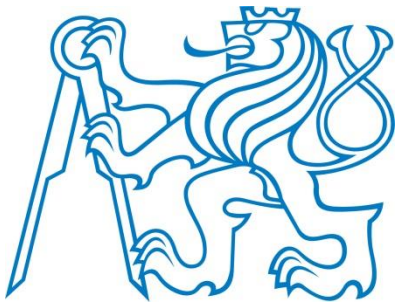
Experimental Data Analysis

in ©MATLAB

Lecture 2:

Introduction to the statistics, probability distributions, and plotting statistical data

Jan Rusz
Czech Technical University in Prague



Motivation

Why to analyze data?

- We want to make good decisions
  - E.g. Patient has 39°C → He has got a fever!
    - Yes, we know the range of fever from medical books.
    - But, how did the authors of medical books know it?
- Good decisions are based on reality
  - Authors did measure temperature of many peoples, analyzed the data and found that temperatures higher than ≈ 38°C are very rare and are related to unhealthy physical state

Motivation

Why statistical inference?

Blind faith vs. science

> "There is a fundamental difference between religion, which is based on authority, and science, which is based on observation and reason. Science will win because it works."

Hawking, S. (2010). Stephen Hawking on Religion: 'Science Will Win' on ABC World News.

How confidently can we trust to our decision?

> *"Probability is common sense reduced to calculation"*

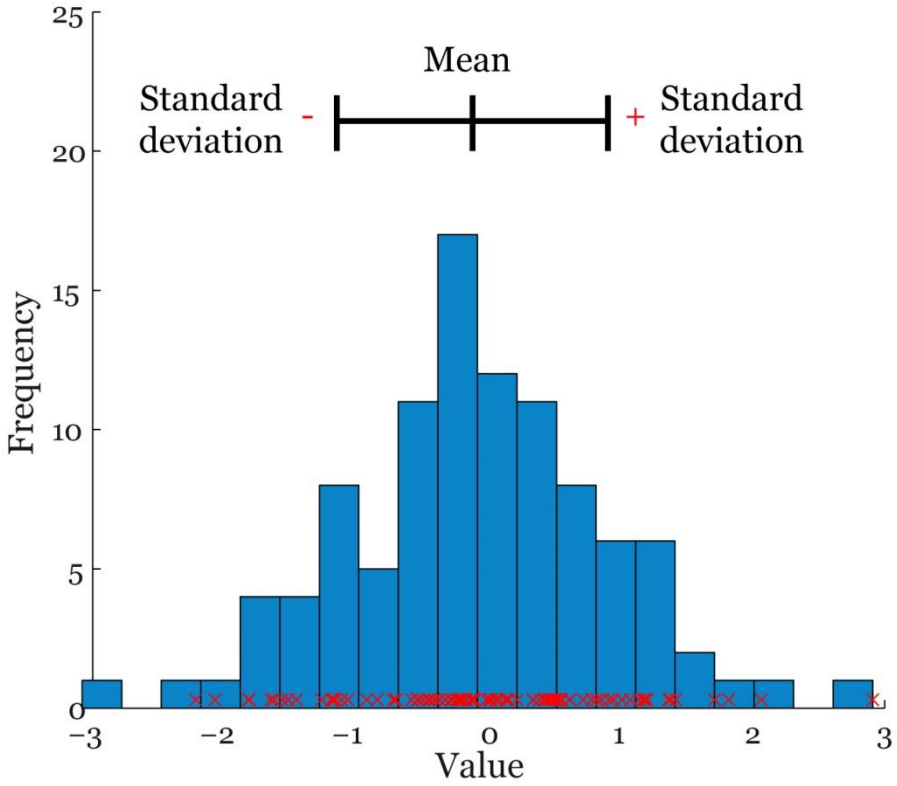
Laplace, P. S. (1814). Essai philosophique sur les probabilités.

Motivation

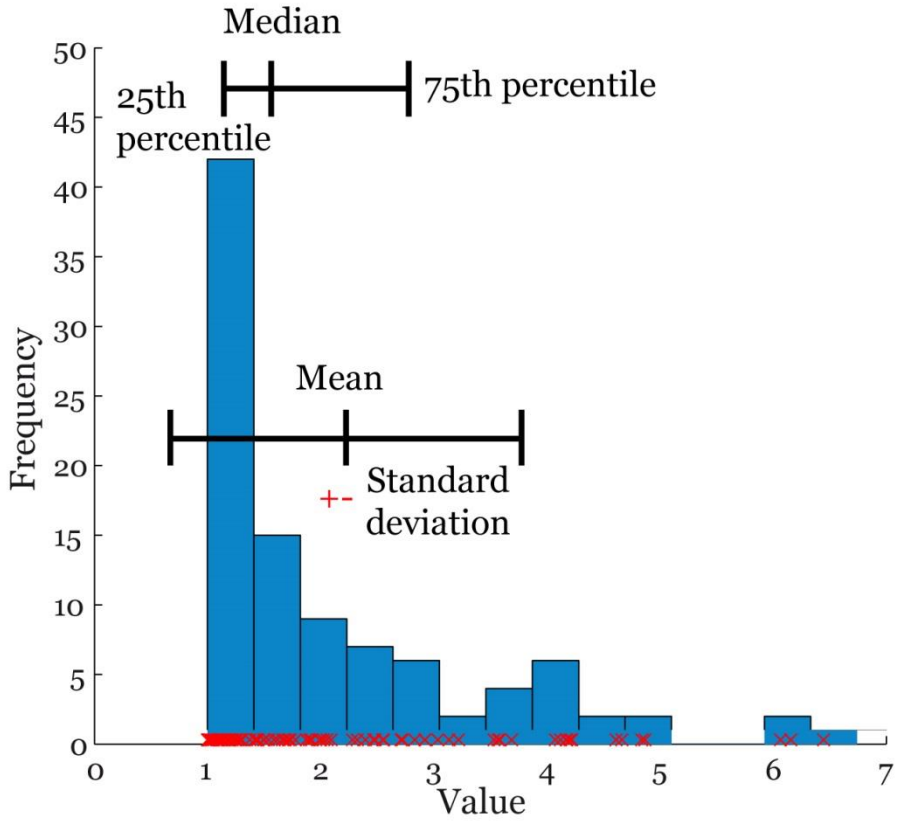
How to analyze the data?

- Direct approach
  - Look at the numbers:
    - 36.8°C, 36.7°C, 36.5°C, 36.6°C, 36.9°C, 36.7°C, … it sucks!
  - Visualize
    - We all love pictures, but what if there is too many data?!
- Statistics
  - Empirical models
    - Summarize and describe data in convenient way
  - Statistical models
    - Fit the data into something more simple e.g. equation of probability distribution function

Histogram of normal data



Histogram of non-normal data



Mean: $\mu(x) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Standard deviation: $\sigma(x) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Median: is the value separating the higher half of a data sample, from the lower half.

