Week 1: Introduction & Numerical EDA

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-01-06

Course Overview

Tutorials

Quercus

Grading Scheme

Tutorials

Learning Activities

Week 2

Week 4

Week 6

Week 9

Week 11

Week 12

Quizzes

Week 3

Week 5

Week 10

Week 13

Tutorials

Learning activities

Grouped in pairs from the same tutorial section

Computer-based activities on Crowdmark

Require one laptop or tablet with internet connection per group

***Open Book**: you can refer to your notes, lecture slides, and course Quercus site*

Quizzes

Individual

Closed book

You may use a non-programmable electronic calculator

You must bring your TCard or a valid photo ID

The quizzes will be based on lectures and weekly R modules

Grading Scheme

- 1% Syllabus Scavenger Hunt
- 12% Tutorial Learning Activities
- 27% Tutorial Quizzes (Best 3 out of 4)
- 20% Midterm (Feb 27th, 3pm-5pm)
- 40% Final Exam

What happens if I miss an assessment?

Reference materials

- Dekking et al. A modern introduction to probability and statistics: Understanding why and how (2005, First Edition)
- Devore et al. Modern mathematical statistics with applications (2021, Third Edition)
- Johnson et al. Bayes Rules! An introduction to applied Bayesian modeling (2021)
- McElreath. Statistical rethinking: A Bayesian course with examples in R and Stan (2020, Second Edition)
- Gibbs and Stringer. STA238 Supplementary material (2021)

Quercus will be used to post information about the course

- Lecture/Tutorial materials
- Announcements and Updates

Other resources

Course Syllabus

Office Hours

Discussion Boards

Request an accommodation

Request a regrading

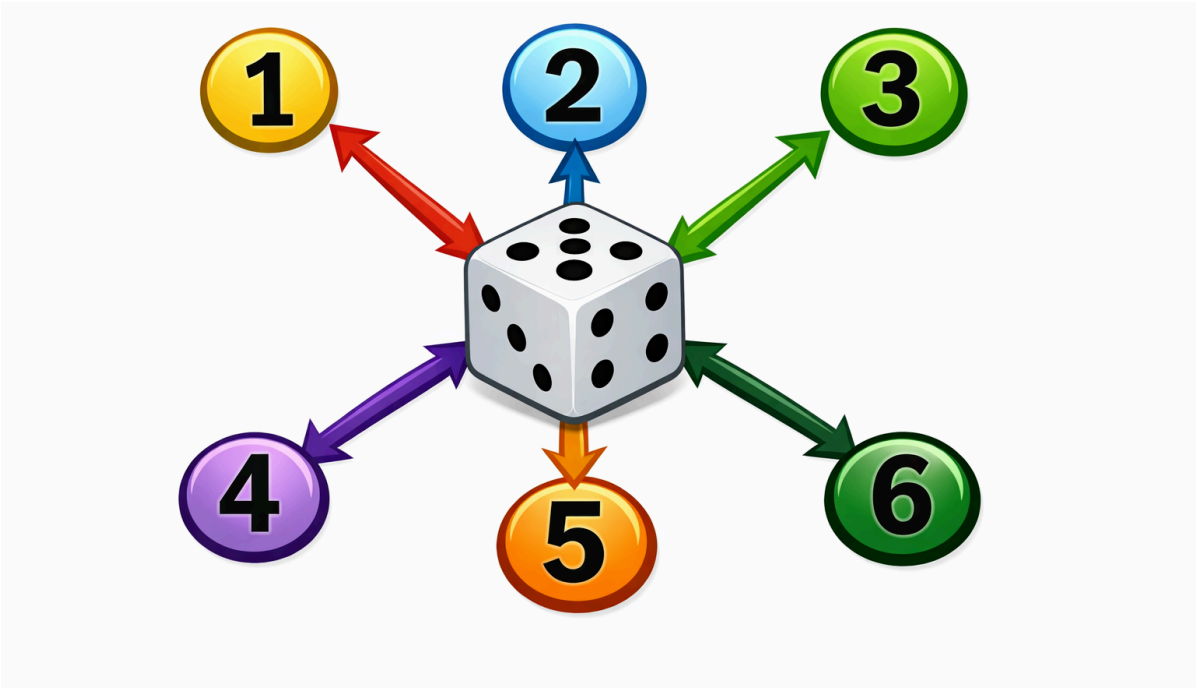
Sample Study Schedule

Mental Health Resources

Data and Random Variables

Recap: Random variables

> Given a sample space Ω, a random variable X is a measurable function onto $\mathbb{R}$
>
> $$X : \Omega \to \mathbb{R}.$$



Example: Streetcar failure time

Example: Streetcar failure time

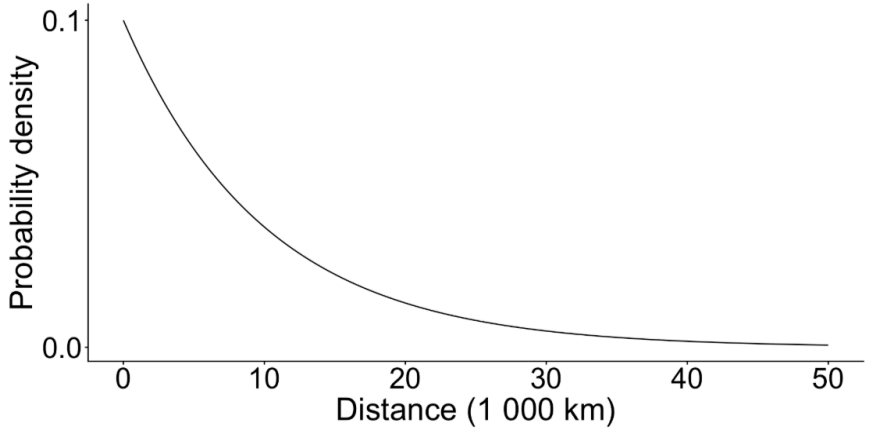
Suppose each TTC streetcar fails every 10 000 km on average and the failures occur according to a Poisson process.

Then, the time between failures (interarrival times) for a TTC streetcar follows _____.

Example: Streetcar failure time

Suppose each TTC streetcar fails every 10 000 km on average and the failures occur according to a Poisson process.

