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

Siamak Taati

Chapter 3
Measures of center and variation
Part 2

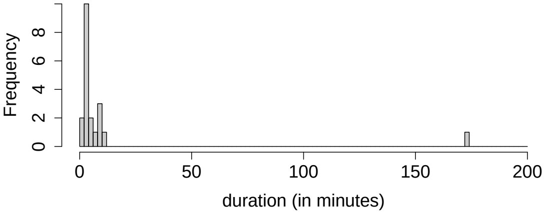
Measures of center and variation

Outliers

Example

Researchers collect data on a sample of 20 phone calls.
Duration of the sampled calls (in minutes):

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21

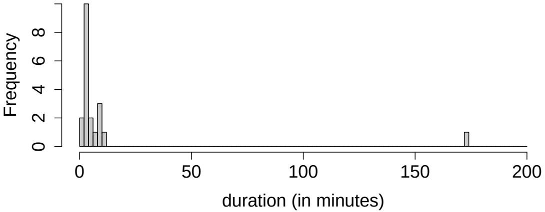


Outliers

Example

Researchers collect data on a sample of 20 phone calls.
Duration of the sampled calls (in minutes):

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21

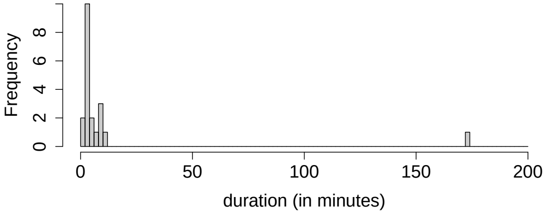


Outliers

Example

Researchers collect data on a sample of 20 phone calls.
Duration of the sampled calls (in minutes):

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21



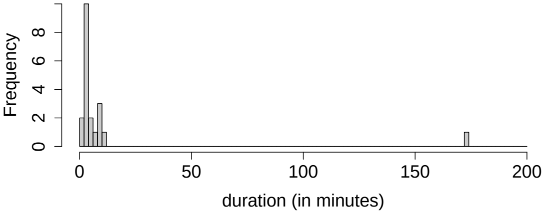
The case with 173.10 min duration is an example of an outlier.

Outliers

Example

Researchers collect data on a sample of 20 phone calls.
Duration of the sampled calls (in minutes):

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21



The case with 173.10 min duration is an example of an outlier.

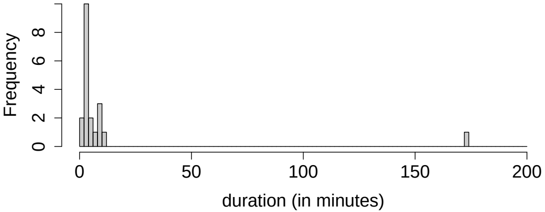
Q What is the reason behind this case?

Outliers

Example

Researchers collect data on a sample of 20 phone calls.
Duration of the sampled calls (in minutes):

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21



The case with 173.10 min duration is an example of an outlier.

Q What is the reason behind this case?

Q Should the researchers keep it or remove it?

Outliers

An outlier in a data set is a case which is exceptionally different from the rest of the data set (i.e., has values which are either much larger, or much smaller than the rest).

Outliers

An outlier in a data set is a case which is exceptionally different from the rest of the data set (i.e., has values which are either much larger, or much smaller than the rest).

Ⓠ Where do outliers come from?

Outliers

An outlier in a data set is a case which is exceptionally different from the rest of the data set (i.e., has values which are either much larger, or much smaller than the rest).

Ⓠ Where do outliers come from?

$\longrightarrow$ Outliers could occur due to various reasons (e.g., error in measurement or data collection, unusual cases in the population, or sheer chance).

Outliers

An outlier in a data set is a case which is exceptionally different from the rest of the data set (i.e., has values which are either much larger, or much smaller than the rest).

Ⓠ Where do outliers come from?

⟶ Outliers could occur due to various reasons (e.g., error in measurement or data collection, unusual cases in the population, or sheer chance).

Ⓠ What to do with the outliers?

