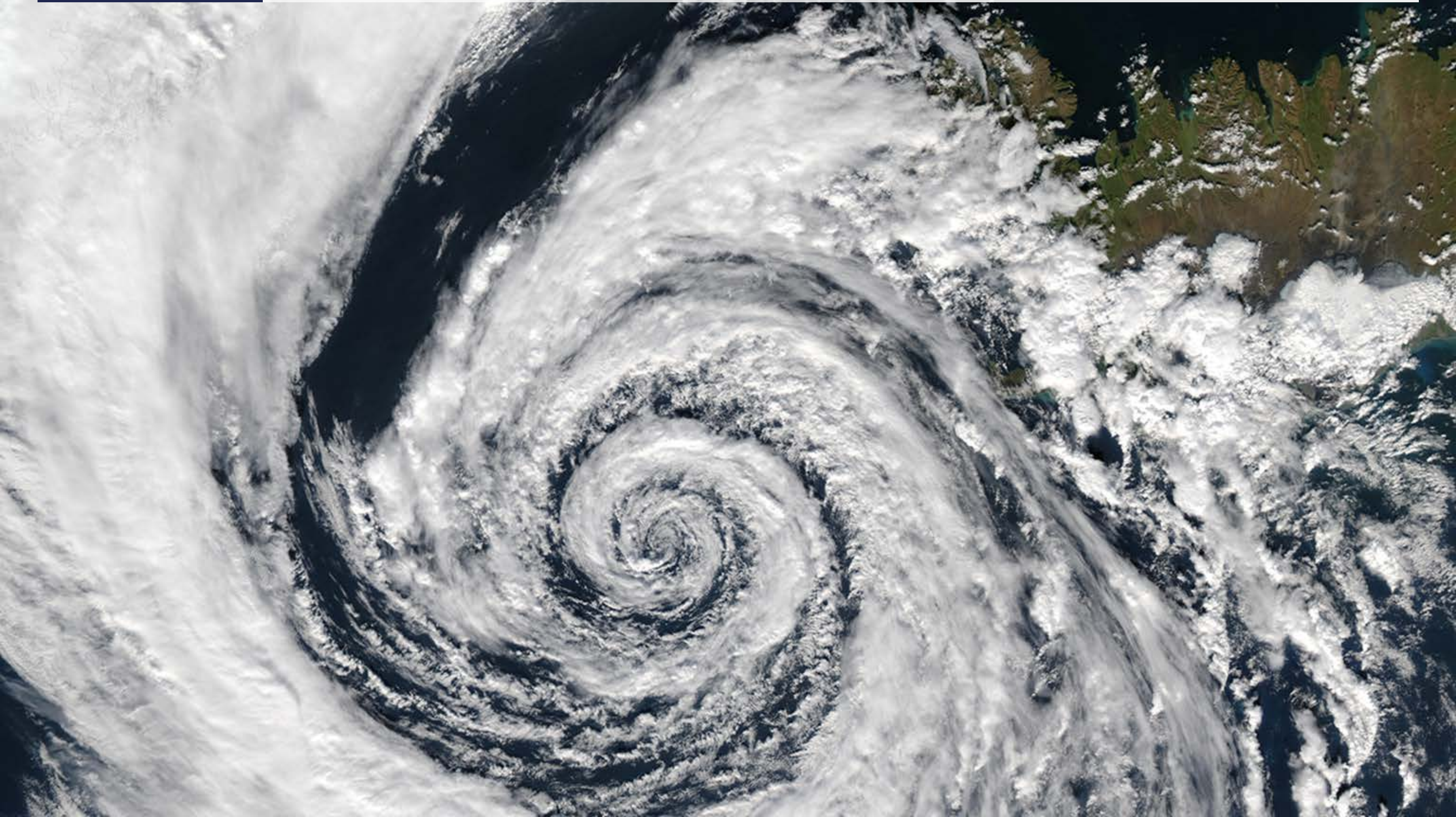
Statistics

Lecture 6.1: Statistical models

Name teacher

TU Delft



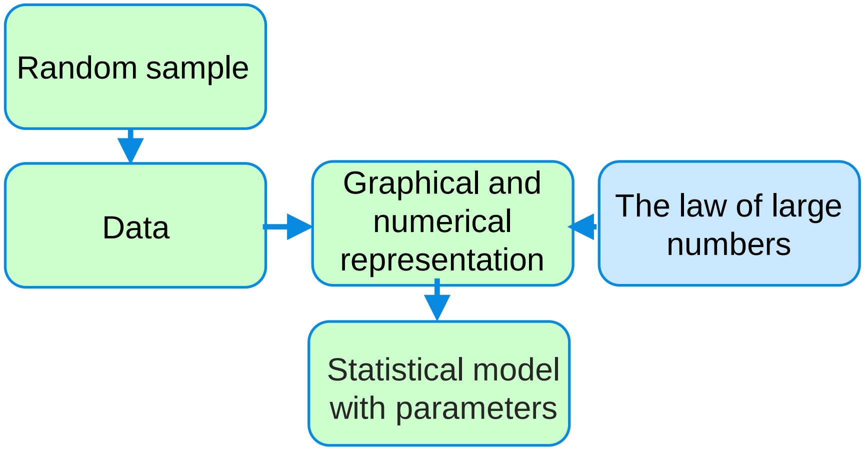
Learning objective

After this class you can:

- Represent data graphically
- Represent data numerically
- Draw a boxplot
- Interpret a boxplot
- Build a statistical model for repeated measurements
- Make the connection between sample statistics and distribution features

Book: Chapters 15, 16 and 17

Statistics



Random sample
Data
Graphical and numerical representation
The law of large numbers
Statistical model with parameters

Before this class (week 3.6 lesson 1):

Watch pre-lecture videos ‘Graphical representation of data’ and ‘Numerical representation of data’

Book: Section 15.1, 15.2, 15.3, 15.4, 16.1, 16.2, 16.3

Program

- Five-number summary and boxplot
- Exercise
- Random sample, statistical model, model distribution
- Exercise
- Context
- Distribution features and sample statistics
- Exercises

Consider the histogram you drew for the data 1, 2, 3, 3, 3, 5, 6, 9, 10. The height of the histogram above the interval [2, 4) is equal to

A) $\frac{3}{10}$