Then, the time between failures (interarrival times) for a TTC streetcar follows _____.



What we see in real life:

What we see in real life:

Photo by Secondarywaltz - Michael Moon.

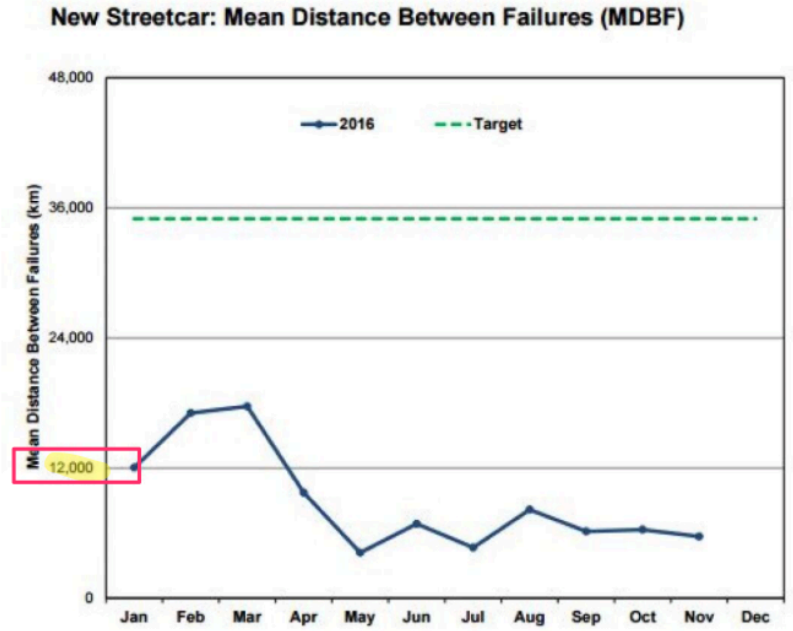
What we see in real life:

Photo by Secondarywaltz - Michael Moon.

According to the report, "failures" were reported every 5,696 kilometres that they in service November 2016. The target distance between "failures" for the new streetcars is 35,000 kilometres.

Mechanical problems include doors that didn't open properly, brake issues of some kind, and intercom systems that didn't work.

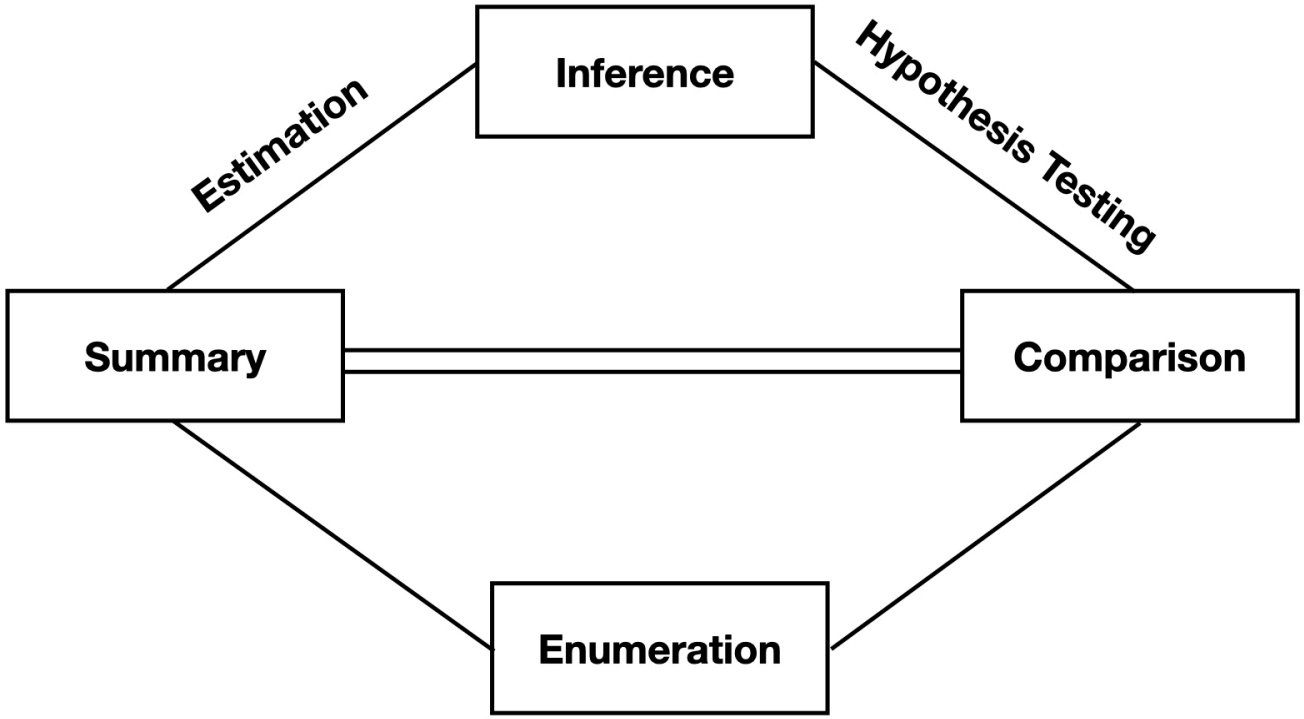
The problems are plotted in a graph entitled, "Mean Distance Between Failures," which means the distance travelled before a mechanical problem occurs and is fixed.



TTC's staff report for January 2017 contains this graph that shows 'mean distance between failures' for its new streetcars. That means the distance travelled before the new streetcars on the road experience mechanical problems. (TTC Chief Executive Officer's Report January 2017)

From a CBC news article by Muriel Draaisma.

Statistics



Four basic operations of statistics from Efron's *Maximum Likelihood and Decision Theory* (1982).