Outliers

An outlier in a data set is a case which is exceptionally different from the rest of the data set (i.e., has values which are either much larger, or much smaller than the rest).

Ⓠ Where do outliers come from?

⟶ Outliers could occur due to various reasons (e.g., error in measurement or data collection, unusual cases in the population, or sheer chance).

Ⓠ What to do with the outliers?

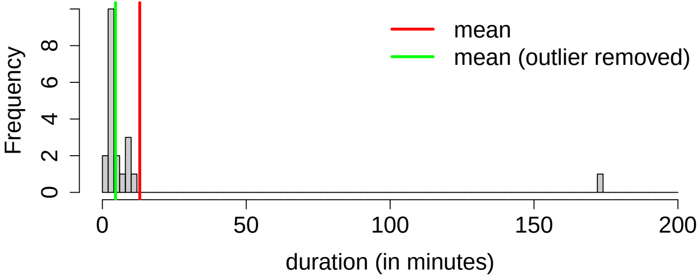
⟶ We should always detect them and examine the source of their occurrence. Depending on the source and the application, we may then decide to keep or delete them from data set.

Sensitivity to outliers

Q How do outliers affect mean and median?

Sensitivity to outliers

Ⓠ How do outliers affect mean and median?



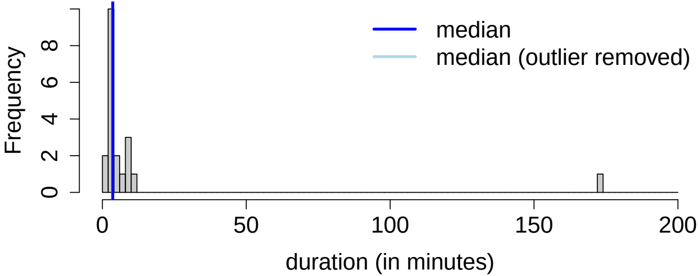
$\longrightarrow$ Mean is sensitve to outliers. [It is "pulled" by the outliers.]

$$\text{mean with outlier} \approx 12.985\,\text{min}$$

$$\text{mean without outlier} \approx 4.557\,\text{min}$$

Sensitivity to outliers

Ⓠ How do outliers affect mean and median?



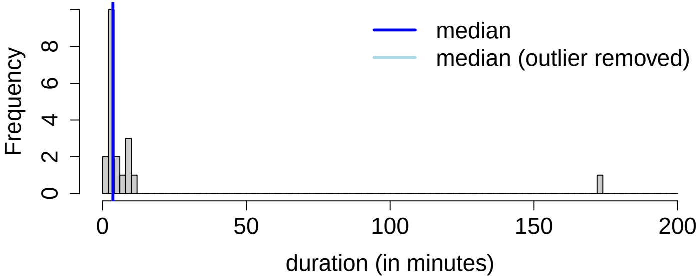
$\longrightarrow$ Median is resistant to outliers.

$$\text{median with outlier} = 3.605\,\text{min}$$

$$\text{median without outlier} = 3.58\,\text{min}$$

Sensitivity to outliers

Ⓠ How do outliers affect mean and median?



⟶ Median is resistant to outliers.

As a measure of the center, median is more robust than mean.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Example

For the duration of the calls,

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21

The mean is (approximately) 12.985 min.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Example

For the duration of the calls,

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21

The mean is (approximately) 12.985 min.
To find the 5%-trimmed mean, note that 5% of 20 is 1. So, we remove the largest value and the smallest value, and take the average of the rest.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Example

For the duration of the calls,

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	173.10	3.29	3.34	1.32	2.75	8.41	3.21	10.51	2.21

The mean is (approximately) 12.985 min.
To find the 5%-trimmed mean, note that 5% of 20 is 1. So, we remove the largest value and the smallest value, and take the average of the rest.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Example

For the duration of the calls,

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	~~173.10~~	3.29	3.34	~~1.32~~	2.75	8.41	3.21	10.51	2.21

The mean is (approximately) 12.985 min.
To find the 5%-trimmed mean, note that 5% of 20 is 1. So, we remove the largest value and the smallest value, and take the average of the rest.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Example

For the duration of the calls,

1.60	4.02	3.58	8.77	9.84	4.23	2.25	3.63	7.12	2.86
3.65	~~173.10~~	3.29	3.34	~~1.32~~	2.75	8.41	3.21	10.51	2.21

The mean is (approximately) 12.985 min.
To find the 5%-trimmed mean, note that 5% of 20 is 1. So, we remove the largest value and the smallest value, and take the average of the rest. The result is (approximately) 4.737 min.