Percentile:

$n = \left\lceil \frac{P}{100} \times N \right\rceil$

P-th percentile of the N ordered values (sorted from least to greatest)

Mean:
Data set {1, 2, 3, 4, 5}
Mean = (1 + 2 + 3 + 4 + 5)/5 = 3

Standard deviation:
Data set {1, 2, 3, 4, 5}
Mean = 3
$(1-3)^2 = 4$ $(2-3)^2 = 1$ $(3-3)^2 = 0$
$(4-3)^2 = 1$ $(5-3)^2 = 4$
Variance (sum of the values/N-1) = (4 + 1 + 0 + 1 + 4)/4 = 2.5
Standard deviation = $\sqrt{2.5} \approx 1.58$

Median:
Data set {9, 5, 1, 4, 11, 2, 8}
Sorted data set {1, 2, 4, 5, 8, 9, 11}
median = 5

Percentile: 25^{th} P = ?
data set {10, 20, 30, 40, 50, 55, 60, 65, 70, 75}
$N = 10$

$n = \left\lceil \frac{25}{100} \times 10 \right\rceil = [2.5] = 3$

data set {10, 20, 30, 40, 50, 55, 60, 65, 70, 75}
25^{th} $P = 30$

Standard deviation:

Data set {1, 2, 3, 4, 5}

Mean = 3

$(1-3)^2 = 4$ $(2-3)^2 = 1$ $(3-3)^2 = 0$

$(4-3)^2 = 1$ $(5-3)^2 = 4$

Variance (sum of the values/N-1) = $(4 + 1 + 0 + 1 + 4)/4 = 2.5$

Standard deviation = $\sqrt{2.5} \approx 1.58$

Median:

Data set {9, 5, 1, 4, 11, 2, 8}

Sorted data set {1, 2, 4, 5, 8, 9, 11}

median = 5

Median absolute deviation (MAD):

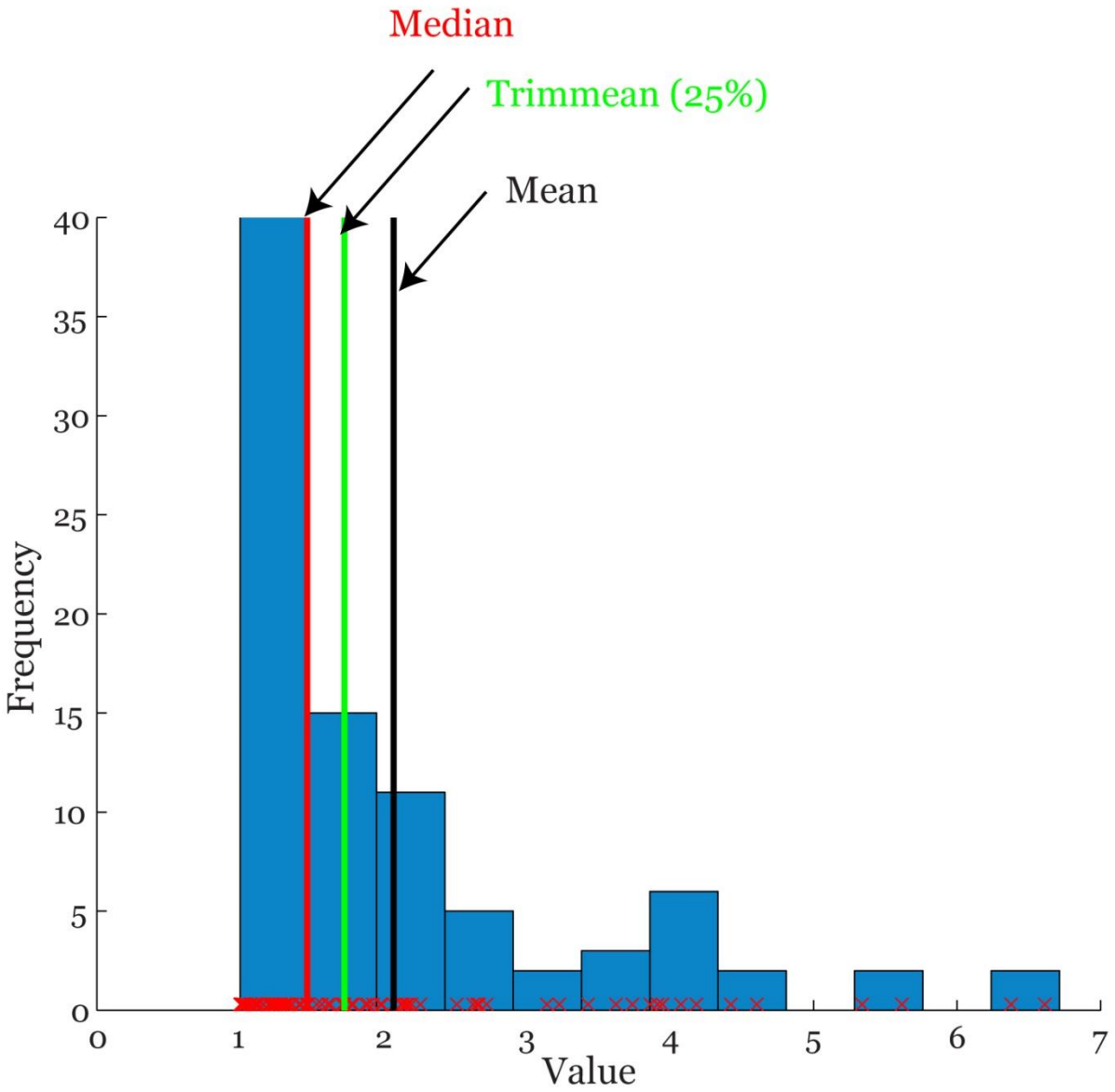
MAD = median($|X_i - \text{median}(X)|$)

