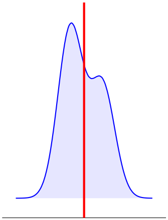
American University of Beirut
STAT 210: Elementary Statistics for Sciences
2022–2023 Fall

Siamak Taati

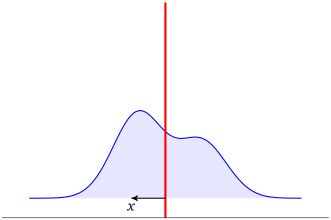
Chapter 3
Measures of center and variation
Part 3

Measures of center and variation

Measures of spread: variance and standard deviation



narrowly spread



widely spreads

The variance and standard deviation measure the "typical" deviation of the data values from the mean.

$$\text{deviation from the population mean} = x - \mu$$
$$\text{deviation from the sample mean} = x - \bar{x}$$

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

population variance of x : $\quad \sigma^2 := \frac{\sum (x - \mu)^2}{N}$

sample variance of x : $\quad s^2 := \frac{\sum (x - \bar{x})^2}{n - 1}$

where N is the population size and n is the sample size.

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

population variance of x : $\sigma^2 := \dfrac{\sum (x-\mu)^2}{N}$

sample variance of x : $s^2 := \dfrac{\sum (x-\bar{x})^2}{n-1}$

where N is the population size and n is the sample size.

An alternative way to calculate the variances:

population variance of x : $\sigma^2 = \dfrac{\sum x^2 - \left(\sum x\right)^2 / N}{N}$

sample variance of x : $s^2 := \dfrac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1}$

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

population variance of x : $$\sigma^2 := \frac{\sum (x-\mu)^2}{N}$$

sample variance of x : $$s^2 := \frac{\sum (x-\bar{x})^2}{n-1}$$

where N is the population size and n is the sample size.

An alternative way to calculate the variances:

population variance of x : $$\sigma^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / N}{N}$$

sample variance of x : $$s^2 := \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1}$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

$$1 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \quad 1 \quad 5 \quad 3 \quad 3$$

value	frequency
1	2
3	2
4	1
5	5

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

$$1 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \quad 1 \quad 5 \quad 3 \quad 3$$

value	frequency
1	2
3	2
4	1
5	5

Q What is the sample variance?

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

$$1 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \quad 1 \quad 5 \quad 3 \quad 3$$

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Ⓠ What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{\begin{gathered}(1-3.7)^2 + (4-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 \\ + (5-3.7)^2 + (1-3.7)^2 + (5-3.7)^2 + (3-3.7)^2 + (3-3.7)^2\end{gathered}}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Ⓠ What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{\begin{array}{c}(1-3.7)^2+(4-3.7)^2+(5-3.7)^2+(5-3.7)^2+(5-3.7)^2\\ +(5-3.7)^2+(1-3.7)^2+(5-3.7)^2+(3-3.7)^2+(3-3.7)^2\end{array}}{10-1} \approx \boxed{2.6778}.$$

Method 2 (definition + frequency table):

$$s^2 = \frac{2\times(1-3.7)^2+2\times(3-3.7)^2+1\times(4-3.7)^2+5\times(5-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{(1-3.7)^2 + (4-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (1-3.7)^2 + (5-3.7)^2 + (3-3.7)^2 + (3-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Method 2 (definition + frequency table):

$$s^2 = \frac{2 \times (1-3.7)^2 + 2 \times (3-3.7)^2 + 1 \times (4-3.7)^2 + 5 \times (5-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{\begin{array}{c}(1-3.7)^2+(4-3.7)^2+(5-3.7)^2+(5-3.7)^2+(5-3.7)^2\\ +(5-3.7)^2+(1-3.7)^2+(5-3.7)^2+(3-3.7)^2+(3-3.7)^2\end{array}}{10-1} \approx \boxed{2.6778}.$$

Method 2 (definition + frequency table):

$$s^2 = \frac{2\times(1-3.7)^2+2\times(3-3.7)^2+1\times(4-3.7)^2+5\times(5-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{\begin{array}{c}(1-3.7)^2 + (4-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 \\ + (5-3.7)^2 + (1-3.7)^2 + (5-3.7)^2 + (3-3.7)^2 + (3-3.7)^2\end{array}}{10-1} \approx \boxed{2.6778}.$$