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \dots, x_n$$

is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Q How do outliers affect trimmed mean?

Sensitivity to outliers

The 5%-trimmed mean of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

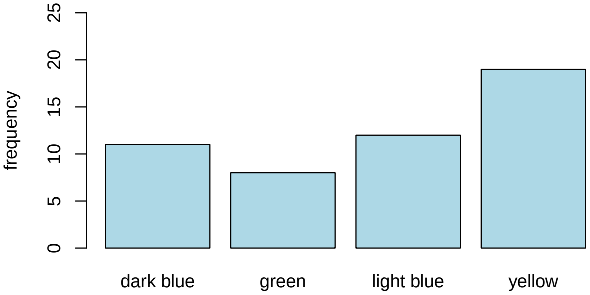
is their mean once we remove the upper 5% and the lower 5% of the data values. In other words, once we sort the values from smallest to largest, the 5%-trimmed mean is mean of the $\frac{90}{100} \times n$ middle values.

Ⓠ How do outliers affect trimmed mean?

$\longrightarrow$ Unlike mean, trimmed mean is resistant to outliers.

Measures of center: mode

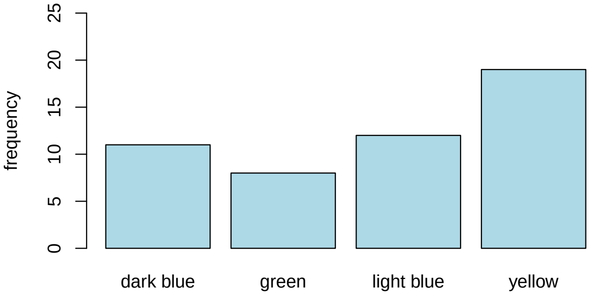
Example (Sample of 50 birds from a species)



Distribution of the color of feather

Measures of center: mode

Example (Sample of 50 birds from a species)



Distribution of the color of feather

The most common feather color, yellow, is called the mode of the feather color within the sampled birds.

Measures of center: mode

The mode of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is the value that occurs with the highest frequency.

Measures of center: mode

The mode of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is the value that occurs with the highest frequency.

Example

What is the mode of the following list of values?

$$2, 3, 2, 5, 1, 5, 4, 4, 2, 2$$

Measures of center: mode

The mode of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is the value that occurs with the highest frequency.

Example

What is the mode of the following list of values?

$$2, 3, 2, 5, 1, 5, 4, 4, 2, 2$$

Answer: 2

Measures of center: mode

The mode of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is the value that occurs with the highest frequency.

Example

What is the mode of the following list of values?

$$2, 3, 2, 5, 1, 5, 4, 4, 2, 2$$

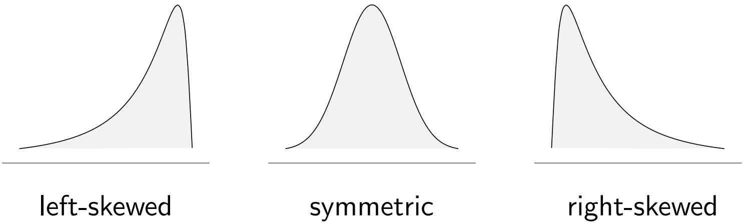
Answer: 2

Advantage and disadvantage of mode

- Disadvantage: meaningful only for unimodal distributions
- Advantage: meaningful also for categorical variables

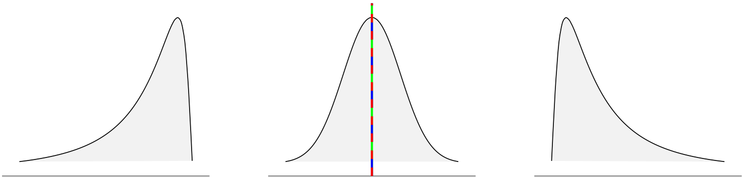
Mean, median, mode in unimodal distributions

Ⓠ What are the relative positions of the mean, median and mode in each of the following histograms?



