HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 3 — Types of data. Descriptive statistics

Jorge N. Tendeiro
Hiroshima University

Today

- Types of data:
  $\hookrightarrow$ *variable types*.
- Descriptive statistics:
  $\hookrightarrow$ *how to summarize data*.

Types of data

Statistical data

Household data in household survey (Unit: 1000 yen per month)

Household	**Food**	**Education**	**Recreation**	**Size of household**	**Working family***
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

*Working family: 1 = Worker households; 0 = other households.

This is a data set with 100 rows and 5 columns, with a total of $100 \times 5 = 500$ entries.

Statistical data

Household data in household survey (Unit: 1000 yen per month)

Household	**Food**	**Education**	**Recreation**	**Size of household**	**Working family***
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

*Working family: 1 = Worker households; 0 = other households.

This is a data set with 100 rows and 5 columns, with a total of $100 \times 5 = 500$ entries.

Rows:
Each row is a unit of measurement (a *household* in this example).
The total number of rows is called the sample size.

Statistical data

Household data in household survey (Unit: 1000 yen per month)

Household	**Food**	**Education**	**Recreation**	**Size of household**	**Working family***
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

*Working family: 1 = Worker households; 0 = other households.

This is a data set with 100 rows and 5 columns, with a total of $100 \times 5 = 500$ entries.

Rows:
Each row is a unit of measurement (a *household* in this example).
The total number of rows is called the sample size.

Columns:
Each column is a variable.
Variables are features of the units that we are interested in studying.
The total number of columns is called the dimension of the data.

Statistical data

Household data in household survey (Unit: 1000 yen per month)

Household	**Food**	**Education**	**Recreation**	**Size of household**	**Working family***
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

*Working family: 1 = Worker households; 0 = other households.

This is a data set with 100 rows and 5 columns, with a total of $100 \times 5 = 500$ entries.

Rows:
Each row is a unit of measurement (a *household* in this example).
The total number of rows is called the sample size.

Columns:
Each column is a variable.
Variables are features of the units that we are interested in studying.
The total number of columns is called the dimension of the data.

Cells:
Each cell is an observation or datum.

Variable types

There are two main types of variables.

Variable types

There are two main types of variables.

Quantitative variables:
Numerical variables, we can quantify features.

Example: Height, weight, annual income, etc.

Variable types

There are two main types of variables.

Quantitative variables:
Numerical variables, we can quantify features.

Example: Height, weight, annual income, etc.

Qualitative variables:
Based on text or labels, we can identify features.

Example: Having a car, academic background, nationality, etc.

Variable types

There are two main types of variables.

Quantitative variables:
Numerical variables, we can quantify features.

Example: Height, weight, annual income, etc.

Qualitative variables:
Based on text or labels, we can identify features.

Example: Having a car, academic background, nationality, etc.

	Quantitative variables				*Qualitative variables*
Household	**Food**	**Education**	**Recreation**	**Size of household**	**Working family***
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

*Working family: 1 = Worker households; 0 = other households.

Variable types — another example

Which variables are qualitative or quantitative?

House Prices and Environmental Conditions data (Takahashi et al., 2000)

ID	Price	Period	Area	Size	JR	Time	Age	Distance
54	3730	0	148	103	0	9	0	13
55	1480	1	110	67	0	3	23	8
64	3590	1	105	100	1	8	8	8

*Working family: 1 = Worker households; 0 = other households.

- **Price**
  Single house price (in $¥10,000$).
- **Period**
  Year (0 = 1998; 1 = 1999).
- **Area**
  Land area (in m^2).
- **Size**
  House are (in m^2).
- **JR**
  Type of nearest station (0 = Not JR; 1 = JR).
- **St. time**
  Time to walk from nearest station/bus stop (in min).
- **Age**
  Time since it was built (in years).
- **Distance**
  Distance from the center of Hiroshima city (in km).

Variable types — another example

Which variables are qualitative or quantitative?

	Quantitative variables					*Qualitative variables*		
ID	**Price**	**Period**	**Area**	**Size**	**JR**	**Time**	**Age**	**Distance**
54	3730	0	148	103	0	9	0	13
55	1480	1	110	67	0	3	23	8
64	3590	1	105	100	1	8	8	8

*Working family: 1 = Worker households; 0 = other households.

Scales of measurement

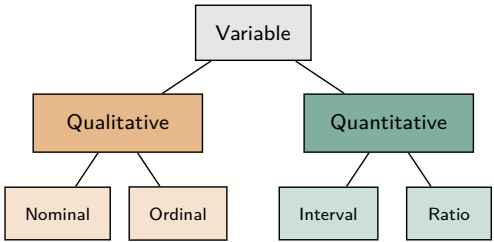
Variables can be of various types.
Scales of measurement help distinguishing between different types of variables.

Scales of measurement

Variables can be of various types.
Scales of measurement help distinguishing between different types of variables.

There are four main variable types (Stevens, 1946):

- Nominal scale variables.
- Ordinal scale variables.
- Interval scale variables.
- Ratio scale variables.



Scales of measurement: Nominal scale variables

Also known as categorical scale variables.

There is no particular relationship between the different values (like in size or quality).

Example:

- Eye color
- Gender
- Nationality.

No computations with the values is possible (like sum or mean).

Artwork by @allison_horst, CC BY 4.0, adapted.

Scales of measurement: Ordinal scale variables

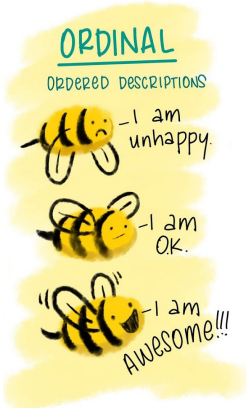
There is a natural, meaningful way to order different values (numbers or labels).

But there is no way to assess the *relative degree of difference* between any two values.

Example:

- Race ranking (1st, 2nd, 3rd, ...).
- Questions with answer options 'completely agree', 'mostly agree', 'mostly disagree', 'completely disagree'. This is known as a Likert item.

No computations with the values is possible (like sum or mean).



Artwork by @allison_horst, CC BY 4.0, adapted.

Scales of measurement: Interval scale variables

Numerical values such that *differences* between numbers are interpretable.
Addition and subtraction is allowed.

However, interval scale variables lack a "natural" zero value.
Hence, we cannot multiply or divide.

Scales of measurement: Interval scale variables

Numerical values such that *differences* between numbers are interpretable.
Addition and subtraction is allowed.

However, interval scale variables lack a "natural" zero value.
Hence, we cannot multiply or divide.

Example: Temperature in degrees celsius.