$|X_i - \text{median}(X)|$ = {4,0,4,1,6,3,3}

MAD = 3

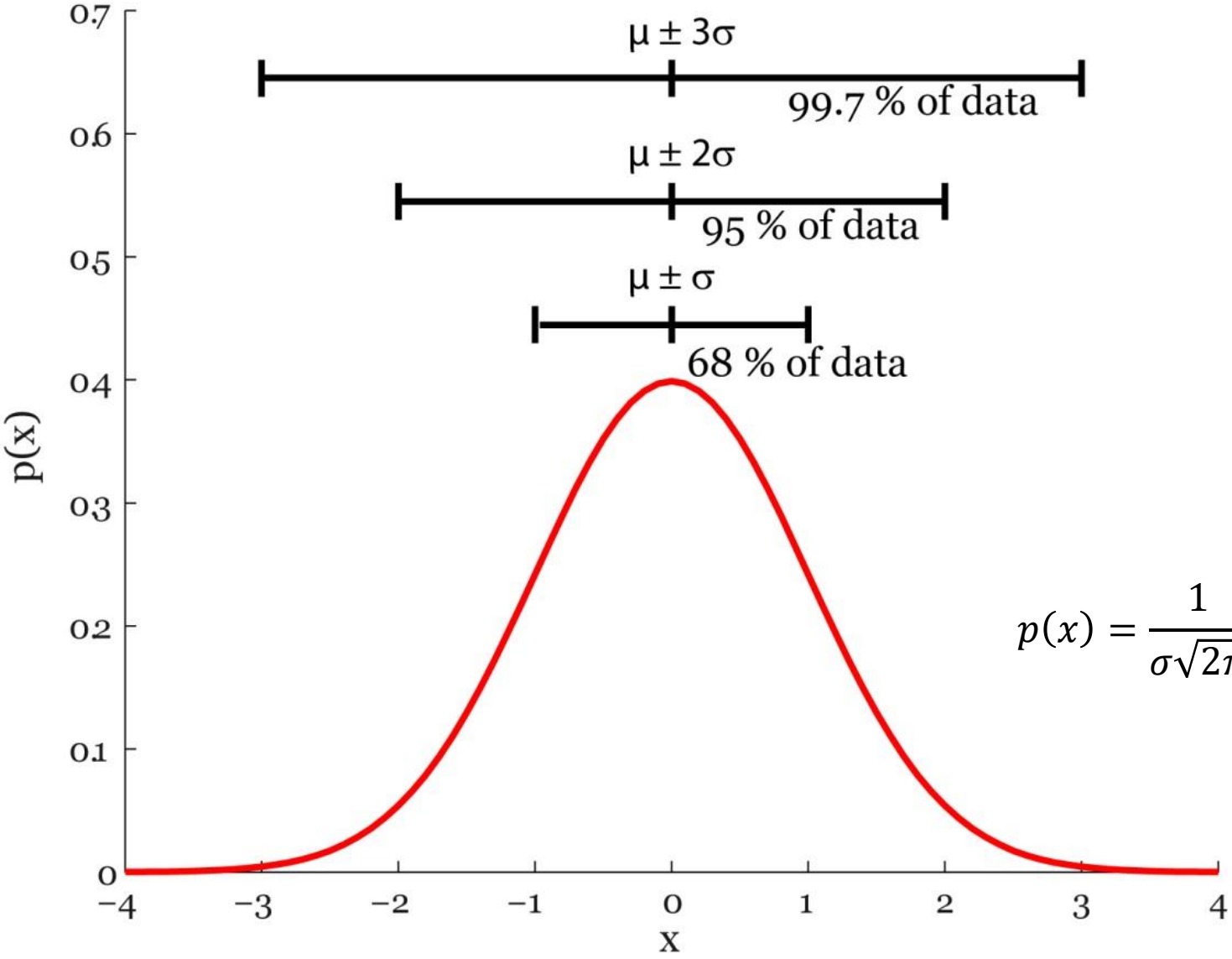
Matlab example **1**

Trimmed mean



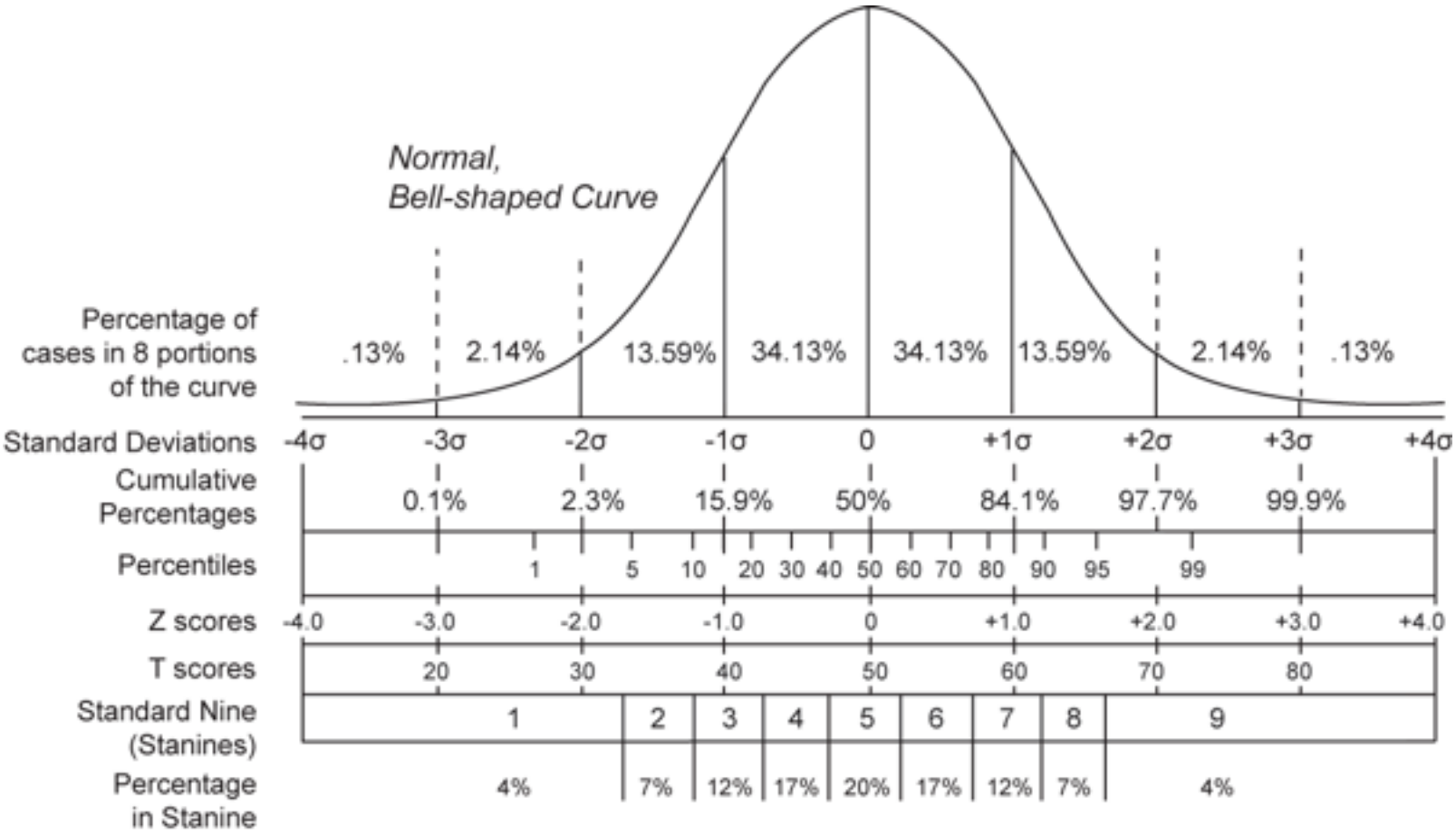
Matlab example **2**

Gaussian probability distribution

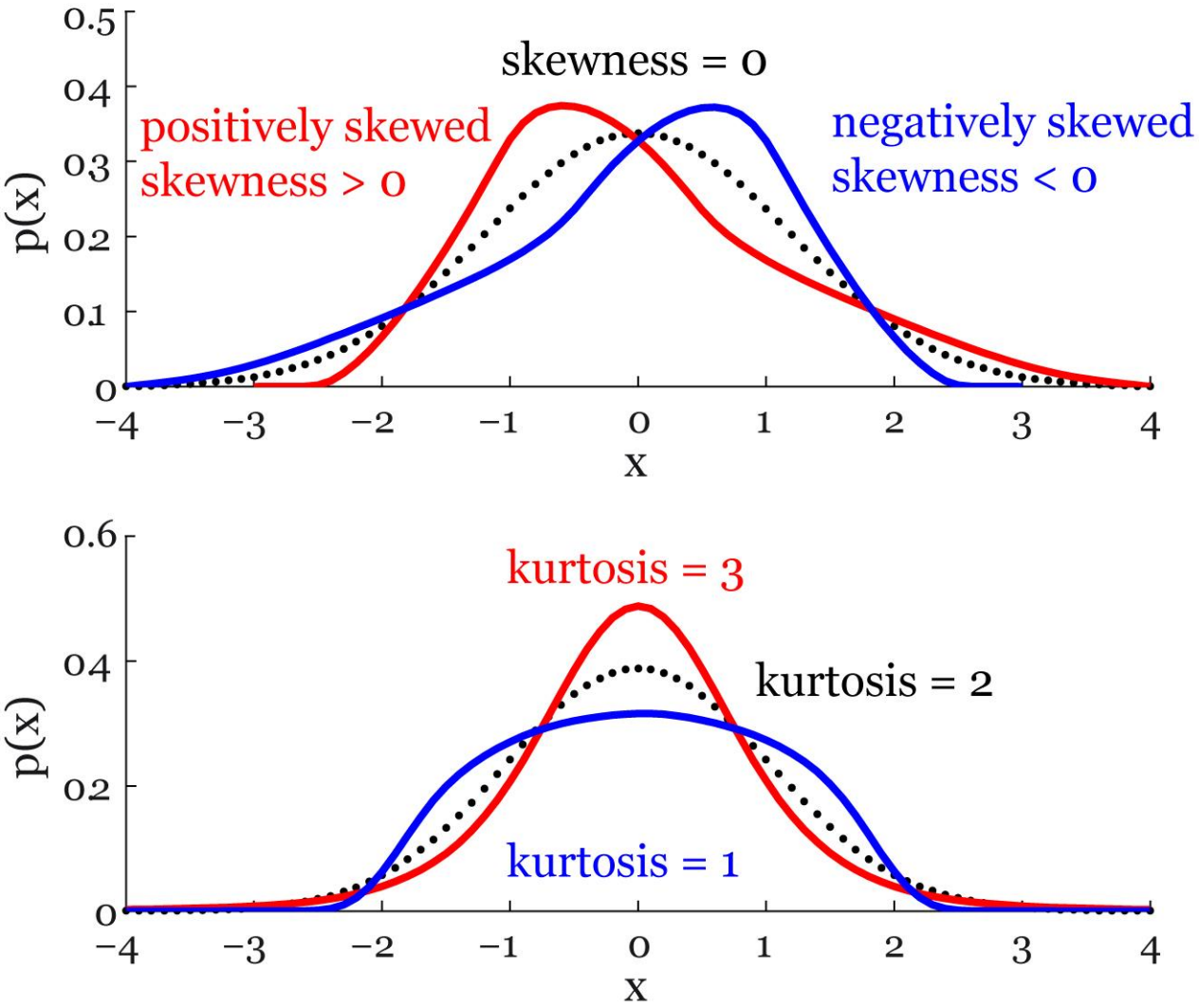


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Matlab example **3**

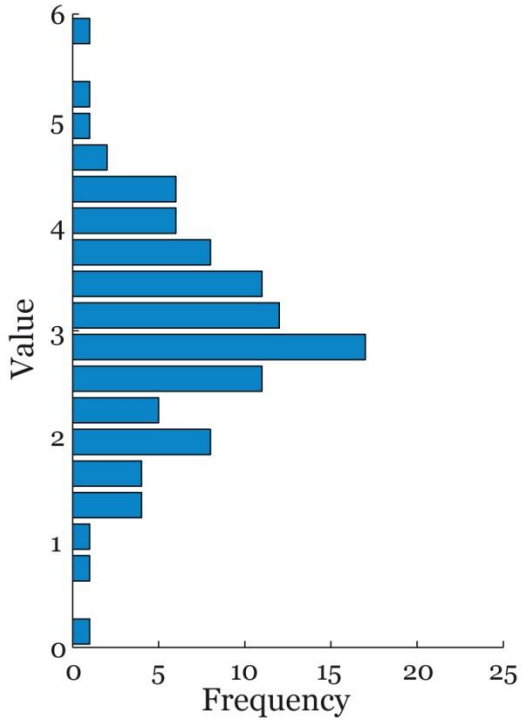


