HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 4 — Descriptive statistics

Jorge N. Tendeiro
Hiroshima University

Today

Descriptive statistics (continuation from previous lecture):

Today

Descriptive statistics (continuation from previous lecture):

Measures of spread:

- Variance
- Standard deviation
- Range
- Quartiles
- Interquartile range

Today

Descriptive statistics (continuation from previous lecture):

Measures of spread:

- Variance
- Standard deviation
- Range
- Quartiles
- Interquartile range

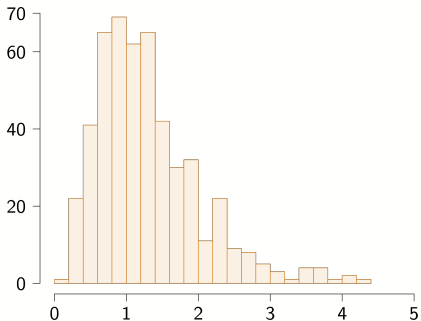
More plots:

- Boxplot
- Scatterplot

Review

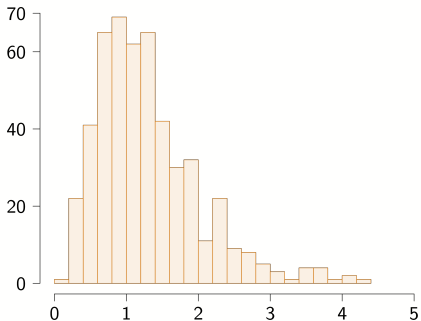
Review

Histogram:
One way to visually represent the data.



Review

Histogram:
One way to visually represent the data.



Descriptive statistics — measures of location:
Mean, median, mode

Mean	Median	Mode
1.3	1.2	0.9

Review

Below are scores of 10 students on three tests (A, B, C) (left), and the location for each test score distribution (right):

Test	Student									
	1	2	3	4	5	6	7	8	9	10
A	0	3	3	5	5	5	5	7	7	10
B	0	1	2	3	5	5	7	8	9	10
C	3	4	4	5	5	5	5	6	6	7

Measure of location	Test		
	A	B	C
Mean			
Median			

1. Fill in the right-hand side table above.
2. Are the three test distributions alike?

Review — ANSWER

Below are scores of 10 students on three tests (A, B, C) (left), and the location for each test score distribution (right):

Test	Student									
	1	2	3	4	5	6	7	8	9	10
A	0	3	3	5	5	5	5	7	7	10
B	0	1	2	3	5	5	7	8	9	10
C	3	4	4	5	5	5	5	6	6	7

Measure of location	Test		
	A	B	C
Mean	5	5	5
Median	5	5	5

1. Fill in the right-hand side table above.
2. Are the three test distributions alike?

Review — ANSWER

Below are scores of 10 students on three tests (A, B, C) (left), and the location for each test score distribution (right):

Test	Student									
	1	2	3	4	5	6	7	8	9	10
A	0	3	3	5	5	5	5	7	7	10
B	0	1	2	3	5	5	7	8	9	10
C	3	4	4	5	5	5	5	6	6	7

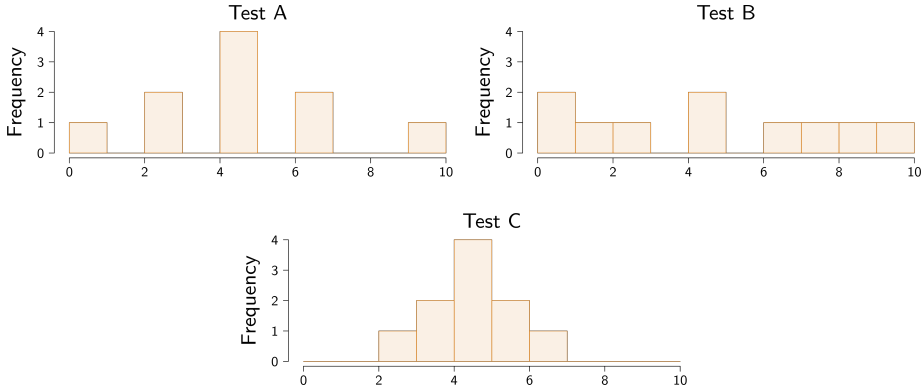
Measure of location	Test		
	A	B	C
Mean	5	5	5
Median	5	5	5

1. Fill in the right-hand side table above.
2. Are the three test distributions alike?

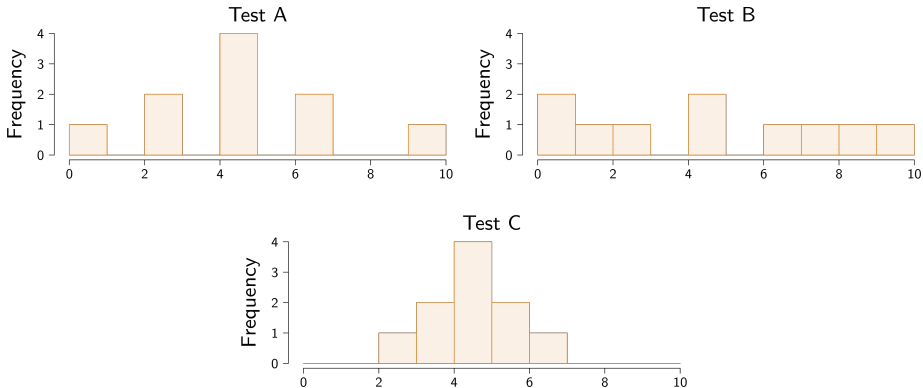
Answer

Even when the location of two or more distributions is the same, their spread may differ.
See the next page.

Spread of a distribution



Spread of a distribution



Q: What is a representative value to describe the spread for each distribution?
A: There are various!, collectively known as measures of spread.

Measures of spread:
Variance, standard deviation

Variance

The variance of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean (recall last lecture).

Variance

The variance of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean (recall last lecture).

The variance describes the spread of the data $(x_1, \ldots, x_n)$ around the mean value $(\overline{x})$.

Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

Note. $n = 5$.

Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

Note. $n = 5$.

Note that

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 65}{5} = \frac{295}{5} = 59.$$

Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

Note. $n = 5$.

Note that

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 65}{5} = \frac{295}{5} = 59.$$

Then:

$$s_x^2 = \frac{(60 - 59)^2 + (52 - 59)^2 + (44 - 59)^2 + (74 - 59)^2 + (65 - 59)^2}{5}$$

$$= \frac{1^2 + 7^2 + 15^2 + 15^2 + 6^2}{5} = \frac{536}{5} = 107.2$$

Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

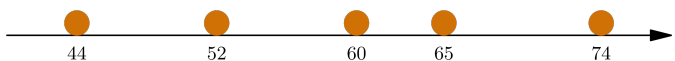
Note. $n = 5$.

Note that

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 65}{5} = \frac{295}{5} = 59.$$

Then:

$$s_x^2 = \frac{(60 - 59)^2 + (52 - 59)^2 + (44 - 59)^2 + (74 - 59)^2 + (65 - 59)^2}{5}$$
$$= \frac{1^2 + 7^2 + 15^2 + 15^2 + 6^2}{5} = \frac{536}{5} = 107.2$$



Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

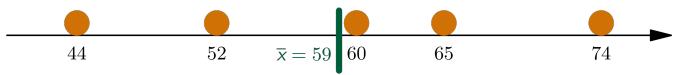
Note. $n = 5$.

