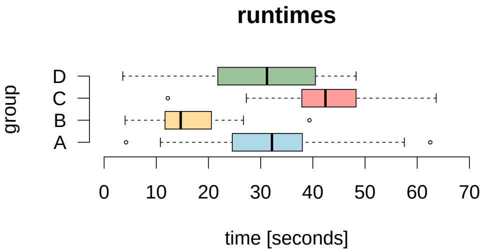
Descriptive Statistics

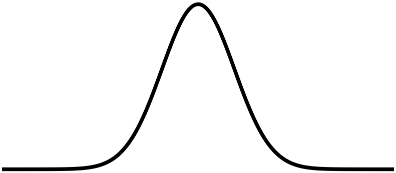


For comments please email: M. Messer, michael.messer@tuwien.ac.at

Overview

We differentiate:

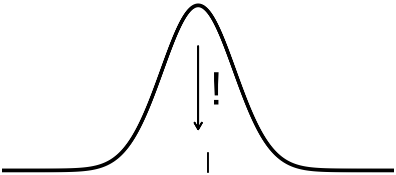
Probability theory
(Stochastics)
$=$
Theory of randomness



Overview

We differentiate:

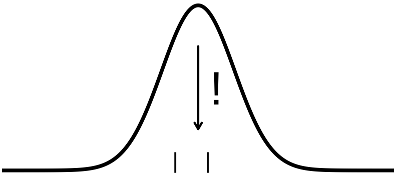
Probability theory
(Stochastics)
$=$
Theory of randomness



Overview

We differentiate:

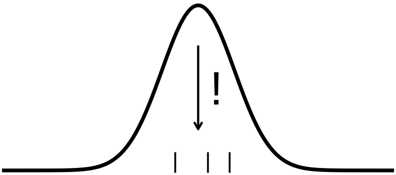
Probability theory
(Stochastics)
$=$
Theory of randomness



Overview

We differentiate:

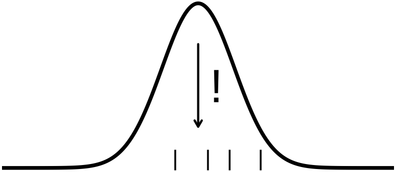
Probability theory
(Stochastics)
$=$
Theory of randomness



Overview

We differentiate:

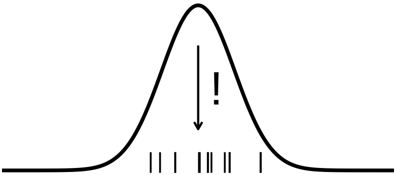
Probability theory
(Stochastics)
=
Theory of randomness



Overview

We differentiate:

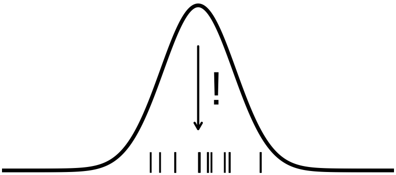
Probability theory
(Stochastics)
$=$
Theory of randomness



Overview

We differentiate:

Probability theory
(Stochastics)
$=$
Theory of randomness



and

Statistics
=
Description of data $\longrightarrow$

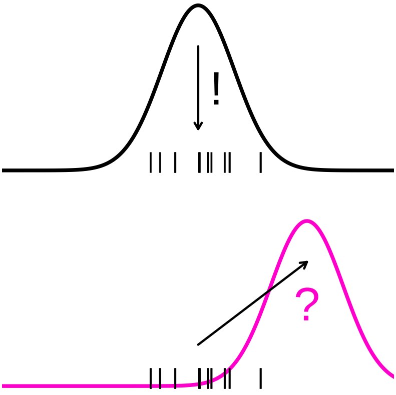
Overview

We differentiate:

Probability theory
(Stochastics)
=
Theory of randomness

and

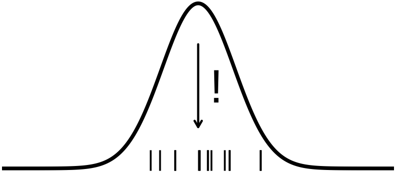
Statistics
=
Description of data $\longrightarrow$
(using stochastic models)



Overview

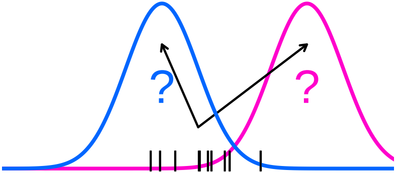
We differentiate:

Probability theory
(Stochastics)
=
Theory of randomness



and

Statistics
=
Description of data $\longrightarrow$
(using stochastic models)



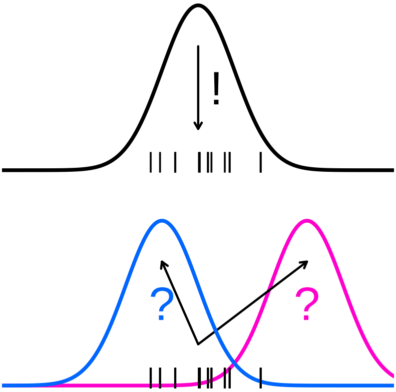
Overview

We differentiate:

Probability theory
(Stochastics)
=
Theory of randomness

and

Statistics
=
Description of data $\longrightarrow$
(using stochastic models)

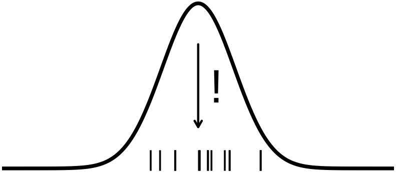


Today: Short excursion to descriptive Statistics
How do data look like? How can they be summarized?

Overview

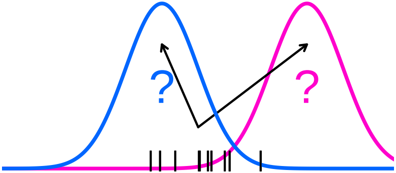
We differentiate:

Probability theory
(Stochastics)
=
Theory of randomness



and

Statistics
=
Description of data $\longrightarrow$
(using stochastic models)



Today: <u>Short excursion to descriptive Statistics</u>
How do data look like? How can they be summarized?

From then on: <u>inferential Statistics</u> (Modelling)
How did the data occur?

Scales

We differentiate scales

- Categorial data (nominal scale, no ordering)
  - Do you drink coffee? yes or no (two categories)
  - What is the color of your hair? blond, brown, black, red, neither (five categories)

Scales

We differentiate scales

- Categorial data (nominal scale, no ordering)
  - Do you drink coffee? yes or no (two categories)
  - What is the color of your hair? blond, brown, black, red, neither (five categories)

- Metric data (Ratio scale, metric distance, 2*3=6, 0=0)
  - How large are you? size in cm
  - How long is the runtime of an algorithm that you implemented? time in seconds

Scales

We differentiate scales

- Categorial data (nominal scale, no ordering)
  - Do you drink coffee? yes or no (two categories)
  - What is the color of your hair? blond, brown, black, red, neither (five categories)
- Ordinal data (order, but no metric distance)
  - How much did you learn in the course? nothing, few, much or very much (four ordered categories)
  - How often do you use Tuwel? never, sometimes, often (three ordered categories)
- Metric data (Ratio scale, metric distance, 2*3=6, 0=0)
  - How large are you? size in cm
  - How long is the runtime of an algorithm that you implemented? time in seconds

Scales

We differentiate scales

- Categorial data (nominal scale, no ordering)
  - Do you drink coffee? yes or no (two categories)
  - What is the color of your hair? blond, brown, black, red, neither (five categories)
- Ordinal data (order, but no metric distance)
  - How much did you learn in the course? nothing, few, much or very much (four ordered categories)
  - How often do you use Tuwel? never, sometimes, often (three ordered categories)
- Metric data (Ratio scale, metric distance, 2*3=6, 0=0)
  - How large are you? size in cm
  - How long is the runtime of an algorithm that you implemented? time in seconds

(Today we stick to metric data)

Data collection

How long is the runtime of an algorithm that you implemented?

Data collection

How long is the runtime of an algorithm that you implemented?

$n = 121$ students requested (same technical setup)

Data collection

How long is the runtime of an algorithm that you implemented?

$n = 121$ students requested (same technical setup)

Results (in seconds):

24.6, 24, 31.4, 29.9, 37.8, 19.9, 46.1, 32.8, 30.3, 29, 47.1, 27.8, 33.8, 30.1, 53.3, 23.8, 32.1, 4.2, 42.8, 25.2, 52.3, 35, 30.1, 43.2, 25.4, 62.5, 35.4, 25.2, 37.6, 37.1, 22.9, 29.5, 44.5, 34.8, 33.3, 21.9, 37.2, 24, 37, 34, 24.1, 10.8, 24.9, 37.2, 52, 30.8, 22, 18.6, 22, 26.8, 52.3, 27, 23.6, 33.5, 30.8, 20.9, 35.6, 37.2, 57.5, 46.2, 36.1, 19.8, 38.1, 36.9, 26.5, 23.6, 30.3, 49.9, 39, 50.2, 35.7, 11.4, 24.1, 27.5, 36.4, 29.8, 49, 42.6, 22.5, 32.7, 34.3, 21.4, 34.7, 47.3, 20.3, 35.4, 41.8, 24.9, 15.2, 42.2, 29.1, 25.1, 22.7, 41, 28.2, 30.3, 25.6, 41.8, 16.6, 38, 43.1, 29.5, 40.3, 20.5, 39.9, 24.5, 33.7, 14.6, 23.3, 36.7, 34.7, 34.9, 39.1, 32.2, 43, 12.1, 19.8, 27.4, 39.3, 35, 46.3

Data collection

How long is the runtime of an algorithm that you implemented?

$n = 121$ students requested (same technical setup)

Results (in seconds):

24.6, 24, 31.4, 29.9, 37.8, 19.9, 46.1, 32.8, 30.3, 29, 47.1, 27.8, 33.8, 30.1, 53.3, 23.8, 32.1, 4.2, 42.8, 25.2, 52.3, 35, 30.1, 43.2, 25.4, 62.5, 35.4, 25.2, 37.6, 37.1, 22.9, 29.5, 44.5, 34.8, 33.3, 21.9, 37.2, 24, 37, 34, 24.1, 10.8, 24.9, 37.2, 52, 30.8, 22, 18.6, 22, 26.8, 52.3, 27, 23.6, 33.5, 30.8, 20.9, 35.6, 37.2, 57.5, 46.2, 36.1, 19.8, 38.1, 36.9, 26.5, 23.6, 30.3, 49.9, 39, 50.2, 35.7, 11.4, 24.1, 27.5, 36.4, 29.8, 49, 42.6, 22.5, 32.7, 34.3, 21.4, 34.7, 47.3, 20.3, 35.4, 41.8, 24.9, 15.2, 42.2, 29.1, 25.1, 22.7, 41, 28.2, 30.3, 25.6, 41.8, 16.6, 38, 43.1, 29.5, 40.3, 20.5, 39.9, 24.5, 33.7, 14.6, 23.3, 36.7, 34.7, 34.9, 39.1, 32.2, 43, 12.1, 19.8, 27.4, 39.3, 35, 46.3

We see: n data: $x_1 = 24.6, x_2 = 24.0, \ldots, x_n = 46.3$

Data collection

How long is the runtime of an algorithm that you implemented?

