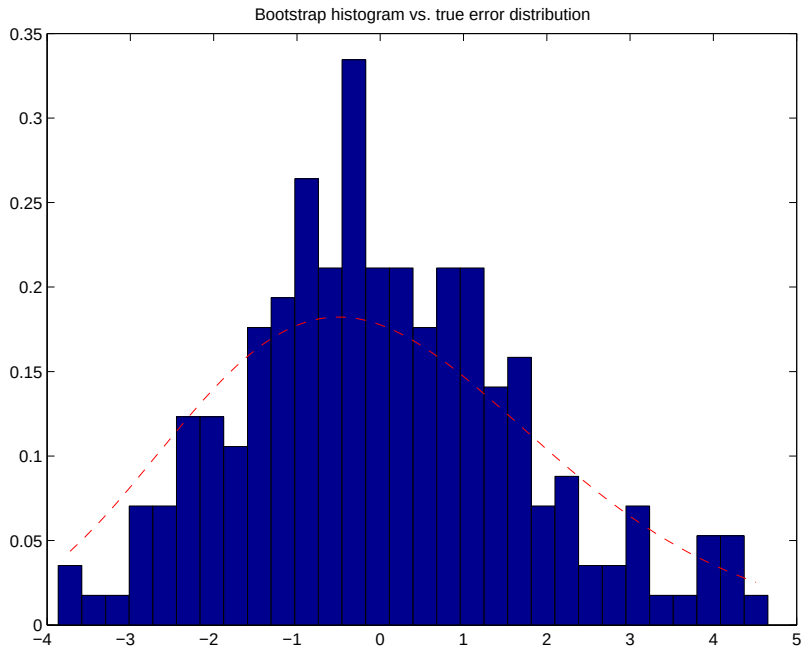
A toy example: MATLAB implementation

- In MATLAB:

```
n = 20;
B = 200;
tau_hat = mean(y);
boot = zeros(1,B);
for b = 1:B, % bootstrap
    I = randsample(n,n,'true',ones(1,n));
    boot(b) = mean(y(I));
end
delta = sort(boot - tau_hat);
alpha = 0.05; % CB level
L = tau_hat - delta(ceil((1 - alpha/2)*B));
U = tau_hat - delta(ceil(alpha*B/2));
```

A toy example: exponential distribution

- This histogram of $(\Delta_b^*)_{b=1}^{200}$ looks like follows:



- The associated confidence bound is (in fact, the y_is were observations of $Y_i \sim \text{Exp}(\theta)$, simulated under $\theta_0 = 10$)

$$\text{I}_{.05} = (7.3, 15.7).$$