Method 2 (definition + frequency table):

$$s^2 = \frac{2 \times (1-3.7)^2 + 2 \times (3-3.7)^2 + 1 \times (4-3.7)^2 + 5 \times (5-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 1 (definition):

$$s^2 = \frac{(1-3.7)^2 + (4-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (5-3.7)^2 + (1-3.7)^2 + (5-3.7)^2 + (3-3.7)^2 + (3-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Method 2 (definition + frequency table):

$$s^2 = \frac{2 \times (1-3.7)^2 + 2 \times (3-3.7)^2 + 1 \times (4-3.7)^2 + 5 \times (5-3.7)^2}{10-1} \approx \boxed{2.6778}.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 3 (alternative expression):

$$\sum x =$$

$$(\sum x)^2 =$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n-1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 3 (alternative expression):

$$\sum x = 1 + 4 + 5 + 5 + 5 + 5 + 1 + 5 + 3 + 3 = 37$$

$$\left(\sum x\right)^2 =$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n - 1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 3 (alternative expression):

$$\sum x = 1+4+5+5+5+5+1+5+3+3 = 37$$

$$(\sum x)^2 = 37^2 = 1369$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 3 (alternative expression):

$$\sum x = 1 + 4 + 5 + 5 + 5 + 5 + 1 + 5 + 3 + 3 = 37$$

$$(\sum x)^2 = 37^2 = 1369$$

$$\sum x^2 = 1^2 + 4^2 + 5^2 + 5^2 + 5^2 + 5^2 + 1^2 + 5^2 + 3^2 + 3^2 = 161$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n - 1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 3 (alternative expression):

$$\sum x = 1+4+5+5+5+5+1+5+3+3 = 37$$

$$\left(\sum x\right)^2 = 37^2 = 1369$$

$$\sum x^2 = 1^2+4^2+5^2+5^2+5^2+5^2+1^2+5^2+3^2+3^2 = 161$$

$$s^2 = \frac{\sum x^2 - \left(\sum x\right)^2/n}{n-1} = \frac{161 - 1369/10}{9} \approx \boxed{2.6778}\,.$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Ⓠ What is the sample variance?

Method 4 (alternative expression+ frequencies):

$$\sum x =$$

$$(\sum x)^2 =$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Ⓠ What is the sample variance?

Method 4 (alternative expression+ frequencies):

$$\sum x = 2 \times 1 + 2 \times 3 + 1 \times 4 + 5 \times 5 = 37$$

$$(\sum x)^2 =$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n-1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 4 (alternative expression+ frequencies):

$$\sum x = 2 \times 1 + 2 \times 3 + 1 \times 4 + 5 \times 5 = 37$$

$$\left(\sum x\right)^2 = 37^2 = 1369$$

$$\sum x^2 =$$

$$s^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Q What is the sample variance?

Method 4 (alternative expression+ frequencies):

$$\sum x = 2 \times 1 + 2 \times 3 + 1 \times 4 + 5 \times 5 = 37$$

$$(\sum x)^2 = 37^2 = 1369$$

$$\sum x^2 = 2 \times 1^2 + 2 \times 3^2 + 1 \times 4^2 + 5 \times 5^2 = 161$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n - 1} =$$

Computing variance via frequencies

Example

A list of values and its corresponding frequency table:

1 4 5 5 5 5 1 5 3 3

value	frequency
1	2
3	2
4	1
5	5

We calculated before: $\boxed{\text{mean} = 3.7}$.

Ⓠ What is the sample variance?

Method 4 (alternative expression+ frequencies):

$$\sum x = 2 \times 1 + 2 \times 3 + 1 \times 4 + 5 \times 5 = 37$$

$$(\sum x)^2 = 37^2 = 1369$$

$$\sum x^2 = 2 \times 1^2 + 2 \times 3^2 + 1 \times 4^2 + 5 \times 5^2 = 161$$

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n-1} = \frac{161 - 1369/10}{9} \approx \boxed{2.6778}.$$

Coefficient of variation

Scenario

A group of marine biologists are studying two species of fish. The population mean and standard deviation of the length of the fish in the two species are:

Species 1	Species 2
$\mu_1 = 5\,\text{cm}$	$\mu_2 = 20\,\text{cm}$
$\sigma_1 = 0.7\,\text{cm}$	$\sigma_2 = 2\,\text{cm}$

Coefficient of variation

Scenario

A group of marine biologists are studying two species of fish. The population mean and standard deviation of the length of the fish in the two species are:

Species 1	Species 2
$\mu_1 = 5\,\text{cm}$	$\mu_2 = 20\,\text{cm}$
$\sigma_1 = 0.7\,\text{cm}$	$\sigma_2 = 2\,\text{cm}$

Q The length of which species of fish has more variations?