$n = 121$ students requested (same technical setup)

Results (in seconds):

24.6, 24, 31.4, 29.9, 37.8, 19.9, 46.1, 32.8, 30.3, 29, 47.1, 27.8, 33.8, 30.1, 53.3, 23.8, 32.1, 4.2, 42.8, 25.2, 52.3, 35, 30.1, 43.2, 25.4, 62.5, 35.4, 25.2, 37.6, 37.1, 22.9, 29.5, 44.5, 34.8, 33.3, 21.9, 37.2, 24, 37, 34, 24.1, 10.8, 24.9, 37.2, 52, 30.8, 22, 18.6, 22, 26.8, 52.3, 27, 23.6, 33.5, 30.8, 20.9, 35.6, 37.2, 57.5, 46.2, 36.1, 19.8, 38.1, 36.9, 26.5, 23.6, 30.3, 49.9, 39, 50.2, 35.7, 11.4, 24.1, 27.5, 36.4, 29.8, 49, 42.6, 22.5, 32.7, 34.3, 21.4, 34.7, 47.3, 20.3, 35.4, 41.8, 24.9, 15.2, 42.2, 29.1, 25.1, 22.7, 41, 28.2, 30.3, 25.6, 41.8, 16.6, 38, 43.1, 29.5, 40.3, 20.5, 39.9, 24.5, 33.7, 14.6, 23.3, 36.7, 34.7, 34.9, 39.1, 32.2, 43, 12.1, 19.8, 27.4, 39.3, 35, 46.3

We see: n data: $x_1 = 24.6$, $x_2 = 24.0, \ldots, x_n = 46.3$

We understand: nothing?

Data collection

How long is the runtime of an algorithm that you implemented?

$n = 121$ students requested (same technical setup)

Results (in seconds):

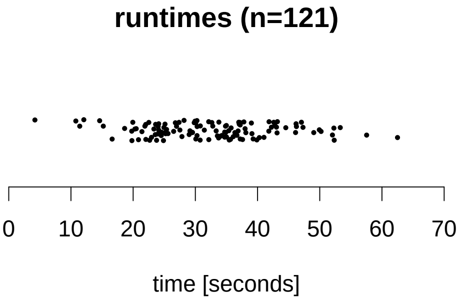
24.6, 24, 31.4, 29.9, 37.8, 19.9, 46.1, 32.8, 30.3, 29, 47.1, 27.8, 33.8, 30.1, 53.3, 23.8, 32.1, 4.2, 42.8, 25.2, 52.3, 35, 30.1, 43.2, 25.4, 62.5, 35.4, 25.2, 37.6, 37.1, 22.9, 29.5, 44.5, 34.8, 33.3, 21.9, 37.2, 24, 37, 34, 24.1, 10.8, 24.9, 37.2, 52, 30.8, 22, 18.6, 22, 26.8, 52.3, 27, 23.6, 33.5, 30.8, 20.9, 35.6, 37.2, 57.5, 46.2, 36.1, 19.8, 38.1, 36.9, 26.5, 23.6, 30.3, 49.9, 39, 50.2, 35.7, 11.4, 24.1, 27.5, 36.4, 29.8, 49, 42.6, 22.5, 32.7, 34.3, 21.4, 34.7, 47.3, 20.3, 35.4, 41.8, 24.9, 15.2, 42.2, 29.1, 25.1, 22.7, 41, 28.2, 30.3, 25.6, 41.8, 16.6, 38, 43.1, 29.5, 40.3, 20.5, 39.9, 24.5, 33.7, 14.6, 23.3, 36.7, 34.7, 34.9, 39.1, 32.2, 43, 12.1, 19.8, 27.4, 39.3, 35, 46.3

We see: n data: $x_1 = 24.6, x_2 = 24.0, \ldots, x_n = 46.3$

We understand: nothing?

Thus: descriptive Statistics $\rightarrow$ graphical representation and summary of data

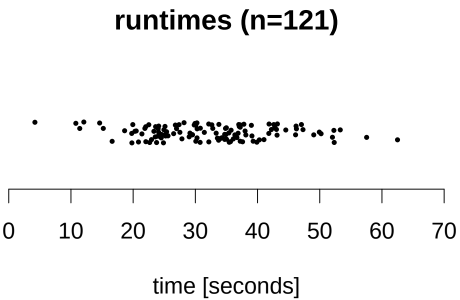
Stripchart



At first sight we understand how the n data distribute:

- Many data lie close to 30 (typical runtime)

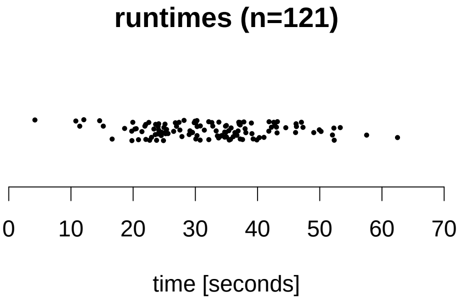
Stripchart



At first sight we understand how the n data distribute:

- Many data lie close to 30 (typical runtime)
- The minimum is about 5 (fastest runtime),
  the maximum is about 65 (slowest runtime)

Stripchart

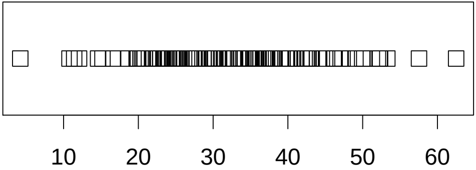


At first sight we understand how the n data distribute:

- Many data lie close to 30 (typical runtime)
- The minimum is about 5 (fastest runtime),
  the maximum is about 65 (slowest runtime)
- Remark.: the y-value has no meaning. The data are 'jittered' along the y-direction for a better overview.

Stripchart in R

```
#Enter data
x <- c(24.6, 24.0, 31.4, 29.9,...,39.3, 35.0, 46.3)
#Create stripchart
stripchart(x)
```

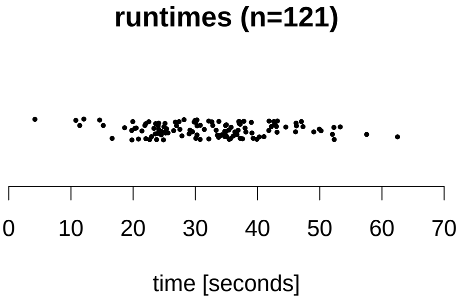


We don't understand too much - points superposed, axes annotations are missing, title is missing etc.

→ customize graphic using additional arguments or lowlevel graphics

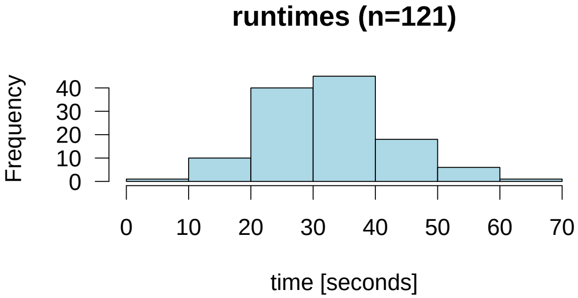
Stripchart in R

```
#Enter data
x <- c(24.6, 24.0, 31.4, 29.9,...,39.3, 35.0, 46.3)
#Create stripchart with additional arguments
stripchart(x,method="jitter",pch=19,cex=0.4,axes=FALSE,
   xlim=c(0,70),main="runtimes␣(n=121)",xlab="time␣[seconds]")
#add x-axis (lowlevelgraphic)
axis(1,at=seq(0,70,10))
```



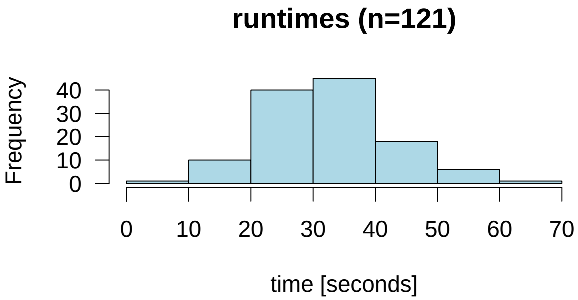
Much more informative!

Histogram



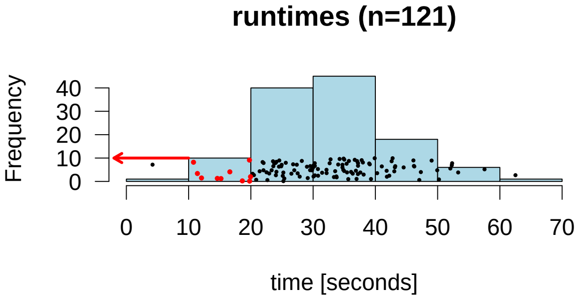
- Description of the distribution of data
  Here: approximately *bell-shaped*, i.e., unimodal and symmetric

Histogram



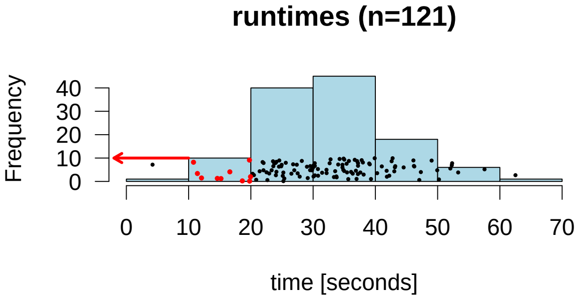
- Description of the distribution of data
  Here: approximately *bell-shaped*, i.e., unimodal and symmetric
- Absolute frequencies in the intervals $\{(10k, 10(k+1)] : k = 0, 1, \ldots, 6\}$ given through the height of the bars

Histogram



- Description of the distribution of data
  Here: approximately *bell-shaped*, i.e., unimodal and symmetric
- Absolute frequencies in the intervals $\{(10k, 10(k+1)] : k = 0, 1, \ldots, 6\}$ given through the height of the bars
  e.g.: 10 data are > 10 and $\leqslant 20$, for short $\sum_{i=1}^{n} \mathbb{1}_{(10,20]}(x_i) = 10$

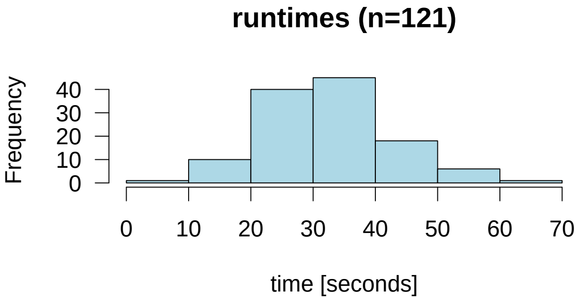
Histogram



- Description of the distribution of data
  Here: approximately *bell-shaped*, i.e., unimodal and symmetric
- Absolute frequencies in the intervals $\{(10k, 10(k+1)] : k = 0, 1, \ldots, 6\}$ given through the height of the bars
  e.g.: 10 data are > 10 and $\leqslant 20$, for short $\sum_{i=1}^{n} \mathbb{1}_{(10,20]}(x_i) = 10$
  Consequence: The sum of the bar heights is $n = 121$

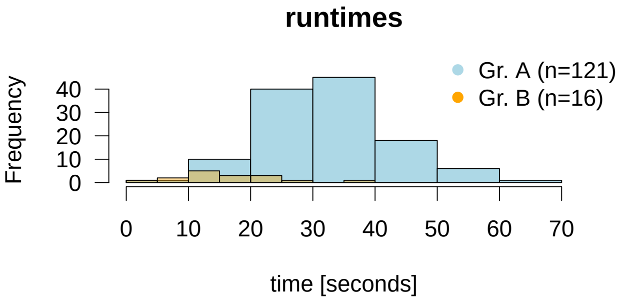
Histogram in R

```r
# Histogram with additional arguments
hist(x,las=1,xlab="time␣[seconds]",ylab="Frequency",
main="runtimes␣(n=121)",col="lightblue")
```



Histogram

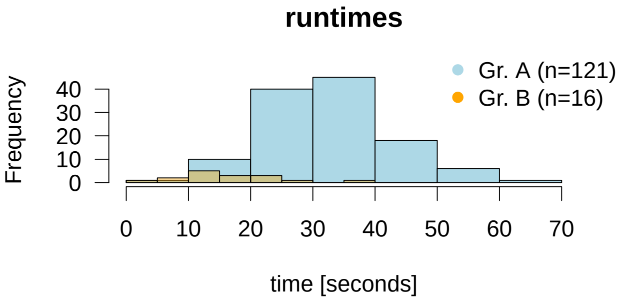
The same algorithm was implemented by 16 other students after they attended a certain programming course (group B)



- Comparison of group A ($n_A = 121$) and group B ($n_B = 16$) inappropriate, because the sizes of the groups differ tremendously.