Mean, median, mode in unimodal distributions

Ⓠ What are the relative positions of the mean, median and mode in each of the following histograms?

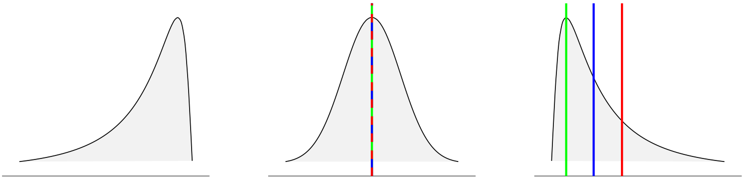


left-skewed symmetric right-skewed

mode: green
median: blue
mean: red

Mean, median, mode in unimodal distributions

Ⓠ What are the relative positions of the mean, median and mode in each of the following histograms?

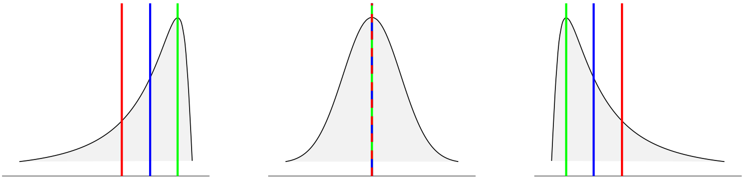


left-skewed symmetric right-skewed

mode:	green
median:	blue
mean:	red

Mean, median, mode in unimodal distributions

Ⓠ What are the relative positions of the mean, median and mode in each of the following histograms?



left-skewed symmetric right-skewed

mode:	green
median:	blue
mean:	red

Computing mean 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

Computing mean 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

Q What is the mean?

Computing mean 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

Q What is the mean?

Method 1 (directly):

$$\text{mean} = \frac{1+4+5+5+5+5+1+5+3+3}{10} = \boxed{3.7}\,.$$

Computing mean 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

Q What is the mean?

Method 1 (directly):

$$\text{mean} = \frac{1+4+5+5+5+5+1+5+3+3}{10} = \boxed{3.7}\,.$$

Method 2 (via the frequency table):

$$\text{mean} = \frac{2\times 1+2\times 3+1\times 4+5\times 5}{2+2+1+5} = \boxed{3.7}\,.$$

Computing mean via frequencies

Example (Sample of 500 families from a country)

no.children	frequency	relative frequency	percentage
0	282	0.564	56.4
1	108	0.216	21.6
2	48	0.096	9.6
3	26	0.052	5.2
4	14	0.028	2.8
5	4	0.008	0.8
6	6	0.012	1.2
7	6	0.012	1.2
8	3	0.006	0.6
10	2	0.004	0.4
11	1	0.002	0.2

Distribution of the number of children

Computing mean via frequencies

Example (Sample of 500 families from a country)

no.children	frequency	relative frequency	percentage
0	282	0.564	56.4
1	108	0.216	21.6
2	48	0.096	9.6
3	26	0.052	5.2
4	14	0.028	2.8
5	4	0.008	0.8
6	6	0.012	1.2
7	6	0.012	1.2
8	3	0.006	0.6
10	2	0.004	0.4
11	1	0.002	0.2

Distribution of the number of children

Q What is the mean number of children per family?