Skewness & Kurtosis

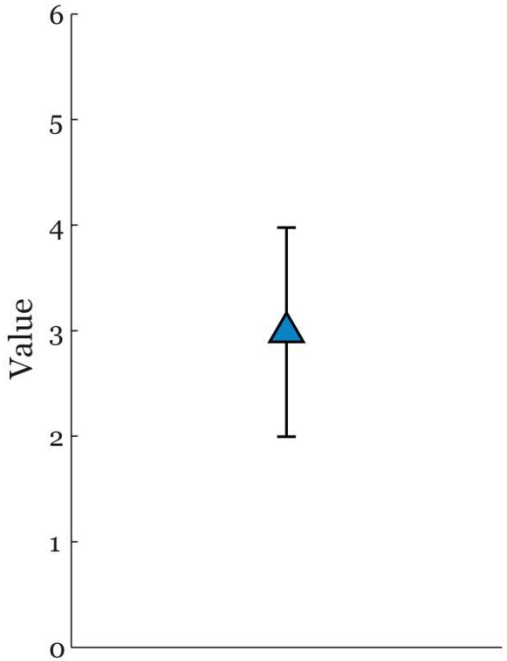


Matlab example 4

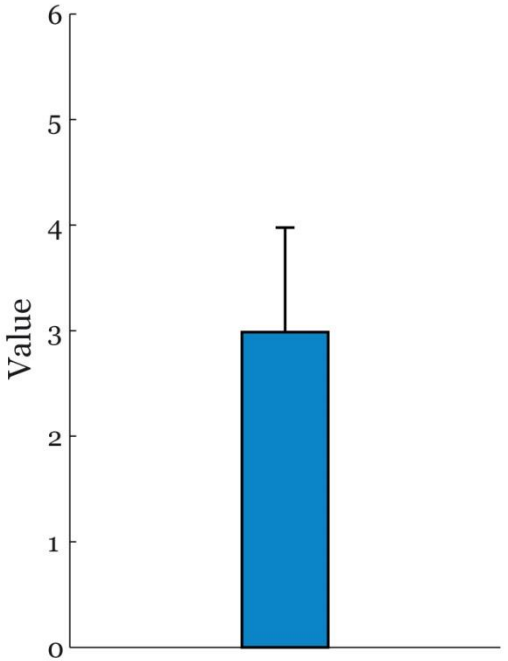
Histogram of normal data

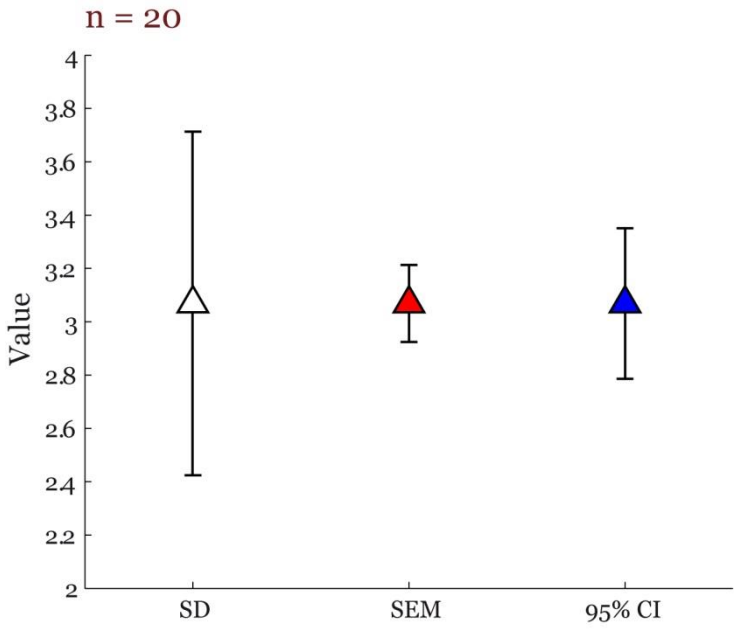


Errorbar



Bar + errorbar

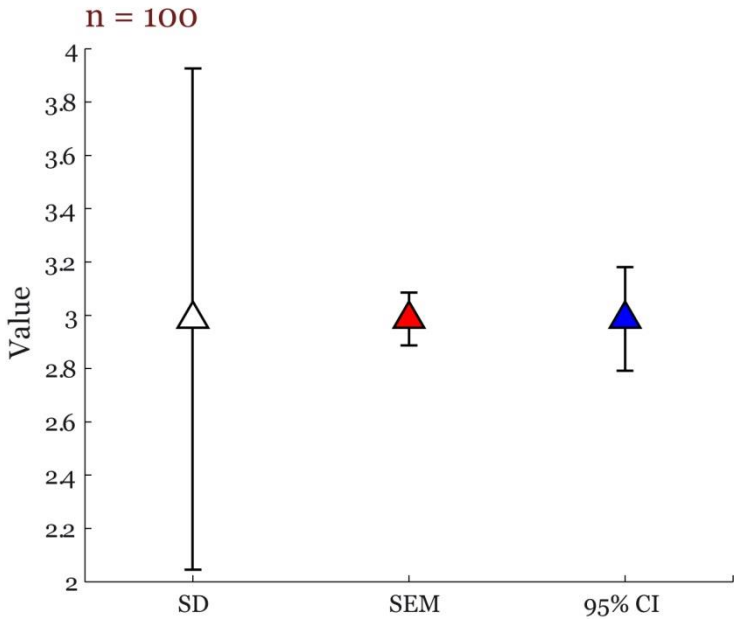




Standard Deviation (SD): $\sigma(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