- We can consider temperature differences, so 15°C is 5°C higher than 10°C.

Scales of measurement: Interval scale variables

Numerical values such that *differences* between numbers are interpretable.
Addition and subtraction is allowed.

However, interval scale variables lack a "natural" zero value.
Hence, we cannot multiply or divide.

Example: Temperature in degrees celsius.

- We can consider temperature differences, so 15°C is 5°C higher than 10°C.
- However, 0°C **cannot** be interpreted as "no temperature".
  So, one cannot say that 20°C is twice as hot as 10°C.

Scales of measurement: Ratio scale variables

These are interval scale variables with a meaningful zero value.

Scales of measurement: Ratio scale variables

These are interval scale variables with a meaningful zero value.

Example: Response time.
This works because:

Scales of measurement: Ratio scale variables

These are interval scale variables with a meaningful zero value.

Example: Response time.
This works because:

- Zero implies an instantaneous reaction (no time lag at all).

Scales of measurement: Ratio scale variables

These are interval scale variables with a meaningful zero value.

Example: Response time.
This works because:

- Zero implies an instantaneous reaction (no time lag at all).
- We can say that 2 seconds is twice as long as 1 second, for instance.

Scales of measurement: Ratio scale variables

These are interval scale variables with a meaningful zero value.

Example: Response time.
This works because:

- Zero implies an instantaneous reaction (no time lag at all).
- We can say that 2 seconds is twice as long as 1 second, for instance.

Other examples:
Height, weight, annual income.

Scales of measurement: Why classify variables like this?

1. The effectiveness of data analysis methods varies depending on the type of variable used.

Scales of measurement: Why classify variables like this?

1. The effectiveness of data analysis methods varies depending on the type of variable used.

   Example: The *average value* of a qualitative variable is meaningless.

Scales of measurement: Why classify variables like this?

1. The effectiveness of data analysis methods varies depending on the type of variable used.

   Example: The *average value* of a qualitative variable is meaningless.

2. Statistical methods are typically quite specific about the type of variables that they allow:
   *In the analysis method ***, use variable(s) of the type ***.*

Scales of measurement: Why classify variables like this?

1. The effectiveness of data analysis methods varies depending on the type of variable used.

   Example: The *average value* of a qualitative variable is meaningless.

2. Statistical methods are typically quite specific about the type of variables that they allow:
   *In the analysis method ***, use variable(s) of the type ****.

3. Variables can be ordered based on how easy it is to manipulate them:
   $$\text{Ratio} > \text{Interval} > \text{Ordinal} > \text{Nominal}$$

Descriptive statistics

Descriptive statistics

Descriptive statistics are useful to summarize data.

Descriptive statistics

Descriptive statistics are useful to summarize data.

As an example, consider the following data:

Housing price data of 158 houses (in ¥10,000) (Takahashi et al., 2000):

3150	3150	3500	6500	3800	2890	3170	3390	3650	3500	3880	3950	4400	4600	5980	2450	3480	3880	2980	800
2000	2670	2980	2980	3980	4400	4570	3480	3800	2450	4400	2300	2880	2980	3300	3380	3660	3180	3750	4180
1380	2690	1680	2580	2780	2890	3080	2980	3050	2980	2780	3880	3680	3730	1480	2000	2050	2170	2180	2290
2300	2380	1380	3590	2630	2680	2880	3080	3200	3550	3790	5300	3190	2780	4100	5580	1280	2480	2580	2880
2900	3180	3280	3280	3280	3300	3680	3850	4290	7800	1500	2580	3180	3190	3100	2980	3280	3070	3190	3680
2980	4480	3730	3980	3440	3460	3700	3440	3390	4380	1490	3300	3180	2480	3980	1980	2880	4000	4580	4580
3500	3550	3400	3550	2000	2700	2480	1620	2580	2300	850	2290	3180	2050	2580	3780	5780	2680	4380	5977
2280	3090	4400	1950	1980	3200	3380	4280	5980	4180	4350	3730	3980	4000	4650	4680	4680	1580		

Descriptive statistics

Descriptive statistics are useful to summarize data.

As an example, consider the following data:

Housing price data of 158 houses (in ¥10,000) (Takahashi et al., 2000):

3150	3150	3500	6500	3800	2890	3170	3390	3650	3500	3880	3950	4400	4600	5980	2450	3480	3880	2980	800
2000	2670	2980	2980	3980	4400	4570	3480	3800	2450	4400	2300	2880	2980	3300	3380	3660	3180	3750	4180
1380	2690	1680	2580	2780	2890	3080	2980	3050	2980	2780	3880	3680	3730	1480	2000	2050	2170	2180	2290
2300	2380	1380	3590	2630	2680	2880	3080	3200	3550	3790	5300	3190	2780	4100	5580	1280	2480	2580	2880
2900	3180	3280	3280	3280	3300	3680	3850	4290	7800	1500	2580	3180	3190	3100	2980	3280	3070	3190	3680
2980	4480	3730	3980	3440	3460	3700	3440	3390	4380	1490	3300	3180	2480	3980	1980	2880	4000	4580	4580
3500	3550	3400	3550	2000	2700	2480	1620	2580	2300	850	2290	3180	2050	2580	3780	5780	2680	4380	5977
2280	3090	4400	1950	1980	3200	3380	4280	5980	4180	4350	3730	3980	4000	4650	4680	4680	1580		

There are **too many** houses... It's difficult to 'read' the information.
We need to find ways to summarize the information in the data.

Descriptive statistics

Descriptive statistics are useful to summarize data.

As an example, consider the following data:

Housing price data of 158 houses (in ¥10,000) (Takahashi et al., 2000):

3150	3150	3500	6500	3800	2890	3170	3390	3650	3500	3880	3950	4400	4600	5980	2450	3480	3880	2980	800
2000	2670	2980	2980	3980	4400	4570	3480	3800	2450	4400	2300	2880	2980	3300	3380	3660	3180	3750	4180
1380	2690	1680	2580	2780	2890	3080	2980	3050	2980	2780	3880	3680	3730	1480	2000	2050	2170	2180	2290
2300	2380	1380	3590	2630	2680	2880	3080	3200	3550	3790	5300	3190	2780	4100	5580	1280	2480	2580	2880
2900	3180	3280	3280	3280	3300	3680	3850	4290	7800	1500	2580	3180	3190	3100	2980	3280	3070	3190	3680
2980	4480	3730	3980	3440	3460	3700	3440	3390	4380	1490	3300	3180	2480	3980	1980	2880	4000	4580	4580
3500	3550	3400	3550	2000	2700	2480	1620	2580	2300	850	2290	3180	2050	2580	3780	5780	2680	4380	5977
2280	3090	4400	1950	1980	3200	3380	4280	5980	4180	4350	3730	3980	4000	4650	4680	4680	1580		

There are **too many** houses... It's difficult to 'read' the information.
We need to find ways to summarize the information in the data.