Coefficient of variation

Scenario

A group of marine biologists are studying two species of fish. The population mean and standard deviation of the length of the fish in the two species are:

Species 1	Species 2
$\mu_1 = 5\,\text{cm}$	$\mu_2 = 20\,\text{cm}$
$\sigma_1 = 0.7\,\text{cm}$	$\sigma_2 = 2\,\text{cm}$

Q The length of which species of fish has more variations?

$\longrightarrow$ The absolute standard deviation of the length is larger for the 2nd species,

Coefficient of variation

Scenario

A group of marine biologists are studying two species of fish. The population mean and standard deviation of the length of the fish in the two species are:

Species 1	Species 2
$\mu_1 = 5\,\text{cm}$	$\mu_2 = 20\,\text{cm}$
$\sigma_1 = 0.7\,\text{cm}$	$\sigma_2 = 2\,\text{cm}$

Q The length of which species of fish has more variations?

$\longrightarrow$ The absolute standard deviation of the length is larger for the 2nd species, but relative to the mean length, there is more variation within the 1st species:

$$\frac{\sigma_1}{\mu_1} = \frac{0.7}{5} = \boxed{14\%} \qquad \frac{\sigma_2}{\mu_2} = \frac{2}{20} = \boxed{10\%}$$

Coefficient of variation

The coefficient of variation of a variable x is the ratio of its standard deviation over its mean:

$$\text{population CV} := \frac{\sigma}{\mu}$$
$$\text{sample CV} := \frac{s}{\bar{x}}$$

and is often written in percentage format.

Coefficient of variation

The coefficient of variation of a variable x is the ratio of its standard deviation over its mean:

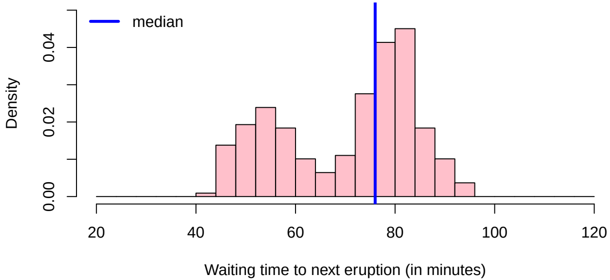
$$\text{population CV} := \frac{\sigma}{\mu}$$
$$\text{sample CV} := \frac{s}{\overline{x}}$$

and is often written in percentage format.

$$\text{population CV for species 1} = 14\%$$
$$\text{population CV for species 2} = 10\%$$

Measures of position: quartiles

Recall that median of a list of data values divides the values into two halves: those larger than the median and those smaller than the median.

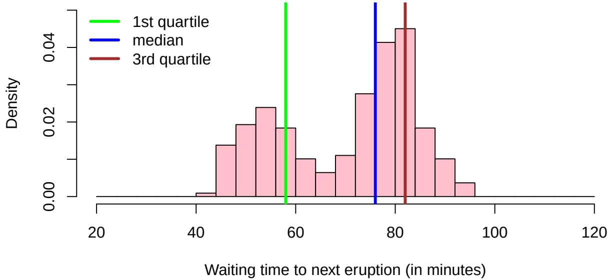


Old Faithful geyser

Measures of position: quartiles

Recall that median of a list of data values divides the values into two halves: those larger than the median and those smaller than the median.

The quartiles further divide the data values into quarters.



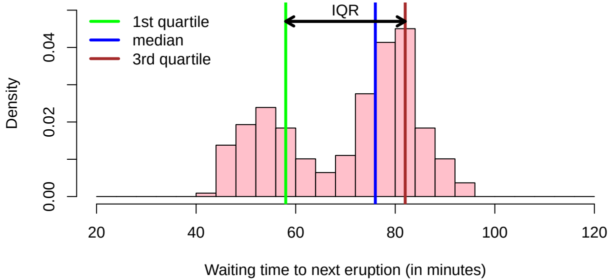
Old Faithful geyser

Measures of position: quartiles

Recall that median of a list of data values divides the values into two halves: those larger than the median and those smaller than the median.