Histogram

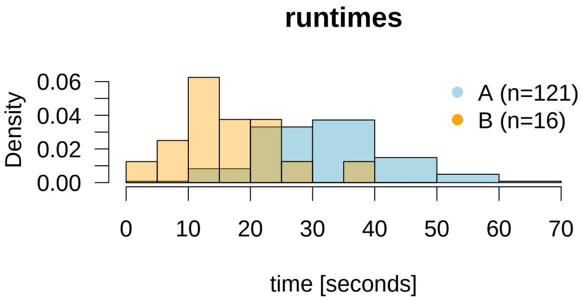
The same algorithm was implemented by 16 other students after they attended a certain programming course (group B)



- Comparison of group A ($n_A = 121$) and group B ($n_B = 16$) inappropriate, because the sizes of the groups differ tremendously.
- Idea: Norm the areas $\rightarrow$ total area of 1 each

Histogram

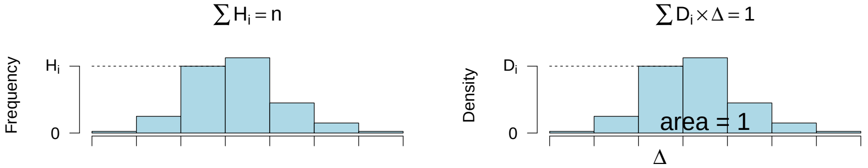
The same algorithm was implemented by 16 other students after they attended a certain programming course (group B)



- Comparison of group A ($n_A = 121$) and group B ($n_B = 16$) inappropriate, because the sizes of the groups differ tremendously.
- Idea: Norm the areas $\rightarrow$ total area of 1 each
  The distributions are now nicely visible:
  shifted against each other and about bell-shaped each.

Histogram

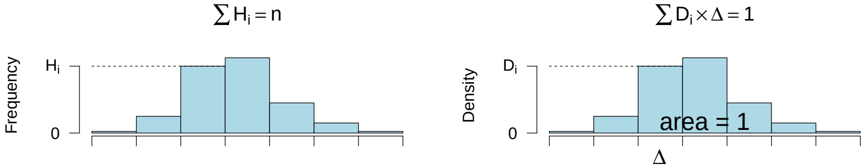
What happens when norming?



- Same 'picture', but different *y*-axis

Histogram

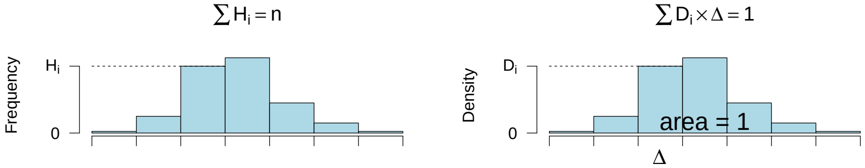
What happens when norming?



- Same 'picture', but different y-axis
  Search D_i such that total area $\sum D_i \cdot \Delta \overset{!}{=} 1$

Histogram

What happens when norming?



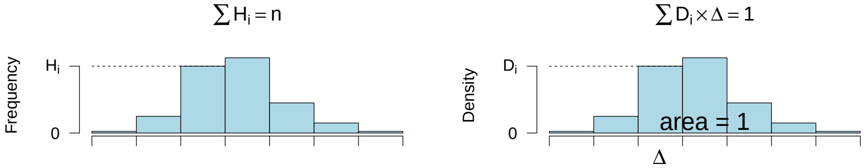
- Same 'picture', but different y-axis

  Search D_i such that total area $\sum D_i \cdot \Delta \stackrel{!}{=} 1$

  $\sum H_i = n$

Histogram

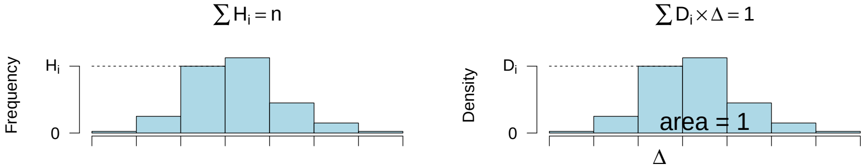
What happens when norming?



- Same 'picture', but different y-axis
  Search D_i such that total area $\sum D_i \cdot \Delta \overset{!}{=} 1$
  $\sum H_i = n \Leftrightarrow 1 = \sum \frac{H_i}{n}$

Histogram

What happens when norming?

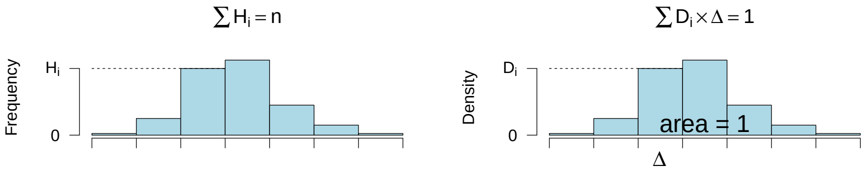


- Same 'picture', but different y-axis

  Search D_i such that total area $\sum D_i \cdot \Delta \stackrel{!}{=} 1$
  $\sum H_i = n \Leftrightarrow 1 = \sum \frac{H_i}{n} = \sum \frac{H_i}{n \cdot \Delta} \cdot \Delta$

Histogram

What happens when norming?



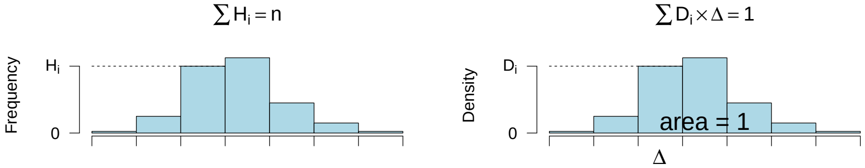
- Same 'picture', but different y-axis

  Search D_i such that total area $\sum D_i \cdot \Delta \stackrel{!}{=} 1$

  $\sum H_i = n \Leftrightarrow 1 = \sum \frac{H_i}{n} = \sum \frac{H_i}{n \cdot \Delta} \cdot \Delta$, hence $D_i = \frac{H_i}{n \cdot \Delta}$

Histogram

What happens when norming?

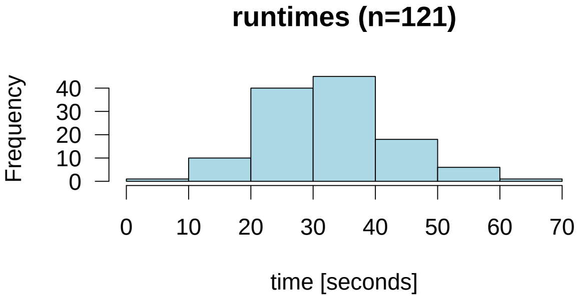


- Same 'picture', but different y-axis

  Search D_i such that total area $\sum D_i \cdot \Delta \overset{!}{=} 1$

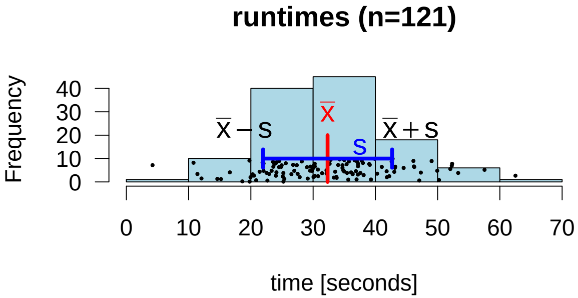
  $\sum H_i = n \Leftrightarrow 1 = \sum \frac{H_i}{n} = \sum \frac{H_i}{n \cdot \Delta} \cdot \Delta$, hence $D_i = \frac{H_i}{n \cdot \Delta}$
- R normes automatically via `hist(...,prob=TRUE)`

Mean and empirical standard deviation



If the data distribute approximately bell-shaped, then they can be summarized nicely by two prominent *statistics*, i.e., functions of the data:

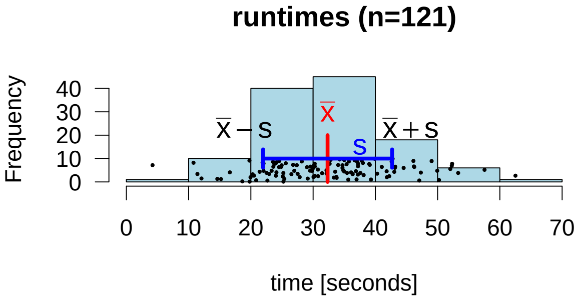
Mean and empirical standard deviation



If the data distribute approximately bell-shaped, then they can be summarized nicely by two prominent *statistics*, i.e., functions of the data:

- 1. the mean $\bar{x}$ $\rightarrow$ where? (location)

Mean and empirical standard deviation



If the data distribute approximately bell-shaped, then they can be summarized nicely by two prominent *statistics*, i.e., functions of the data:

- 1. the mean $\bar{x}$ $\rightarrow$ where? (location)
- 2. the (empirical) standard deviation s $\rightarrow$ how variable? (dispersion)

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

- The mean is

$$\boxed{\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i}$$

(center of mass of the data)

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

- The mean is

$$\boxed{\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i}$$

(center of mass of the data)

- The (empirical) variance is

$$\boxed{s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

'the mean squared deviation of the data from the mean'

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

- The mean is

$$\boxed{\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i}$$

  (center of mass of the data)

- The (empirical) variance is

$$\boxed{s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

  ’the mean squared deviation of the data from the mean’

- The (empirical) standard deviation is

$$s = \sqrt{s^2}$$

  ’the square root of the variance’

Mean and empirical standard deviation

Data $x_1, x_2, \dots, x_n$

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

Remark:

- The factor $n-1$ in s^2 (instead of e.g., n) has technical reasons

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

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

Remark:

- The factor $n-1$ in s^2 (instead of e.g., n) has technical reasons
  We speak about the *corrected* empirical variance, while for large n this correction has no practical relevance.