Computing mean via frequencies

Example (Sample of 500 families from a country)

no.children	frequency	relative frequency	percentage
0	282	0.564	56.4
1	108	0.216	21.6
2	48	0.096	9.6
3	26	0.052	5.2
4	14	0.028	2.8
5	4	0.008	0.8
6	6	0.012	1.2
7	6	0.012	1.2
8	3	0.006	0.6
10	2	0.004	0.4
11	1	0.002	0.2

Distribution of the number of children

Q What is the mean number of children per family?

$$\text{mean} = \frac{282 \times 0 + 108 \times 1 + 48 \times 2 + \cdots + 1 \times 11}{282 + 108 + 48 + \cdots + 1} = \boxed{0.982}.$$

Computing mean via frequencies

Example (Sample of 500 families from a country)

no.children	frequency	relative frequency	percentage
0	282	0.564	56.4
1	108	0.216	21.6
2	48	0.096	9.6
3	26	0.052	5.2
4	14	0.028	2.8
5	4	0.008	0.8
6	6	0.012	1.2
7	6	0.012	1.2
8	3	0.006	0.6
10	2	0.004	0.4
11	1	0.002	0.2

Distribution of the number of children

Q What is the mean number of children per family?

$$\text{mean} = \frac{282 \times 0 + 108 \times 1 + 48 \times 2 + \cdots + 1 \times 11}{282 + 108 + 48 + \cdots + 1} = \boxed{0.982}.$$

$$= 0.564 \times 0 + 0.216 \times 1 + 0.096 \times 2 + \cdots + 0.002 \times 11 = \boxed{0.982}.$$

Computing mean via frequencies

Example (Sample of 50 birds from a species)

weight	frequency	relative frequency
75 to <80	1	0.02
80 to <85	1	0.02
85 to <90	6	0.12
90 to <95	14	0.28
95 to <100	20	0.40
100 to <105	7	0.14
105 to <110	1	0.02

Distribution of the weight (in grams)

Computing mean via frequencies

Example (Sample of 50 birds from a species)

weight	frequency	relative frequency
75 to <80	1	0.02
80 to <85	1	0.02
85 to <90	6	0.12
90 to <95	14	0.28
95 to <100	20	0.40
100 to <105	7	0.14
105 to <110	1	0.02

Distribution of the weight (in grams)

Q What is the mean weight among the sampled birds?

Computing mean via frequencies

Example (Sample of 50 birds from a species)

weight	frequency	relative frequency
75 to <80	1	0.02
80 to <85	1	0.02
85 to <90	6	0.12
90 to <95	14	0.28
95 to <100	20	0.40
100 to <105	7	0.14
105 to <110	1	0.02

Distribution of the weight (in grams)

Q What is the mean weight among the sampled birds?

$$\text{mean} \approx \frac{1 \times 77.5 + 1 \times 82.5 + 6 \times 87.5 + \cdots + 1 \times 107.5}{1 + 1 + 6 + \cdots + 1} = \boxed{95.1\,\mathrm{g}}.$$

Computing mean via frequencies

Example (Sample of 50 birds from a species)

weight	frequency	relative frequency
75 to <80	1	0.02
80 to <85	1	0.02
85 to <90	6	0.12
90 to <95	14	0.28
95 to <100	20	0.40
100 to <105	7	0.14
105 to <110	1	0.02

Distribution of the weight (in grams)

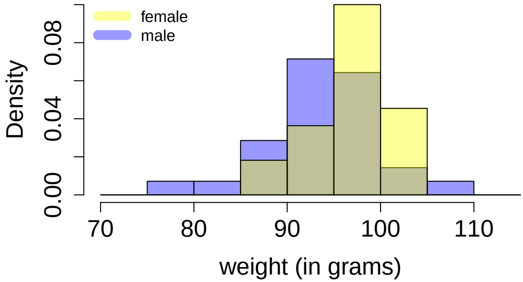
Q What is the mean weight among the sampled birds?

$$\text{mean} \approx \frac{1 \times 77.5 + 1 \times 82.5 + 6 \times 87.5 + \cdots + 1 \times 107.5}{1 + 1 + 6 + \cdots + 1} = \boxed{95.1\,\mathrm{g}}.$$

This is an approximation. The exact value (calculated using the ungrouped data) is 95.09 g.

