# Stochastic Simulation
# The Bootstrap method


Bo Friis Nielsen

Institute of Mathematical Modelling

Technical University of Denmark

2800 Kgs. Lyngby – Denmark

Email: bfni@dtu.dk

# The Bootstrap method

- A technique for estimating the variance (etc) of an estimator.
- Based on sampling from the empirical distribution.
- Non-parametric technique

# Recall the simple situation

- We have $n$ observations $x_i$, $i = 1, \ldots, n$.
- If we want to estimate the mean value of the underlying distribution, we (typically) just use the estimator $\bar{x} = \sum x_i / n$.
- This estimator has the variance $\frac{1}{n}\mathrm{Var}(X)$. To estimate this, we (typically) just use the sample variance.

# A not-so-simple-situation

- Assume we want to estimate the median, rather than the mean.
- (This makes much sense w.r.t. robustness)
- The natural estimator for the median is the sample median.
- But what is the variance of the estimator?

# The variance of the sample median

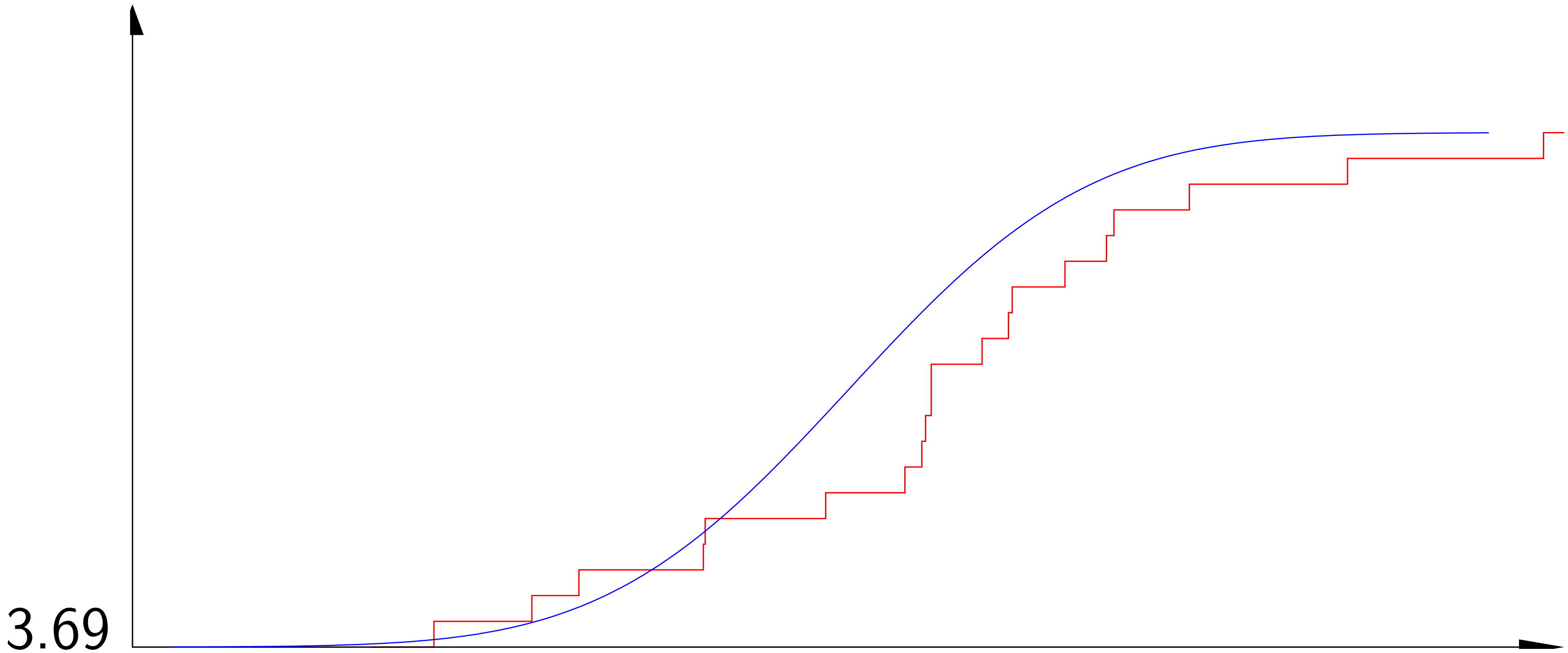
If we had access to the “true” underlying distribution, we could

1. Simulate a number of data sets like the one we had.
2. For each simulated data set, compute the median.
3. Finally report the variance among these medians.

We don’t have the true distribution. But we have the **empirical** distribution!

# Empirical distribution

20 $N(0,1)$ variates (sorted): -2.20, -1.68, -1.43, -0.77, -0.76, -0.12, 0.30, 0.39, 0.41, 0.44, 0.44, 0.71, 0.85, 0.87, 1.15, 1.37, 1.41, 1.81, 2.65,



# Empirical distribution

$X_i$ iid random variables with $F(x) = \mathsf{P}(X \leq x)$

Each leads to a (simple) random function $F_{e,i}(x) = \mathbf{1}_{\{\mathbf{X_i \leq x}\}}$

leading to $F_e(x) = \frac{1}{n}\sum_{i=1}^{n} F_{e,i}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{X_i \leq x}\}}$

$\mathsf{E}\left(F_e(x)\right) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{X_i \leq x}\}}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(\mathbf{1}_{\{\mathbf{X_i \leq x}\}}\right) = F(x)$

Once we have sample $x_i, i = 1, 2, \ldots, n$ we have a realised version of the empirical distribution function

$$F_e(x) = \frac{1}{n}\sum_{i=1}^{n} F_{e,i}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\{x_i \leq x\}}$$

where $\delta$ is Kroneckers delta-function

# The Bootstrap Algorithm for the variance of a parameter estimator

- Given a data set with $n$ observations.
- Simulate $r$
- (e.g., $r = 100$)
- data sets,
- each with $n$ "observations"
- sampled form the empirical distribution $F_e$.
- (To simulate such one data set, simply take $n$ samples from the original data set *with* replacement)
- For each simulated data set, estimate the parameter of interest (e.g., the median). This is a *bootstrap replicate* of the estimate.
- Finally report the variance among the bootstrap replicates.

# Advantages of the Bootstrap method

- Does not require the distribution in parametric form.
- Easily implemented.
- Applies also to estimators which cannot easily be analysed.
- Generalizes e.g. to confidence intervals.

# Exercise 8

1. Exercise 13 in Chapter 8 of Ross (P.152).
2. Exercise 15 in Chapter 8 of Ross (P.152).
3. Write a subroutine that takes as input a "data" vector of observed values, and which outputs the median as well as the bootstrap estimate of the variance of the median, based on $r = 100$ bootstrap replicates. Simulate $N = 200$ Pareto distributed random variates with $\beta = 1$ and $k = 1.05$.
   (a) Compute the mean and the median (of the sample)
   (b) Make the bootstrap estimate of the variance of the sample mean.
   (c) Make the bootstrap estimate of the variance of the sample median.
   (d) Compare the precision of the estimated median with the precision of the estimated mean.