Mean and empirical standard deviation

Data $x_1, x_2, \ldots, x_n$

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

Random variable X (here discrete)

$$\mathbb{E}[X] := \sum x \cdot \mathbb{P}(X = x) \qquad \mathbb{V}ar(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] \qquad \sigma_X := \sqrt{\mathbb{V}ar(X)}$$

Remark:

- The factor $n-1$ in s^2 (instead of e.g., n) has technical reasons
  We speak about the *corrected* empirical variance, while for large n this correction has no practical relevance.
- Analogy to the 'universe of randomness': mean $\leftrightarrow$ expectation

Mean and empirical standard deviation

Excursion: Analogy to the 'universe of randomness'. Reminder

Lemma: Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[|X_1|^4] < \infty$.
Set $\mu := \mathbb{E}[X_1]$, $\sigma^2 := \mathbb{V}ar(X_1)$ and $\nu^2 := \mathbb{E}[(X_1 - \mu)^4] - \sigma^4$.
Then for

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

it holds

unbiasedness:

$$\mathbb{E}[\bar{X}] = \mu \qquad \mathbb{E}[S^2] = \sigma^2 \quad (\forall n \geqslant 2) \tag{1}$$

Ideas for proofs: (1) linearity of the expectation (correction $n-1$ yields unbiasedness of S^2),

Mean and empirical standard deviation

Excursion: Analogy to the 'universe of randomness'. Reminder

<u>Lemma:</u> Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[|X_1|^4] < \infty$.
Set $\mu := \mathbb{E}[X_1]$, $\sigma^2 := \mathbb{V}ar(X_1)$ and $\nu^2 := \mathbb{E}[(X_1 - \mu)^4] - \sigma^4$.
Then for

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

it holds

unbiasedness:

$$\mathbb{E}[\bar{X}] = \mu \qquad \mathbb{E}[S^2] = \sigma^2 \qquad (\forall n \geqslant 2) \qquad (1)$$

(strong) consistency:

$$\bar{X} \xrightarrow{a.s.} \mu \qquad S^2 \xrightarrow{a.s.} \sigma^2 \qquad (n \to \infty) \qquad (2)$$

'$\xrightarrow{a.s.}$' denotes convergence with probability 1, i.e., 'almost surely'. Throughout the course we implicitly consider all random variables to derive from a single probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

<u>Ideas for proofs:</u> (1) linearity of the expectation (correction $n-1$ yields unbiasedness of S^2),
(2) Strong law of large numbers,

Mean and empirical standard deviation

Excursion: Analogy to the 'universe of randomness'. Reminder

<u>Lemma:</u> Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[|X_1|^4] < \infty$.
Set $\mu := \mathbb{E}[X_1]$, $\sigma^2 := \mathbb{V}ar(X_1)$ and $\nu^2 := \mathbb{E}[(X_1 - \mu)^4] - \sigma^4$.
Then for

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

it holds

unbiasedness:

$$\mathbb{E}[\bar{X}] = \mu \qquad \mathbb{E}[S^2] = \sigma^2 \qquad (\forall n \geqslant 2) \tag{1}$$

(strong) consistency:

$$\bar{X} \xrightarrow{a.s.} \mu \qquad S^2 \xrightarrow{a.s.} \sigma^2 \qquad (n \to \infty) \tag{2}$$

asymptotic normality:

$$\sqrt{n}[\bar{X} - \mu] \xrightarrow{d} N(0, \sigma^2) \qquad \sqrt{n}[S^2 - \sigma^2] \xrightarrow{d} N(0, \nu^2) \qquad (n \to \infty) \tag{3}$$

'$\xrightarrow{a.s.}$' denotes convergence with probability 1, i.e., 'almost surely'. Throughout the course we implicitly consider all random variables to derive from a single probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

'$\xrightarrow{d}$' denotes convergence in distribution

<u>Ideas for proofs:</u> (1) linearity of the expectation (correction $n-1$ yields unbiasedness of S^2), (2) Strong law of large numbers, (3) Central limit theorem / delta method

Notation

<u>Convention:</u>
We use *capital letters* for random variables, e.g.,

$$X_1, X_2, \ldots, X_n \qquad\qquad (\text{'random'})$$

and *lowercase letters* for data or realizations of the random variables

$$x_1, x_2, \ldots, x_n \qquad\qquad (\text{'non-random'})$$

Notation

<u>Convention:</u>
We use *capital letters* for random variables, e.g.,

$$X_1, X_2, \ldots, X_n \qquad \text{('random')}$$

and *lowercase letters* for data or realizations of the random variables

$$x_1, x_2, \ldots, x_n \qquad \text{('non-random')}$$

Outlook:

The main idea of statistical modelling:

Treat data $x_1, x_2, \ldots, x_n$ ('real world')

as realizations of random variables $X_1, X_2, \ldots, X_n$ ('universe of randomness')

Notation

<u>Convention:</u>
We use *capital letters* for random variables, e.g.,

$$X_1, X_2, \ldots, X_n \qquad\qquad (\text{'random'})$$

and *lowercase letters* for data or realizations of the random variables

$$x_1, x_2, \ldots, x_n \qquad\qquad (\text{'non-random'})$$

Outlook:

The main idea of statistical modelling:

Treat data $x_1, x_2, \ldots, x_n$ ('real world')

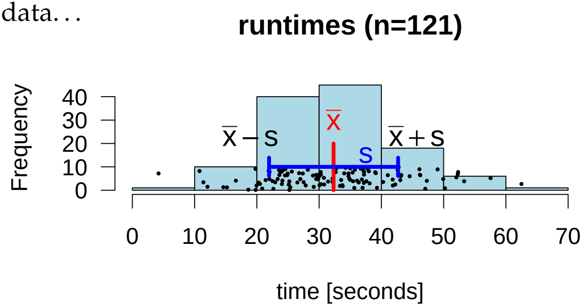
as realizations of random variables $X_1, X_2, \ldots, X_n$ ('universe of randomness')

Note that we evaluate *statistics* either on data, e.g., $\bar{x} = (1/n) \sum^n x_i$ ($\rightarrow$ non-random), or on random variables $\overline{X} = (1/n) \sum^n X_i$ ($\rightarrow$ random)

more on modeling in the following sessions

Mean and empirical standard deviation

Back to the data…

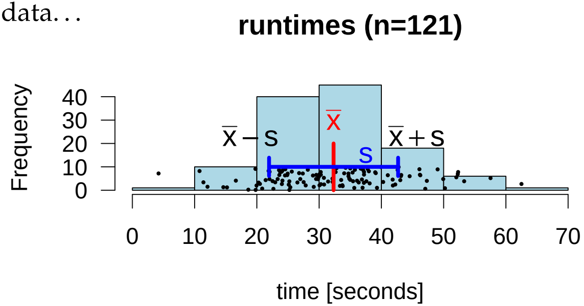


Data $x_1, x_2, \dots, x_n$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

Mean and empirical standard deviation

Back to the data...



Data $x_1, x_2, \ldots, x_n$

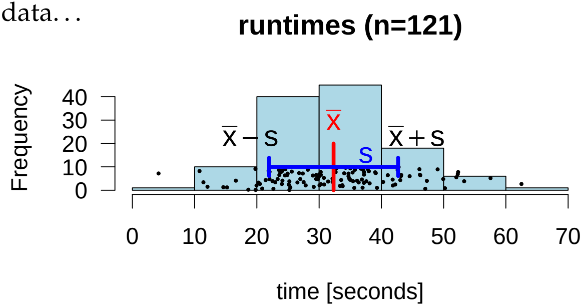
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

Evaluation

$$\bar{x} \approx 32.3 \qquad s^2 \approx 107.4 \qquad s \approx 10.4$$

Mean and empirical standard deviation

Back to the data…



Data $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

Evaluation

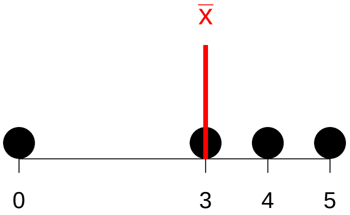
$$\bar{x} \approx 32.3 \qquad s^2 \approx 107.4 \qquad s \approx 10.4$$

in R via

`mean(x)` `var(x)` `sd(x)`

Mean and empirical standard deviation

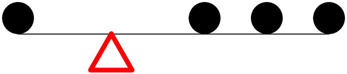
Geometrical interpretation of the mean $\bar{x}$



- Numerically: $\bar{x} = (0 + 3 + 4 + 5)/4 = 3$

Mean and empirical standard deviation

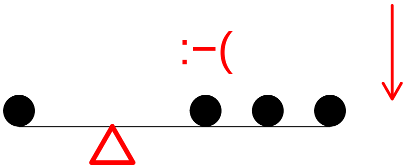
Geometrical interpretation of the mean $\bar{x}$



- Numerically: $\bar{x} = (0 + 3 + 4 + 5)/4 = 3$
- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?

Mean and empirical standard deviation

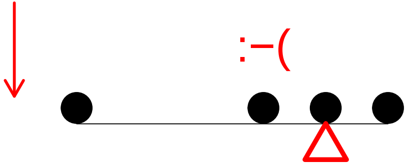
Geometrical interpretation of the mean $\bar{x}$



- Numerically: $\bar{x} = (0 + 3 + 4 + 5)/4 = 3$
- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?

Mean and empirical standard deviation

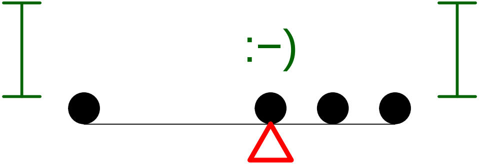
Geometrical interpretation of the mean $\bar{x}$



- Numerically: $\bar{x} = (0 + 3 + 4 + 5)/4 = 3$
- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?

Mean and empirical standard deviation

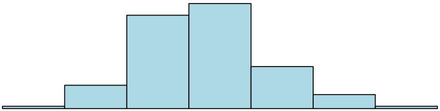
Geometrical interpretation of the mean $\bar{x}$



- Numerically: $\bar{x} = (0 + 3 + 4 + 5)/4 = 3$
- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?

Mean and empirical standard deviation

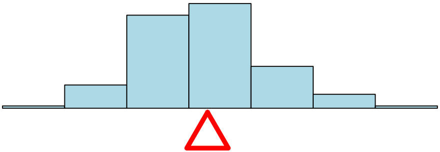
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic

Mean and empirical standard deviation

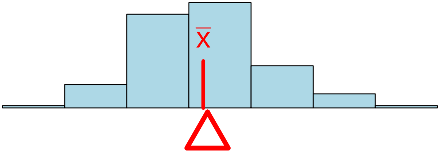
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic

Mean and empirical standard deviation

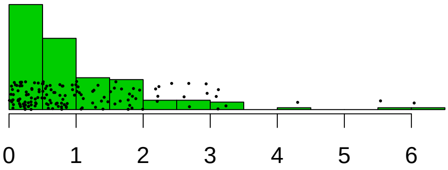
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic

Mean and empirical standard deviation

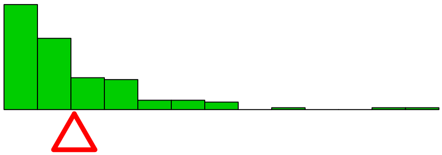
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic
  Distribution not bell-shaped but *asymmetric*
  few large values 'pull' $\bar{x}$ to the right

Mean and empirical standard deviation

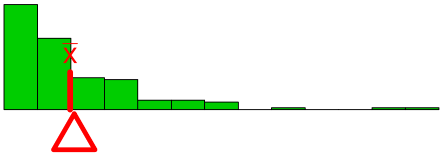
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic
  Distribution not bell-shaped but *asymmetric*
  few large values 'pull' $\bar{x}$ to the right

Mean and empirical standard deviation

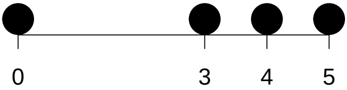
Geometrical interpretation of the mean $\bar{x}$



- Geometrically: Center of mass
  points of same mass on a balance
  Where is the center of rotation Δ, such that the balance is in equilibrium?
- Consequence: Naive estimation from graphic
  Distribution not bell-shaped but *asymmetric*
  few large values 'pull' $\bar{x}$ to the right

Mean and empirical standard deviation

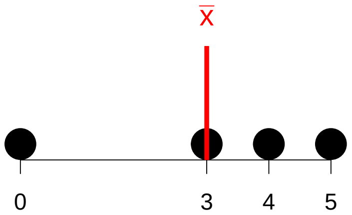
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Mean and empirical standard deviation

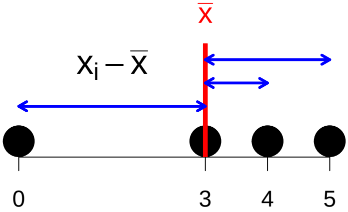
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Mean and empirical standard deviation

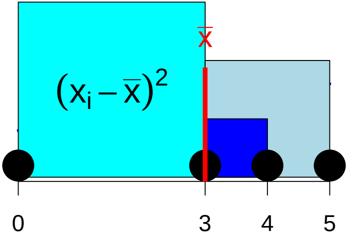
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Mean and empirical standard deviation

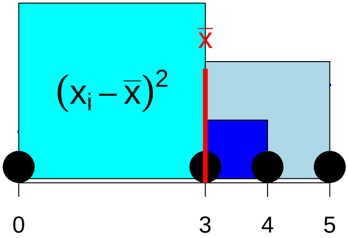
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Large deviations from the mean have a large impact (squaring)

Mean and empirical standard deviation

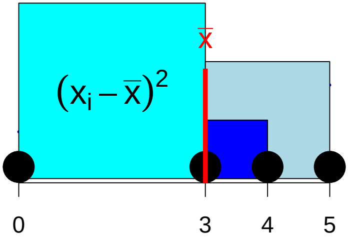
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{3}(3^2 + 0^2 + 1^2 + 2^2) = \frac{14}{3}$
- Large deviations from the mean have a large impact (squaring)

Mean and empirical standard deviation

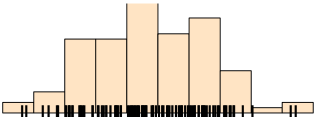
For the standard deviation s



- numerically: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{3}(3^2 + 0^2 + 1^2 + 2^2) = \frac{14}{3} \to s = \sqrt{\frac{14}{3}}$
- Large deviations from the mean have a large impact (squaring)

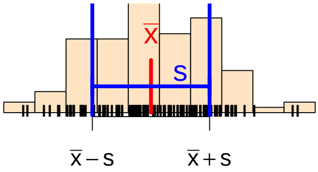
Mean and empirical standard deviation

Naive estimation of s (only for bell-shaped distributions!)