B) $\frac{4}{10}$

C) $\frac{4}{18}$

D) $\frac{3}{18}$

FeedbackFruits

Consider the empirical distribution function you drew for the data 1, 2, 3, 3, 3, 5, 6, 9, 10. The value of this function in the point 7 is

A) $F_n(7) = \frac{8}{10}$

B) $F_n(7) = \frac{8}{9}$

C) $F_n(7) = \frac{7}{9}$

D) $F_n(7) = 0$

From the pre-lecture video:
Give the mean, median, sd and MAD of the following data set:
90, 83, 99, 93, 104, 89, 88, 95, 82, 100.

A) $(92.3,\ 91.5,\ 21.8,\ 5.5)$

B) $(92.3,\ 91.5,\ 5.5,\ 21.8)$

C) $(91.5,\ 92.3,\ 7.27,\ 5.5)$

D) $(92.3,\ 91.5,\ 7.27,\ 5.5)$

The five-number summary

The five-number summary consists of the following five numbers:

1. Minimum
2. Lower quartile
3. Median
4. Upper quartile
5. Maximum

Example:

For the dataset 0,1,4,4,4,5,5,6,6,10,12:

Min=0, Lower quartile=4, Median=5, Upper quartile=6, Max=12

Quantiles and their computation

Definition:
Let $x_1, \ldots, x_n$ be a dataset. For any $p \in [0, 1]$ the *pth empirical value* is the number $q_n(p)$ such that a proportion p of the dataset is less than $q_n(p)$ and a proportion $1 - p$ is larger than $q_n(p)$.