One possible way to do this is by plotting the data.

Plots

Plots (or graphs):

Visual representations *that allow to quickly grasp the relevant information in the data.*

Plots

Plots (or graphs):

Visual representations that allow to quickly grasp the relevant information in the data.

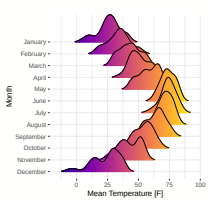
There are many types of plots available!!

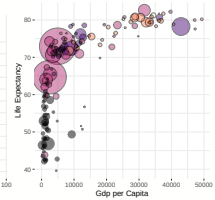
Plots

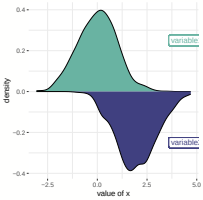
Plots (or graphs):

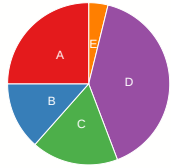
Visual representations that allow to quickly grasp the relevant information in the data.

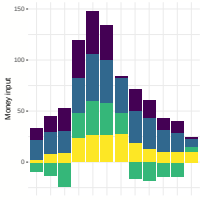
There are many types of plots available!!









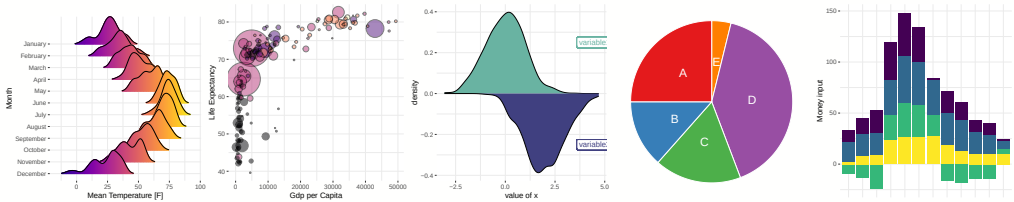


Plots

Plots (or graphs):

Visual representations that allow to quickly grasp the relevant information in the data.

There are many types of plots available!!



We will only learn about a small few, simple, plots.

Histogram

Histograms are used to representing the **distribution** of quantitative variables.

Histogram

Histograms are used to representing the **distribution** of quantitative variables.

Distribution:

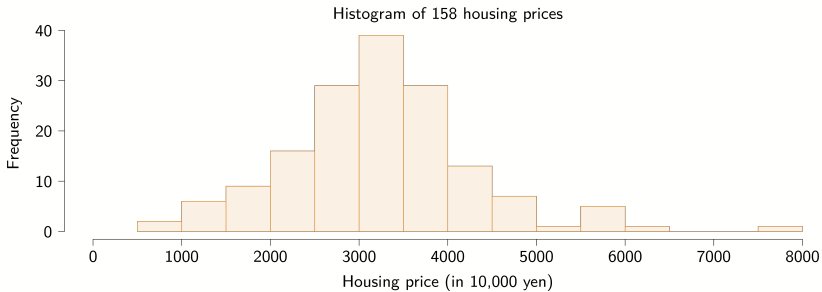
The shape reflecting how the data values are distributed.

Histogram

Histograms are used to representing the **distribution** of quantitative variables.

Distribution:

The shape reflecting how the data values are distributed.

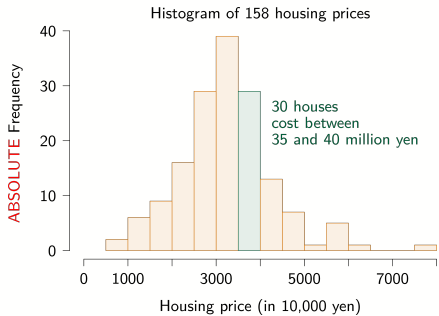


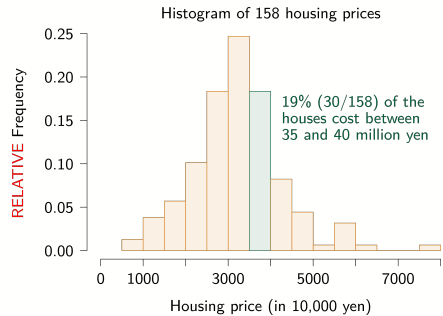
Histogram

The y-axis can be edited to display either the absolute frequencies (i.e., counts) or relative frequencies (i.e., proportions).

Histogram

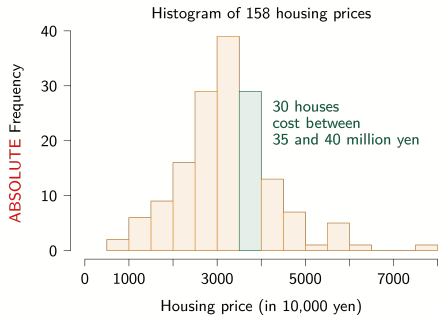
The y-axis can be edited to display either the absolute frequencies (i.e., counts) or relative frequencies (i.e., proportions).

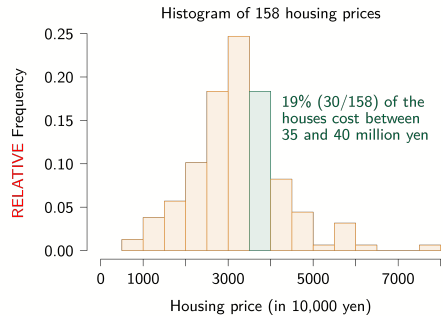




Histogram

The y-axis can be edited to display either the absolute frequencies (i.e., counts) or relative frequencies (i.e., proportions).





bin height $=$ number of observations in each bin

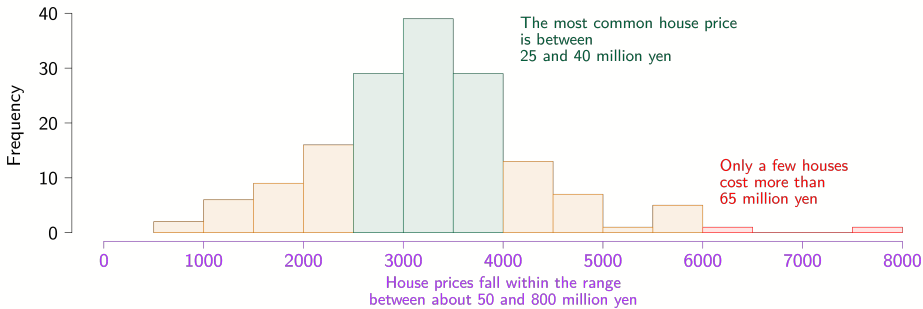
bin height $= \frac{\text{bin count}}{\text{sample size}}$