Enumeration (Obtaining Samples)

How is our sample $X_1, X_2, \ldots, X_n$ *obtained?* (STA304)

Consider the following cases:

Enumeration (Obtaining Samples)

How is our sample $X_1, X_2, \ldots, X_n$ *obtained?* (STA304)

Consider the following cases:

- Patient's sugar level taken every week for n weeks

Enumeration (Obtaining Samples)

How is our sample $X_1, X_2, \ldots, X_n$ *obtained?* (STA304)

Consider the following cases:

- Patient's sugar level taken every week for n weeks
- Price of oil over n consecutive days

Enumeration (Obtaining Samples)

How is our sample $X_1, X_2, \ldots, X_n$ *obtained?* (STA304)

Consider the following cases:

- Patient's sugar level taken every week for n weeks
- Price of oil over n consecutive days
- LDL Cholesterol levels of students in this classroom

(A) Statistic

"Assume we have an (**iid**) sample $X_1, X_2, \ldots, X_n$"

(A) Statistic

"Assume we have an (**iid**) sample $X_1, X_2, \ldots, X_n$"

Often denoted as $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f$

(A) Statistic

"Assume we have an (**iid**) sample $X_1, X_2, \ldots, X_n$"

Often denoted as $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f$

A statistic is a measurable function of our sample $t = h(X_1, \ldots, X_n)$.

(A) Statistic

"Assume we have an (**iid**) sample $X_1, X_2, \ldots, X_n$"

Often denoted as $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f$

A statistic is a measurable function of our sample $t = h(X_1, \ldots, X_n)$.

- Summarize a sample,
- Estimate the value of a parameter,
- Evaluate a hypothesis.

Statistical Process

1. Formulate a research question
2. Design your experiment or study
3. Collect your data
4. Exploratory data analysis
5. Data analysis & inference
6. Draw conclusions

Statistical Process

1. Formulate a research question
2. Design your experiment or study
3. Collect your data
4. **Exploratory data analysis**
5. Data analysis & inference
6. Draw conclusions

Exploratory Data Analysis

Its basic tools include:

- Numerical summaries
- Graphical summaries

Exploratory Data Analysis

Its basic tools include:

- Numerical summaries
- Graphical summaries

These are calculated from a fixed dataset assumed to be the realization of random variables: $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

Exploratory Data Analysis

Its basic tools include:

- Numerical summaries
- Graphical summaries

These are calculated from a fixed dataset assumed to be the realization of random variables: $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

Notation: while $X_1, X_2, \ldots, X_n$ are random variables, their *realizations* $x_1, x_2, \ldots, x_n \in \mathrm{R}$ are real numbers.

Example: Streetcar working distance

Example: Streetcar working distance

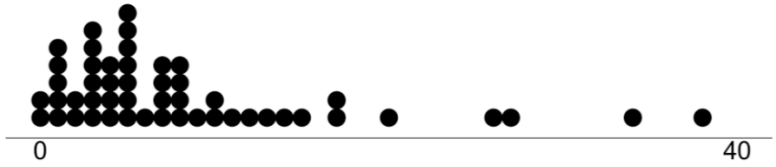
Example: Distances travelled by TTC streetcars between mechanical failures (1 000 km)

3	3	7	5	34
0	1	15	1	8
8	1	5	10	9
5	5	0	17	14
8	4	17	5	12
13	7	3	27	2
4	26	5	3	6
4	5	3	10	4
1	1	3	7	20
8	38	2	11	7

Example: Streetcar working distance

Example: Distances travelled by TTC streetcars between mechanical failures (1 000 km)

3	3	7	5	34
0	1	15	1	8
8	1	5	10	9
5	5	0	17	14
8	4	17	5	12
13	7	3	27	2
4	26	5	3	6
4	5	3	10	4
1	1	3	7	20
8	38	2	11	7



A type of *histogram*, to be seen later in the course.

STA313: Data Visualization.

Summarising data: *Measures of central tendency*

What does a typical value in our data look like?

Summarising data: *Measures of central tendency*

What does a typical value in our data look like?

Summarising data: *Measures of central tendency*

What does a typical value in our data look like?

- Sample mode

Summarising data: *Measures of central tendency*

What does a typical value in our data look like?

- Sample mode
- Sample median

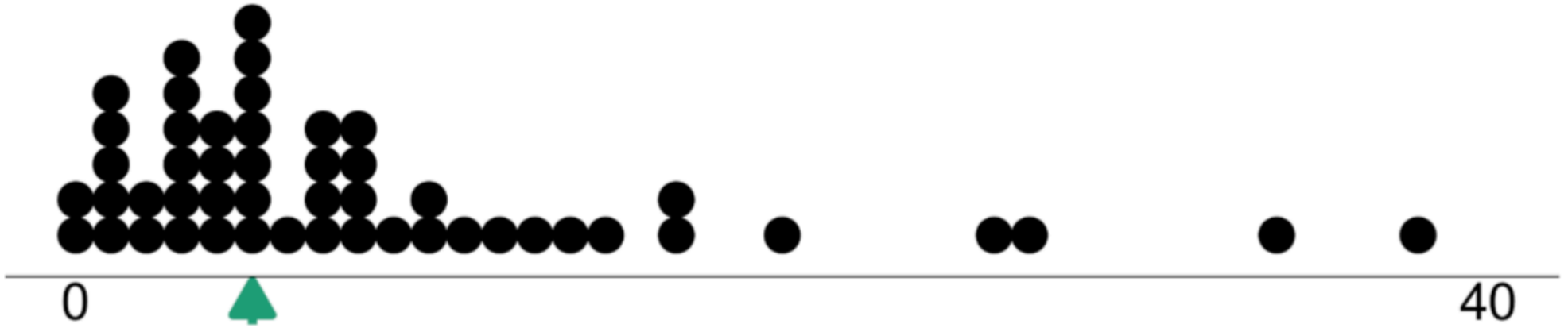
Summarising data: *Measures of central tendency*

What does a typical value in our data look like?