The quartiles further divide the data values into quarters.

The difference between the 1st and the 3rd quartiles is a measure of variation. It is called the inter-quartile range (IQR).



Old Faithful geyser

Measures of position: quartiles

The quartiles partition the data values into three equal parts.

- The 2nd quartile (denoted by Q_2) is the same as the median.
- The 1st quartile (denoted by Q_1) is the median of the data values that are smaller than the median.
- The 3rd quartile (denoted by Q_3) is the median of the data values that are larger than the median.

The inter-quartile range is the difference between the 1st and the 3rd quartiles:

$$\mathrm{IQR} := Q_3 - Q_1 \ .$$

It is a measure of spread.

Measures of position: quartiles

Back to the birds data set . . .

Weight of females

98.8	92.5	87.3	103.5	98.3	99.1	96.9	103.0	96.5	93.4
102.9	101.1	100.9	95.6	96.1	99.3	93.3	95.4	88.5	97.6
92.2	98.0								

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Measures of position: quartiles

Back to the birds data set ...

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

$$Q_2 = \text{median} = \frac{96.9 + 97.6}{2} = \boxed{97.25} \text{ grams}$$

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

$$Q_2 = \text{median} = \frac{96.9 + 97.6}{2} = \boxed{97.25} \text{ grams}$$

$$Q_1 = \boxed{93.4} \text{ grams}$$

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

$$Q_2 = \text{median} = \frac{96.9 + 97.6}{2} = \boxed{97.25}\ \text{grams}$$

$$Q_1 = \boxed{93.4}\ \text{grams}$$

$$Q_3 = \boxed{99.3}\ \text{grams}$$

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

$$Q_2 = \text{median} = \frac{96.9 + 97.6}{2} = \boxed{97.25} \text{ grams}$$

$$Q_1 = \boxed{93.4} \text{ grams}$$

$$Q_3 = \boxed{99.3} \text{ grams}$$

$$\text{IQR} = Q_3 - Q_1 = 99.3 - 93.4 = \boxed{5.9} \text{ grams}$$

Measures of position: quartiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

$$Q_2 = \text{median} = \frac{96.9 + 97.6}{2} = \boxed{97.25} \text{ grams}$$

$$Q_1 = \boxed{93.4} \text{ grams}$$

$$Q_3 = \boxed{99.3} \text{ grams}$$

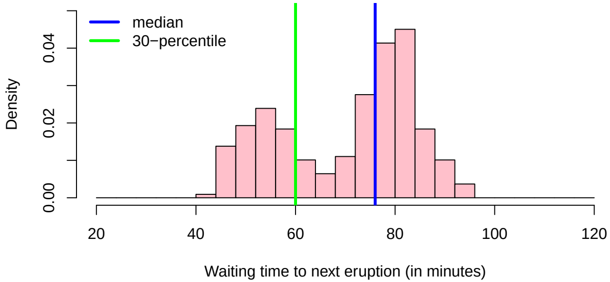
$$\text{IQR} = Q_3 - Q_1 = 99.3 - 93.4 = \boxed{5.9} \text{ grams}$$

Exercise: Compare the IQR for the sampled male and female birds.

Measures of position: percentiles

The percentiles are similar to quartiles, but with other ratios.

The 30-percentile divides the data values into the lower 30% and the upper 70%.

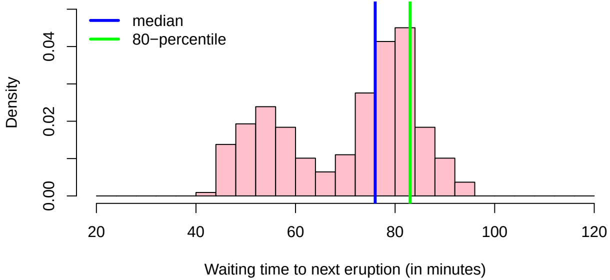


Old Faithful geyser

Measures of position: percentiles

The percentiles are similar to quartiles, but with other ratios.