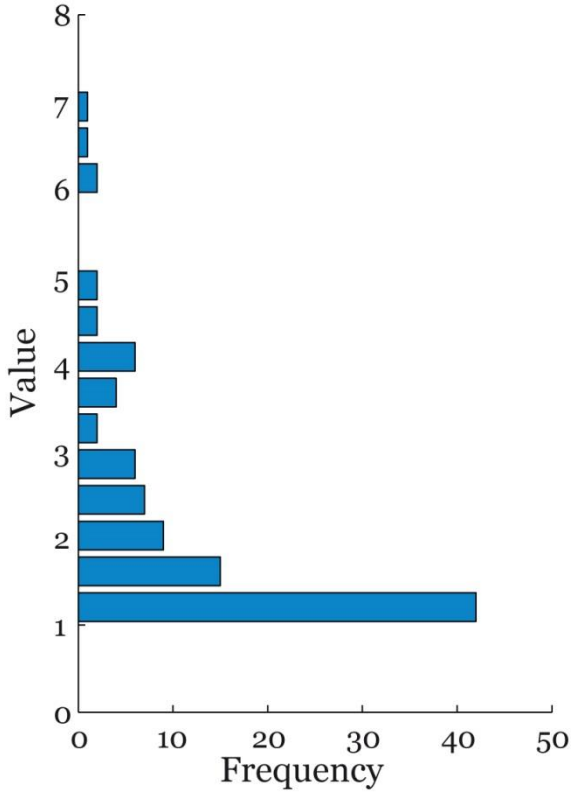
Standard Error of the Mean (SEM): $SEM = \frac{\sigma(x)}{\sqrt{n}}$