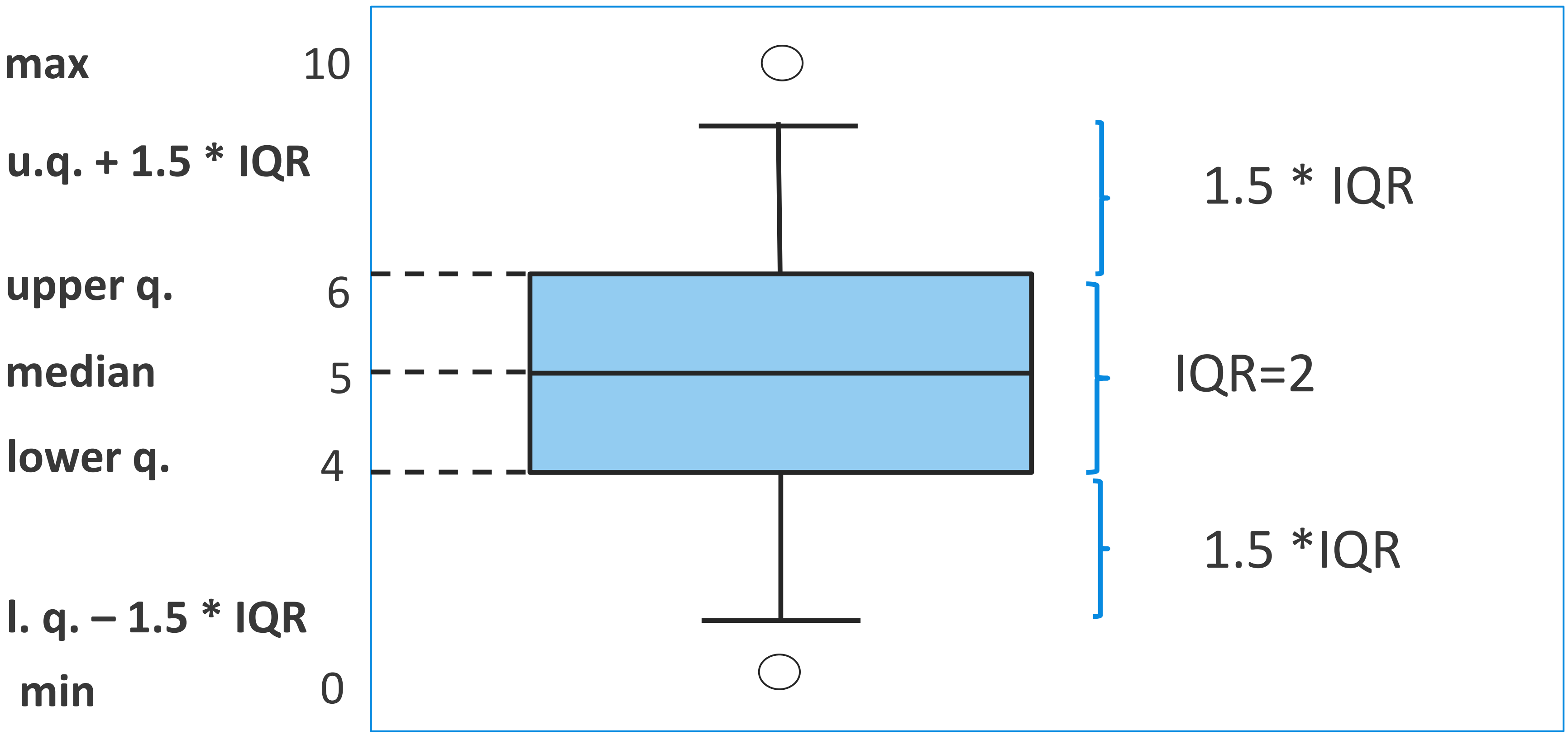
$q_n(0.5)$ is the median and $q_n(0.25)$ is the first quartile.

How to compute quantiles?
Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered dataset. Compute $p(n+1) = k + \alpha$, where $k = \lfloor p(n+1) \rfloor$ is the integer part and $\alpha = p(n+1) - k$ the decimal part of $p(n+1)$.
Then $$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)}).$$

The boxplot: visualising the five-number summary

For the dataset 0,1,4,4,4,5,5,6,6,10,12 :

Min=0, Lower quartile=4, Median=5, Upper quartile=6, Max=12



Exercise

Book: 16.1 and 16.3

Random sample and statistical model

Definition:
A *random sample* is a collection of RV's

$$X_1, X_2, \cdots, X_n$$

that have the same probability distribution and that are mutually independent.

A *dataset* consisting of repeated measurements $x_1, x_2, \ldots, x_n$ is modelled as the *realization* of a random sample $X_1, X_2, \ldots, X_n$.
n is called the *sample size*

Let X_1, X_2 be a random sample from a normal distribution with variance 4. The correlation coefficient of X_1, X_2 is