Histogram

Observe how it is easy to draw information from a histogram:

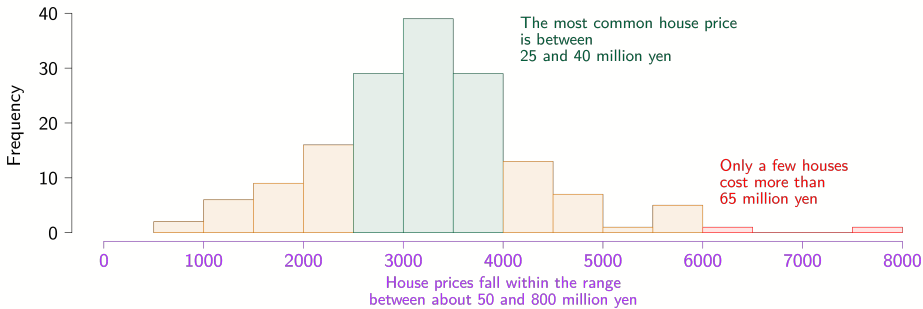
Histogram

Observe how it is easy to draw information from a histogram:



Histogram

Observe how it is easy to draw information from a histogram:



We can tell all this even without looking at the numbers in the data!

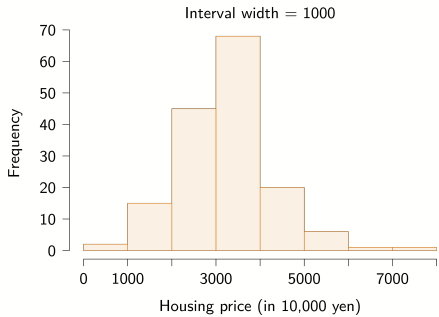
Histogram

Histograms look different according to the number of intervals (or bins).

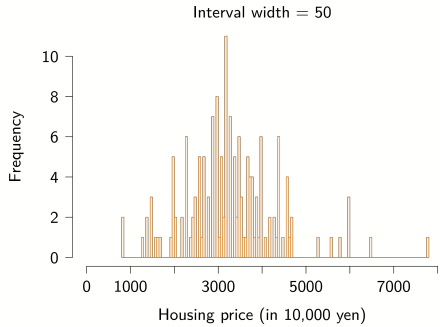
Histogram

Histograms look different according to the number of intervals (or bins).

Too few bins: Too "zoomed out". Avoid!



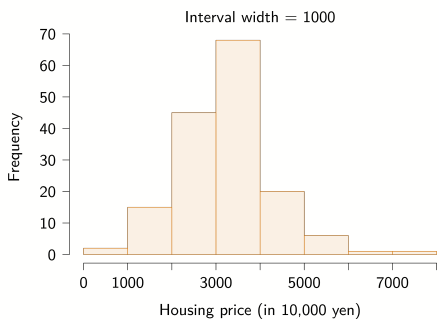
Too many bins: Too "zoomed in". Avoid!



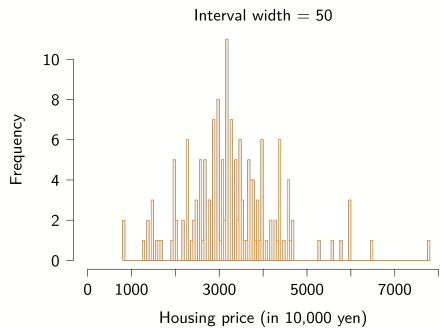
Histogram

Histograms look different according to the number of intervals (or bins).

Too few bins: Too "zoomed out". Avoid!



Too many bins: Too "zoomed in". Avoid!



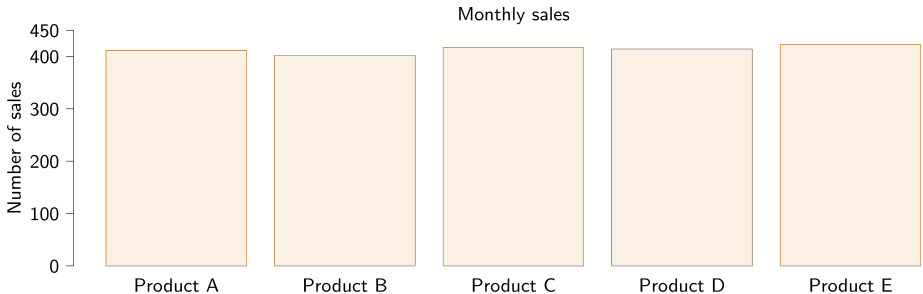
A sensible intermediate number of bins (not too few, not too many) is the best.

Bar chart

For qualitative variables we use bar charts (also known as bar plots) instead of histograms.
With bar charts, one separate bar is drawn for each group.

Bar chart

For qualitative variables we use bar charts (also known as bar plots) instead of histograms.
With bar charts, one separate bar is drawn for each group.

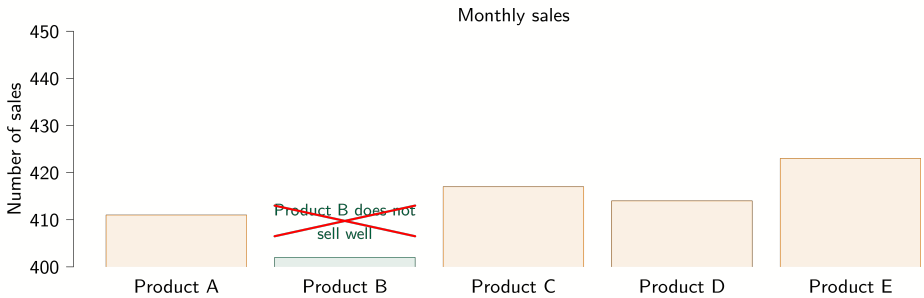


Bar chart — Be careful!