95% Confidence Interval (95% CI): $95\%\ CI = SEM \times 1.96$

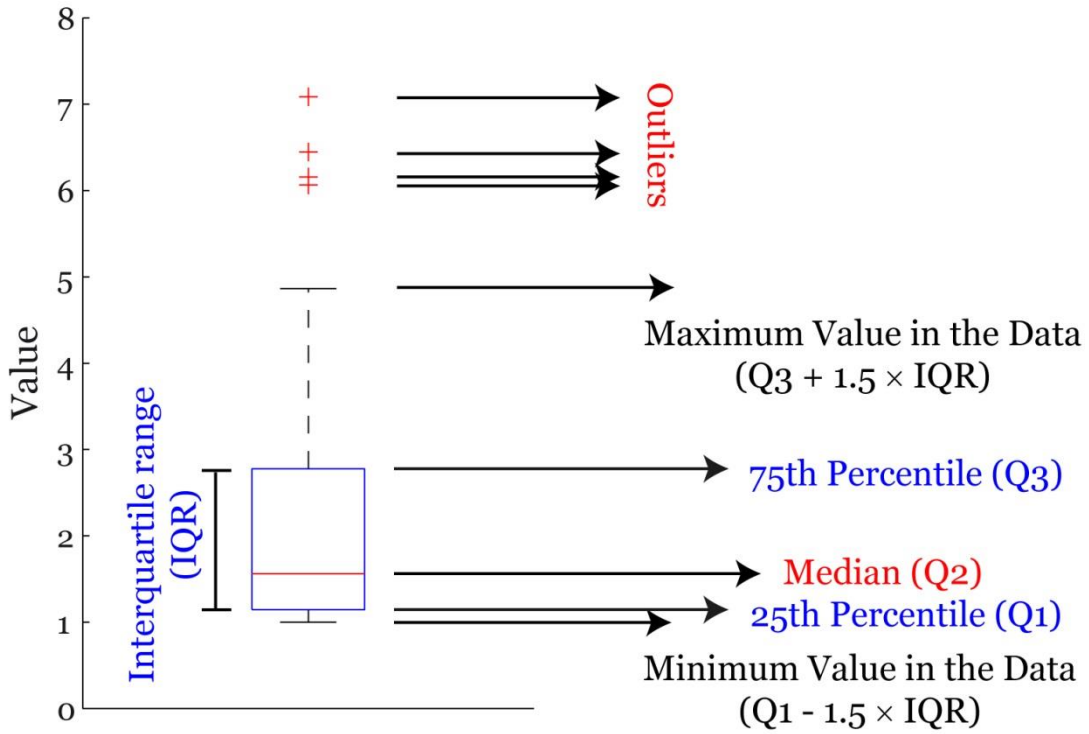


Matlab example **5**

Histogram of non-normal data

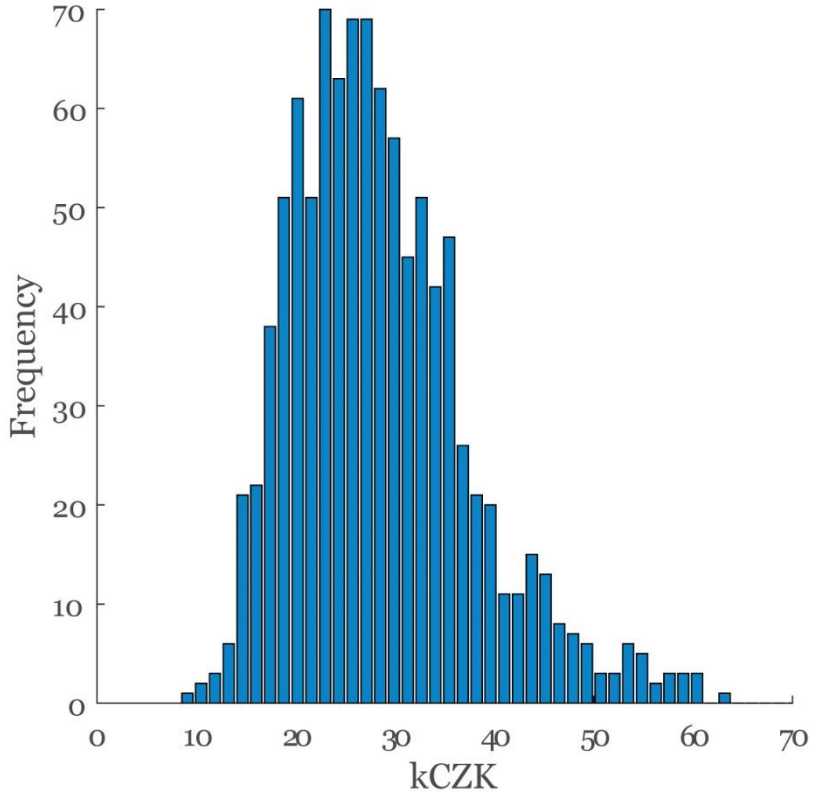


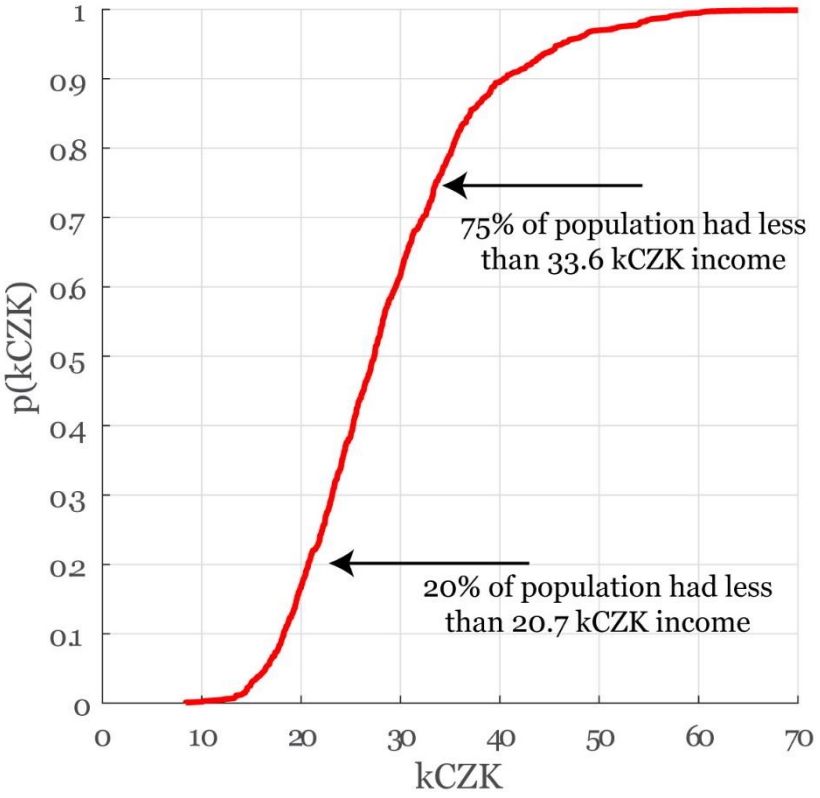
Box plot

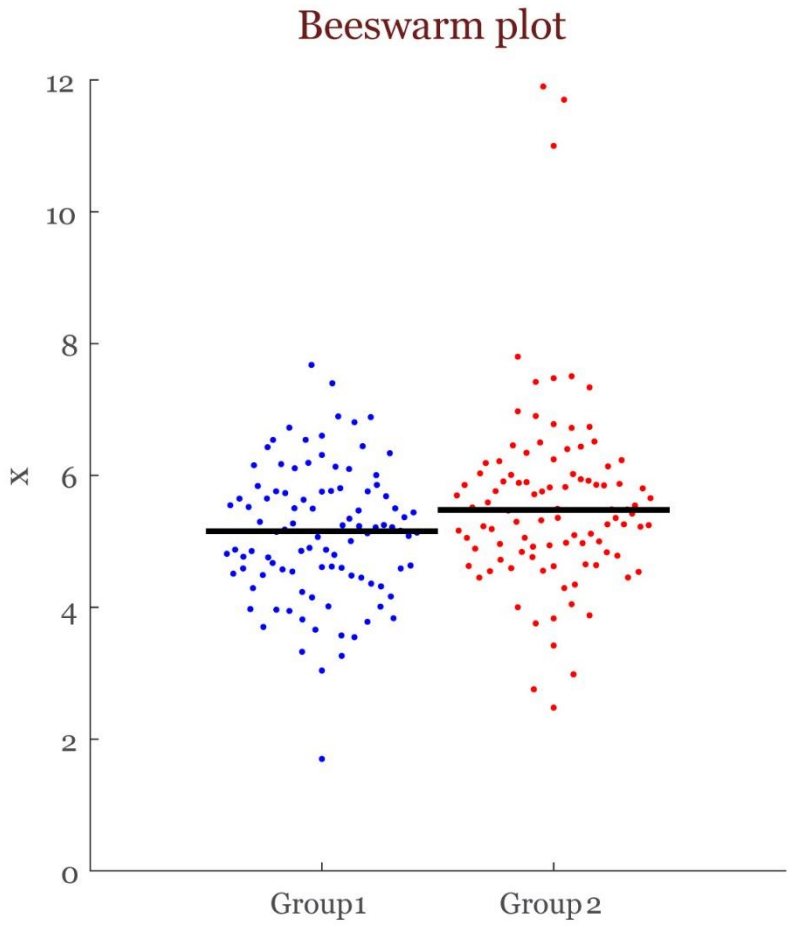


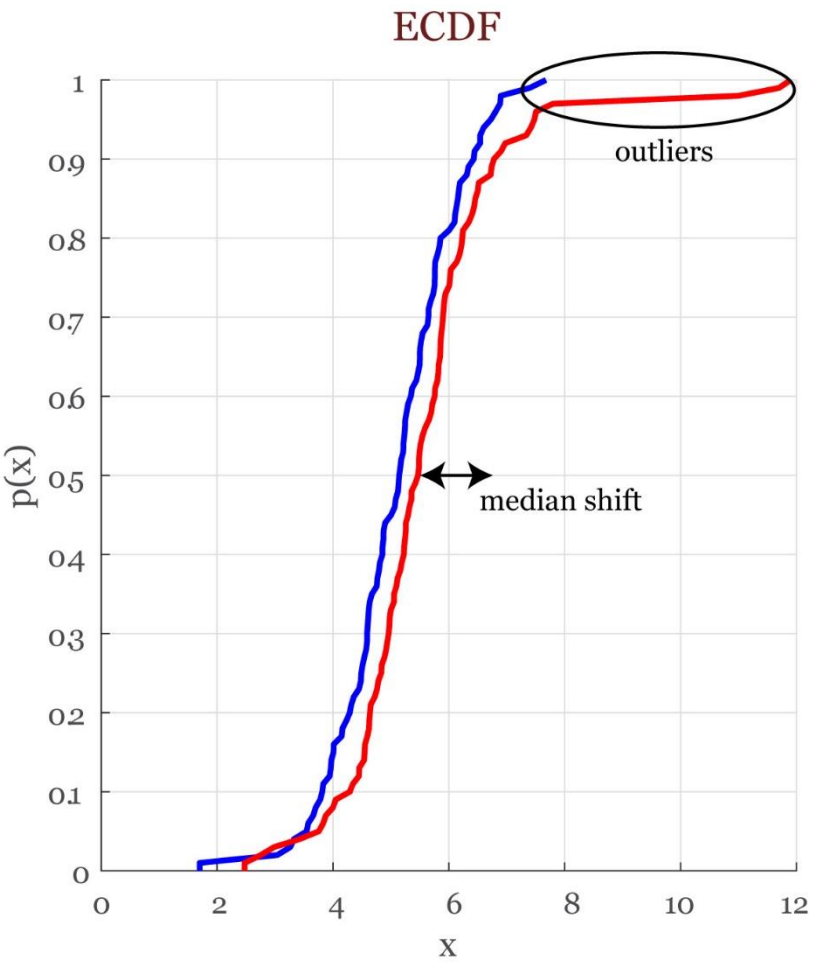