- Sample mode
- Sample median
- Sample mean

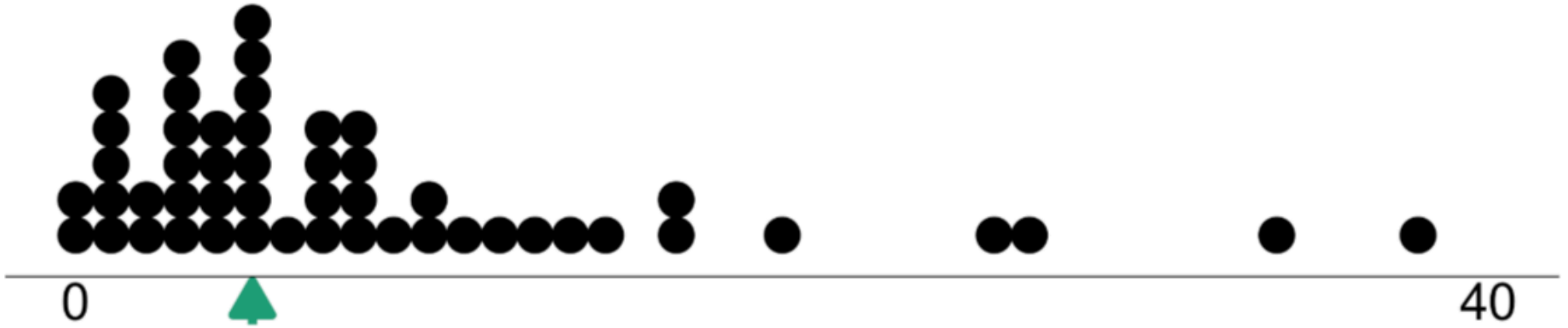
Sample mode

> The sample mode is the most frequently occurring data point in the dataset.



Sample mode

> The sample mode is the most frequently occurring data point in the dataset.



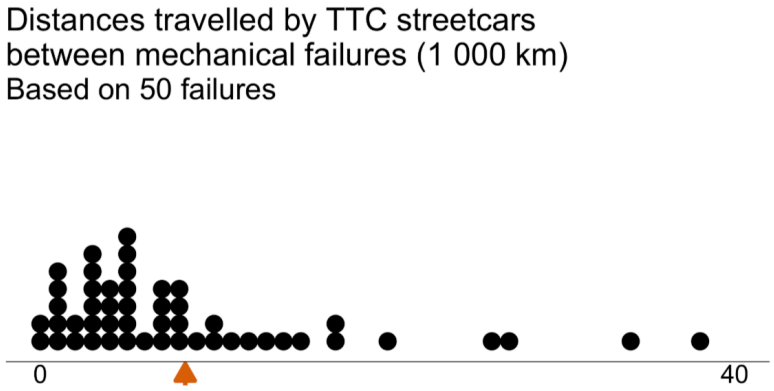

```
X
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 17 20 26 27 34 38
 2  5  2  6  4  7  1  4  4  1  2  1  1  1  1  1  2  1  1  1  1  1
```

```
[1] "5"
```

Is it likely that this will be close to the mode of the population distribution?

Sample mean

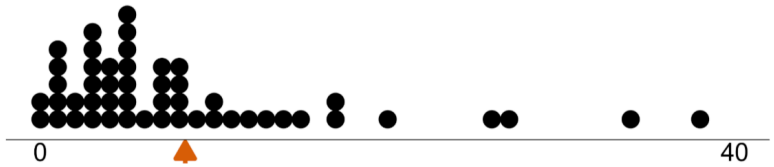
Sample mean




```
[1] 8.34
```

Sample mean

Distances travelled by TTC streetcars
between mechanical failures (1 000 km)
Based on 50 failures



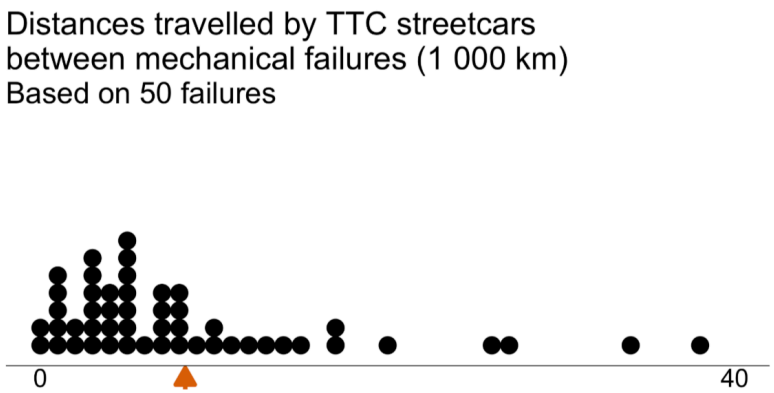

```
[1] 8.34
```

Sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

"Center of mass of the data".

Sample mean




```
[1] 8.34
```

Sample mean

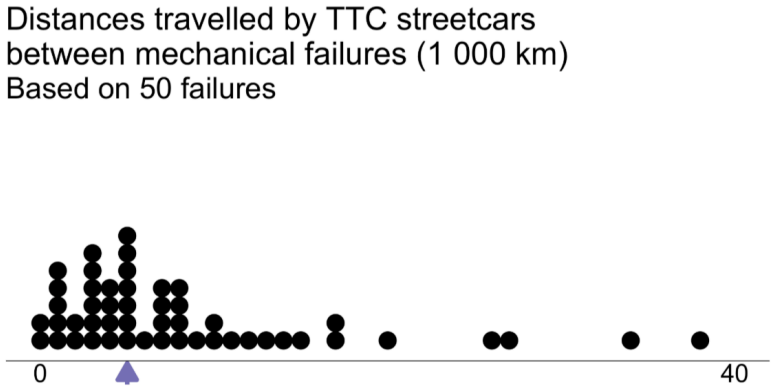
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

"Center of mass of the data".

In what sense is this value representative of the typical datapoint?

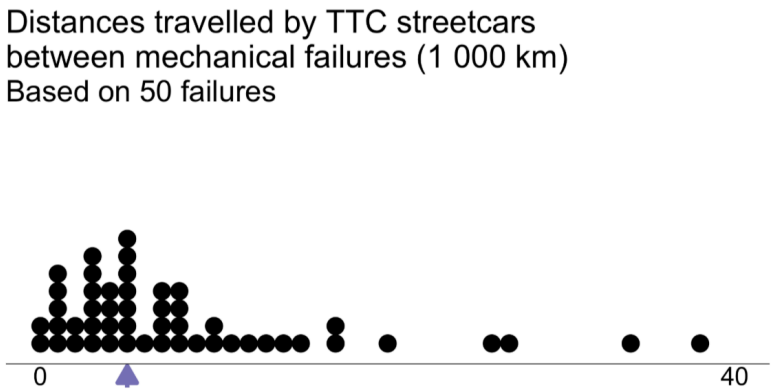
Sample median

Sample median




```
[1] 5
```

Sample median




```
[1] 5
```

For sorted data

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

the sample median is

$$\text{Med}_n(x_1, \ldots, x_n) = x_{(n-1/2)}$$

if n is odd and

$$\frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$$

if n is even.