Note that

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 65}{5} = \frac{295}{5} = 59.$$

Then:

$$s_x^2 = \frac{(60 - 59)^2 + (52 - 59)^2 + (44 - 59)^2 + (74 - 59)^2 + (65 - 59)^2}{5}$$
$$= \frac{1^2 + 7^2 + 15^2 + 15^2 + 6^2}{5} = \frac{536}{5} = 107.2$$



Variance

Individual	Weight (kg)
1	$x_1 = 60$
2	$x_2 = 52$
3	$x_3 = 44$
4	$x_4 = 74$
5	$x_5 = 65$

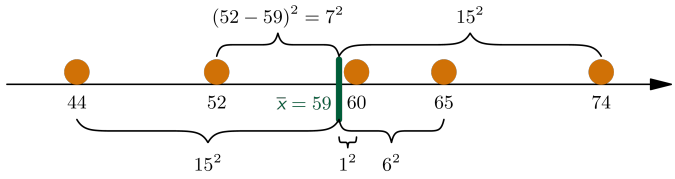
Note. $n = 5$.

Note that

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 65}{5} = \frac{295}{5} = 59.$$

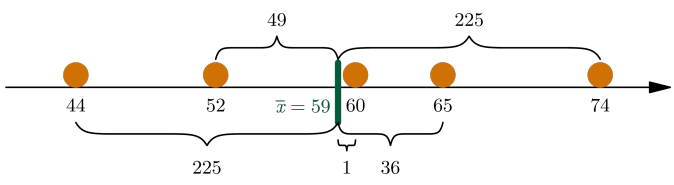
Then:

$$s_x^2 = \frac{(60 - 59)^2 + (52 - 59)^2 + (44 - 59)^2 + (74 - 59)^2 + (65 - 59)^2}{5}$$
$$= \frac{1^2 + 7^2 + 15^2 + 15^2 + 6^2}{5} = \frac{536}{5} = 107.2$$



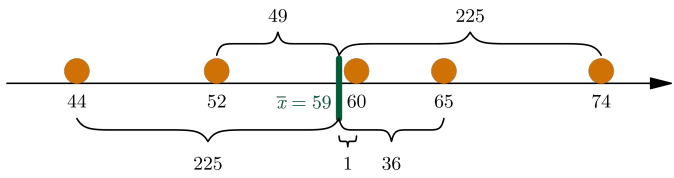
Variance

The further an observation is from the mean (in either direction!), the more it contributes to increase the variance.



Variance

The further an observation is from the mean (in either direction!), the more it contributes to increase the variance.

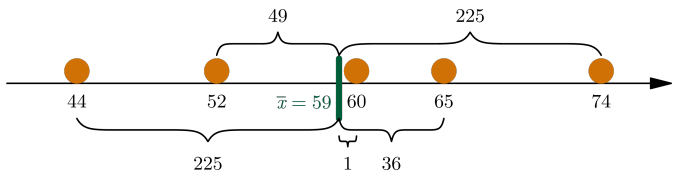


Interpreting a variance:

- The variance value itself is hard to interpret, because it is an average of squared distances to the mean.

Variance

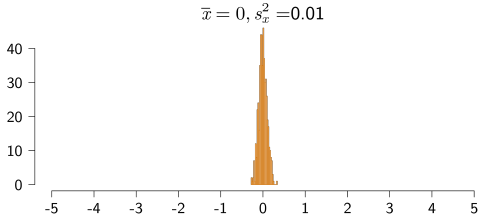
The further an observation is from the mean (in either direction!), the more it contributes to increase the variance.

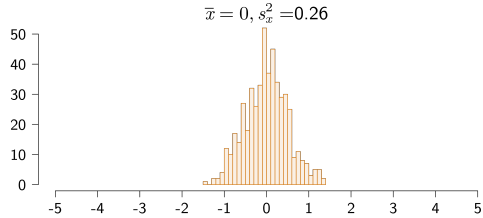


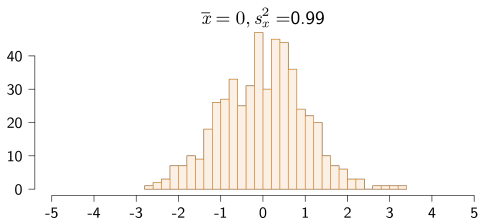
Interpreting a variance:

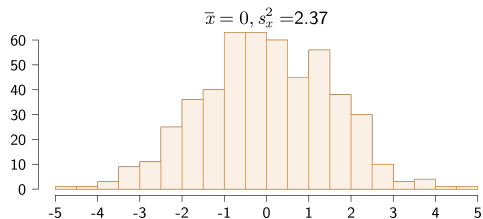
- The variance value itself is hard to interpret, because it is an average of squared distances to the mean.
- However, the magnitude of a variance is easier to interpret:
  - Small variance $\iff$ the data are closely scattered around its mean.
  - Large variance $\iff$ the data are widely scattered around its mean.

Variance

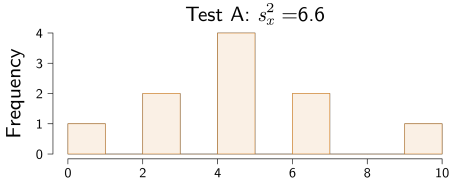

$\overline{x} = 0, s_x^2 = 0.01$
40
30
20
10
0
-5 -4 -3 -2 -1 0 1 2 3 4 5

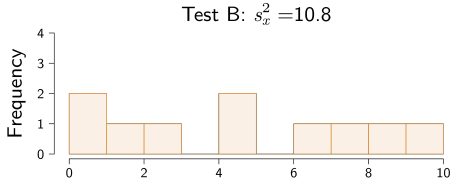

$\overline{x} = 0, s_x^2 = 0.26$
50
40
30
20
10
0
-5 -4 -3 -2 -1 0 1 2 3 4 5

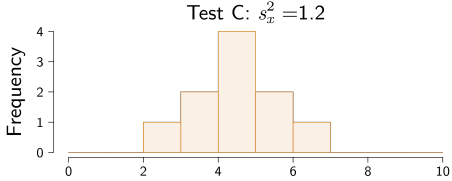

$\overline{x} = 0, s_x^2 = 0.99$
40
30
20
10
0
-5 -4 -3 -2 -1 0 1 2 3 4 5


$\overline{x} = 0, s_x^2 = 2.37$
60
50
40
30
20
10
0
-5 -4 -3 -2 -1 0 1 2 3 4 5

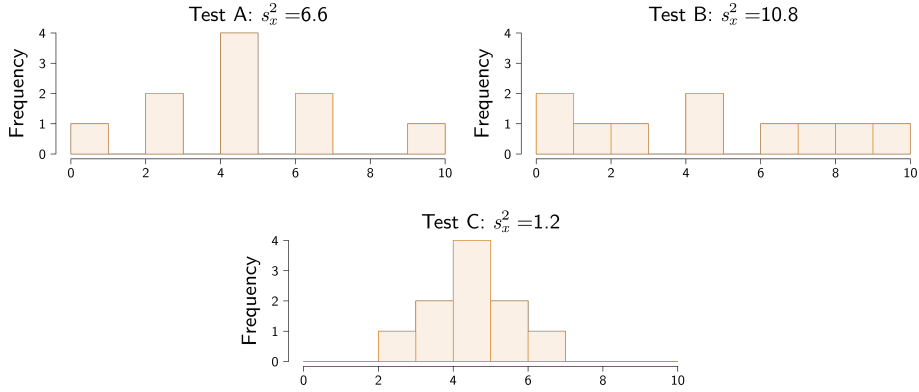
Variance — Tests A, B, C







Variance — Tests A, B, C



The three tests can be ordered in terms of how much the scores scatter around their mean:

$$\text{Test C} < \text{Test A} < \text{Test B}$$

Variance — Some properties

- The variance is equal to 0 *if and only if* all the data have the same value.

  Example: Let $x = \{8, 8, 8, 8, 8, 8\}$.

  - $\overline{x} = \frac{8+8+8+8+8+8}{6} = 8$
  - $s_x^2 = \frac{(8-8)^2+\cdots+(8-8)^2}{6} = 0.$

Variance — Some properties

- The variance is equal to 0 *if and only if* all the data have the same value.