The 30-percentile divides the data values into the lower 30% and the upper 70%.



Old Faithful geyser

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

- Sort all the data values.

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

- Sort all the data values.
- Let k be $\frac{30}{100} \times n$ rounded up to the next whole number.

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

- Sort all the data values.
- Let k be $\frac{30}{100} \times n$ rounded up to the next whole number.
- The 30-percentile is the k-th element in the sorted list.

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

- Sort all the data values.
- Let k be $\frac{30}{100} \times n$ rounded up to the next whole number.
- The 30-percentile is the k-th element in the sorted list.

Remarks

- 25-percentile $\approx Q_1$.
- 50-percentile $\approx Q_2 =$ median.
- 75-percentile $\approx Q_3$.

Measures of position: percentiles

The 30-percentile of a list of data values is a value q with the following property:

- 30-percent of the data values are smaller than (or equal to) q 70-percent of them are larger (or equal to) than q.

The 30-percentile of a list of n data values can be calculated as follows:

- Sort all the data values.
- Let k be $\frac{30}{100} \times n$ rounded up to the next whole number.
- The 30-percentile is the k-th element in the sorted list.

Remarks

- 25-percentile $\approx Q_1$.
- 50-percentile $\approx Q_2 =$ median.
- 75-percentile $\approx Q_3$.
- In our convention, the values might be somewhat different!

Measures of position: percentiles

Back to the birds data set ...

Weight of females

98.8	92.5	87.3	103.5	98.3	99.1	96.9	103.0	96.5	93.4
102.9	101.1	100.9	95.6	96.1	99.3	93.3	95.4	88.5	97.6
92.2	98.0								

Q What is the 30-percentile of the data values?

Measures of position: percentiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

Measures of position: percentiles

Back to the birds data set . . .

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

$\longrightarrow$ $\frac{30}{100} \times 22 = 6.6$, which is rounded up to 7.

Measures of position: percentiles

Back to the birds data set ...

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

$\longrightarrow$ $\frac{30}{100} \times 22 = 6.6$, which is rounded up to 7.

$\longrightarrow$ 30-percentile $= \boxed{95.4}$ grams.

Measures of position: percentiles

Back to the birds data set ...

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

$\longrightarrow$ $\frac{30}{100} \times 22 = 6.6$, which is rounded up to 7.

$\longrightarrow$ 30-percentile $= \boxed{95.4}$ grams.

Q What is the percentile rank of 98.3?

Measures of position: percentiles

Back to the birds data set ...

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

$\longrightarrow$ $\frac{30}{100} \times 22 = 6.6$, which is rounded up to 7.

$\longrightarrow$ 30-percentile $= \boxed{95.4}$ grams.

Q What is the percentile rank of 98.3?

$\longrightarrow$ There are 13 values smaller than 98.3.

Measures of position: percentiles

Back to the birds data set ...

Weight of females (sorted)

87.3	88.5	92.2	92.5	93.3	93.4	95.4	95.6	96.1	96.5
96.9	97.6	98.0	98.3	98.8	99.1	99.3	100.9	101.1	102.9
103.0	103.5								

Q What is the 30-percentile of the data values?

$\longrightarrow$ Sort the values!

$\longrightarrow$ $\frac{30}{100} \times 22 = 6.6$, which is rounded up to 7.

$\longrightarrow$ 30-percentile $= \boxed{95.4}$ grams.

Q What is the percentile rank of 98.3?

$\longrightarrow$ There are 13 values smaller than 98.3.

$\longrightarrow$ percentile rank of 98.3 $= \frac{13}{22} \approx \boxed{59.1\%}$.

Measures of position: percentiles

The percentile rank of a value q in a list of n data values is the percentage of the values in the list that are smaller than q:

$$(\text{percentile rank of } q) := (\text{percentage of data values} < q)$$

Measures of position: percentiles

The percentile rank of a value q in a list of n data values is the percentage of the values in the list that are smaller than q:

$$(\text{percentile rank of } q) := (\text{percentage of data values} < q)$$

In other words, the percentage rank of q is simply the cumulative percentage at q.

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow \frac{2}{10} = \boxed{20\%}$.

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow \frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow$ $\frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

$\longrightarrow$ $\frac{1}{10} = \boxed{10\%}$.