Matlab example **6**

Empirical cumulative distribution function

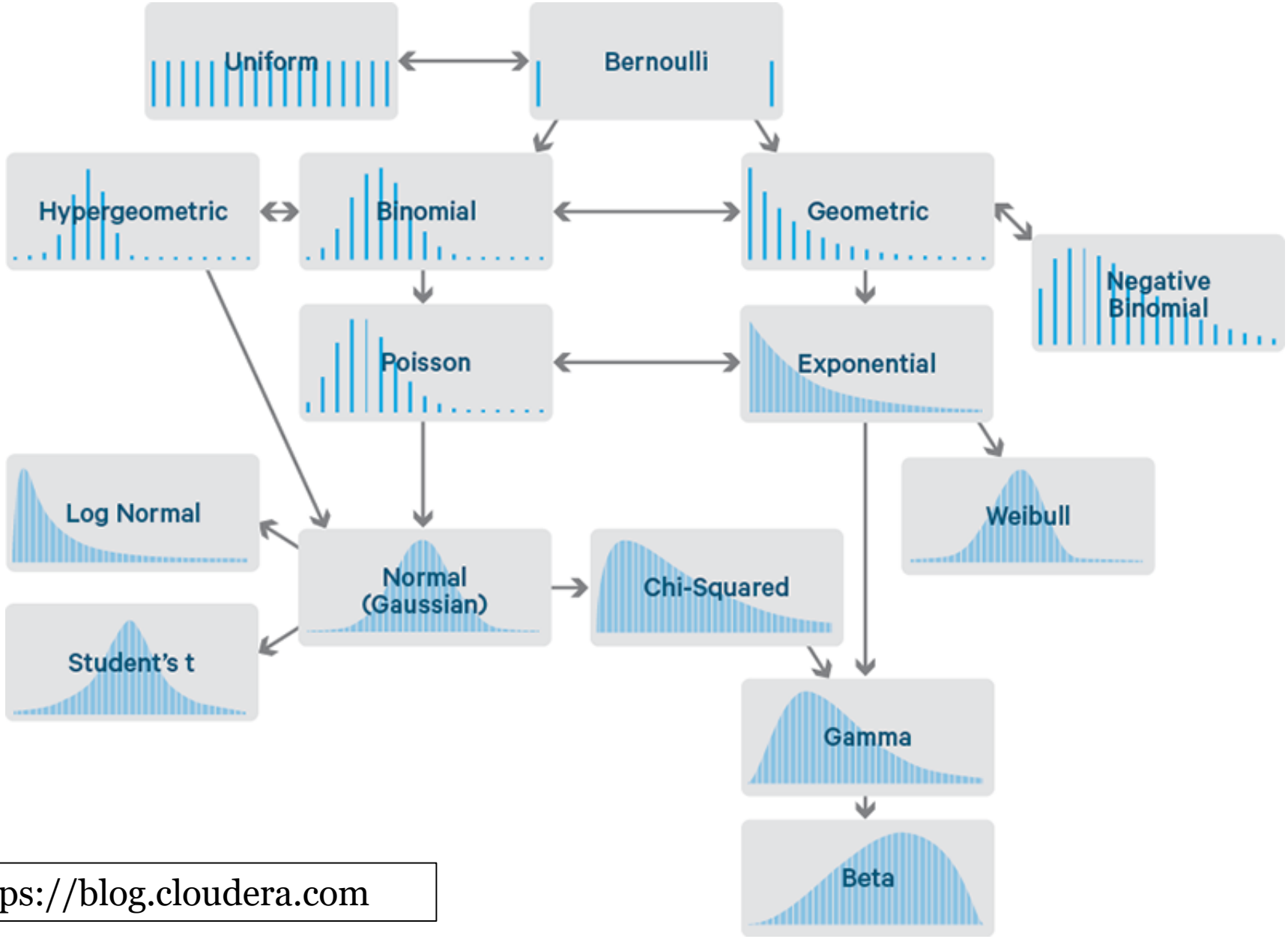




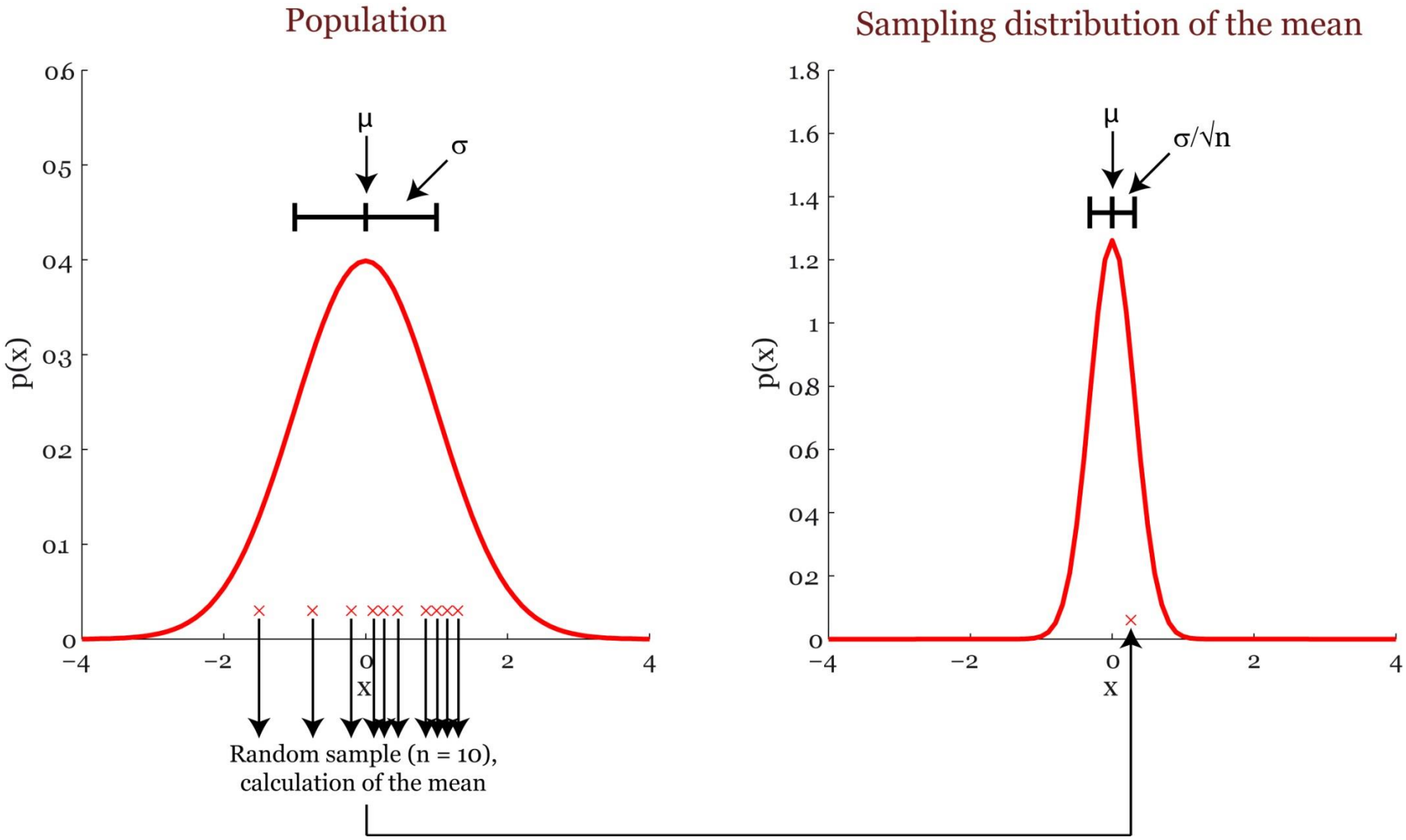

Beeswarm plot
12
10
8
6
4
2
0
x
Group1
Group2


ECDF
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
p(x)
outliers
median shift
0
2
4
6
8
10
12
x