  Example: Let $x = \{8, 8, 8, 8, 8, 8\}$.
  - $\overline{x} = \frac{8+8+8+8+8+8}{6} = 8$
  - $s_x^2 = \frac{(8-8)^2+\cdots+(8-8)^2}{6} = 0.$

- The variance, just like the mean, is sensitive to outliers.

  Example:

Data	Variance
$1, 3, 4, 5, 9$	7.04
$1, 3, 4, 5, 15$	23.84
$1, 3, 4, 5, 100$	1499.44

Variance — Some properties

- The sum of the signed distances to the mean is always 0, for any data:

$$\frac{(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})}{n} = 0$$

  This is one of the reasons why the variance is based on squared distances to the mean.

Variance — Some properties

- The sum of the signed distances to the mean is always 0, for any data:

$$\frac{(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})}{n} = 0$$

This is one of the reasons why the variance is based on squared distances to the mean.

Example: Let $x = \{60, 52, 44, 74, 65\}$. Then $\overline{x} = \frac{60+52+44+74+65}{5} = \frac{295}{5} = 59$ and

$$\frac{(60 - 59) + (52 - 59) + (44 - 59) + (74 - 59) + (65 - 59)}{5} = \frac{1 - 7 - 15 + 15 + 6}{5} = 0.$$

Variance — Some properties

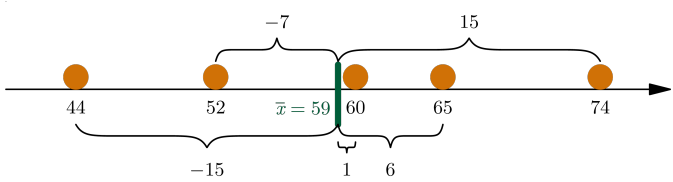
- The sum of the signed distances to the mean is always 0, for any data:

$$\frac{(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})}{n} = 0$$

This is one of the reasons why the variance is based on squared distances to the mean.

Example: Let $x = \{60, 52, 44, 74, 65\}$. Then $\overline{x} = \frac{60+52+44+74+65}{5} = \frac{295}{5} = 59$ and

$$\frac{(60-59) + (52-59) + (44-59) + (74-59) + (65-59)}{5} = \frac{1 - 7 - 15 + 15 + 6}{5} = 0.$$



Variance — Some properties

- We could have used absolute values instead of squared values.

Variance — Some properties

- We could have used absolute values instead of squared values.

  Compare:

$$s_x^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n} \qquad MAD = \frac{|x_1-\overline{x}|+|x_2-\overline{x}|+\cdots+|x_n-\overline{x}|}{n}$$

Variance — Some properties

- We could have used absolute values instead of squared values.

  Compare:

$$s_x^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n} \qquad MAD = \frac{|x_1-\overline{x}|+|x_2-\overline{x}|+\cdots+|x_n-\overline{x}|}{n}$$

  MAD stands for "mean absolute deviation".
  Although it is used in some contexts, it is not nearly as important as the variance.
  The reason is mostly *mathematical*.

Variance — Limitation

We said before that the value of the variance is difficult to interpret.
The main reason is that the variance is the average squared distance to the mean.

Variance — Limitation

We said before that the value of the variance is difficult to interpret.
The main reason is that the variance is the average squared distance to the mean.

Example:
Suppose x is a set of 5 distances in cm:

$$3, 5, 1, 1, 5.$$

Variance — Limitation

We said before that the value of the variance is difficult to interpret.
The main reason is that the variance is the average squared distance to the mean.

Example:
Suppose x is a set of 5 distances in cm:

$$3, 5, 1, 1, 5.$$

Then $\overline{x} = \frac{3+5+1+1+5}{5} = 3$ cm and

$$s_x^2 = \frac{(3-3)^2 + (5-3)^2 + (1-3)^2 + (1-3)^2 + (5-3)^2}{5} = 3.2 \text{ cm}^2.$$

Interpretation — The average squared distance to the mean is 3.2 cm^2...

Variance — Limitation

We said before that the value of the variance is difficult to interpret.
The main reason is that the variance is the average squared distance to the mean.

Example:
Suppose x is a set of 5 distances in cm:

$$3, 5, 1, 1, 5.$$

Then $\overline{x} = \frac{3+5+1+1+5}{5} = 3$ cm and

$$s_x^2 = \frac{(3-3)^2 + (5-3)^2 + (1-3)^2 + (1-3)^2 + (5-3)^2}{5} = 3.2 \text{ cm}^2.$$

Interpretation — The average squared distance to the mean is 3.2 cm^2...

One way to avoid this problem is to use the squared root of the variance.
This leads us to the next measure of spread:

the standard deviation.

Standard deviation

The standard deviation (SD) of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean.

Standard deviation

The standard deviation (SD) of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean.

Like the variance, the standard deviation describes the spread of the data $(x_1, \ldots, x_n)$ around the mean value $(\overline{x})$.
But unlike the variance, the standard deviation's unit of measurement is the same as that from the data.

Standard deviation

The standard deviation (SD) of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean.

Like the variance, the standard deviation describes the spread of the data $(x_1, \ldots, x_n)$ around the mean value $(\overline{x})$.
But unlike the variance, the standard deviation's unit of measurement is the same as that from the data.

Example:
Suppose x is a set of 5 distances in cm:

$$3, 5, 1, 1, 5.$$

Standard deviation

The standard deviation (SD) of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ is the mean.

Like the variance, the standard deviation describes the spread of the data $(x_1, \ldots, x_n)$ around the mean value $(\overline{x})$.
But unlike the variance, the standard deviation's unit of measurement is the same as that from the data.

Example:
Suppose x is a set of 5 distances in cm:

$$3, 5, 1, 1, 5.$$

Then

- Mean $= \overline{x} = \frac{3+5+1+1+5}{5} = 3$ cm.
- Variance $= s_x^2 = \frac{(3-3)^2+(5-3)^2+(1-3)^2+(1-3)^2+(5-3)^2}{5} = 3.2$ cm^2.
- SD $= s_x = \sqrt{s_x^2} = \sqrt{3.2} = 1.79$ cm.

Exercise (2)

Calculate the mean, variance, and standard deviation of the data below:

1. $1, 4, 5, 3, 7$.

2. $4, 4, 4, 4, 4$.

Exercise (2) — ANSWER

Calculate the mean, variance, and standard deviation of the data below:

1. $1, 4, 5, 3, 7$.

Answer

- $\overline{x} = \frac{1+4+5+3+7}{5} = \frac{20}{5} = 4$.
- $s_x^2 = \frac{(1-4)^2+(4-4)^2+(5-4)^2+(3-4)^2+(7-4)^2}{5} = \frac{20}{5} = 4$.
- $s_x = \sqrt{s_x^2} = \sqrt{4} = 2$.

Exercise (2) — ANSWER

Calculate the mean, variance, and standard deviation of the data below:

1. $1, 4, 5, 3, 7$.

Answer

- $\overline{x} = \frac{1+4+5+3+7}{5} = \frac{20}{5} = 4$.
- $s_x^2 = \frac{(1-4)^2+(4-4)^2+(5-4)^2+(3-4)^2+(7-4)^2}{5} = \frac{20}{5} = 4$.
- $s_x = \sqrt{s_x^2} = \sqrt{4} = 2$.

2. $4, 4, 4, 4, 4$.

Answer

- $\overline{x} = \frac{4+4+4+4+4}{5} = \frac{20}{5} = 4$.
- $s_x^2 = \frac{(4-4)^2+(4-4)^2+(4-4)^2+(4-4)^2+(4-4)^2}{5} = \frac{0}{5} = 0$.
- $s_x = \sqrt{s_x^2} = \sqrt{0} = 0$.

Measures of spread:
Range, quartiles, interquartile range

Range

The range of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$\text{Range} R = \max(\text{data}) - \min(\text{data}).$$

Range

The range of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$\text{Range} R = \max(\text{data}) - \min(\text{data}).$$

In other words,

The range is the width of the shortest interval containing all the data.