Mean and empirical standard deviation

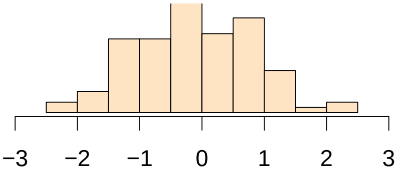
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$

Mean and empirical standard deviation

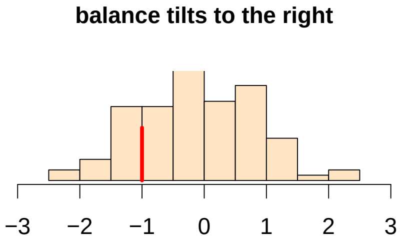
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)

Mean and empirical standard deviation

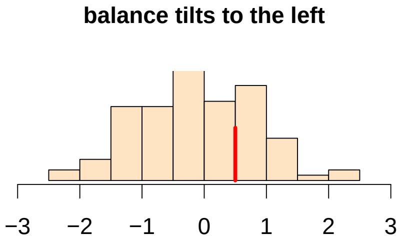
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)

Mean and empirical standard deviation

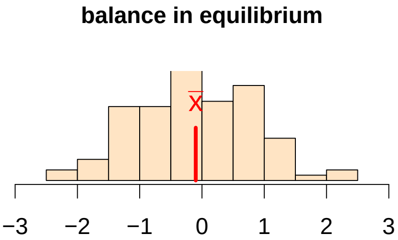
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)

Mean and empirical standard deviation

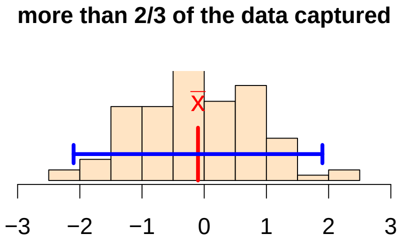
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)

Mean and empirical standard deviation

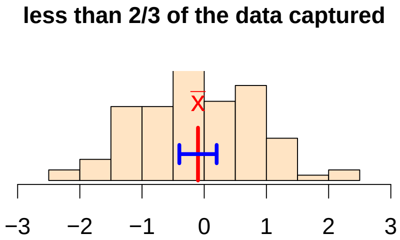
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)
  - Capture 2/3 of the data around $\bar{x}$

Mean and empirical standard deviation

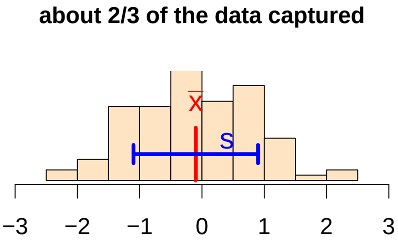
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)
  - Capture 2/3 of the data around $\bar{x}$

Mean and empirical standard deviation

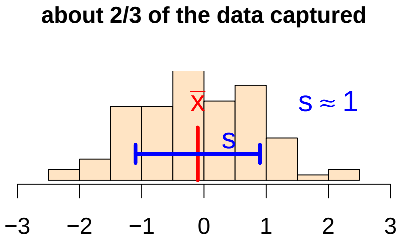
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)
  - Capture 2/3 of the data around $\bar{x}$

Mean and empirical standard deviation

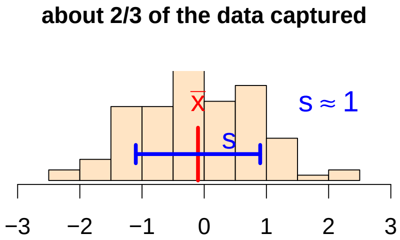
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)
  - Capture 2/3 of the data around $\bar{x}$

Mean and empirical standard deviation

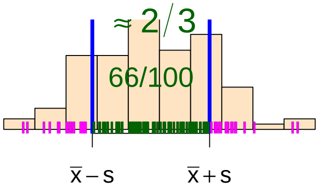
Naive estimation of s (only for bell-shaped distributions!)



- Fact: About 2/3 of the data lie in the s-neighborhood of $\bar{x}$
- Turn the tables
  - Estimate $\bar{x}$ ($\rightarrow$ balance)
  - Capture 2/3 of the data around $\bar{x}$
- Numerically: $\bar{x} \approx -0.1$ and $s \approx 0.94$

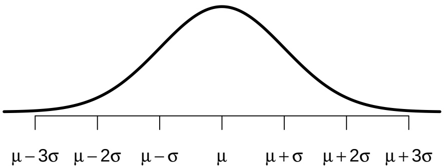
Mean and empirical standard deviation

We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



Mean and empirical standard deviation

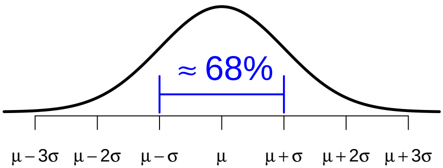
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$

Mean and empirical standard deviation

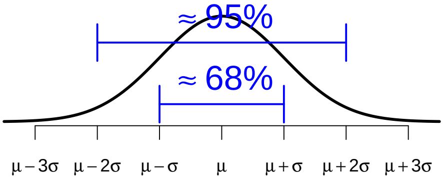
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$

Mean and empirical standard deviation

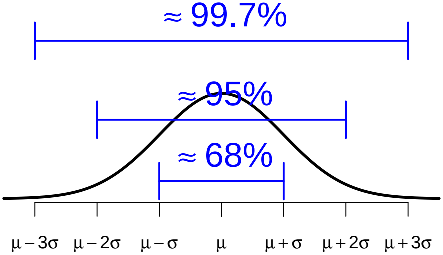
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$

Mean and empirical standard deviation

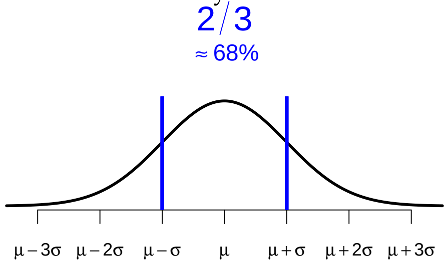
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$

Mean and empirical standard deviation

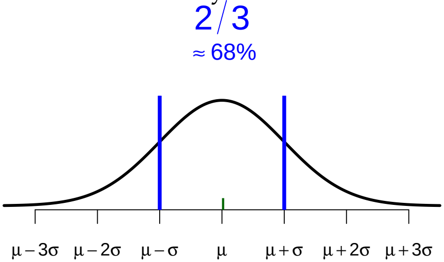
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3

Mean and empirical standard deviation

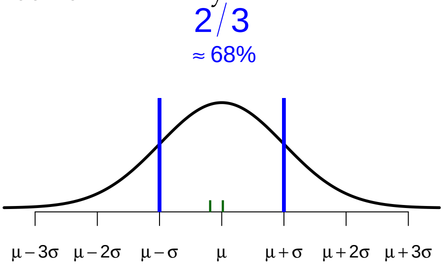
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about $2/3$
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

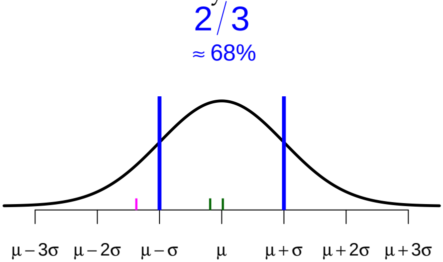
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about $2/3$
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

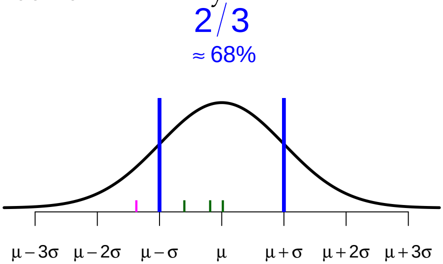
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about $2/3$
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

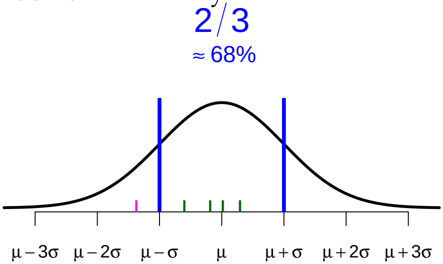
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

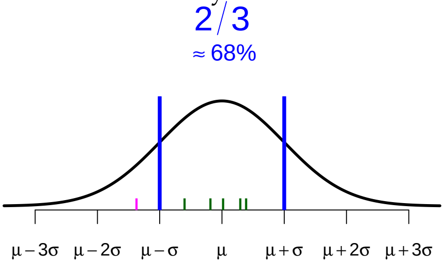
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

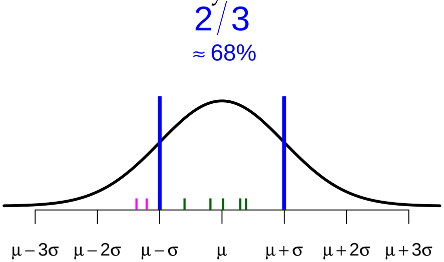
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

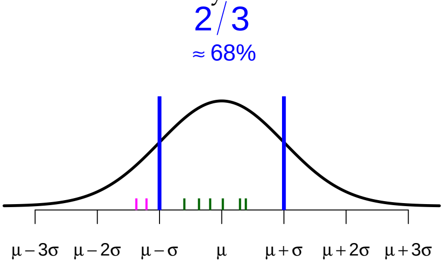
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

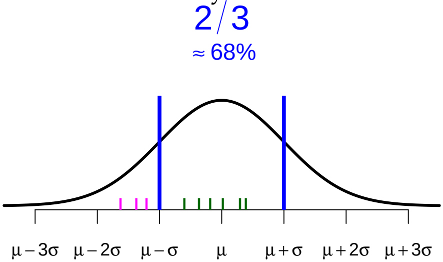
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

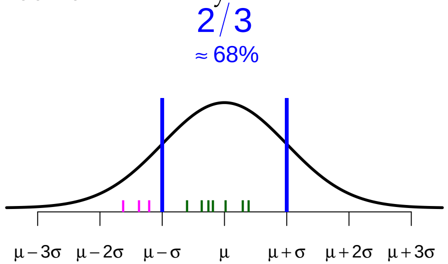
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

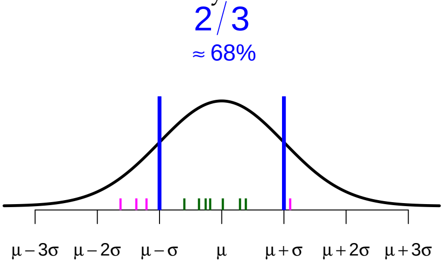
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

