Chapter 4: The Average and the Standard Deviation

Marius Ionescu

09/06/2011 and 09/08/2011
Center and spread of a histogram

A histogram can be used to summarize large amounts of data. Often the histogram is summarized by two numbers: the center and the spread. The center represents the level or position of the distribution, and the spread represents the variation within the population. However, things do not always work so well.
A histogram can be used to summarize large amounts of data.
A histogram can be used to summarize large amounts of data. Often the histogram is summarized by two numbers: the center and the spread.
Fact

- A histogram can be used to summarize large amounts of data.
- Often the histogram is summarized by two numbers: the center and the spread.
- The center represents the “level” or “position” of the distribution.
A histogram can be used to summarize large amounts of data. Often the histogram is summarized by two numbers: the center and the spread. The center represents the “level” or “position” of the distribution. The spread represents the variation within population.
A histogram can be used to summarize large amounts of data.

Often the histogram is summarized by two numbers: the center and the spread.

The center represents the “level” or “position” of the distribution.

The spread represents the variation within population.

However, things do not always work so well.
Measures of the centers

Definition

The average is the sum of all values divided by the number of values.

The median: the value with 50% of the values higher and 50% lower.
Measures of the centers

Definition

- The **average** is the sum of all values divided by the number of values
Measure of the centers

Definition

- The **average** is the sum of all values divided by the number of values.
- The **median**: the value with 50% of the values higher and 50% lower.
Example

Find the average

Example

Add 5 to the last example and find the average

Example

Multiply each number by 5 and find the average

Example

If the average of the day temperature during the last month is 27°F, what is the average in terms of Celsius?
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4

Average

Add 5 to the last example and find the average

Multiply each number by 5 and find the average

Example

If the average of the day temperature during the last month is 27°F, what is the average in terms of Celsius?
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
- Add 5 to the last example and find the average

If the average of the day temperature during the last month is 27°F, what is the average in terms of Celsius?
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
- Add 5 to the last example and find the average
- Multiply each number by 5 and find the average
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
- Add 5 to the last example and find the average
- Multiply each number by 5 and find the average
Example

Find the average

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
- Add 5 to the last example and find the average
- Multiply each number by 5 and find the average

Example

If the average of the day temperature during the last month is 27°F what is the average in terms of Celsius?
Fact
Fact

- $\text{average}(x+5) = \text{average}(x) + 5$
Rules

Fact

- \( \text{average}(x+5) = \text{average}(x) + 5 \)
- \( \text{average}(x \cdot 5) = \text{average}(x) \cdot 5 \)
Example

Find the median for each of the following sequence of numbers:
Example

Find the median for each of the following sequence of numbers:

- 1, 1, 1, 1, 2, 2, 2, 2, 15
Find the median for each of the following sequence of numbers:

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
Example

Find the median for each of the following sequence of numbers:

- 1, 1, 1, 1, 2, 2, 2, 2, 15
- 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4
- 8, 10, 15, 20
Average and median in Excel

Fact

To find the average of the cells A1:A10 you need to write

=average(A1:A10)

To find the median of the cells A1:A10 you need to write

=median(A1:A10)
To find the average of the cells A1:A10 you need to write:

=average(A1:A10)
Fact

- *To find the average of the cells A1:A10 you need to write*  
  \[ \text{=average(A1:A10)} \]

- *To find the median of the cells A1:A10 you need to write*  
  \[ \text{=median(A1:A10)} \]
Example

Which histogram has higher average? Which histogram has higher median?
Average, median, and the histogram

**Fact**

- Average is the point at which the distribution balances.
- Median is the point for which $1/2$ of the area is on the left and $1/2$ is on the right.
- Median describes a “middle” individual, a typical subject.
Example

For income in US, which would you expect to be larger? The median or the income?

Answer: In 2008 the median income was $61,521 and the average income was $79,634.
Example

For income in US, which would you expect to be larger? The median or the income?
Answer: In 2008
- the median income was $61,521
Example

For income in US, which would you expect to be larger? The median or the income?
Answer: In 2008

- the median income was $61,521
- the average income was $79,634
Fact

**Standard deviation (SD)** is a common way of measuring the spread around the average.
The Root-mean-square

Definition

Root-mean-square = the square root of averages of square
The Root-mean-square

**Definition**

Root-mean-square = the square root of averages of squares

**Example**

The root-mean-square of $5, -5, 0, 6$ is

$$\sqrt{\frac{5^2 + (-5)^2 + 0^2 + 6^2}{4}} = 4.6368.$$
Definition

$SD = \text{root-mean-square of distance to the average.}$
Definition

SD = root-mean-square of distance to the average.

Example

Find the standard deviation of 20, 10, 15, 15.

\[ \text{Avg} = \frac{20 + 10 + 15 + 15}{4} = 15. \]

\[ \text{SD} = \sqrt{\frac{(20-15)^2 + (10-15)^2 + (15-15)^2 + (15-15)^2}{4}} \approx 3.5355. \]
Definition

SD = root-mean-square of distance to the average.