Examples

Examples

Find the sample mean and sample median of the following sets:

- $\{5, 10, 8, 7\}$
- $\{7, 8, 8, 1\}$
- $\{9, 1, 2, 6, 8, 7\}$

Examples

Find the sample mean and sample median of the following sets:

- $\{5, 10, 8, 7\}$
- $\{7, 8, 8, 1\}$
- $\{9, 1, 2, 6, 8, 7\}$

```r
ex1 <- c(5, 10, 8, 7)
mean(ex1)
```

```
[1] 7.5
```

```r
median(ex1)
```

```
[1] 7.5
```

```r
ex2 <- c(7, 8, 8, 1)
mean(ex2)
```

```
[1] 6
```

```r
median(ex2)
```

```
[1] 7.5
```

```r
ex3 <- c(9, 1, 2, 6, 8, 7)
mean(ex3)
```

```
[1] 5.5
```

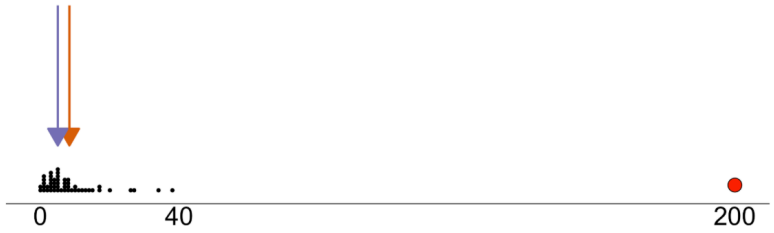
```r
median(ex3)
```

```
[1] 6.5
```

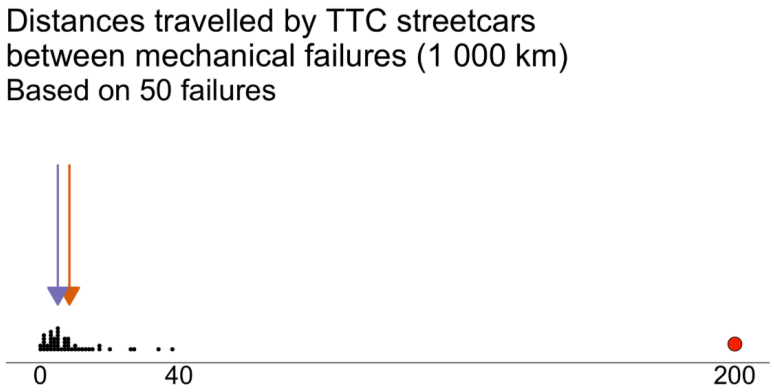
Example: Streetcar working distance, X

Example: Streetcar working distance, X

Distances travelled by TTC streetcars
between mechanical failures (1 000 km)
Based on 50 failures



Example: Streetcar working distance, X



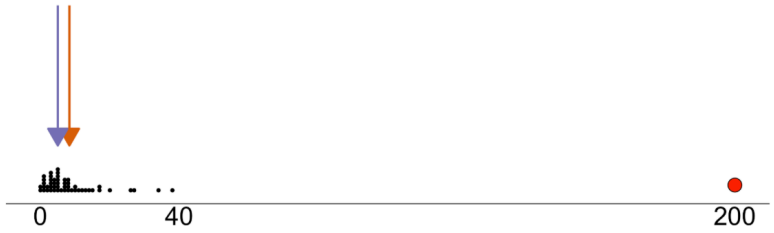
Suppose you collect data from another vehicle and it travelled for 200 000 km before failing.

How does that change your opinion about the “typical” lifetime of a vehicle?

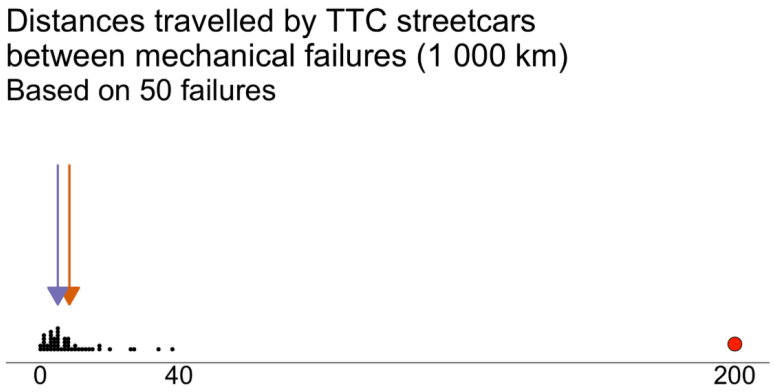
Example:

Example:

Distances travelled by TTC streetcars
between mechanical failures (1 000 km)
Based on 50 failures



Example:




```r
newX <- c(X, 200) # newX contains the updated data set
mean(X)
```

```
[1] 8.34
```

```r
mean(newX)
```

```
[1] 12.09804
```

```r
median(X)
```

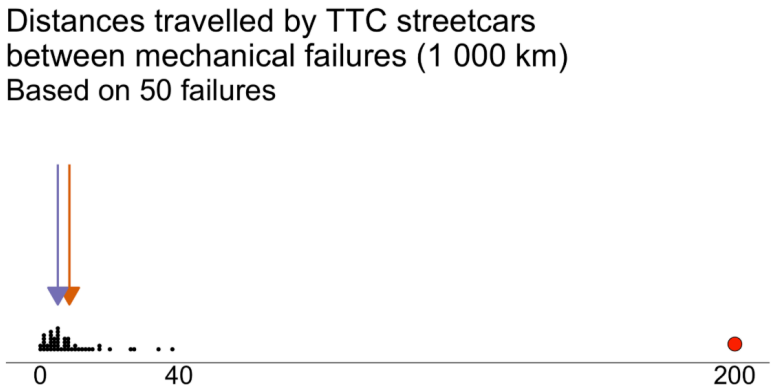
```
[1] 5
```

```r
median(newX)
```

```
[1] 5
```