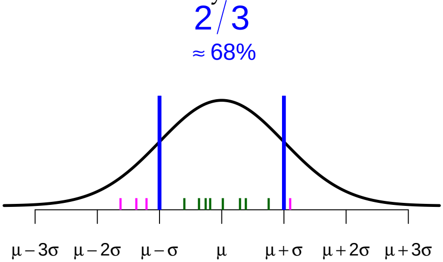
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

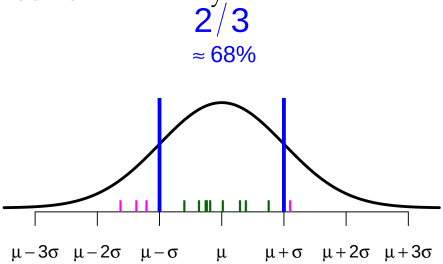
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

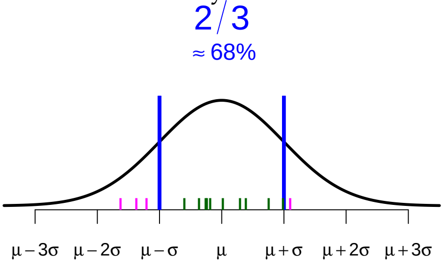
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

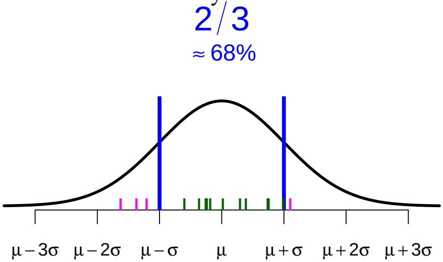
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

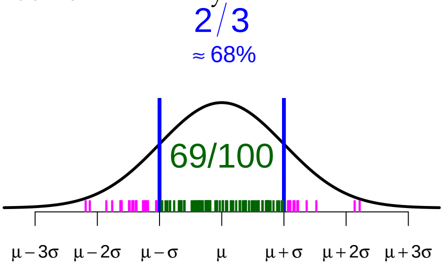
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about $2/3$
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped

Mean and empirical standard deviation

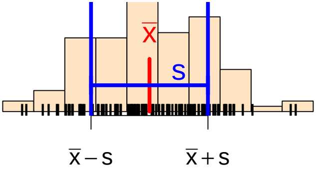
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped
  - The *proportion* within $\mu \pm \sigma$ lies close to 2/3 ($\rightarrow$ Law of large numbers)
    $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[\mu \pm \sigma]}(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbb{1}_{[\mu \pm \sigma]}(X_1)] = \mathbb{P}(X_1 \in [\mu \pm \sigma]) \approx 2/3$, as $n \rightarrow \infty$

Mean and empirical standard deviation

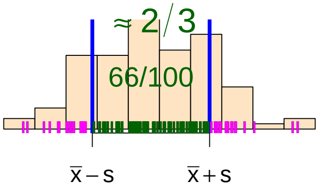
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped
  - The *proportion* within $\mu \pm \sigma$ lies close to 2/3 ($\to$ Law of large numbers)
    $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[\mu \pm \sigma]}(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbb{1}_{[\mu \pm \sigma]}(X_1)] = \mathbb{P}(X_1 \in [\mu \pm \sigma]) \approx 2/3$, as $n \to \infty$
  - $\bar{X}$ and S consistently *estimate* μ and σ ($\to$ Law of large numbers)
    $\bar{X} \xrightarrow{\text{a.s.}} \mu$ and $S \xrightarrow{\text{a.s.}} \sigma$, as $n \to \infty$

Mean and empirical standard deviation

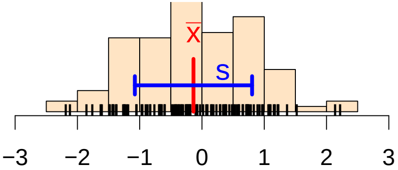
We used: For a bell-shaped distribution about 2/3 of the data lie in the s-neighborhood of $\bar{x}$. But why?



- Recall: Normal distribution $N(\mu, \sigma^2)$
  - Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68 \approx 2/3$
    X falls in the σ-neighborhood of μ with probability about 2/3
  - Consider data $n = 100$ independent copies $X_1, \ldots, X_n$ of X
    data $x_1, \ldots, x_n$ are interpreted as realizations of $X_1, \ldots, X_n$, reasonable as data is approx bell-shaped
  - The *proportion* within $\mu \pm \sigma$ lies close to 2/3 ($\rightarrow$ Law of large numbers)
    $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[\mu \pm \sigma]}(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbb{1}_{[\mu \pm \sigma]}(X_1)] = \mathbb{P}(X_1 \in [\mu \pm \sigma]) \approx 2/3$, as $n \to \infty$
  - $\bar{X}$ and S consistently *estimate* μ and σ ($\rightarrow$ Law of large numbers)
    $\bar{X} \xrightarrow{\text{a.s.}} \mu$ and $S \xrightarrow{\text{a.s.}} \sigma$, as $n \to \infty$
  - Also the *proportion* within $\bar{X} \pm S$ is close to 2/3
    $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[\bar{X} \pm S]}(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[\mathbb{1}_{[\mu \pm \sigma]}(X_1)] \approx 2/3$, as $n \to \infty$ (note that the CDF of X is continuous)

Mean and empirical standard deviation

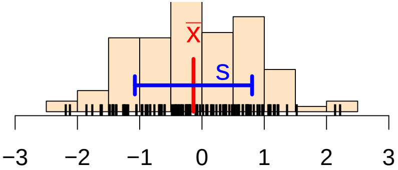
Interpretation (only for bell-shaped distributions of data)



- $\bar{x}$ is interpreted as a *typical observation*

Mean and empirical standard deviation

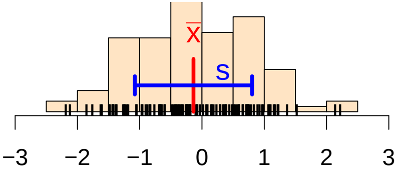
Interpretation (only for bell-shaped distributions of data)



- $\bar{x}$ is interpreted as a *typical observation*
- s is interpreted as the *typical deviation* of an observation (from the mean)

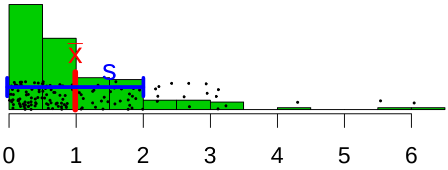
Mean and empirical standard deviation

Interpretation (only for bell-shaped distributions of data)



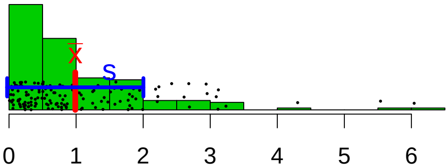
- $\bar{x}$ is interpreted as a *typical observation*
- s is interpreted as the *typical deviation* of an observation (from the mean)
- These two statistics (only two!) suitably summarize the whole set of data (many!)

Mean and empirical standard deviation



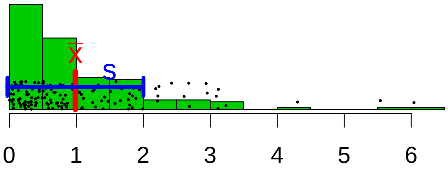
- If the data are not distributed approximately bell-shaped, then this interpretation is not useful

Mean and empirical standard deviation



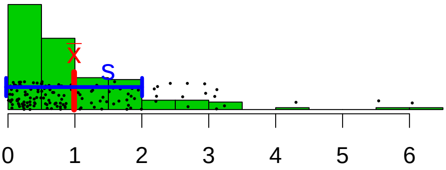
- If the data are not distributed approximately bell-shaped, then this interpretation is not useful
- Here $\bar{x}$ is not a typical observation. Much more data lie left of $\bar{x}$ than right of it

Mean and empirical standard deviation



- If the data are not distributed approximately bell-shaped, then this interpretation is not useful
- Here $\bar{x}$ is not a typical observation. Much more data lie left of $\bar{x}$ than right of it
- s does not describe the typical deviation of $\bar{x}$. Almost all of the data lie within the s-neighborhood of $\bar{x}$, only few outliers lie outside of it

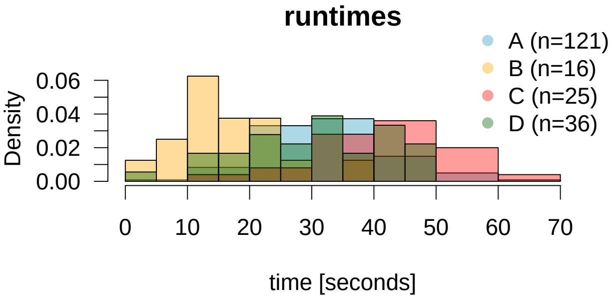
Mean and empirical standard deviation



- If the data are not distributed approximately bell-shaped, then this interpretation is not useful
- Here $\bar{x}$ is not a typical observation. Much more data lie left of $\bar{x}$ than right of it
- s does not describe the typical deviation of $\bar{x}$. Almost all of the data lie within the s-neighborhood of $\bar{x}$, only few outliers lie outside of it
- $\bar{x}$ and s should not be used for the description of the location and the dispersion of the data

Boxplot

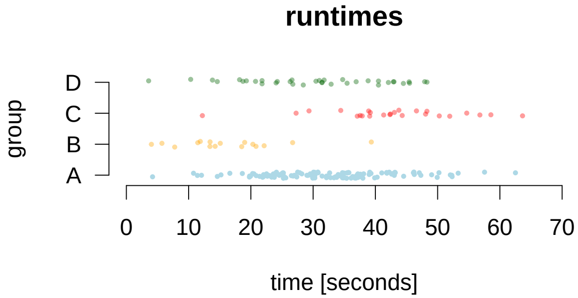
Comparison of four groups A, B, C and D



- Histograms overplotted

Boxplot

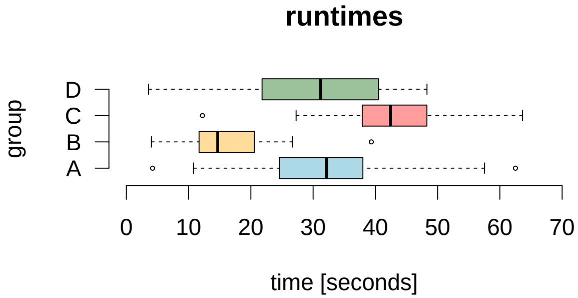
Comparison of four groups A, B, C and D



- Histograms overplotted
  Could represent the data in a stripchart

Boxplot

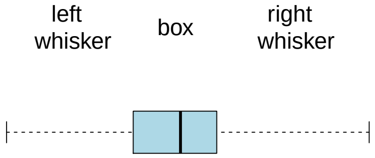
Comparison of four groups A, B, C and D



- Histograms overplotted
  Could represent the data in a stripchart
  Other possibility: the *box and whisker plot*, short boxplot

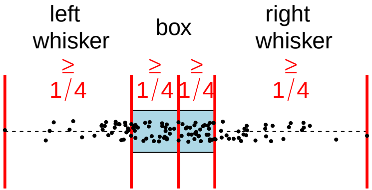
Boxplot

Boxplot



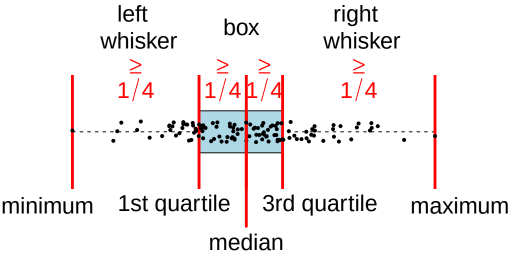
- Consists of a box and two whisker ('Schnurrhaare', meow!)