A) -1

B) 4

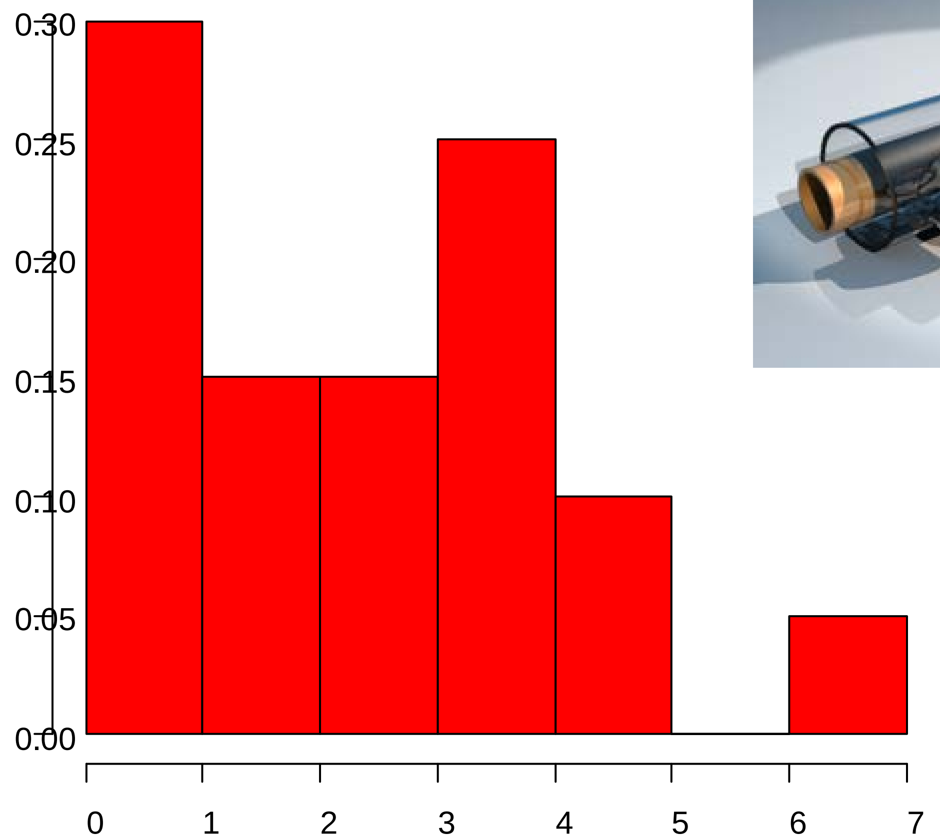
C) 0

D) 1

Cycles of heater switches

Number of cycles 20 heater switches make after an overload voltage (in ten thousands)

0.0100
0.0340
0.1940
0.5670
0.6010
0.7120
1.2910
1.3670
1.9490
2.3700
2.4110
2.8750
3.1620
3.2800
3.4910
3.6860
3.8540
4.2110
4.3970
6.4730



Kapur, K. C. and Lamberson, L. R. (1977), *Reliability in Engineering Design*

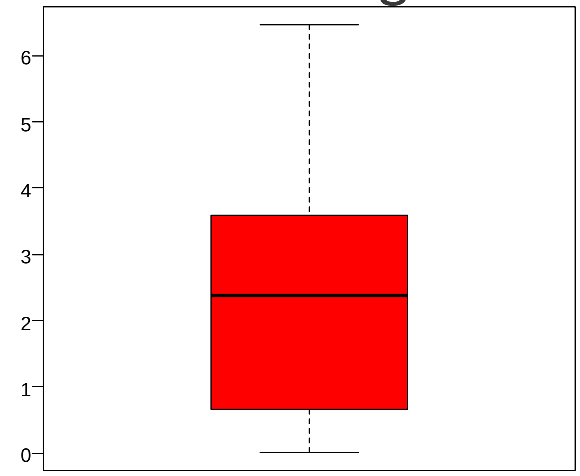
Cycles of heater switches

Number of cycles 20 heater switches make after an overload voltage (in ten thousands)

0.0100
0.0340
0.1940
0.5670
0.6010
0.7120
1.2910
1.3670
1.9490
2.3700
2.4110
2.8750
3.1620
3.2800
3.4910
3.6860
3.8540
4.2110
4.3970
6.4730

The five-number summary consists of the following five numbers:

1. Minimum: 0.01000
2. Lower quartile: 0.68425
3. Median: 2.39050
4. Upper quartile: 3.53975
5. Maximum: 6.47300



Can we find a mechanism that mimics the data generating process?

Build a stochastic model for the data generating process

Model distribution

Definition:
The probability distribution of each RV from a random sample is called the *model distribution.*

Definition:
The RV $h(X_1, \ldots, X_n)$, which depends only on the random sample $X_1, \ldots, X_n$ is called a *sample statistic*.