Key concepts:

Example:




```r
newX <- c(X, 200) # newX contains the updated data set
mean(X)
```

```
[1] 8.34
```

```r
mean(newX)
```

```
[1] 12.09804
```

```r
median(X)
```

```
[1] 5
```

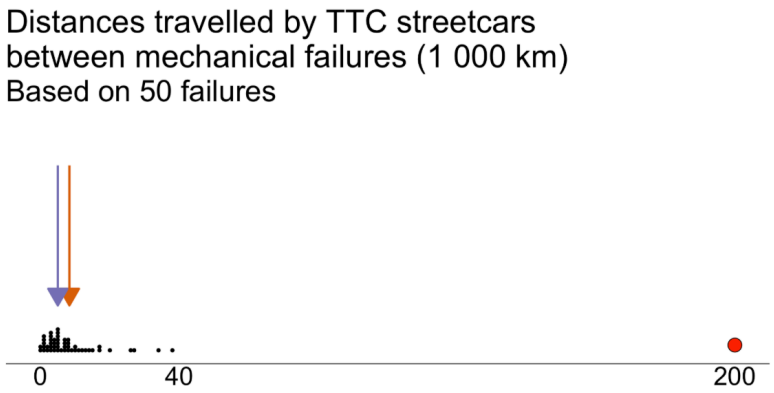
```r
median(newX)
```

```
[1] 5
```

Key concepts:

- **Robustness**

Example:




```r
newX <- c(X, 200) # newX contains the updated data set
mean(X)
```

```
[1] 8.34
```

```r
mean(newX)
```

```
[1] 12.09804
```

```r
median(X)
```

```
[1] 5
```

```r
median(newX)
```

```
[1] 5
```

Key concepts:

- **Robustness**
- **Outliers**

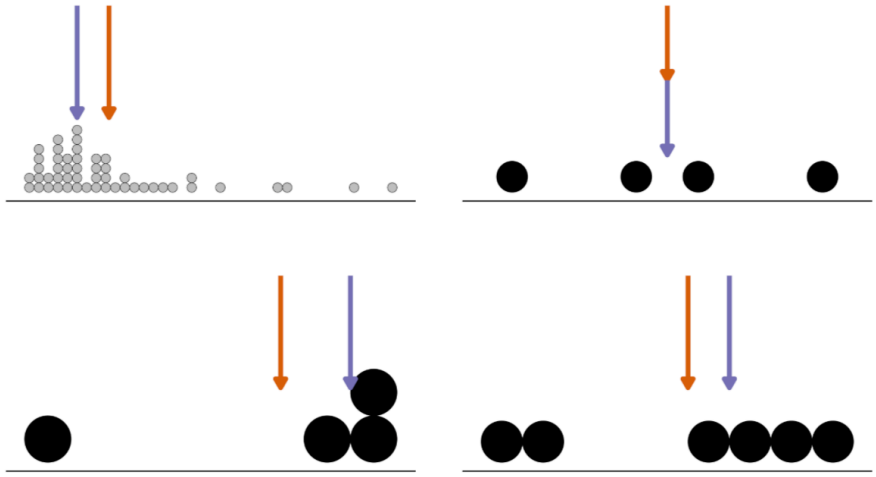
Sample mean vs. sample median

Sample mean vs. sample median

- Both quantities serve as a notion of where a typical value falls in the data.
- The sample mean is the expectation of the empirical probability distribution
- The sample median is less sensitive, or more **robust** to outliers.

Sample mean vs. sample median

- Both quantities serve as a notion of where a typical value falls in the data.
- The sample mean is the expectation of the empirical probability distribution
- The sample median is less sensitive, or more **robust** to outliers.



Example: Comparing two samples

Example: Comparing two samples

Suppose you are interested in which method is faster to get to Queen's Park from Yonge/Bloor Station.

Walk across College St.

13 15 15 12 16

Take TTC to Queen's Park Station.

10 13 19 13 11

Times in minutes.

Example: Comparing two samples

Suppose you are interested in which method is faster to get to Queen's Park from Yonge/Bloor Station.

Walk across College St.

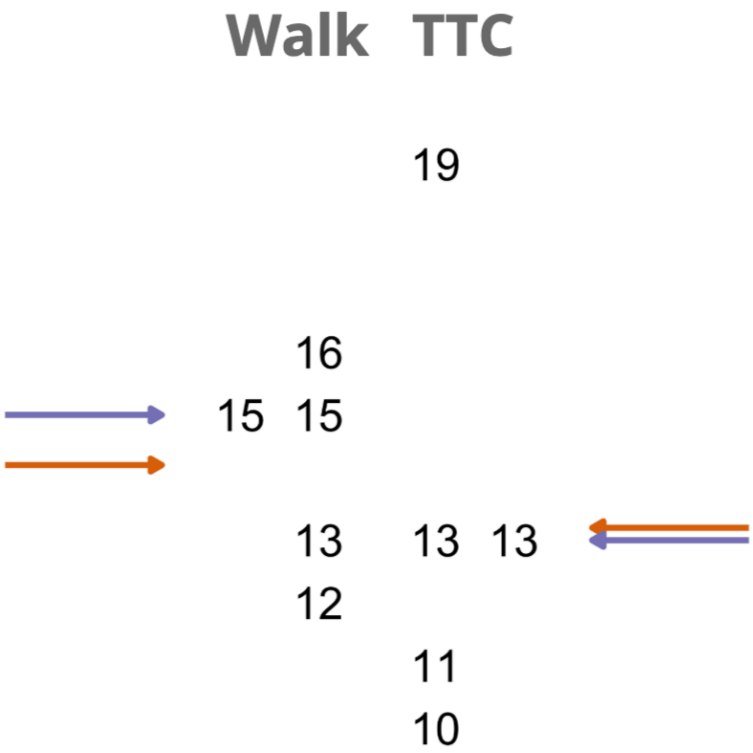
13 15 15 12 16

Take TTC to Queen's Park Station.