Boxplot



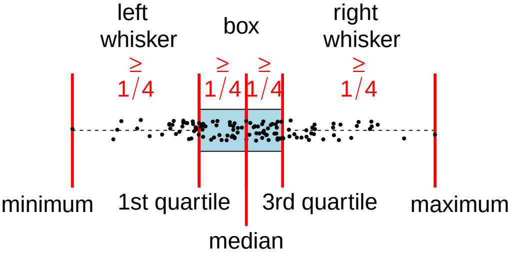
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data

Boxplot



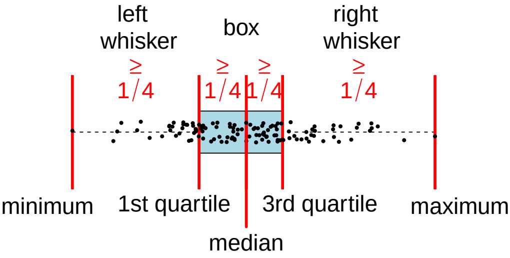
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation

Boxplot



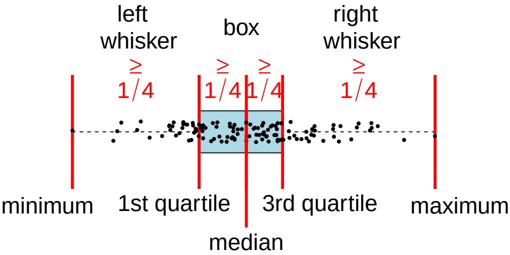
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation
  - *Median* (m), at least 50% of the data $\geqslant m$ and at least 50% are $\leqslant m$

Boxplot



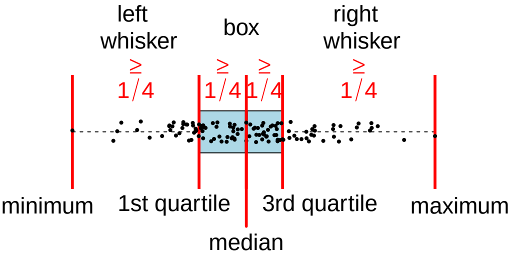
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation
  - *Median* (m), at least 50% of the data $\geqslant m$ and at least 50% are $\leqslant m$
  - *1st qua<u>r</u>tile* ($q_{1/4}$), at least 25% are $\leqslant q_{1/4}$ and at least 75% are $\geqslant q_{1/4}$

Boxplot



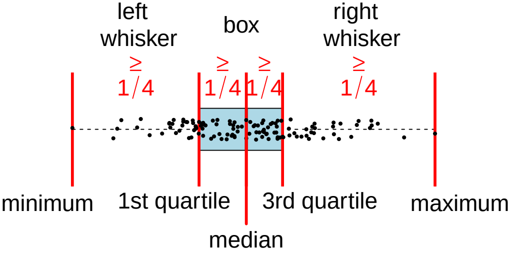
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least $1/4$ of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation
  - *Median* (m), at least 50% of the data $\geqslant m$ and at least 50% are $\leqslant m$
  - *1st quartile* ($q_{1/4}$), at least 25% are $\leqslant q_{1/4}$ and at least 75% are $\geqslant q_{1/4}$
  - *3rd quartile* ($q_{3/4}$), at least 75% are $\leqslant q_{3/4}$ and at least 25% are $\geqslant q_{3/4}$

Boxplot



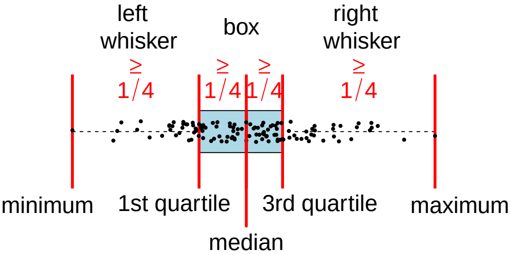
- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation
  - *Median* (m), at least 50% of the data $\geqslant m$ and at least 50% are $\leqslant m$
  - *1st quartile* ($q_{1/4}$), at least 25% are $\leqslant q_{1/4}$ and at least 75% are $\geqslant q_{1/4}$
  - *3rd quartile* ($q_{3/4}$), at least 75% are $\leqslant q_{3/4}$ and at least 25% are $\geqslant q_{3/4}$
- Interpretation:
  - Median m is a measure for the location of the observations ($\rightarrow$ where?)

Boxplot



- Consists of a box and two whisker ('Schnurrhaare', meow!)
- Four sections, contain at least 1/4 of the data
- $\rightarrow$ five statistics:
  - *Minimum*, smallest observation
  - *Maximum*, largest observation
  - *Median* (m), at least 50% of the data $\geqslant m$ and at least 50% are $\leqslant m$
  - *1st quartile* ($q_{1/4}$), at least 25% are $\leqslant q_{1/4}$ and at least 75% are $\geqslant q_{1/4}$
  - *3rd quartile* ($q_{3/4}$), at least 75% are $\leqslant q_{3/4}$ and at least 25% are $\geqslant q_{3/4}$
- Interpretation:
  - Median m is a measure for the location of the observations ($\rightarrow$ where?)
  - Interquartile range $q_{3/4} - q_{1/4}$ (width of the box) is a measure for the dispersion of the data ($\rightarrow$ how variable?)

Boxplot



left whisker
box
right whisker
≥ 1/4
≥ 1/4
≥ 1/4
≥ 1/4
minimum
1st quartile
3rd quartile
maximum
median

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0, 1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-quan\underline{n}tile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1 - p$.

Empirical quantile (general)

- <u>Definition:</u> Given n data $x_1, \dots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-qua\underline{n}tile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

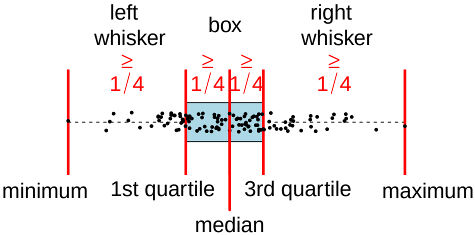
$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) *$p-$quantile*, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty, q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p, \infty)}(x_i) \geqslant 1-p$$

- We already know three prominant candidates (with their own name):
  a median is a 50%-quantile ($p = 1/2$)

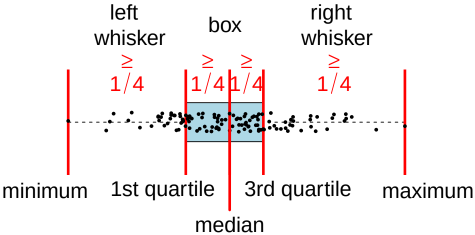


Empirical quantile (general)

- <u>Definition:</u> Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-$*quan<u>n</u>tile*, if
i. the proportion of the data that are smaller or equal q_p is at least p and
ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

- We already know three prominant candidates (with their own name):
a median is a 50%-qua<u>n</u>tile ($p = 1/2$)
a 1st qua<u>r</u>tile is a 25%-qua<u>n</u>tile ($p = 1/4$)

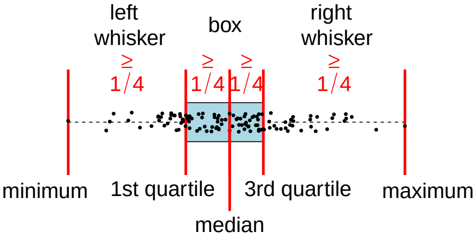


Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-$*quantile*, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

- We already know three prominant candidates (with their own name):
  a median is a 50%-quantile ($p = 1/2$)
  a 1st quartile is a 25%-quantile ($p = 1/4$)
  a 3rd quartile is a 75%-quantile ($p = 3/4$)



Empirical quantile (general)

- Definition: Given n data $x_1, \dots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p - quan\underline{n}tile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty, q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p, \infty)}(x_i) \geqslant 1-p$$

- Example: Four observations $x = (1,2,3,4)^t$

superscript t denotes the transpose

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-quan\underline{n}tile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

- Example: Four observations $x = (1,2,3,4)^t$ superscript t denotes the transpose
  - Many medians: Every number in the interval $[2,3]$ is a median

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-quantile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty, q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p, \infty)}(x_i) \geqslant 1-p$$

- Example: Four observations $x = (1,2,3,4)^t$ — superscript t denotes the transpose
  - Many medians: Every number in the interval $[2,3]$ is a median
  - Often: Define the *unique* median as the mean value of the bounds, here 2.5

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-quan\underline{n}tile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

- Example: Four observations $x = (1,2,3,4)^t$ superscript t denotes the transpose
  - Many medians: Every number in the interval $[2,3]$ is a median
  - Often: Define the *unique* median as the mean value of the bounds, here 2.5
  - Analog: Every number in $[1,2]$ is 1/4-quantile, the unique quartile is 1.5

Empirical quantile (general)

- Definition: Given n data $x_1, \ldots, x_n$. Let $p \in (0,1)$. A number $q_p \in \mathbb{R}$ is called an (empirical) $p-quantile$, if
  i. the proportion of the data that are smaller or equal q_p is at least p and
  ii. the proportion of the data that are larger or equal q_p is at least $1-p$.
  In formulas:

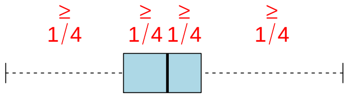
$$i.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,q_p]}(x_i) \geqslant p \qquad \text{and} \qquad ii.: \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[q_p,\infty)}(x_i) \geqslant 1-p$$

- Example: Four observations $x = (1,2,3,4)^t$ superscript t denotes the transpose
  - Many medians: Every number in the interval $[2,3]$ is a median
  - Often: Define the *unique* median as the mean value of the bounds, here 2.5
  - Analog: Every number in $[1,2]$ is 1/4-quantile, the unique quartile is 1.5
  - Many quantiles equal: The number 2 is a p-quantile for every p of $[0.25, 0.5]$

Empirical quantile (general)

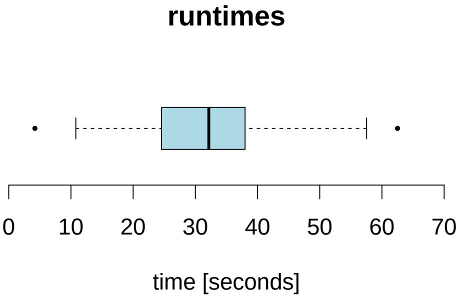
- Example: Four observations $x = (1, 2, 3, 4)^t$ superscript t denotes the transpose
  - Many medians: Every number in the interval $[2, 3]$ is a median
  - Often: Define the *unique* median as the mean value of the bounds, here 2.5
  - Analog: Every number in $[1, 2]$ is 1/4-quantile, the unique quartile is 1.5
  - Many quantiles equal: The number 2 is a p-quantile for every p of $[0.25, 0.5]$
- Remark.: These kind of 'exotic' messages may support the understanding of the definition of a quantile. The main message however is, that the boxplot appropriately summarizes many data using only five simple statistics

  Take home: Many data $\rightarrow$ at first sight: "1/4, 1/4, 1/4, 1/4"



Boxplot in R

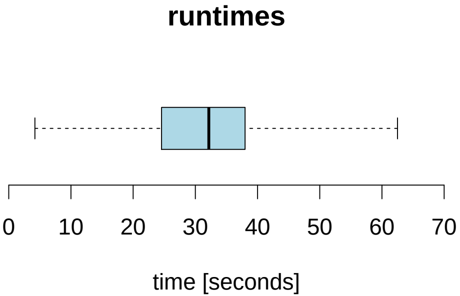
```
#Boxplot, horizontal representation
boxplot(x,horizontal=TRUE,...)
```



Attention: per default a whisker ranges to the observartion which is most far away from the box, but does not exceed 1.5 times the interquartile range. Extreme values ('outliers') are plotted seperately.