We need **extreme care** when drawing plots. Otherwise we may draw nonsense conclusions!
For example, suppose we truncate the y-axis below 400:

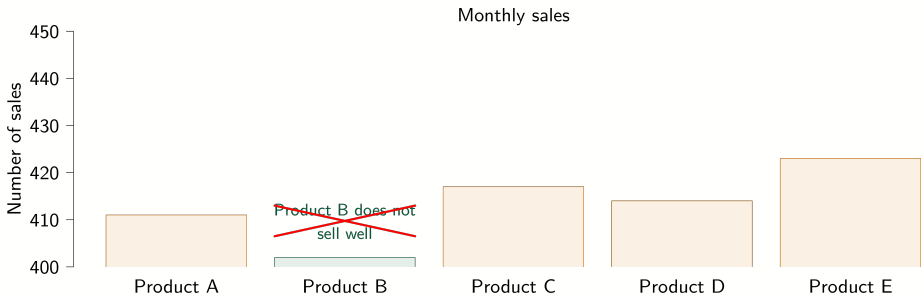
Bar chart — Be careful!

We need **extreme care** when drawing plots. Otherwise we may draw nonsense conclusions!
For example, suppose we truncate the y-axis below 400:



Bar chart — Be careful!

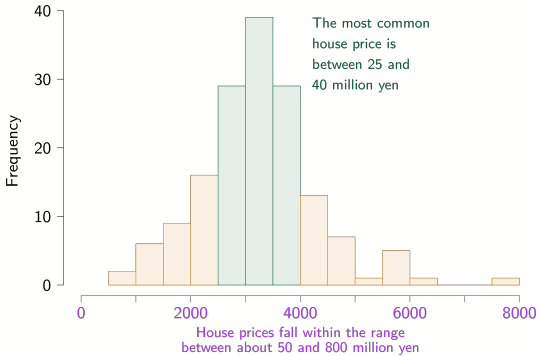
We need **extreme care** when drawing plots. Otherwise we may draw nonsense conclusions!
For example, suppose we truncate the y-axis below 400:



Changing the appearance of the y-axis changed the heights of the bars.
This should **not** be done! Be aware that some advertisement companies do use such tricks...

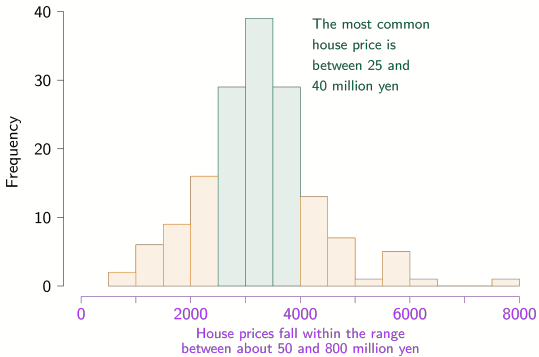
Descriptive statistics — Using numbers

Plotting your data is great!



Descriptive statistics — Using numbers

Plotting your data is great!



But plots provide only a broad overview.

To draw more precise information, we need numbers.

Descriptive statistics — Using numbers

Descriptive statistic:

A value that represents a particular characteristic of the data distribution.
Easy to communicate and objective.

Descriptive statistics — Using numbers

Descriptive statistic:

> *A value that represents a particular characteristic of the data distribution.*
> *Easy to communicate and objective.*

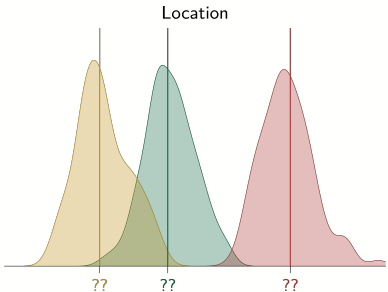
Common characteristics of a distribution that we are interested in include the location ('center') and the spread.

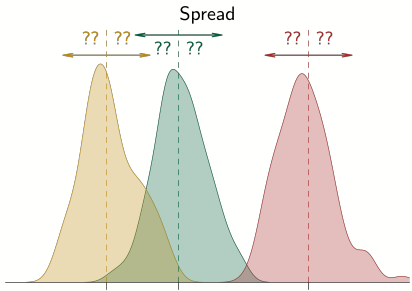
Descriptive statistics — Using numbers

Descriptive statistic:

A value that represents a particular characteristic of the data distribution.
Easy to communicate and objective.

Common characteristics of a distribution that we are interested in include the location ('center') and the spread.





Descriptive statistics — Using numbers

There are *many* measures of location and spread available.
We will only discuss a few.

Descriptive statistics — Using numbers

There are *many* measures of location and spread available.
We will only discuss a few.

Location: (today)

- Mean
- Median
- Mode

Spread: (next lecture)

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

Measure of location: Mean

The sample mean of n data points $x_1, \ldots, x_n$ from variable x is

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

Individual	Variable x
1	x_1
2	x_2
$\vdots$	$\vdots$
$n-1$	x_{n-1}
n	x_n

Measure of location: Mean

The sample mean of n data points $x_1, \ldots, x_n$ from variable x is

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

Individual	Variable x
1	x_1
2	x_2
$\vdots$	$\vdots$
$n-1$	x_{n-1}
n	x_n

In words,

Sample mean = add all the data up, and divide by the sample size.

Measure of location: Mean

The sample mean of n data points $x_1, \ldots, x_n$ from variable x is

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

Individual	Variable x
1	x_1
2	x_2
$\vdots$	$\vdots$
$n-1$	x_{n-1}
n	x_n

In words,

Sample mean = add all the data up, and divide by the sample size.

Example:

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

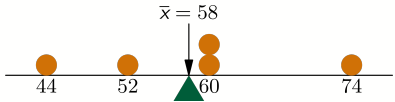
- $n = 5$
- Variable $x =$ weight in Kg
- $x_1 = 60$, $x_2 = 52$, ..., $x_5 = 60$

$$\overline{x} = \frac{60+52+44+74+60}{5} = 58$$

Measure of location: Mean

The mean is…

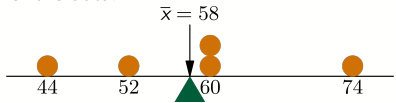
- …actually the center of gravity for the data:



Measure of location: Mean

The mean is…

- …actually the center of gravity for the data:

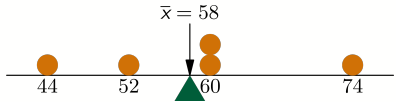


- …a suitable representative value for the location of a quantitative variables.
  For example: height, weight, annual income, year, temperature (°C).

Measure of location: Mean

The mean is…

- …actually the center of gravity for the data:



- …a suitable representative value for the location of a quantitative variables.
  For example: height, weight, annual income, year, temperature (°C).

- …**not** a suitable representative value for the location of a qualitative variable.
  For example, what would be the mean of "blood type" (A, B, AB, O)?

  Also for numeric qualitative variables, the mean makes no sense.
  For example, what would be the mean of "blood type" (1 = A, 2 = B, 3 = AB, 4 = O)?
  Here, the numbers are just *labels*. They have **no** numerical meaning.