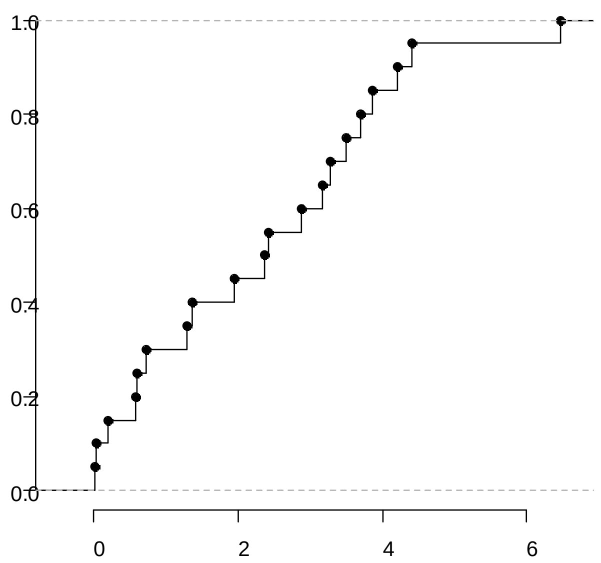
Statistics:

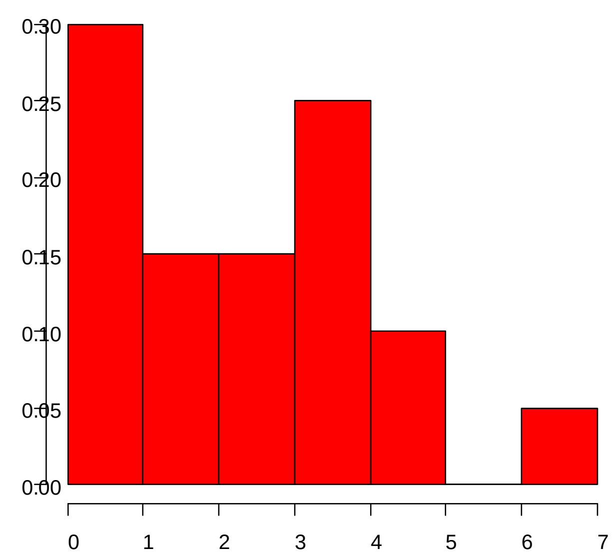
1. estimate features of the model distribution using sample statistics from the data set
2. which model distribution fits a particular dataset best?

Distribution features and sample statistics

1. Sample mean vs expectation
2. Empirical distribution function vs model distribution function
3. Histogram vs density of the model distribution
4. Sample median vs $F^{inv}(0.5)$

The expectation of the random variable is ***approximately*** 2.34675





$F^{inv}(0.5)$ is ***approximately*** equal to 2.39050

Estimating features of the "true" distribution

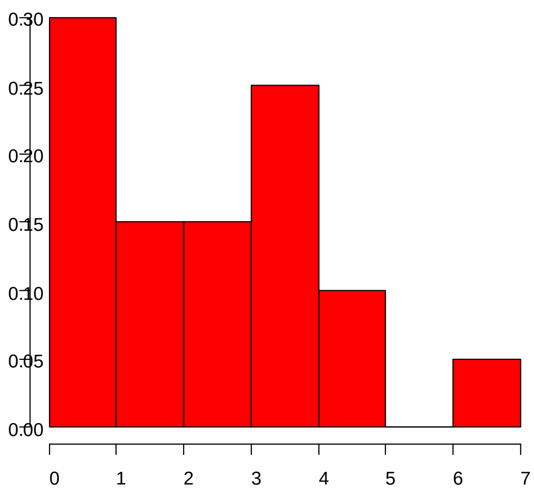
Sample statistic	Distribution feature
Graphical	
Empirical distribution function F_n	Distribution function F
Kernel density estimate $f_{n,h}$ and histogram	Probability density f
(Number of X_i equal to a)$/n$	Probability mass function $p(a)$
Numerical	
Sample mean $\bar{X}_n$	Expectation μ
Sample median $\mathrm{Med}(X_1, X_2, \ldots, X_n)$	Median $q_{0.5} = F^{\mathrm{inv}}(0.5)$
pth empirical quantile $q_n(p)$	$100p$th percentile $q_p = F^{\mathrm{inv}}(p)$
Sample variance S_n^2	Variance σ^2
Sample standard deviation S_n	Standard deviation σ
$\mathrm{MAD}(X_1, X_2, \ldots, X_n)$	$F^{\mathrm{inv}}(0.75) - F^{\mathrm{inv}}(0.5)$, for symmetric F