Range

Interpretation:

- Small range $\Longrightarrow$ the data are closely spread.
- Large range $\Longrightarrow$ the data are widely spread.

Range

Interpretation:

- Small range $\Longrightarrow$ the data are closely spread.
- Large range $\Longrightarrow$ the data are widely spread.

Important
The range is sensitive to outliers.

For example:

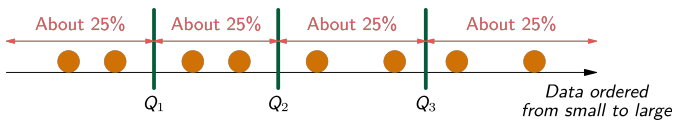
Data	Range
$1, 3, 4, 5, 9$	$R = 9 - 1 = 8$
$1, 3, 4, 5, 15$	$R = 15 - 1 = 14$
$1, 3, 4, 5, 100$	$R = 100 - 1 = 99$

Quartiles

The quartiles of variable x with n observations $x_1, x_2, \ldots, x_n$ are boundary values that divide the data in roughly four equal parts:

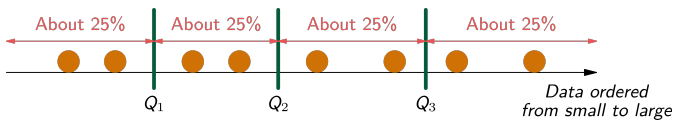
Quartiles

The quartiles of variable x with n observations $x_1, x_2, \ldots, x_n$ are boundary values that divide the data in roughly four equal parts:



Quartiles

The quartiles of variable x with n observations $x_1, x_2, \ldots, x_n$ are boundary values that divide the data in roughly four equal parts:



- Q_1: 1^{st} quartile.
  The median of the lowest 50% of the data.
- Q_2: 2^{nd} quartile.
  The median.
- Q_3: 3^{rd} quartile.
  The median of the highest 50% of the data.

Quartiles — computation

There are various formulas to compute quartiles (and *quantiles* in general).
There are at least nine formulas (Hyndman and Fan, 1996).

You usually use software to compute such statistics.
Below you can see two possible algorithms.

Quartiles — computation

There are various formulas to compute quartiles (and *quantiles* in general).
There are at least nine formulas (Hyndman and Fan, 1996).

You usually use software to compute such statistics.
Below you can see two possible algorithms.

Example
When the number of observations is even:

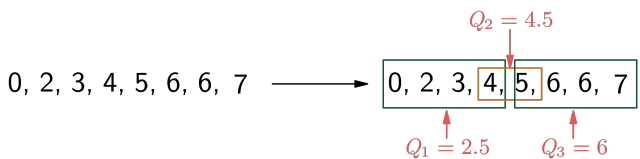
0, 2, 3, 4, 5, 6, 6, 7

Quartiles — computation

There are various formulas to compute quartiles (and *quantiles* in general).
There are at least nine formulas (Hyndman and Fan, 1996).

You usually use software to compute such statistics.
Below you can see two possible algorithms.

Example
When the number of observations is even:



- $Q_1 =$ median of $\{0, 2, 3, 4\}$
- $Q_2 =$ median
- $Q_3 =$ median of $\{5, 6, 6, 7\}$

Quartiles — computation

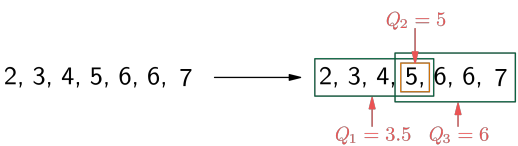
Example
When the number of observations is odd:

Including Q_2 (Hinge's method).

Quartiles — computation

Example
When the number of observations is odd:

Including Q_2 (Hinge's method).

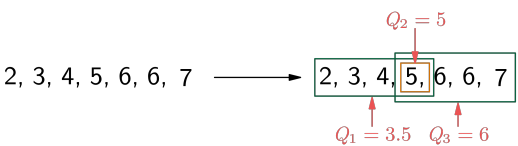


- $Q_1 =$ median of $\{2, 3, 4, 5\}$
- $Q_2 =$ median
- $Q_3 =$ median of $\{5, 6, 6, 7\}$

Quartiles — computation

Example
When the number of observations is odd:

Including Q_2 (Hinge's method).



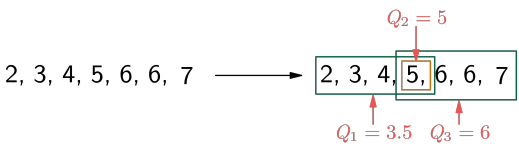
- $Q_1 =$ median of $\{2, 3, 4, 5\}$
- $Q_2 =$ median
- $Q_3 =$ median of $\{5, 6, 6, 7\}$

Excluding Q_2.

Quartiles — computation

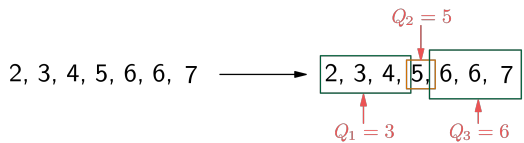
Example
When the number of observations is odd:

Including Q_2 (Hinge's method).



- Q_1 = median of $\{2, 3, 4, 5\}$
- Q_2 = median
- Q_3 = median of $\{5, 6, 6, 7\}$

Excluding Q_2.



- Q_1 = median of $\{2, 3, 4\}$
- Q_2 = median
- Q_3 = median of $\{6, 6, 7\}$

Interquartile range

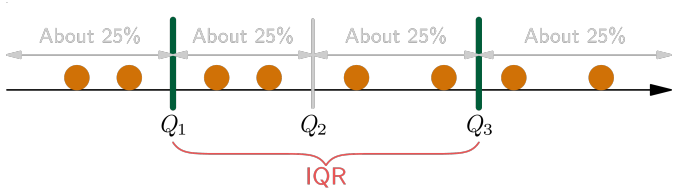
The interquartile range of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

$$\text{IQR} = Q_3 - Q_1.$$

Interquartile range

The interquartile range of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

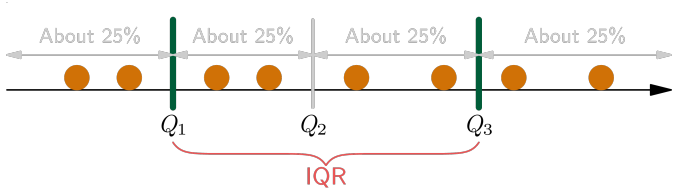
$$\text{IQR} = Q_3 - Q_1.$$



Interquartile range

The interquartile range of variable x with n observations $x_1, x_2, \ldots, x_n$ is given by

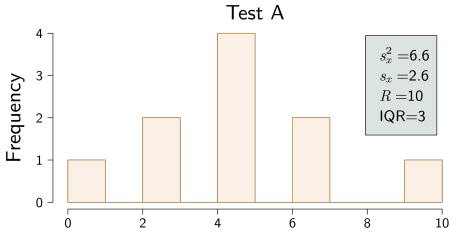
$$\text{IQR} = Q_3 - Q_1.$$

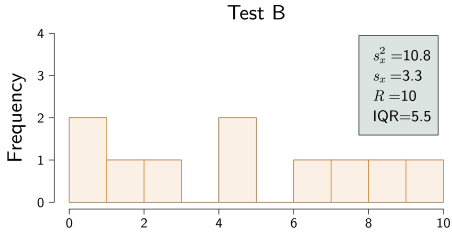


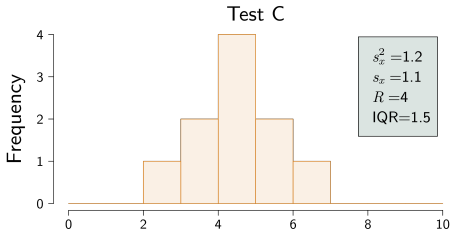
The IQR…

- …is equal to the width of the central part (about 50%) of the data.
- …describes how much the data are scattered around the median.
- …is less sensitive to outliers than the range.
- …can be computed and reported together with the median.