Measure of location: Median

The median of a set of observations is the middle value.

Measure of location: Median

The median of a set of observations is the middle value.

To find the median, do this:

- Sort the values from smallest to largest.
- If there is an **odd number** of values, the median is the middle value.
- If there is an **even number** of values, the median is the mean of the two most middle values.

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?
A: First, sort the values in increasing order:

$$0, 1, \mathbf{3}, 5, 5.$$

The middle value is 3, thus the median is 3.

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?
A: First, sort the values in increasing order:

$$0, 1, \mathbf{3}, 5, 5.$$

The middle value is 3, thus the median is 3.

Example — even number of values.

Q: What is the median of the values $\{-2, 5, 4, 1\}$?

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?
A: First, sort the values in increasing order:

$$0, 1, \mathbf{3}, 5, 5.$$

The middle value is 3, thus the median is 3.

Example — even number of values.

Q: What is the median of the values $\{-2, 5, 4, 1\}$?
A: First, sort the values in increasing order:

$$-2, \mathbf{1}, \mathbf{4}, 5.$$

The two most middle values are 1 and 4 and their mean is $\frac{1+4}{2} = 2.5$.
Thus the median is 2.5.

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?
A: First, sort the values in increasing order:

$$0, 1, \mathbf{3}, 5, 5.$$

The middle value is 3, thus the median is 3.

Example — even number of values.

Q: What is the median of the values $\{-2, 5, 4, 1\}$?
A: First, sort the values in increasing order:

$$-2, \mathbf{1}, \mathbf{4}, 5.$$

The two most middle values are 1 and 4 and their mean is $\frac{1+4}{2} = 2.5$.
Thus the median is 2.5.

Example:

Measure of location: Median

Example — odd number of values.

Q: What is the median of the values $\{5, 1, 5, 0, 3\}$?
A: First, sort the values in increasing order:

$$0, 1, \mathbf{3}, 5, 5.$$

The middle value is 3, thus the median is 3.

Example — even number of values.

Q: What is the median of the values $\{-2, 5, 4, 1\}$?
A: First, sort the values in increasing order:

$$-2, \mathbf{1}, \mathbf{4}, 5.$$

The two most middle values are 1 and 4 and their mean is $\frac{1+4}{2} = 2.5$.
Thus the median is 2.5.

Example:

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

Sort the values in increasing order:

$$44, 52, \mathbf{60}, 60, 74.$$

The middle value is 60, thus the median is 60.

Measure of location: Median

The median is a suitable representative value for the location of a quantitative variable.

And, interestingly,

> *The median is also suitable for the location of an ordinal variable.*

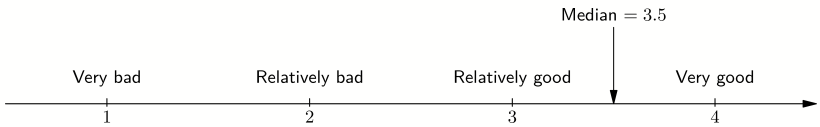
Measure of location: Median

The median is a suitable representative value for the location of a quantitative variable.

And, interestingly,

The median is also suitable for the location of an ordinal variable.

Example:
A question with four answer options (1 = very bad; 2 = relatively bad; 3 = relatively good; 4 = very good) was given to a group of people. The median score in the sample is 3.5.



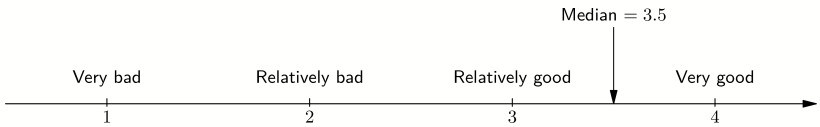
Measure of location: Median

The median is a suitable representative value for the location of a quantitative variable.

And, interestingly,

The median is also suitable for the location of an ordinal variable.

Example:
A question with four answer options (1 = very bad; 2 = relatively bad; 3 = relatively good; 4 = very good) was given to a group of people. The median score in the sample is 3.5.



We can then say that the middle of the data is in between *Relatively good* and *Very good*.

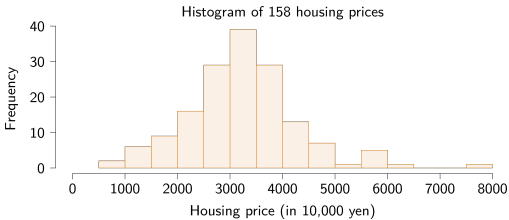
Exercises (1)

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

1. Calculate the mean and median of six people when a sixth observation, 61 (kg), is added to the table above.
2. Calculate the mean and median of six people when the sixth observation was accidentally added as 1000 (kg) to the table above.
3. Consider the difference between the mean and the median from the results of questions 1 and 2.

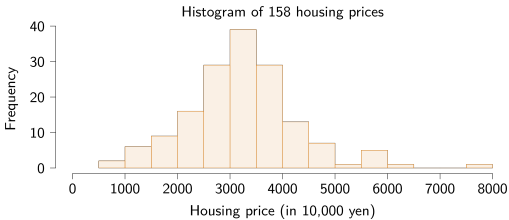
Measure of location: Mean vs median



For the *housing price* data we have that:

- Mean $= 3291.12$.
- Median $= 3195$.

Measure of location: Mean vs median

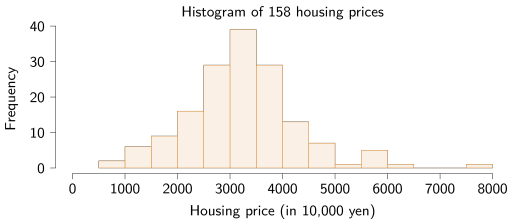


In this case, the mean and the median are *almost* the same.
But, they are indeed different.

For the *housing price* data we have that:

- Mean $= 3291.12$.
- Median $= 3195$.

Measure of location: Mean vs median



For the *housing price* data we have that:

- Mean $= 3291.12$.
- Median $= 3195$.

In this case, the mean and the median are *almost* the same.
But, they are indeed different.

Which value should be preferred?
What are the differences between the mean and the median?

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\bar{x} = 58$ kg
median $= 60$ kg

1. Calculate the mean and median of six people when a sixth observation, 61 (kg), is added to the table above.

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

1. Calculate the mean and median of six people when a sixth observation, 61 (kg), is added to the table above.

Answer

- $\overline{x} = \frac{60+52+44+74+60+61}{6} = 58.5.$

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

1. Calculate the mean and median of six people when a sixth observation, 61 (kg), is added to the table above.

Answer

- $\overline{x} = \frac{60+52+44+74+60+61}{6} = 58.5$.
- Ordered data: $44, 52, 60, 60, 61, 74$.
  Median $= \frac{60+60}{2} = 60$.