Descriptive measures

Ⓠ Is there a difference in weight between male and female birds?



Weight of sampled birds

Among the sampled birds,

⟶ females seem to be typically heavier. [center]

⟶ the weight seems to vary more among males. [variation/spread]

Ⓠ Can we quantitatively measure these distinctions?

Measures of center: range

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								

Weight of males (sorted)

79.7	83.4	86.2	87.0	88.5	89.7	90.8	91.1	91.1	92.3
92.3	92.3	92.6	93.3	93.7	94.6	95.8	95.9	96.4	96.5
96.6	96.8	97.0	98.9	99.2	101.8	102.8	108.0		

Measures of center: range

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								

Weight of males (sorted)

79.7	83.4	86.2	87.0	88.5	89.7	90.8	91.1	91.1	92.3
92.3	92.3	92.6	93.3	93.7	94.6	95.8	95.9	96.4	96.5
96.6	96.8	97.0	98.9	99.2	101.8	102.8	108.0		

Measures of center: range

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								

range of weight among sampled female birds

$$= 103.5 - 87.3 = \boxed{16.2} \text{ grams}$$

Weight of males (sorted)

79.7	83.4	86.2	87.0	88.5	89.7	90.8	91.1	91.1	92.3
92.3	92.3	92.6	93.3	93.7	94.6	95.8	95.9	96.4	96.5
96.6	96.8	97.0	98.9	99.2	101.8	102.8	108.0		

Measures of center: range

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								

range of weight among sampled female birds

$$= 103.5 - 87.3 = \boxed{16.2} \text{ grams}$$

Weight of males (sorted)

79.7	83.4	86.2	87.0	88.5	89.7	90.8	91.1	91.1	92.3
92.3	92.3	92.6	93.3	93.7	94.6	95.8	95.9	96.4	96.5
96.6	96.8	97.0	98.9	99.2	101.8	102.8	108.0		

range of weight among sampled male birds

$$= 108.0 - 79.7 = \boxed{28.3} \text{ grams}$$

Measures of spread: range

The range of a list of data values

$$x_1, x_2, x_3, \ldots, x_n$$

is the difference between the largest and the smallest values.

Measures of spread: range

The range of a list of data values

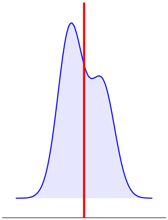
$$x_1, x_2, x_3, \ldots, x_n$$

is the difference between the largest and the smallest values.

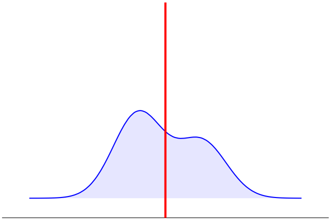
Disadvantages

- Highly sensitive to outliers
- The population range and the sample range can be dramatically different.

Measures of spread: variance and standard deviation



narrowly spread

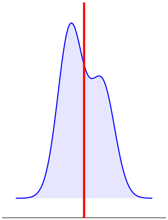


widely spreads

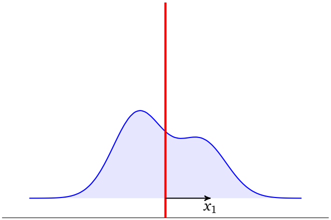
The variance and standard deviation measure the "typical" deviation of the data values from the mean.

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

Measures of spread: variance and standard deviation



narrowly spread

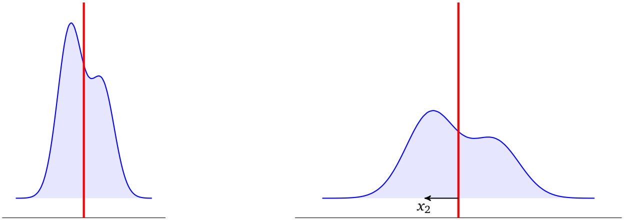


widely spreads

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

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.

Remark

The usage of $n-1$ instead of n is to compensate for a bias caused by using $\bar{x}$ instead of μ.

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

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

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

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

Remark

The usage of $n-1$ instead of n is to compensate for a bias caused by using $\bar{x}$ instead of μ.