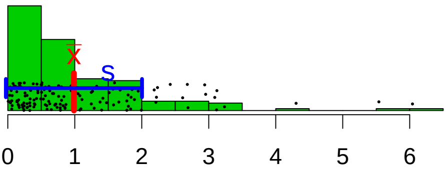
Boxplot in R

```
#Boxplot, Whisker range to extreme values
boxplot(x,horizontal=TRUE,range=0,...)
```



Through the argument `range=0` the whiskers are extended to the extreme values

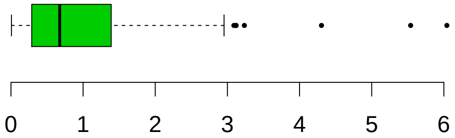
Boxplot



Reminder

- due to the asymmetric distribution of the data, $\bar{x}$ and s should not be used for the description of the location and the dispersion

Boxplot



Reminder

- due to the asymmetric distribution of the data, $\bar{x}$ and s should not be used for the description of the location and the dispersion
- The five statistics of the boxplot are more appropriate for the description of the data

Most important message today

3

Most important message today

2

Most important message today

1

Most important message today

Always graphically visualize your data first

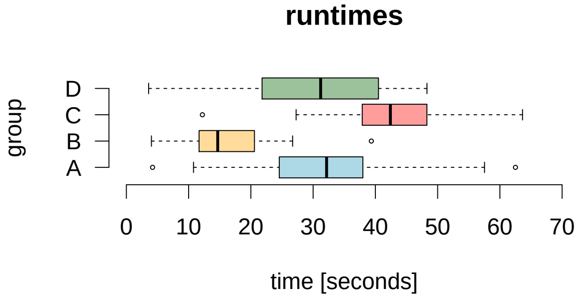
Most important message today

Always graphically visualize your data first

(and start computing afterwards)

Questions

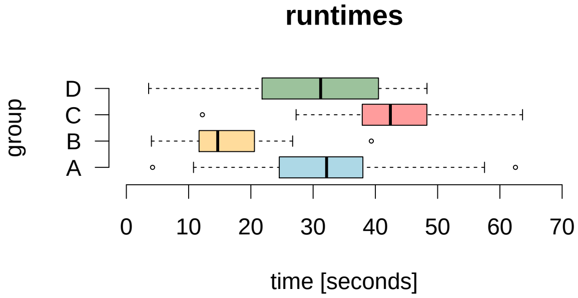
Comparison of four groups A, B, C und D



- The slowest runtime in C was about?

Questions

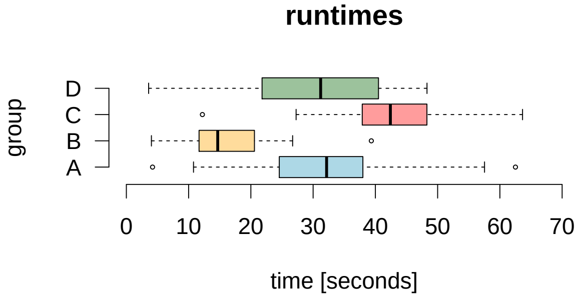
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65

Questions

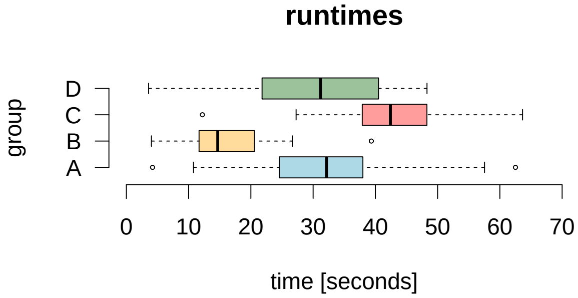
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about?

Questions

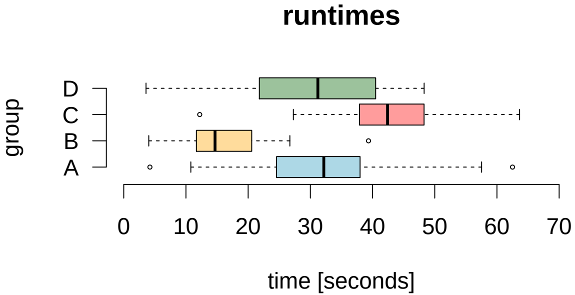
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5

Questions

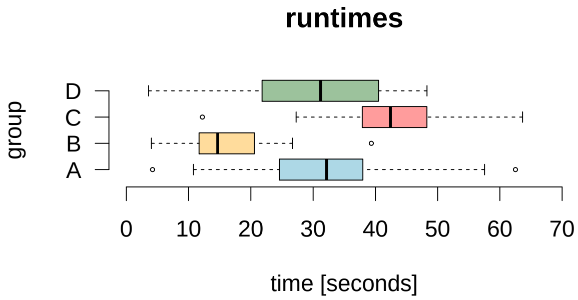
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about?

Questions

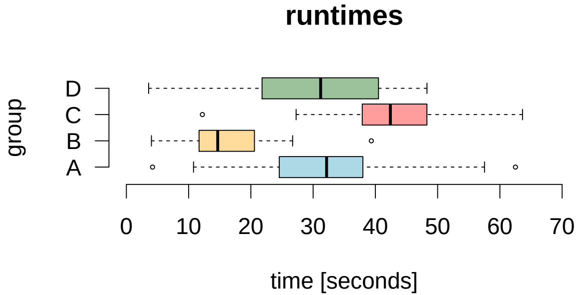
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30

Questions

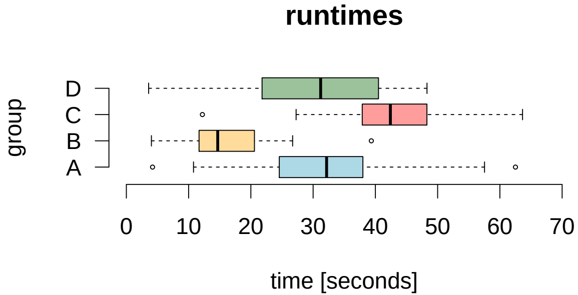
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20?

Questions

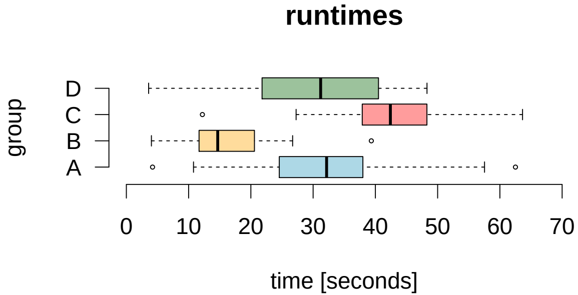
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20? about 75%

Questions

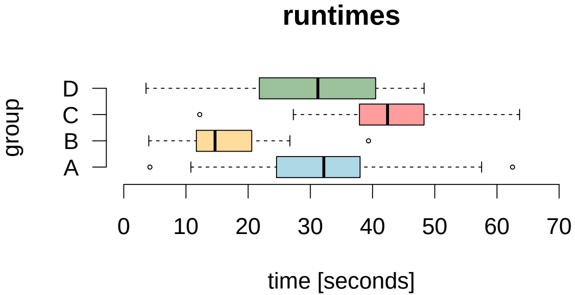
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20? about 75%
- Were 50% of the runtimes in A faster than 75% of the times in C?

Questions

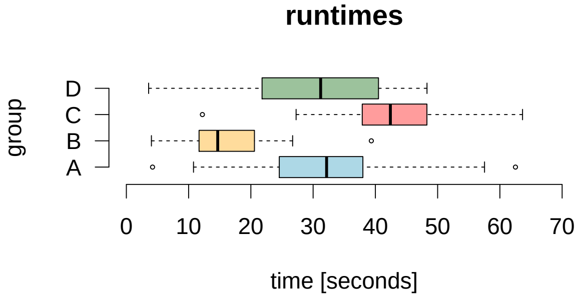
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20? about 75%
- Were 50% of the runtimes in A faster than 75% of the times in C? yes

Questions

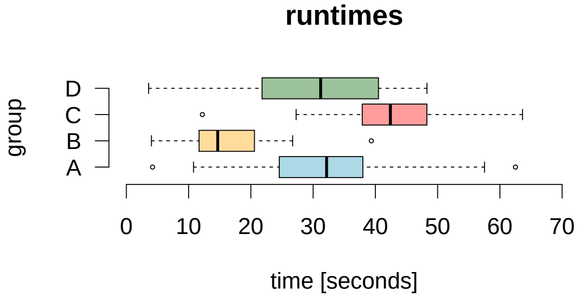
Comparison of four groups A, B, C und D



- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20? about 75%
- Were 50% of the runtimes in A faster than 75% of the times in C? yes
- In group B, apart from a single runtime all others were faster than half of those of group A, half of those of C and half of those of D.

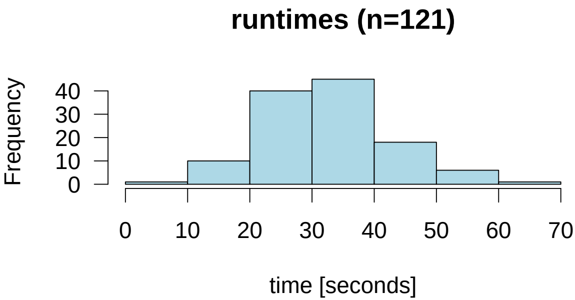
Questions

Comparison of four groups A, B, C und D



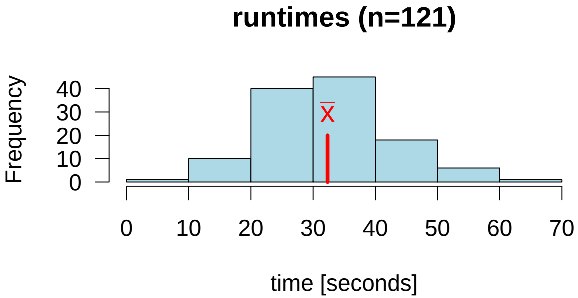
- The slowest runtime in C was about? 65
- The fastest runtime in A is about? 5
- The median runtime in D is about? 30
- What is the percentage of runtimes in group B that are smaller than 20? about 75%
- Were 50% of the runtimes in A faster than 75% of the times in C? yes
- In group B, apart from a single runtime all others were faster than half of those of group A, half of those of C and half of those of D. Correct

Questions



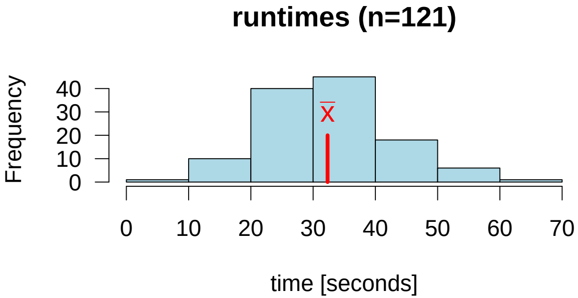
- What is the mean runtime?

Questions



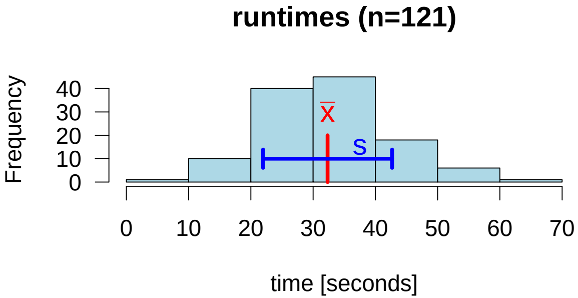
- What is the mean runtime? about 32

Questions



- What is the mean runtime? about 32
- The standard deviation of the runtimes is about?

Questions



- What is the mean runtime? about 32
- The standard deviation of the runtimes is about? 10

Thank you!