Spread of a distribution







Example

Below are the results of a questionnaire aimed at measuring customers' service satisfaction (1 = highly unsatisfied, 2 = unsatisfied, 3 = normal, 4 = satisfied, 5 = highly satisfied):

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Example

Below are the results of a questionnaire aimed at measuring customers' service satisfaction (1 = highly unsatisfied, 2 = unsatisfied, 3 = normal, 4 = satisfied, 5 = highly satisfied):

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Q: Which service had a higher satisfaction?

Example

Below are the results of a questionnaire aimed at measuring customers' service satisfaction (1 = highly unsatisfied, 2 = unsatisfied, 3 = normal, 4 = satisfied, 5 = highly satisfied):

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Q: Which service had a higher satisfaction?

Answer

Service	Descriptive statistic	
	Mean	Variance
A	3.6	2.04
B	3.7	0.81

Service B has the highest mean and the lowest variance.
Therefore, service B is better!

Example

Below are the results of a questionnaire aimed at measuring customers' service satisfaction (1 = highly unsatisfied, 2 = unsatisfied, 3 = normal, 4 = satisfied, 5 = highly satisfied):

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Q: Which service had a higher satisfaction?

Answer — WRONG!

Service	Descriptive statistic	
	Mean	Variance
A	3.6	2.04
B	3.7	0.81

For an ordinal variable, the mean and variance are **not** suitable descriptive statistics!

Exercise (3)

Below are the results of a questionnaire aimed at measuring customers' service satisfaction (1 = highly unsatisfied, 2 = unsatisfied, 3 = normal, 4 = satisfied, 5 = highly satisfied):

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

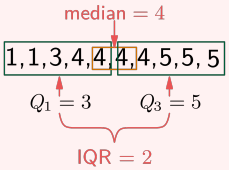
Q: Calculate the median and IQR for each service to determine which service has: (a) a higher satisfaction across the sample, and (b) more stable (i.e., less varying) results.

Exercise (3) — ANSWER

Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Answer

Service A:

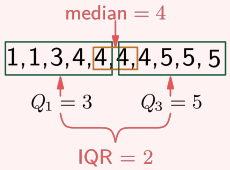


Exercise (3) — ANSWER

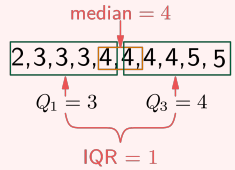
Service	Customer									
	1	2	3	4	5	6	7	8	9	10
A	3	4	4	5	5	1	1	5	4	4
B	5	2	3	4	4	5	4	3	4	3

Answer

Service A:



Service B:

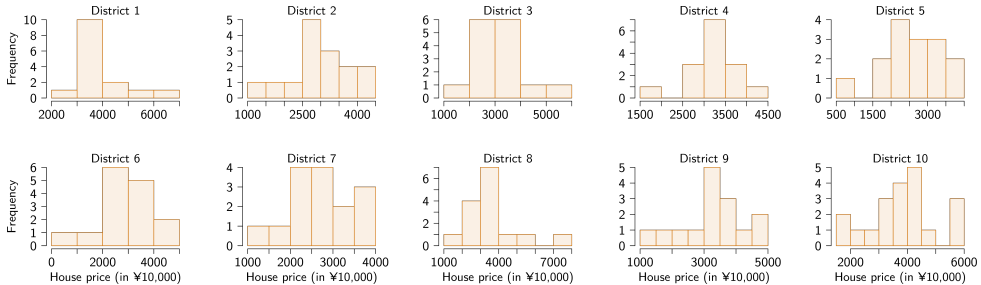


Both services have the same median satisfaction.
However, the scores of Service B are more stable (i.e., vary less) than those of Service A, based on the IQR statistic.

Boxplot

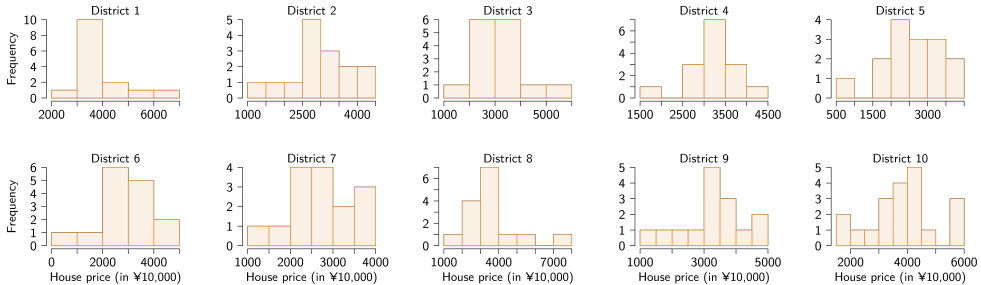
Boxplot

Consider the following plot, summarizing the house prices in ten different districts:



Boxplot

Consider the following plot, summarizing the house prices in ten different districts:



The information is not easy to "read off" from the figure:

- What is the center of each distribution?
- What is the spread of each distribution?
- How can we compare the prices across all districts?

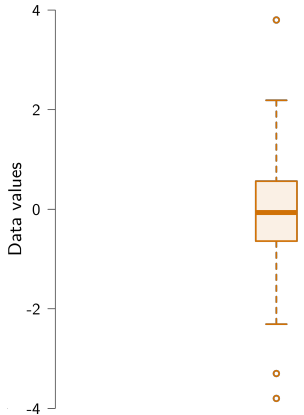
There is a very neat plot that can help us here!

Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!

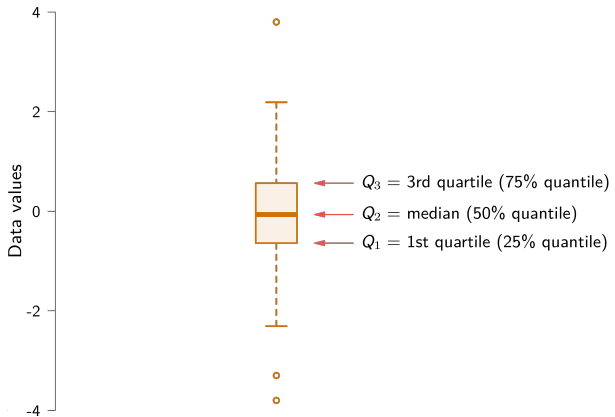
Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!



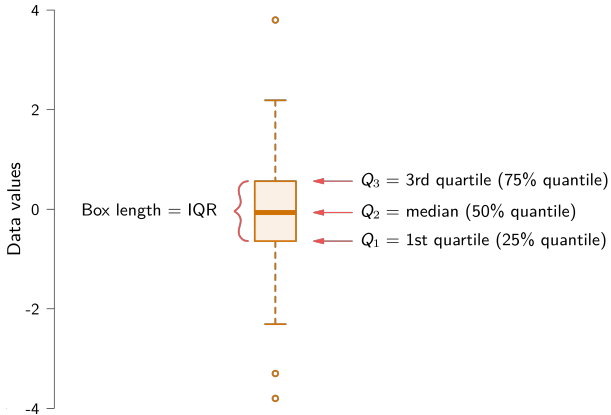
Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!



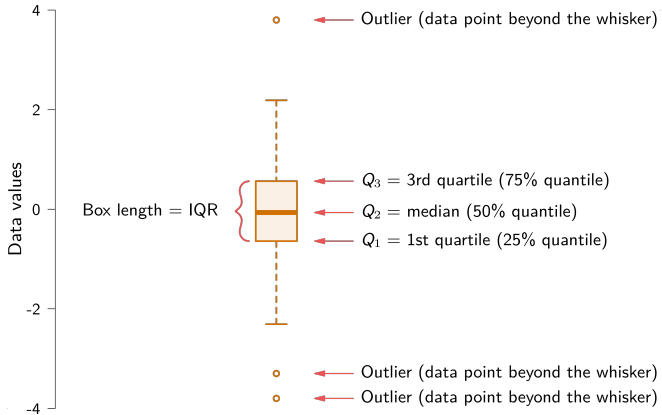
Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!