Measures of spread: variance

Sampled female birds

Weight:

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 spread: variance

Sampled female birds

Weight:

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								

Weight (deviation from the mean): [sample mean ≈ 96.827 grams]

1.973	−4.327	−9.527	6.673	1.473	2.273	0.073	6.173	−0.327	−3.427
6.073	4.273	4.073	−1.227	−0.727	2.473	−3.527	−1.427	−8.327	0.773
−4.627	1.173								

Measures of spread: variance

Sampled female birds

Weight:

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								

Weight (deviation from the mean): [sample mean $\approx$ 96.827 grams]

1.973	−4.327	−9.527	6.673	1.473	2.273	0.073	6.173	−0.327	−3.427
6.073	4.273	4.073	−1.227	−0.727	2.473	−3.527	−1.427	−8.327	0.773
−4.627	1.173								

Weight (squared deviation from the mean):

3.89	18.73	90.77	44.53	2.17	5.17	0.01	38.10	0.11	11.75
36.88	18.26	16.59	1.51	0.53	6.11	12.44	2.04	69.34	0.60
21.41	1.38								

Measures of spread: variance

Sampled female birds

Weight:

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								

Weight (deviation from the mean): [sample mean ≈ 96.827 grams]

1.973	−4.327	−9.527	6.673	1.473	2.273	0.073	6.173	−0.327	−3.427
6.073	4.273	4.073	−1.227	−0.727	2.473	−3.527	−1.427	−8.327	0.773
−4.627	1.173								

Weight (squared deviation from the mean):

3.89	18.73	90.77	44.53	2.17	5.17	0.01	38.10	0.11	11.75
36.88	18.26	16.59	1.51	0.53	6.11	12.44	2.04	69.34	0.60
21.41	1.38								

sample variance of weight for female birds

$$= \frac{3.89 + 18.73 + 90.77 + \cdots + 1.38}{22 - 1} \approx \boxed{19.1564}\,\text{grams}^2$$

Measures of spread: variance

Sampled birds

sample variance of weight for female birds $\approx \boxed{19.156}$ grams^2

sample variance of weight for male birds $\approx \boxed{34.8049}$ grams^2

Measures of spread: variance

Sampled birds

$$\text{sample variance of weight for female birds} \approx \boxed{19.156}\ \text{grams}^2$$

$$\text{sample variance of weight for male birds} \approx \boxed{34.8049}\ \text{grams}^2$$

A disadvantage of variance

- While the weight is measured in grams, its variance is measured in grams^2.

Measures of spread: standard deviation

Sample vs. population variance

For a numerical variable x,

- The population standard deviation σ is the square root of the population variance.
- The sample standard deviation s is the square root of the sample variance.

Measures of spread: standard deviation

Sample vs. population variance

For a numerical variable x,

- The population standard deviation σ is the square root of the population variance.
- The sample standard deviation s is the square root of the sample variance.

Remarks

- The standard deviation of a variable x is measured in the same units as x itself.

Measures of spread: standard deviation

Sample vs. population variance

For a numerical variable x,

- The population standard deviation σ is the square root of the population variance.
- The sample standard deviation s is the square root of the sample variance.

Remarks

- The standard deviation of a variable x is measured in the same units as x itself.
- The variance and standard deviation are never negative.

Measures of spread: standard deviation

Sample vs. population variance

For a numerical variable x,

- The population standard deviation σ is the square root of the population variance.
- The sample standard deviation s is the square root of the sample variance.

Remarks

- The standard deviation of a variable x is measured in the same units as x itself.
- The variance and standard deviation are never negative.
- Both variance and standard deviation are sensitive to outliers.

Measures of spread: standard deviation

Sample vs. population standard deviation

- The population standard deviation σ (a.k.a. the true standard deviation) is a parameter of the population. It is often unknown to us, unless we have access to the entire population (e.g., the census data).
- The sample standard deviation s is a statistic (i.e., a quantity calculated based on the sample). In particular, the sample standard deviation varies from sample to sample.