Example

Find the standard deviation of 20, 10, 15, 15.

\[ \text{Avg} = \frac{20 + 10 + 15 + 15}{4} = 15. \]
Standard Deviation

**Definition**

SD = root-mean-square of distance to the average.

**Example**

Find the standard deviation of 20, 10, 15, 15.

\[
\text{Avg} = \frac{20 + 10 + 15 + 15}{4} = 15.
\]

\[
SD = \sqrt{\frac{5^2 + (-5)^2 + 0^2 + 0^2}{4}} = 3.5355
\]
Fact

The SD says how far away the numbers on a list are from their average. Most entries on the list will be somewhere around one SD away from the average.
Fact

Enter 1 and −1 in your calculator and compute SD. What number do you get?

If the answer is 1, then your calculator is computing SD. If the answer is 1.41..., then your calculator is computing SD +.

To obtain SD from SD + you need to use the following formula:

$$SD = \sqrt{\frac{\text{sum of squares}}{\text{number of entries} - 1} - \left(\frac{\text{sum of entries}}{\text{number of entries}}\right)^2}$$

To compute the standard deviation in Excel of a sequence of numbers in cells A1:A10 you would write

=STDEV.P(A1:A10)
Fact

Enter 1 and −1 in your calculator and compute SD. What number do you get?

- If the answer is 1 then your calculator is computing SD.
Is your calculator computing SD or SD⁺?

Fact

Enter 1 and \(-1\) in your calculator and compute SD. What number do you get?

- **If the answer is 1 then your calculator is computing SD.**
- **If the answer is 1.41..., then your calculator is computing something called SD⁺.**
Is your calculator computing SD or $SD^+$?

**Fact**

*Enter 1 and $-1$ in your calculator and compute SD. What number do you get?*

- If the answer is 1 then your calculator is computing SD.
- If the answer is 1.41..., then your calculator is computing something called $SD^+$.  
- To find SD from $SD^+$ you need to use the following formula

$$SD = \sqrt{\frac{\# \text{ of entries} - 1}{\# \text{ of entries}}} \cdot SD^+.$$
Is your calculator computing SD or SD⁺?

Fact

Enter 1 and −1 in your calculator and compute SD. What number do you get?

- If the answer is 1 then your calculator is computing SD.
- If the answer is 1.41..., then your calculator is computing something called SD⁺.
- To find SD from SD⁺ you need to use the following formula:

\[ SD = \sqrt{\frac{\# \text{ of entries} - 1}{\# \text{ of entries}}} \cdot SD^+. \]

- To compute the standard deviation in Excel of a sequence of numbers in cells A1:A10 you would write =stdevp(A1:A10)
Example

Which of the following histograms has the highest SD and which one has the smallest SD?
**Definition**

$n\%$: sort the numbers and then find which is bigger than $n\%$ of the others.
Percentiles

**Definition**

$n\%$: sort the numbers and then find which is bigger than $n\%$ of the others.

Example

Find the 75th percentile of 1, 2, 3, 4.

Find the 10th, 50th, 90th percentile of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
**Definition**

$n\%$: sort the numbers and then find which is bigger than $n\%$ of the others.

**Example**
Percentiles

Definition

\(n\%: \) sort the numbers and then find which is bigger than \(n\%\) of the others.

Example

- Find the 75th percentile of 1, 2, 3, 4.
Definition

\( n \% \)\: sort the numbers and then find which is bigger than \( n \% \) of the others.

Example

- Find the 75th\%ile of 1,2,3,4.
- Find the 10th, 50th, 90th percentile of 1,2,3,4,5,6,7,8,9,10.
Percentiles in Excel

Fact

To find percentiles in Excel, you should use the percentile function. For example, to find the 75th percentile of a sequence of numbers in A1:A10, you would enter:

=PERCENTILE(A1:A10, 0.75)

Notice that you need to enter 0.75 and not 75!
To find percentiles in Excel you should use the \textit{percentile} function.
Fact

To find percentiles in Excel you should use the `percentile` function.

For example to find the 75th %ile of a sequence of numbers in A1:A10 you would enter `=percentile(A1:A10,0.75)`.
To find percentiles in Excel you should use the *percentile* function.

For example to find the 75th %ile of a sequence of numbers in A1:A10 you would enter =*percentile*(A1:A10,0.75)

*Notice that you need to enter 0.75 and not 75!*
Example

Try to match the following histograms to the following data from a survey of adults in the San Francisco Bay Area:

1. people’s height
2. people’s weight
3. the distance from a person’s home to San Francisco
4. the distance from a person’s home to the nearest airport.

You can use the same diagram more than once or not at all.
Example

The following two histograms are for decibel reading at a basketball game and a hockey game. What does the histogram for the combined data look like (for both sporting events)?