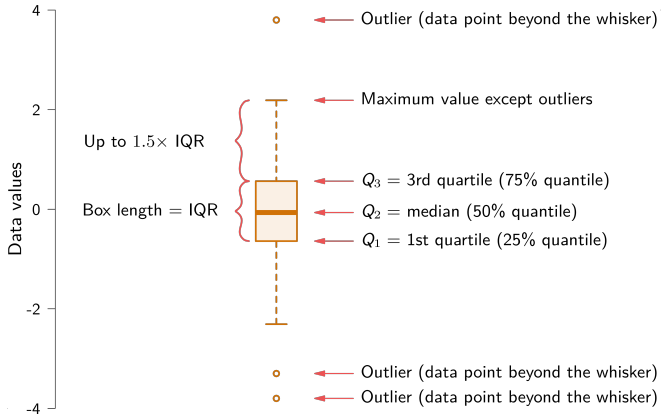
Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!



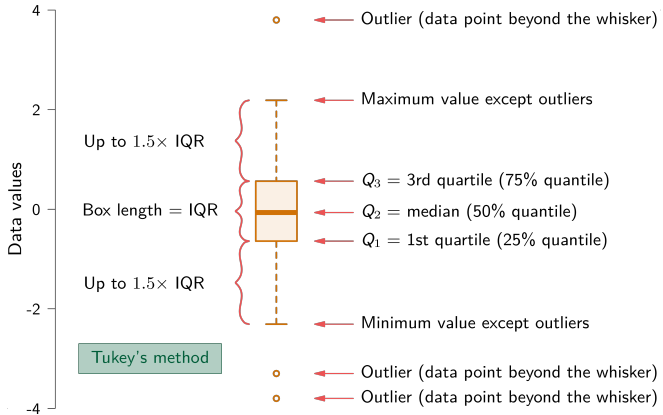
Boxplot

Boxplots are plots from which we can easily infer the location and spread of the data!

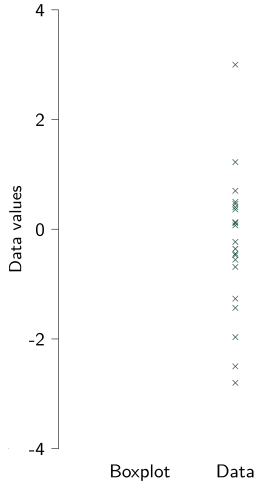


Boxplot

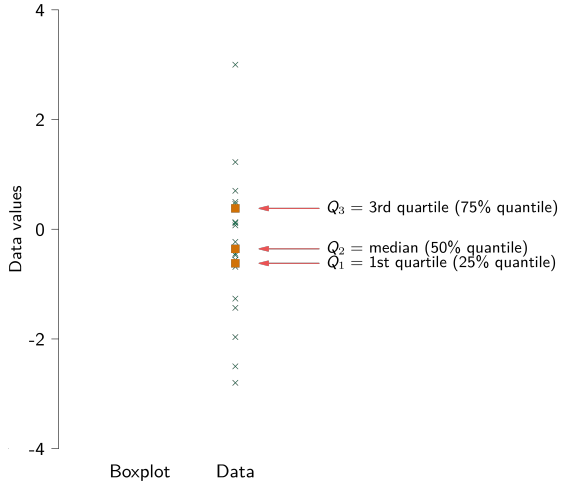
Boxplots are plots from which we can easily infer the location and spread of the data!



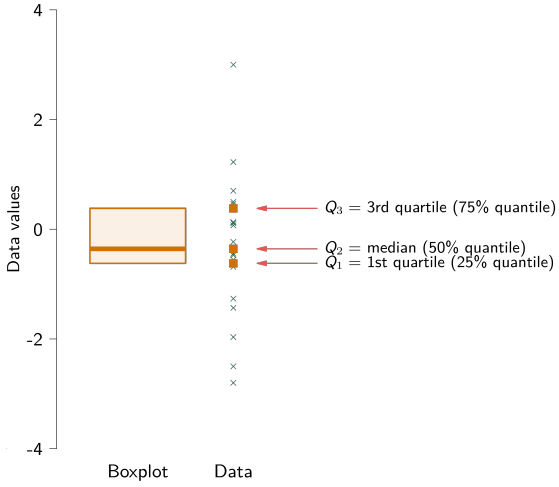
Boxplot



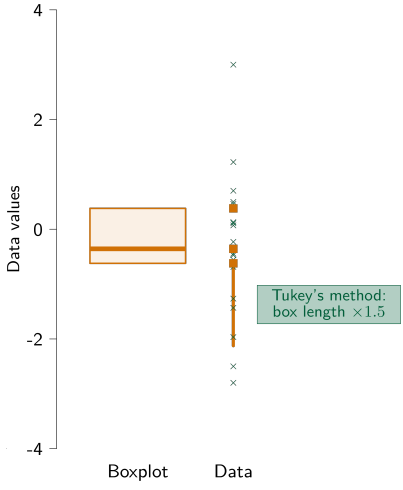
Boxplot


4
2
0
-2
-4
Data values
Q_3 = 3rd quartile (75% quantile)
Q_2 = median (50% quantile)
Q_1 = 1st quartile (25% quantile)
Boxplot
Data

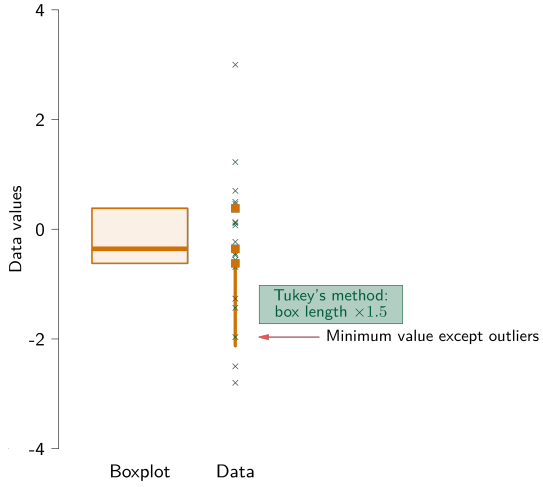
Boxplot


4
2
0
-2
-4
Data values
Q_3 = 3rd quartile (75% quantile)
Q_2 = median (50% quantile)
Q_1 = 1st quartile (25% quantile)
Boxplot
Data

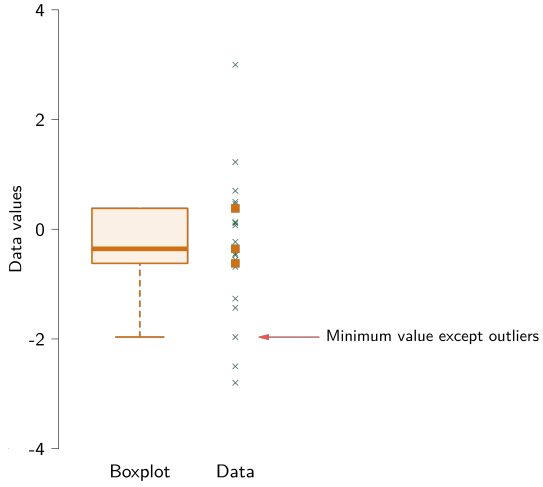
Boxplot


4
2
0
-2
-4
Data values
Tukey's method:
box length $\times 1.5$
Boxplot
Data

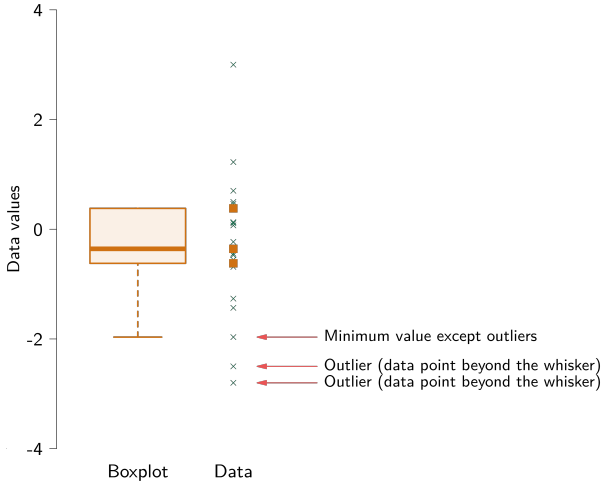
Boxplot


4
2
0
-2
-4
Data values
Tukey's method:
box length $\times 1.5$
Minimum value except outliers
Boxplot
Data

