Universiti Brunei Darussalam

SM-1402 Basic Statistics

Chapter 1: Presentation of data *[handout version]*

Dr. Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Semester II 2022/23

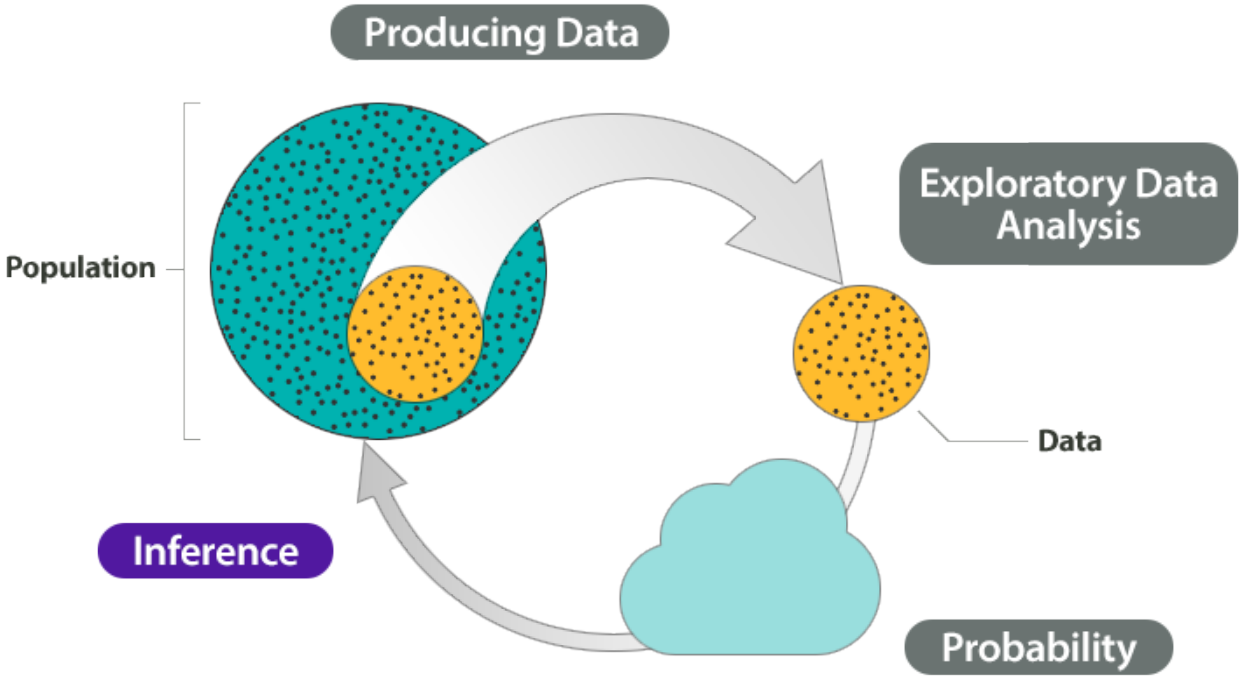
Learning outcomes

- Know which best visualisation tool to use for a particular data type.
- Construct and draw a histogram for continuous data.
- Summarise data in the form of tables for concisely presenting information.

Required reading

- Madsen (2016) Chapters 1 & 2.

Population vs sample



The *population* is the actual entity of interest. Certain reasons constrain us to analyse the *sample*, a subset of the population, instead.

- It may be too costly, time-consuming, or downright infeasible to collect the population data.
- In other cases, the population is "theoretical" e.g. arrival times at hospital A&E.

Running example

Fitness Club

Fitness Club has a number of sports facilities. This includes facilities for strength training, weight loss and cardiovascular workout. Fitness Club wants to understand the needs of their young customers, kids of age 12–17 years. The club wants to know, how satisfied these kids are with the sports facilities. They also want to obtain information about their health in order to better customize the sports facilities for the various types of training.

- **Population of interest:** Kids aged 12–17 years.
- **Sample:** $n = 30$ kids.
- Careful consideration of the organisation of this survey. If interested, read Chapter 6.

Data set

No.	Sex	Age (years)	Height (cm)	Weight (kg)
1	M	12	157	66
2	F	14	151	41
3	M	13	174	58
4	M	13	171	52
5	M	15	198	77
6	F	12	145	59
7	F	13	166	59
8	M	13	141	47
9	M	13	166	45
10	F	13	160	39
⋮	⋮	⋮	⋮	⋮

Source: `Fitness Club Example.xlsx` (`Charts` sheet).