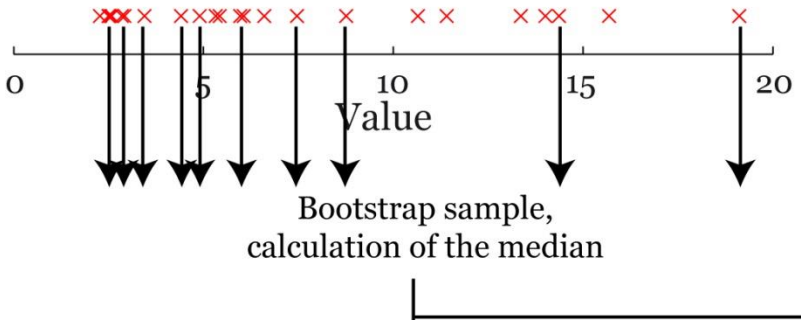
Common probability distributions



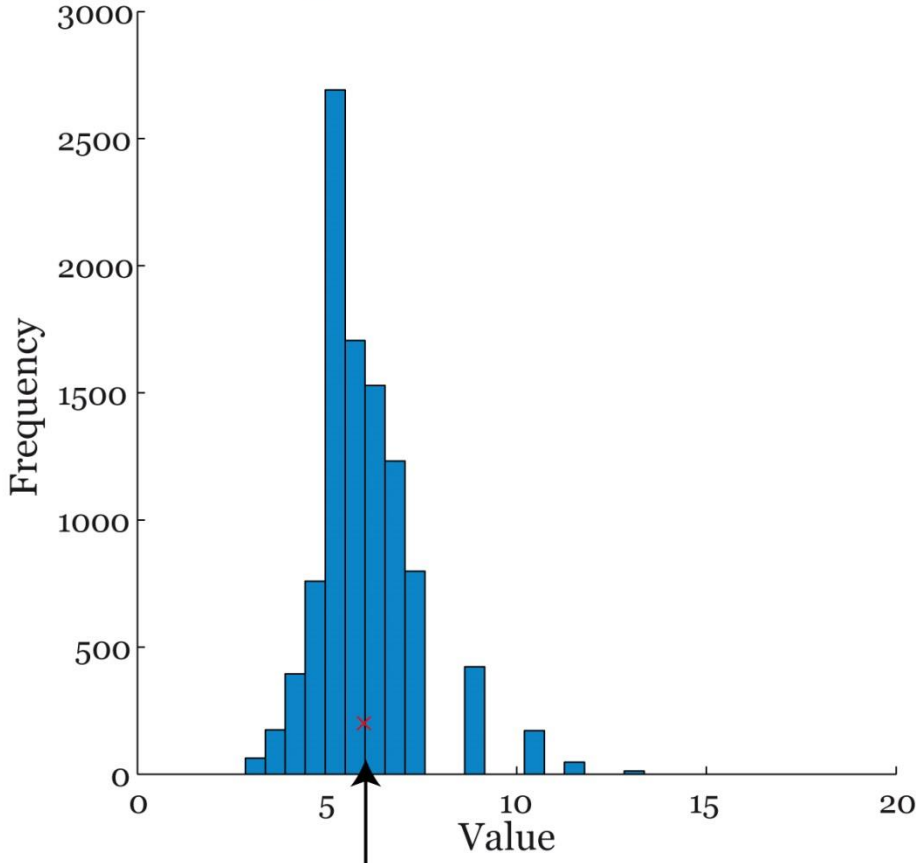
https://blog.cloudera.com


Population
Sampling distribution of the mean
μ
σ
σ/√n
p(x)
x
0.6
0.5
0.4
0.3
0.2
0.1
0
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
−4
−2
0
2
4
Random sample (n = 10),
calculation of the mean

Bootstrapping data

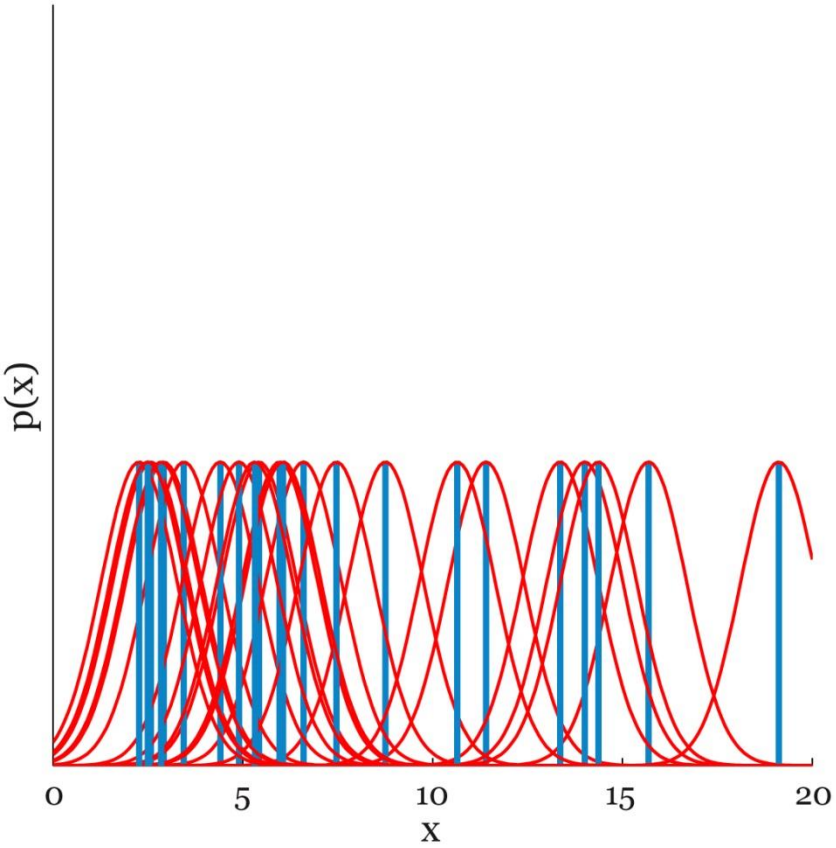


Bootstrap distribution of the median

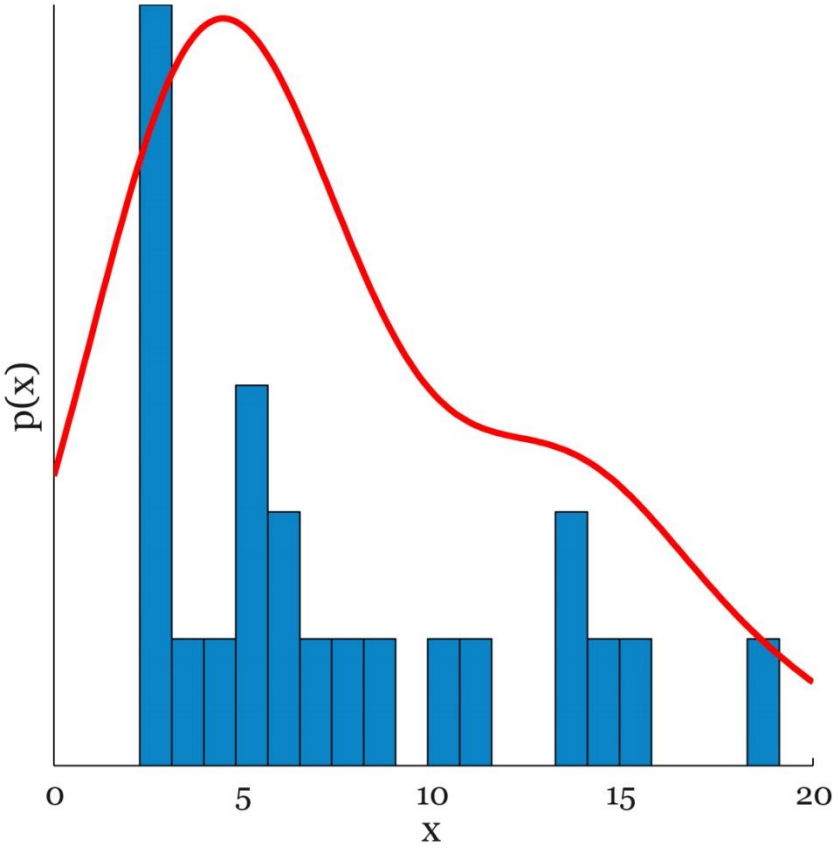


Matlab example 7

Probability distribution



Probability distribution
via kernel density estimation



Matlab example **8**

You always **have to** report your descriptive statistics with mean & standard deviation considered as a minimum !!!

Table 1. Clinical characteristics of PreHD subjects.

n = 28 (14 men)	Mean (SD)	Range
Age (years)	37.1 (9.3)	20–55
UHDRS motor score	2.2 (2.4)	0–8
Cognitive score	337 (44)	242–411
Tapping ζ	189 (23)	142–229
Pegboard ψ	4492 (805)	3469–7519
Disease burden score	251 (82)	116–413
Years to onset (years)	16.7 (8.2)	5–36