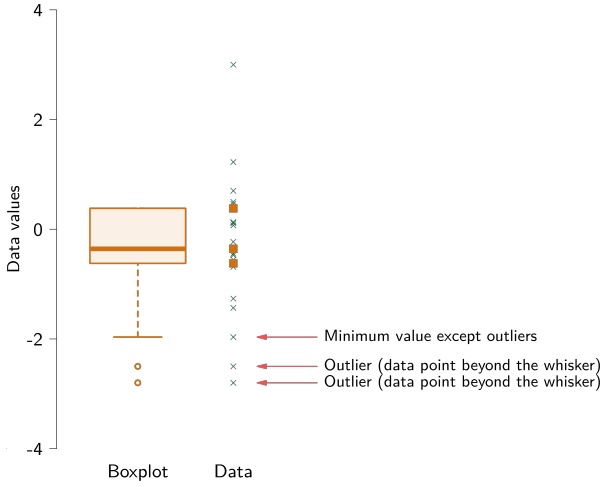
Boxplot


4
2
0
-2
-4
Data values
Minimum value except outliers
Boxplot
Data

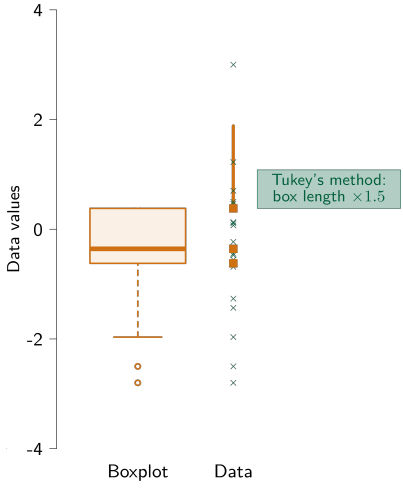
Boxplot


4
2
0
-2
-4
Data values
Minimum value except outliers
Outlier (data point beyond the whisker)
Outlier (data point beyond the whisker)
Boxplot
Data

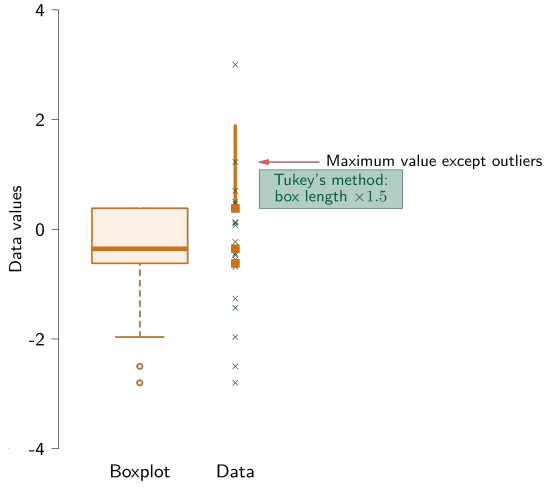
Boxplot


4
2
0
-2
-4
Data values
Minimum value except outliers
Outlier (data point beyond the whisker)
Outlier (data point beyond the whisker)
Boxplot
Data

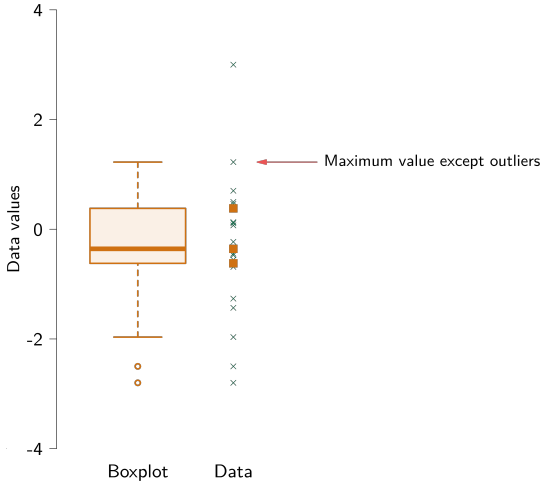
Boxplot


4
2
0
-2
-4
Data values
Tukey's method:
box length $\times 1.5$
Boxplot
Data

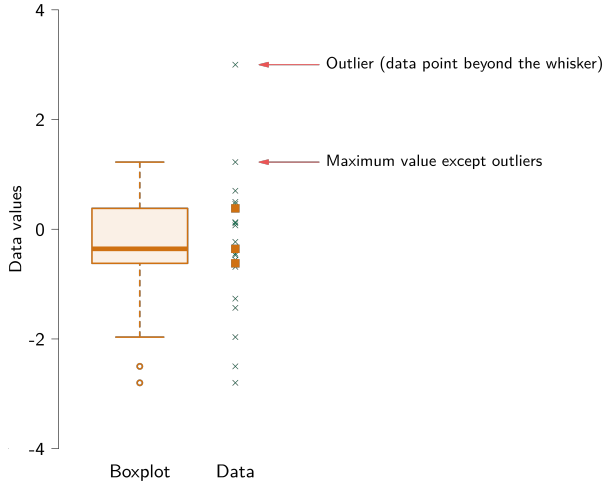
Boxplot


4
2
0
-2
-4
Data values
Maximum value except outliers
Tukey's method:
box length $\times 1.5$
Boxplot
Data

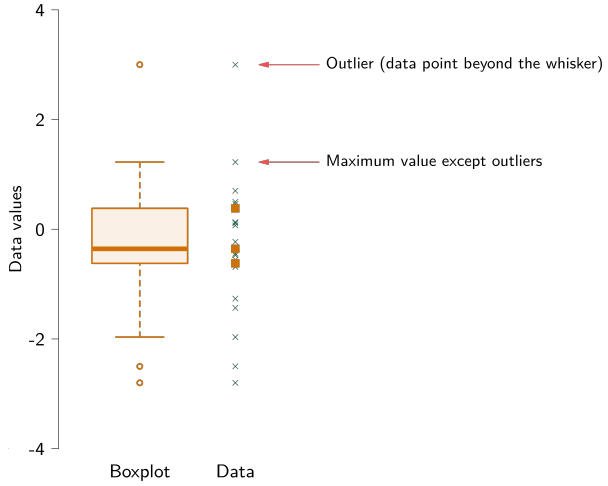
Boxplot


4
2
0
-2
-4
Data values
Maximum value except outliers
Boxplot
Data

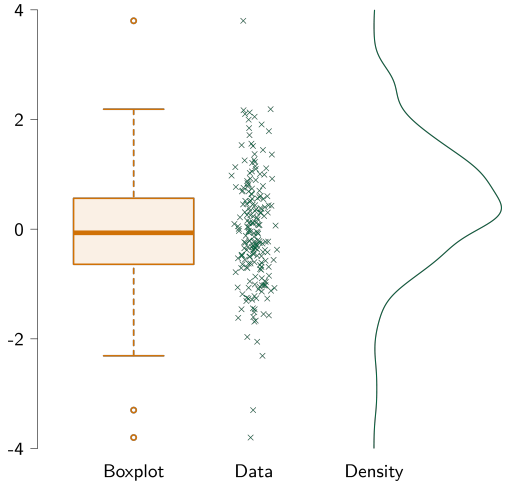
Boxplot


4
2
0
-2
-4
Data values
Outlier (data point beyond the whisker)
Maximum value except outliers
Boxplot
Data