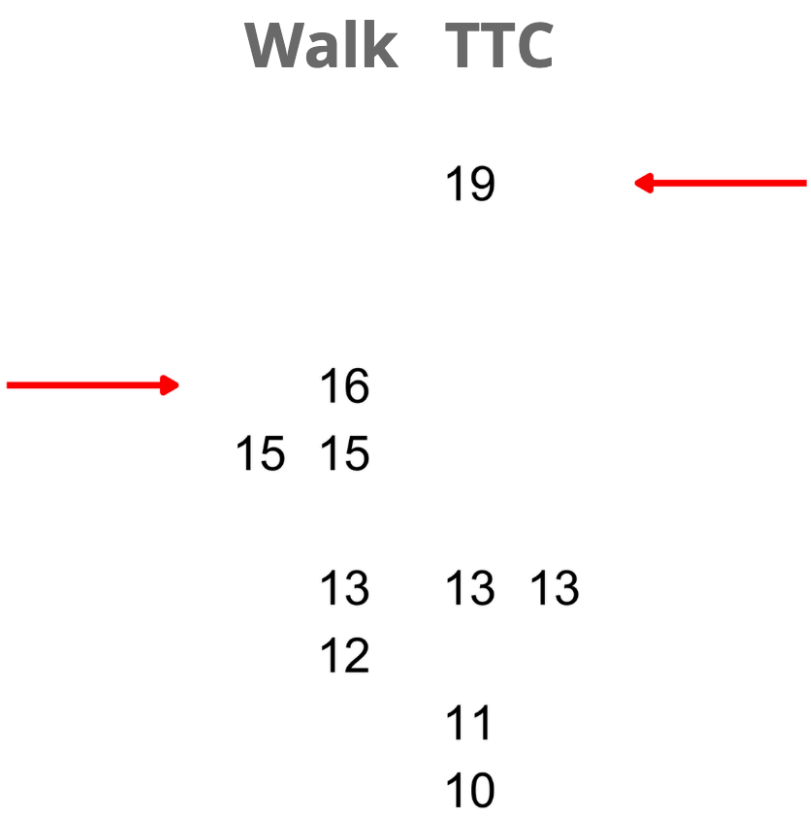
10 13 19 13 11

Times in minutes.

Example: Comparing two samples



Example: Comparing two samples



If what we are interested in is to plan for the *worst-case* scenario, the statistic of interest should be the *maximum*.

Order Statistics

Order Statistics

If we consider sorting our data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, then

$$\min \{x_1, \ldots, x_n\} = x_{(1)},$$
$$\max \{x_1, \ldots, x_n\} = x_{(n)}.$$

Recall the median was expressed in terms of order statistics:

$$\text{Med} = x_{((n+1)/2)} \quad \text{or} \quad \text{Med} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}).$$

Order Statistics

If we consider sorting our data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, then

$$\min \{x_1, \ldots, x_n\} = x_{(1)},$$
$$\max \{x_1, \ldots, x_n\} = x_{(n)}.$$

Recall the median was expressed in terms of order statistics:

$$\text{Med} = x_{((n+1)/2)} \quad \text{or} \quad \text{Med} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}).$$

Then we can define statistics such as the *range*:

$$x_{(n)} - x_{(1)}.$$

(Is this a robust statistic?)

Or the inter-quartile range...

Empirical quantiles

Empirical quantiles

The empirical cumulative distribution function (eCDF) is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} = \frac{\#\{\text{number of datapoints} \leq x\}}{n}$$

Empirical quantiles

The empirical cumulative distribution function (eCDF) is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} = \frac{\#\{\text{number of datapoints} \leq x\}}{n}$$

The pth **empirical quantile** for $p \in [0, 1]$, denoted

$$q_n(p) = F_n^{-1}(p)$$

is the number such that the proportion of the data set below $q_n(p)$ is p.

Empirical quantiles

The empirical cumulative distribution function (eCDF) is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} = \frac{\#\{\text{number of datapoints} \leq x\}}{n}$$

The pth **empirical quantile** for $p \in [0, 1]$, denoted

$$q_n(p) = F_n^{-1}(p)$$

is the number such that the proportion of the data set below $q_n(p)$ is p.

This definition is ambiguous (**Exercise.** Why?).

Important empirical quantiles:

- Median: $q_n(0.5)$
- Quartiles: $q_n(0.25),\ q_n(0.75)$
- Percentiles: $q_n(i/100)$ for $i = \{1, 2, \ldots, 99, 100\}$ is the ith percentile.

Important empirical quantiles:

- Median: $q_n(0.5)$
- Quartiles: $q_n(0.25),\ q_n(0.75)$
- Percentiles: $q_n(i/100)$ for $i = \{1, 2, \ldots, 99, 100\}$ is the ith percentile.

The quartiles are used to define the Interquartile Range

$$\text{IQR} = q_n(0.75) - q_n(0.25),$$

How does the robustness of this statistic compare with that of the range?

Computing empirical quantiles

In terms of the order statistics, we have

$$x_{(k)} = q_n\left(\frac{k}{n+1}\right), \quad k \in \{1, 2, \ldots, n\}.$$

An option to calculate quantiles is:

Computing empirical quantiles

In terms of the order statistics, we have

$$x_{(k)} = q_n\left(\frac{k}{n+1}\right), \quad k \in \{1, 2, \ldots, n\}.$$

An option to calculate quantiles is:

$$q_n(p) = \inf_x \{x \in \mathrm{R} : p \leq F(x)\}$$

Another option to calculate quantiles at any $p \in [0, 1]$ is

$$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$

with $k = \lfloor p(n+1) \rfloor$, $\alpha = p(n+1) - k$.