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

2. Calculate the mean and median of six people when the sixth observation was accidentally added as 1000 (kg) to the table above.

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

2. Calculate the mean and median of six people when the sixth observation was accidentally added as 1000 (kg) to the table above.

Answer

- $\overline{x} = \frac{60+52+44+74+60+1000}{6} = 215.$

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

2. Calculate the mean and median of six people when the sixth observation was accidentally added as 1000 (kg) to the table above.

Answer

- $\overline{x} = \frac{60+52+44+74+60+1000}{6} = 215.$
- Ordered data: $44, 52, 60, 60, 74, 1000.$
  Median $= \frac{60+60}{2} = 60.$

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

3. Consider the difference between the mean and the median from the results of questions 1 and 2.

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

3. Consider the difference between the mean and the median from the results of questions 1 and 2.

Answer

		One more data value	
	Original data for five people	Question 1 (61 kg)	Question 2 (1000 kg)
Mean	58	58.5	215
Median	60	60	60

The mean is much more sensitive to **extreme values** than the median.

Exercises (1) — ANSWER

Individual	Weight (kg)
1	60
2	52
3	44
4	74
5	60

$\overline{x} = 58$ kg
median $= 60$ kg

3. Consider the difference between the mean and the median from the results of questions 1 and 2.

Answer

		One more data value	
	Original data for five people	Question 1 (61 kg)	Question 2 (1000 kg)
Mean	58	58.5	215
Median	60	60	60

The mean is much more sensitive to **extreme values** than the median.

Extreme values are also known as **outliers**.
In case you want to avoid the effect of outliers, it's better to use the median instead of the mean!!

Notes about outliers

Example:
Consider the following exam scores from five students:

$$40, 40, 40, 40, 100.$$

Choosing between the mean or the median also depends on what *question* one is trying to answer.

Notes about outliers

Example:
Consider the following exam scores from five students:

$$40, 40, 40, 40, 100.$$

Choosing between the mean or the median also depends on what *question* one is trying to answer.

Question	Answer
"What is the middle of the ordered scores?"	Median = 40
"What is the average score?"	Mean = $\frac{40+40+40+40+100}{5} = 52$

Notes about outliers

Example:
Consider the following exam scores from five students:

$$40, 40, 40, 40, 100.$$

Choosing between the mean or the median also depends on what *question* one is trying to answer.

Question	Answer
"What is the middle of the ordered scores?"	Median = 40
"What is the average score?"	Mean = $\frac{40+40+40+40+100}{5} = 52$

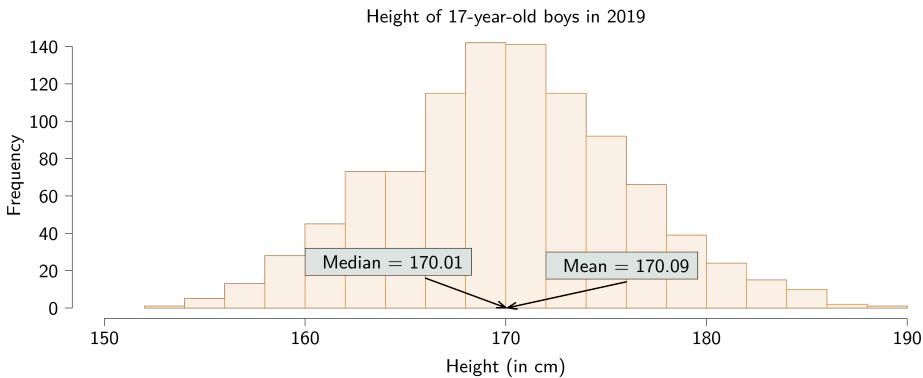
Even with outliers, choosing between the mean or the median may depend based on what kind of "middle" score we want to know.

Mean vs median — symmetric distributions

For symmetric distributions, the mean and the median are very close to each other.

Mean vs median — symmetric distributions

For symmetric distributions, the mean and the median are very close to each other.

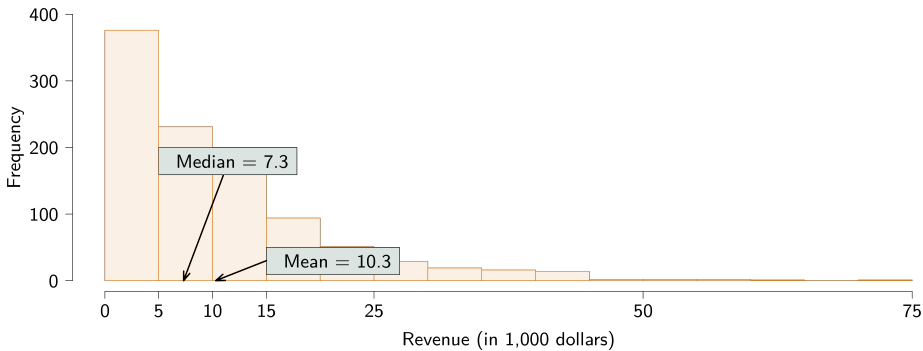


Mean vs median — asymmetric distributions

For asymmetric distributions, the mean and the median may differ greatly.
Typically, the mean is pulled towards the tail of the distribution.

Mean vs median — asymmetric distributions

For asymmetric distributions, the mean and the median may differ greatly.
Typically, the mean is pulled towards the tail of the distribution.



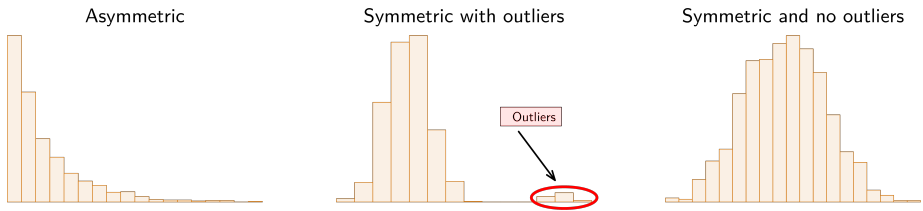
Measure of location: Mean vs median — Which one?

Ordinal variable:
Use the median (the mean is not defined in this case).

Measure of location: Mean vs median — Which one?

Ordinal variable:
Use the median (the mean is not defined in this case).