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow$ $\frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

$\longrightarrow$ $\frac{1}{10} = \boxed{10\%}$.

Q What is the 20-percentile?

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow$ $\frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

$\longrightarrow$ $\frac{1}{10} = \boxed{10\%}$.

Q What is the 20-percentile?

$\longrightarrow$ $\frac{20}{100} \times 10 = 2$, which does not require rounding.

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow$ $\frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

$\longrightarrow$ $\frac{1}{10} = \boxed{10\%}$.

Q What is the 20-percentile?

$\longrightarrow$ $\frac{20}{100} \times 10 = 2$, which does not require rounding.

$\longrightarrow$ Hence, the 20-percentile is $\boxed{-4}$.

Measures of position: percentiles

Remark (slight inconsistency)

Consider the following (sorted) list of 10 data values:

$$-5 \quad -4 \quad -3 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 4$$

Q What is the percentile rank of -3?

$\longrightarrow$ $\frac{2}{10} = \boxed{20\%}$.

Q What is the percentile rank of -4?

$\longrightarrow$ $\frac{1}{10} = \boxed{10\%}$.

Q What is the 20-percentile?

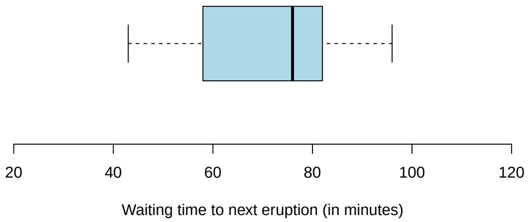
$\longrightarrow$ $\frac{20}{100} \times 10 = 2$, which does not require rounding.

$\longrightarrow$ Hence, the 20-percentile is $\boxed{-4}$.

This inconsistency is partly due to our (simplifying) convention, and partly inevitable.

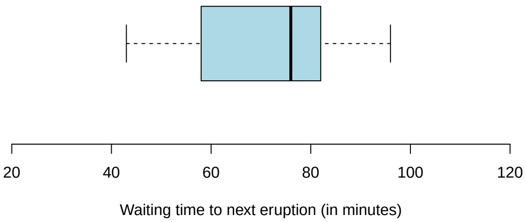
Summarizing and visualizing data: box plots

Old Faithful geyser



Summarizing and visualizing data: box plots

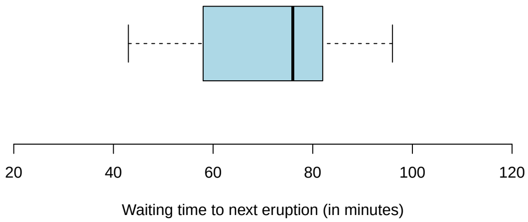
Old Faithful geyser



Ⓠ What does the thick vertical line stand for?

Summarizing and visualizing data: box plots

Old Faithful geyser

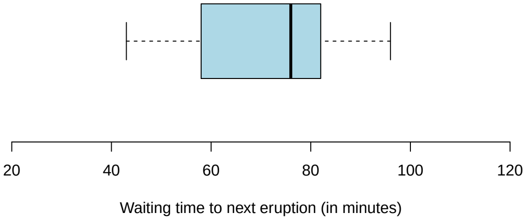


Ⓠ What does the thick vertical line stand for?

$\longrightarrow$ Median

Summarizing and visualizing data: box plots

Old Faithful geyser



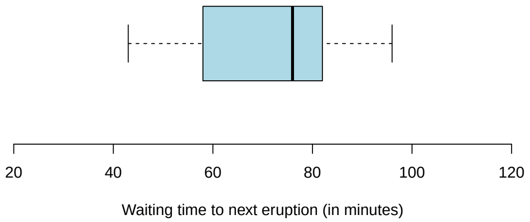
Ⓠ What does the thick vertical line stand for?

$\longrightarrow$ Median

Ⓠ What does the box stand for?

Summarizing and visualizing data: box plots

Old Faithful geyser



Ⓠ What does the thick vertical line stand for?

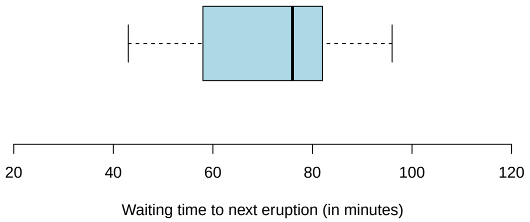
$\longrightarrow$ Median

Ⓠ What does the box stand for?