Five-number summary

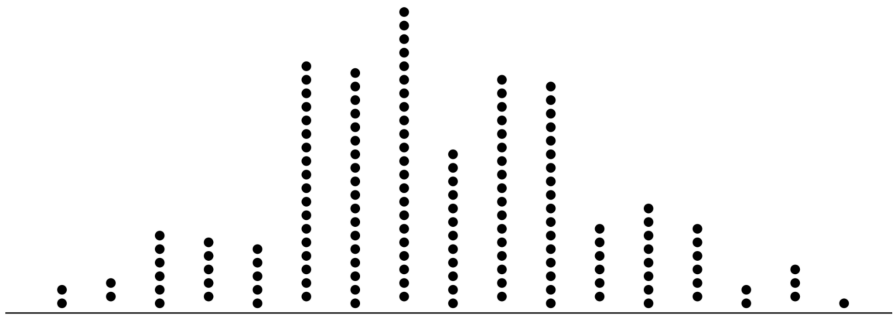
1. Minimum
2. Lower quartile
3. Median
4. Upper quartile
5. Maximum

Summarising data: Measures of Dispersion

How *spread out* are our samples?

- Range
- Interquartile range
- Sample variance
- Sample standard deviation
- Sample mean absolute deviation

Going beyond the range/IQR



- The simplest dispersion measure is the range $x_{(n)} - x_{(1)}$.
- Can we find a notion of range of *typical values*?

Variance

For probability distributions, the essential measure of dispersion of a random variable X is the **variance**

$$\mathbf{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Variance

For probability distributions, the essential measure of dispersion of a random variable X is the **variance**

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

This assigns an equal weight to the squared distance of each observed value from the sample mean.

How can we find such a quantity for our sample data?

Sample variance

An option would be to take the required expectations with respect to the empirical distribution F_n:

$$\begin{aligned}\mathbb{E}_{F_n}\left[(X - \mathbb{E}_{F_n}[X])^2\right] &= \mathbb{E}_{F_n}\left[(X - \bar{X})^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.\end{aligned}$$

Sample variance

An option would be to take the required expectations with respect to the empirical distribution F_n:

$$\begin{aligned} \mathrm{E}_{F_n}[(X - \mathrm{E}_{F_n}[X])^2] &= \mathrm{E}_{F_n}[(X - \bar{X})^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{aligned}$$

However, statisticians tend to prefer something else!

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We will see why this is the case when we introduce estimators.

Sample standard deviation

Sample standard deviation

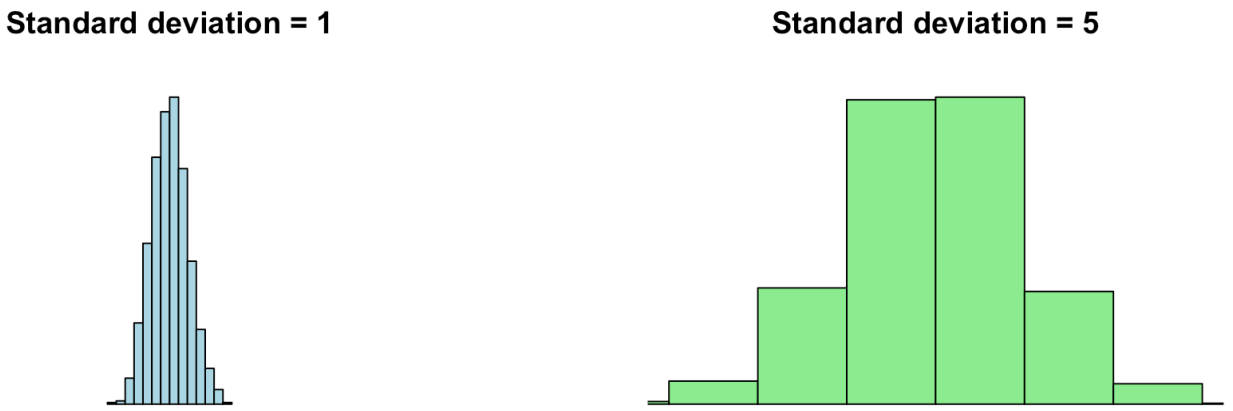
Note that the variance is given in squared units. The sample standard deviation gives us the average squared deviation in the appropriate units.

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

Sample standard deviation

Note that the variance is given in squared units. The sample standard deviation gives us the average squared deviation in the appropriate units.

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$



Mean/Median Absolute Deviation (MAD)

Mean/Median Absolute Deviation (MAD)

The variance/standard deviation is not the only option

$$\mathrm{MAD}_{\mathrm{mean}}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} |X_i - m(X)|,$$

where $m(X)$ is a measure of central tendency, typically the sample mean $\bar{X}$ or the sample median Med.

Mean/Median Absolute Deviation (MAD)

The variance/standard deviation is not the only option

$$\mathrm{MAD}_{\mathrm{mean}}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} |X_i - m(X)|,$$

where $m(X)$ is a measure of central tendency, typically the sample mean $\bar{X}$ or the sample median Med.

It is also common to consider the median deviation:

$$\mathrm{MAD}_{\mathrm{median}}(X_1, \ldots, X_n) = \mathrm{Med}\{|X_i - \mathrm{Med}|\},$$

Mean/Median Absolute Deviation (MAD)

The variance/standard deviation is not the only option

$$\text{MAD}_{\text{mean}}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} |X_i - m(X)|,$$

where $m(X)$ is a measure of central tendency, typically the sample mean $\bar{X}$ or the sample median Med.

It is also common to consider the median deviation:

$$\text{MAD}_{\text{median}}(X_1, \ldots, X_n) = \text{Med}\{|X_i - \text{Med}|\},$$

> **Exercise.** Show that if a distribution is symmetric, then the median absolute deviation is equal to $\frac{1}{2}\text{IQR}$.

Examples

$X = \{2, 2, 3, 4, 14\}$

```r
X <- c(2, 2, 3, 4, 14)

# Variance/Standard deviation
sample_mean <- mean(X)
sample_mean
```

```
[1] 5
```

```r
sum((X - sample_mean)^2)/(length(X)-1)
```

```
[1] 26
```

```r
var(X)
```

```
[1] 26
```

```r
sd(X)
```

```
[1] 5.09902
```

```r
# MAD with mean vs. median
sample_med <- median(2, 2, 3, 4, 14)
sample_med
```

```
[1] 2
```

```r
mean(abs(X - sample_med))
```

```
[1] 3
```

```r
median(abs(X - sample_med))
```

```
[1] 1
```

Summary

- Summarising a dataset provides useful and interpretable information.
- Reducing a whole dataset down to a single number doesn’t tell us the full picture of a data set.
- Different summaries can reveal di!erent characteristics about the observed data.