Boxplot


4
2
0
-2
-4
Data values
Outlier (data point beyond the whisker)
Maximum value except outliers
Boxplot
Data

Boxplot

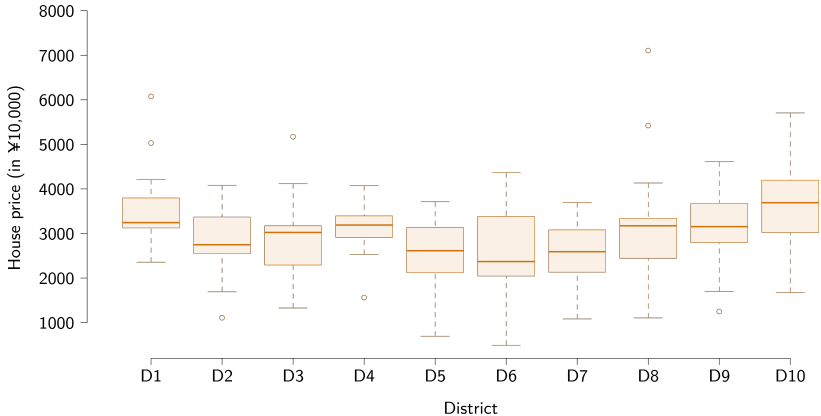


Boxplot

Boxplots can be used to compare multiple groups simultaneously.

Boxplot

Boxplots can be used to compare multiple groups simultaneously.



Scatterplot

Scatterplot

Scatterplots are useful for data with two continuous variables.

Scatterplot

Scatterplots are useful for data with two continuous variables.

Example:
Height (in cm) and weight (in kg) of a sample of students.

Student	$x =$ **height (cm)**	$y =$ **weight (kg)**
1	157	53
2	172	77
⋮	⋮	⋮
57	166	59

Scatterplot

Scatterplots are useful for data with two continuous variables.

Example:
Height (in cm) and weight (in kg) of a sample of students.

Student	$x =$ **height (cm)**	$y =$ **weight (kg)**
1	157	53
2	172	77
⋮	⋮	⋮
57	166	59

In general, there are n pairs of values:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

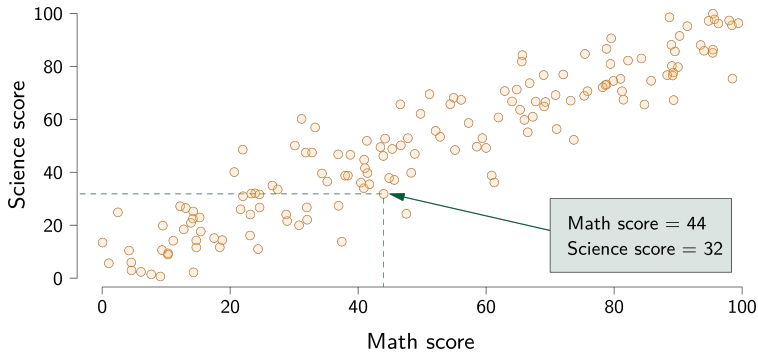
Scatterplot

A scatterplot allows visualizing associations between two variables.

Scatterplot

A scatterplot allows visualizing associations between two variables.

Example:
Math score (x-axis) versus science score (y-axis), for a set of 153 students.
Each dot corresponds to one student.

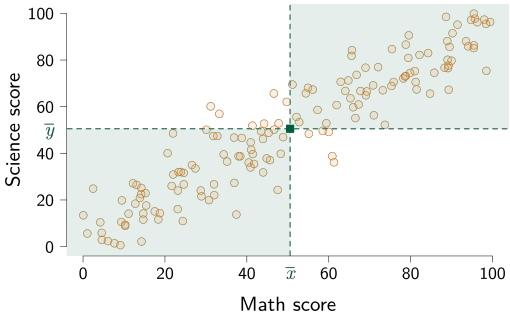


Scatterplot — Interpretation

1. Divide the scatterplot into the four sections determined by the mean scores of the two variables.
2. If most points lie in the…
   - …top-right and bottom-left panels $\Longrightarrow$ positive linear association.
   - …top-left and bottom-right panels $\Longrightarrow$ negative linear association.

Scatterplot — Interpretation

1. Divide the scatterplot into the four sections determined by the mean scores of the two variables.
2. If most points lie in the…
   - …top-right and bottom-left panels $\Longrightarrow$ positive linear association.
   - …top-left and bottom-right panels $\Longrightarrow$ negative linear association.



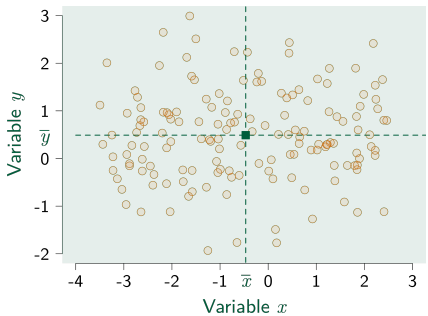
Positive linear association:

On average, increasing math scores are linearly associated with increasing science scores.

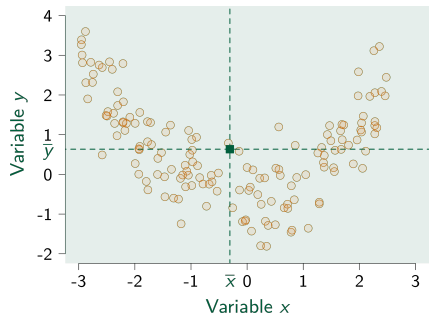
Scatterplot — Interpretation

In some cases the points are scattered across all four sections.
The reason may be:

Lack of linear association:



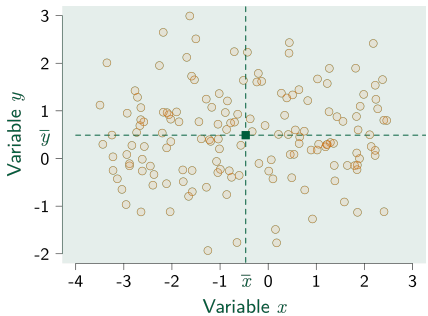
Non-linear association:



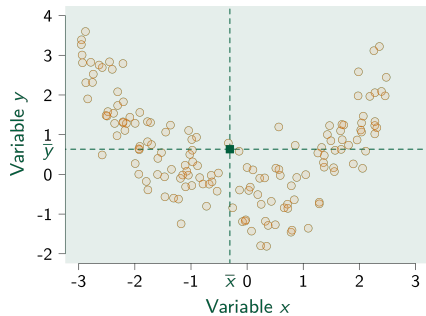
Scatterplot — Interpretation

In some cases the points are scattered across all four sections.
The reason may be:

Lack of linear association:



Non-linear association:



This is why plotting your data is so important:
You may be able to find information in the plot that numerical summaries **don't show**!

Summary

Today we learned about:

Descriptive statistics — measures of spread:

- Variance
- Standard deviation
- Range
- Quartiles
- Interquartile range

Plotting data:

- Boxplot
- Scatterplot (for two variables)

Summary

Today we learned about:

Descriptive statistics — measures of spread:

- Variance
- Standard deviation
- Range
- Quartiles
- Interquartile range

Plotting data:

- Boxplot
- Scatterplot (for two variables)

Use **both** descriptive statistics and plots to *learn*, *communicate*, and *compare* distributions!

To do by the next lecture

Before we start with Lecture 5, make sure you do the following:

- Create a folder called "Stat" on your desktop.
- Download the data file called `seiseki.csv` from *Moodle*, "Lecture 5".
- Save the data file in folder "Stat" on your desktop.
- Register your free account at RStudio Cloud (Posit Cloud).
  See Lecture 1 for instructions.
- Do watch the following clip so that you learn the basics from Posit Cloud:
  https://vimeo.com/913207949