Cycles of heater switches: estimation problem

Number of cycles 20 heater switches make
after an overload voltage (in ten thousands)

Data
0.0100
0.0340
0.1940
0.5670
0.6010
0.7120
1.2910
1.3670
1.9490
2.3700
2.4110
2.8750
3.1620
3.2800
3.4910
3.6860
3.8540
4.2110
4.3970
6.4730

Find model distribution that has features close to the sample features



Shape of the distribution function
Shape of the density function
Expectation
Median
Quartiles
…

Model distribution can be used for summarizing, prediction, simulation, …

Consider the following numerical summaries of the size of datapackages. What can you infer about the model distribution?

FeedbackFruits

Sample mean:	185
Sample median:	15
First quartile:	0
Third quartile:	512
Sample standard deviation:	237

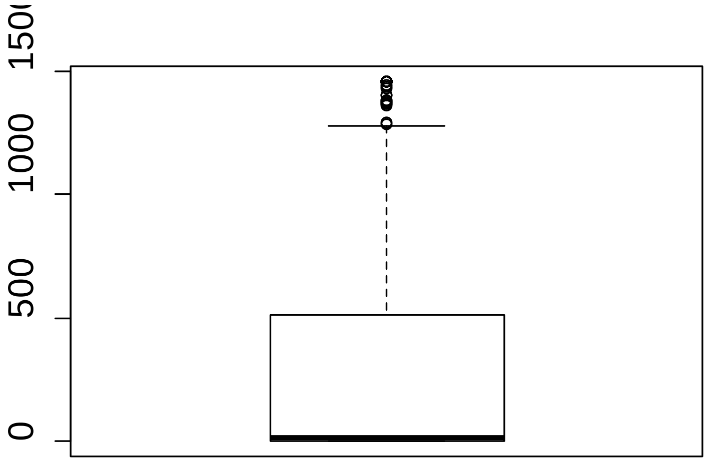
A) Symmetric and unimodal

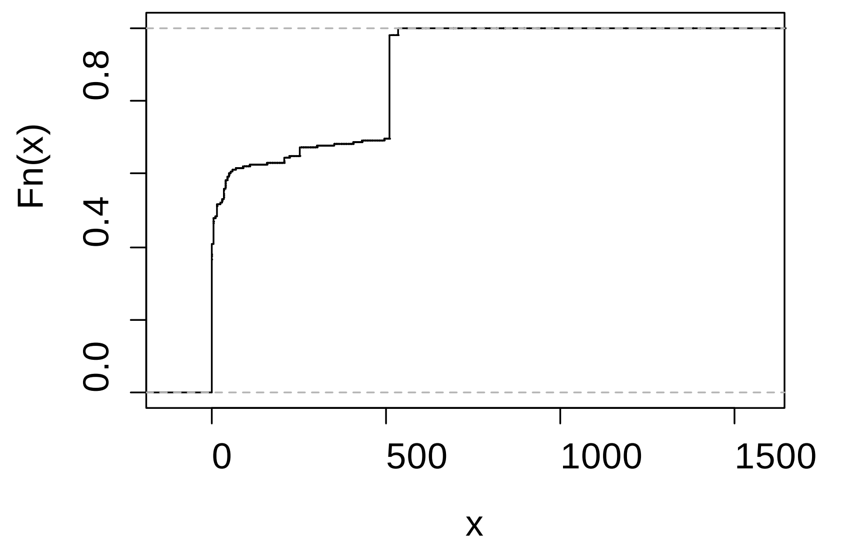
B) Symmetric and bimodal

C) Asymmetric and unimodal

D) Asymmetric and bimodal

Consider the following boxplot and empirical distribution function. What can you infer about the model distribution?





A) Symmetric and unimodal

B) Symmetric and bimodal

C) Asymmetric and unimodal

D) Asymmetric and bimodal

Exercises

Book: 17.1, 17.2, 17.4

For next class (week 3.6 lesson 2):

Complete MyStatlab assignments and book exercises 15.4, 15.5, 15.6, 16.4, 16.6, 17.5 and 17.6

Watch prelectures *‘Unbiased Estimators’*

Book: Section: 19.1, 19.2, 19.3

After this class you can:

- work with estimators and compute estimates
- check whether an estimator is (un)biased
- compare estimators

Statistics

Good luck!