$\longrightarrow$ The 1st and the 3rd quartiles

Summarizing and visualizing data: box plots

Old Faithful geyser



Ⓠ What does the thick vertical line stand for?

$\longrightarrow$ Median

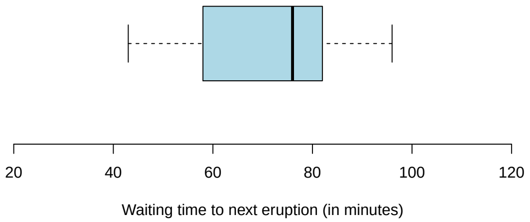
Ⓠ What does the box stand for?

$\longrightarrow$ The 1st and the 3rd quartiles

Ⓠ What do the whiskers stand for?

Summarizing and visualizing data: box plots

Old Faithful geyser



Ⓠ What does the thick vertical line stand for?

$\longrightarrow$ Median

Ⓠ What does the box stand for?

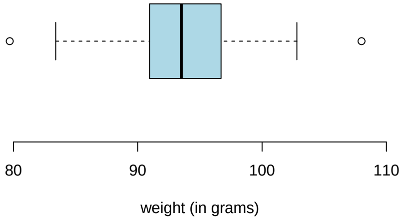
$\longrightarrow$ The 1st and the 3rd quartiles

Ⓠ What do the whiskers stand for?

$\longrightarrow$ The minimum and the maximum

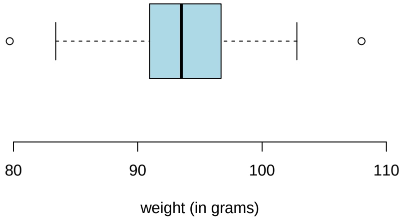
Summarizing and visualizing data: box plots

Sampled male birds



Summarizing and visualizing data: box plots

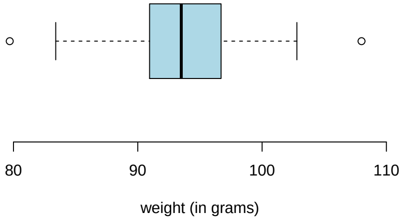
Sampled male birds



Q What do the small circles stand for?

Summarizing and visualizing data: box plots

Sampled male birds

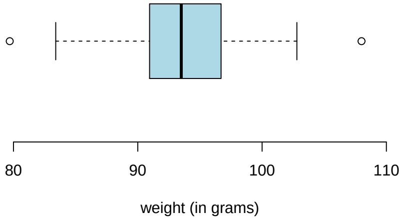


Q What do the small circles stand for?

$\longrightarrow$ The outliers [Mild outliers; Tukey's convention]

Summarizing and visualizing data: box plots

Sampled male birds



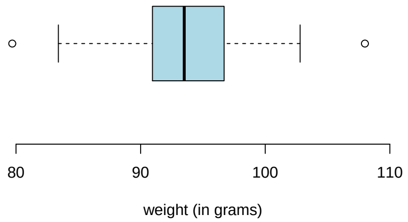
Ⓠ What do the small circles stand for?

⟶ The outliers [Mild outliers; Tukey's convention]

Ⓠ What do the whiskers stand for?

Summarizing and visualizing data: box plots

Sampled male birds



Ⓠ What do the small circles stand for?

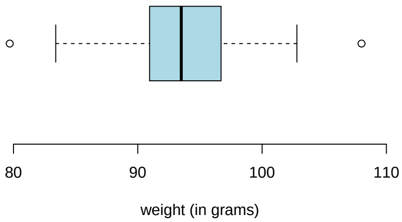
⟶ The outliers [Mild outliers; Tukey's convention]

Ⓠ What do the whiskers stand for?

⟶ The minimum and the maximum excluding the outliers

Summarizing and visualizing data: box plots

Sampled male birds

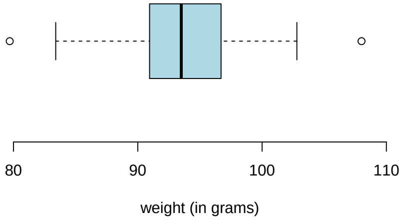


Identifying outliers (Tukey's convension)

- Outliers are those observations that are farther than 1.5 IQR from the box.