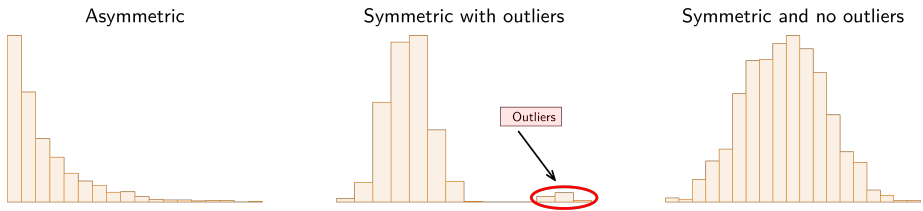
Quantitative variable:
The best advice is to plot the scores and check for symmetry and/or outliers.



Measure of location: Mean vs median — Which one?

Ordinal variable:
Use the median (the mean is not defined in this case).

Quantitative variable:
The best advice is to plot the scores and check for symmetry and/or outliers.



- Distribution asymmetric *or* there are outliers: Use the median (*but you can still also report the mean*).
- Distribution roughly symmetric and with no outliers: Use the mean (*but you can still also report the median*). In this case the mean and the median will be very similar.

Otherwise, the decision to choose between the mean on the median may depend on the question at hand.

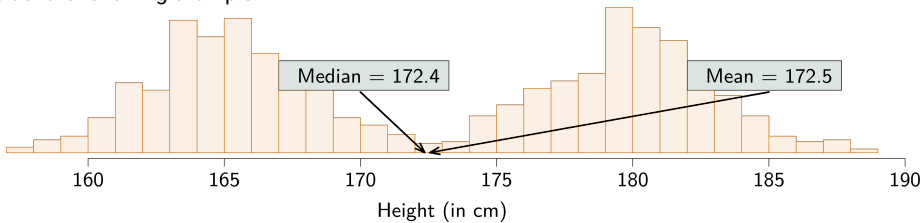
Measure of location: Mean vs median — Neither

For some distributions **neither** the mean **nor** the median are suitable representatives of the location of a distribution…

Measure of location: Mean vs median — Neither

For some distributions **neither** the mean **nor** the median are suitable representatives of the location of a distribution…

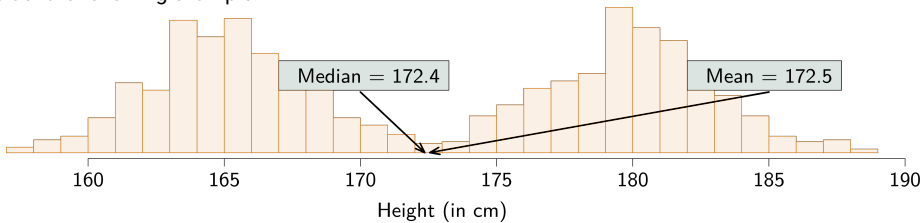
Consider the following example:



Measure of location: Mean vs median — Neither

For some distributions **neither** the mean **nor** the median are suitable representatives of the location of a distribution…

Consider the following example:

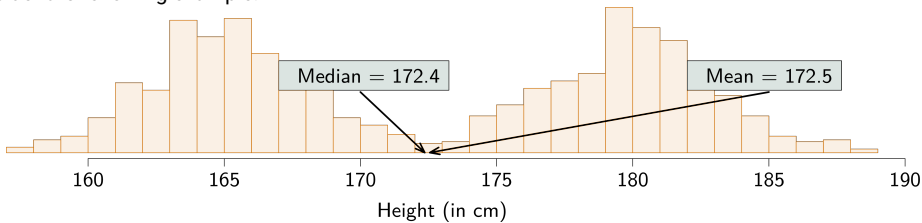


This may be an example of a mixture of two distributions with different characteristics.
Example: Height data with women (left side) and men (right side).

Measure of location: Mean vs median — Neither

For some distributions **neither** the mean **nor** the median are suitable representatives of the location of a distribution…

Consider the following example:



This may be an example of a mixture of two distributions with different characteristics.
Example: Height data with women (left side) and men (right side).

In such cases it would be better to study each distribution separately:

Neither the mean nor the median offer a good representation of the location of this distribution.

Measure of location: Mode

Mode:

The mode is the most frequently observed value.

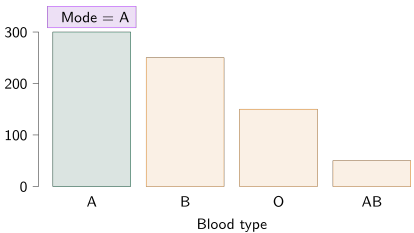
Measure of location: Mode

Mode:

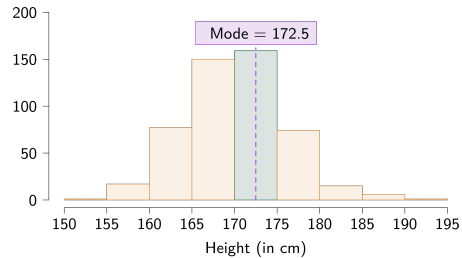
The mode is the most frequently observed value.

The mode can be found as follows:

For qualitative variables:
Find the category with the largest count.



For quantitative variables:
Look at the histogram and…
…find the midpoint of the highest bar in the histogram, or
…find the mean value of the data in the highest bar.



Measure of location: Mode

Note that:

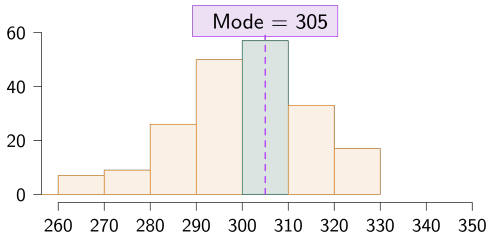
- The mode is a representative value of the location of the distribution, where 'representative' = most frequent.
- The mode can be found for either qualitative or quantitative variables.

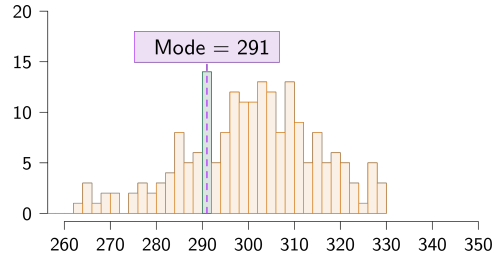
Measure of location: Mode

Note that:

- The mode is a representative value of the location of the distribution, where 'representative' = most frequent.
- The mode can be found for either qualitative or quantitative variables.

However and for quantitative variables, how to set the interval width of the histogram can influence the mode:





Summary

To describe the characteristics of our data we can:

- Use numerical summaries (e.g., mean, median, mode).
- Create plots.

Summary

To describe the characteristics of our data we can:

- Use numerical summaries (e.g., mean, median, mode).
- Create plots.

This is not an either/or decision:

You can, and should, use both!