Measures of spread: standard deviation

Sample vs. population standard deviation

- The population standard deviation σ (a.k.a. the true standard deviation) is a parameter of the population. It is often unknown to us, unless we have access to the entire population (e.g., the census data).
- The sample standard deviation s is a statistic (i.e., a quantity calculated based on the sample). In particular, the sample standard deviation varies from sample to sample.
- Nevertheless, the sample standard deviation can be used to estimate the population standard deviation. The larger the sample, the more accurate the estimate (i.e., the less likely that the error is large).

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

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

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

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

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:

$$\sigma^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / N}{N}, \qquad s^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n - 1}.$$

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

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

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

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

An alternative way to calculate the variances:

$$\sigma^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / N}{N}, \qquad s^2 = \frac{\sum x^2 - \left(\sum x\right)^2 / n}{n-1}.$$

Measures of spread: variance

Sample vs. population variance

For a numerical variable x,

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

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:

$$\sigma^2 = \frac{\sum x^2 - \left(\sum x\right)^2/N}{N}, \qquad s^2 = \frac{\sum x^2 - \left(\sum x\right)^2/n}{n-1}.$$

Warning

- $\sum x^2$ is different from $\left(\sum x\right)^2$.

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:

$$\sigma^2 = \frac{\sum x^2 - \left(\sum x\right)^2/N}{N}, \qquad s^2 = \frac{\sum x^2 - \left(\sum x\right)^2/n}{n-1}.$$

Warning

- $\sum x^2$ is different from $\left(\sum x\right)^2$.
- As much as possible, leave the rounding to the very last step.

Measures of spread: variance

Sampled female birds

Weight:

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 spread: variance

Sampled female birds

Weight:

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								

Weight (squared):

9761.44	8556.25	7621.29	10712.25	9662.89	9820.81	9389.61	10609.00	9312.25	8723.56
10588.41	10221.21	10180.81	9139.36	9235.21	9860.49	8704.89	9101.16	7832.25	9525.76
8500.84	9604.00								

Measures of spread: variance

Sampled female birds

Weight:

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								

Weight (squared):

9761.44	8556.25	7621.29	10712.25	9662.89	9820.81	9389.61	10609.00	9312.25	8723.56
10588.41	10221.21	10180.81	9139.36	9235.21	9860.49	8704.89	9101.16	7832.25	9525.76
8500.84	9604.00								

$$\sum x = 98.8 + 92.5 + 87.3 + \cdots + 98.0 = 2130.2 \text{ grams}$$

Measures of spread: variance

Sampled female birds

Weight:

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								

Weight (squared):

9761.44	8556.25	7621.29	10712.25	9662.89	9820.81	9389.61	10609.00	9312.25	8723.56
10588.41	10221.21	10180.81	9139.36	9235.21	9860.49	8704.89	9101.16	7832.25	9525.76
8500.84	9604.00								

$$\sum x = 98.8 + 92.5 + 87.3 + \cdots + 98.0 = 2130.2\,\text{grams}$$

$$(\sum x)^2 = 2130.2^2 = 4537752.04\,\text{grams}^2$$

Measures of spread: variance

Sampled female birds

Weight:

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								

Weight (squared):

9761.44	8556.25	7621.29	10712.25	9662.89	9820.81	9389.61	10609.00	9312.25	8723.56
10588.41	10221.21	10180.81	9139.36	9235.21	9860.49	8704.89	9101.16	7832.25	9525.76
8500.84	9604.00								

$$\sum x = 98.8 + 92.5 + 87.3 + \cdots + 98.0 = 2130.2\,\text{grams}$$

$$(\sum x)^2 = 2130.2^2 = 4537752.04\,\text{grams}^2$$

$$\sum x^2 = 9761.44 + 8556.25 + 7621.29 + \cdots + 9604.00 = 206663.70\,\text{grams}^2$$

sample variance of weight for female birds

$$= \frac{206663.70 - 4537752.04/22}{22 - 1} \approx \boxed{19.1545}\,\text{grams}^2$$