Summarizing and visualizing data: box plots

Sampled male birds

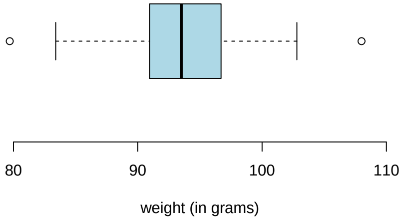


Identifying outliers (Tukey's convension)

- Outliers are those observations that are farther than 1.5 IQR from the box.
- The observations that are farther than 3 IQR from the box are considered extreme outliers.

Summarizing and visualizing data: box plots

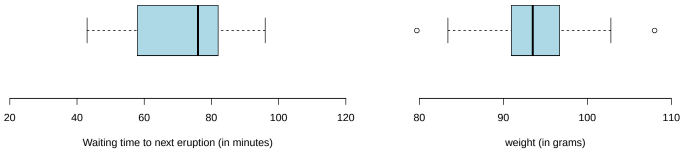
Sampled male birds



Identifying outliers (Tukey's convension)

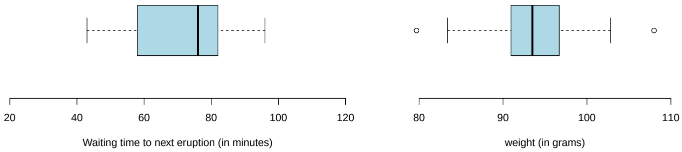
- Outliers are those observations that are farther than 1.5 IQR from the box.
- The observations that are farther than 3 IQR from the box are considered extreme outliers.
- The non-extreme outliers are considered mild outliers.

Summarizing and visualizing data: box plots



The box plot, visually summarizes:

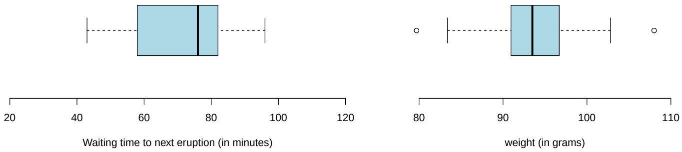
Summarizing and visualizing data: box plots



The box plot, visually summarizes:

- The center (median)

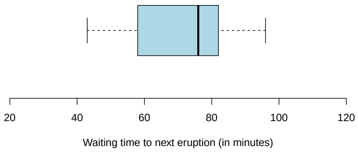
Summarizing and visualizing data: box plots

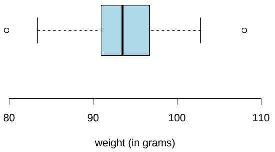


The box plot, visually summarizes:

- The center (median)
- The spread (IQR)

Summarizing and visualizing data: box plots

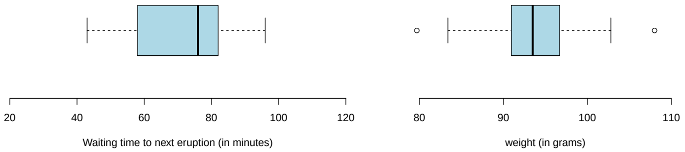




The box plot, visually summarizes:

- The center (median)
- The spread (IQR)
- The skew

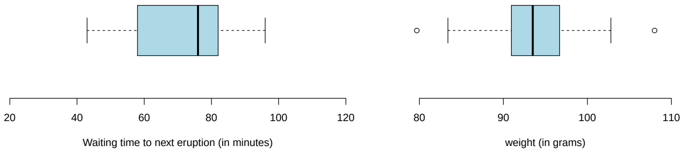
Summarizing and visualizing data: box plots



The box plot, visually summarizes:

- The center (median)
- The spread (IQR)
- The skew
  - $Q_3 - Q_2$ vs. $Q_2 - Q_1$
  - Distances of the left and right whiskers from the box

Summarizing and visualizing data: box plots



The box plot, visually summarizes:

- The center (median)
- The spread (IQR)
- The skew
  - $Q_3 - Q_2$ vs. $Q_2 - Q_1$
  - Distances of the left and right whiskers from the box
- The outliers