Data types

Typically in a survey, two kinds of questions are asked:

1. Background questions (demographic questions).
    - Sex, age, marital status, residence, education, employment, annual income, etc.
2. Study questions; formulated by researchers to answer the research objective.
    - Closed questions (limited categories)
    - Open questions (unlimited categories/responses)
    - Assessment questions (on some scale)

We can categorise the data collected into several types:

a. Nominal data (takes limited values/categories)
b. Ordinal data (categorical, but in order)
c. Integer data (counts)
d. Continuous interval data (may be bounded or unbounded)
e. Continuous ratio data (has a true zero)

Data

Visualisations

- Bar charts
- Histograms
- Pie charts
- Scatter plots
- Line charts
- Bubble plots

Tables

Visualisations

Human brains are terribly adapted to processing raw data. Processing the data, for example, by way of visuals, allows us to:

- get a feel of patterns, structures, trends and relationships in data; and
- reveal unlikely data values (outliers)

We shall look at

1. Bar charts
2. Histograms
3. Pie charts
4. Scatter plots
5. Line charts

Alternatively, tables are great at summarising data too.

Bar charts

Definition 1 (Bar charts)

Bar charts represent numerical values of variables as the height (or length[1]) of lines or rectangles of equal width.

The two components to think about are

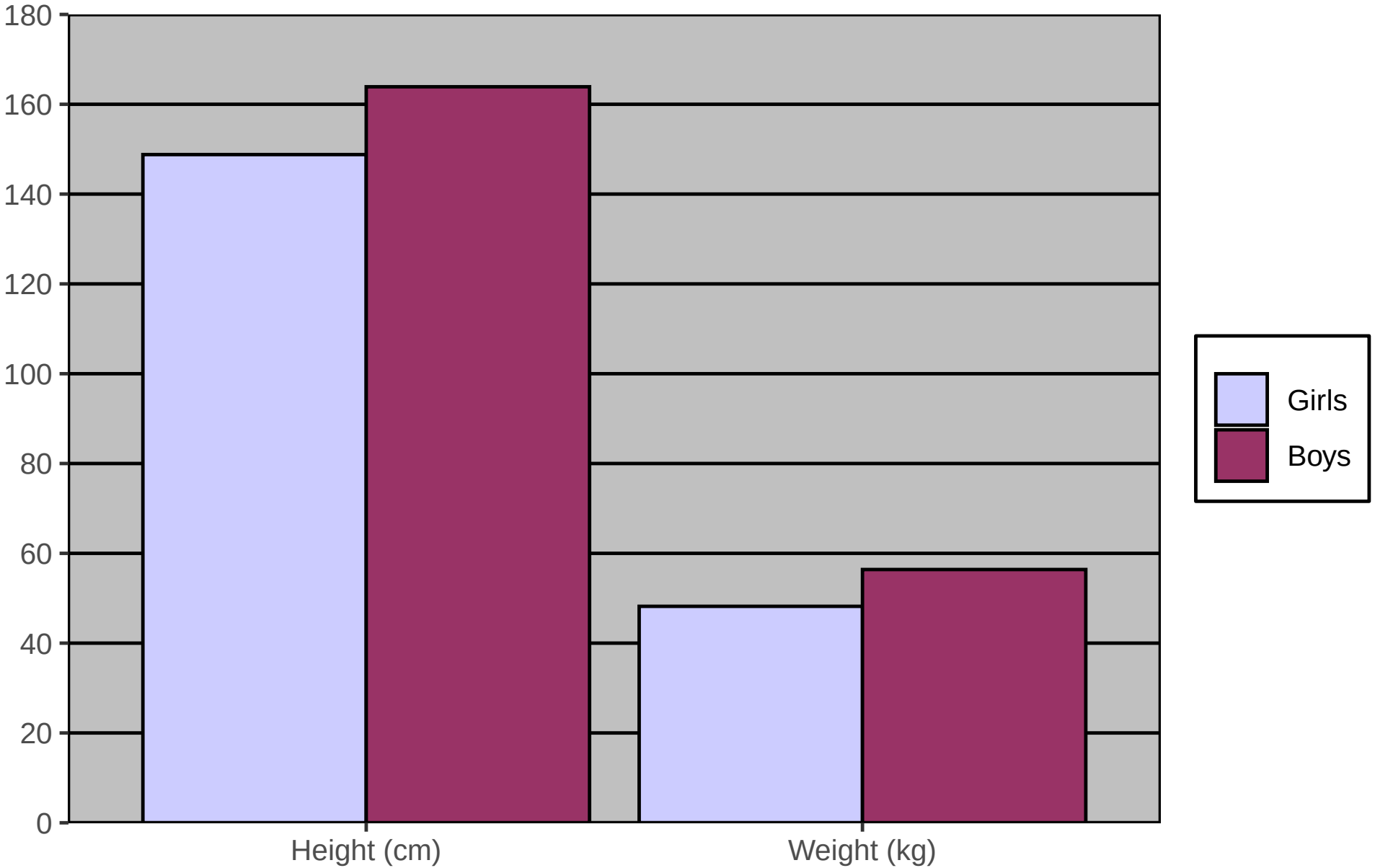
- Horizontal axis: Grouped information (usually demographic clusters)
- Vertical axis: Numeric value (usually averages)

In our Fitness Club example, suppose we wished to look at the height and weight data of girls and boys. A meaningful comparison would the average:

Sex	Height (cm)	Weight (kg)
Girls	148.8	48.2
Boys	163.9	56.4

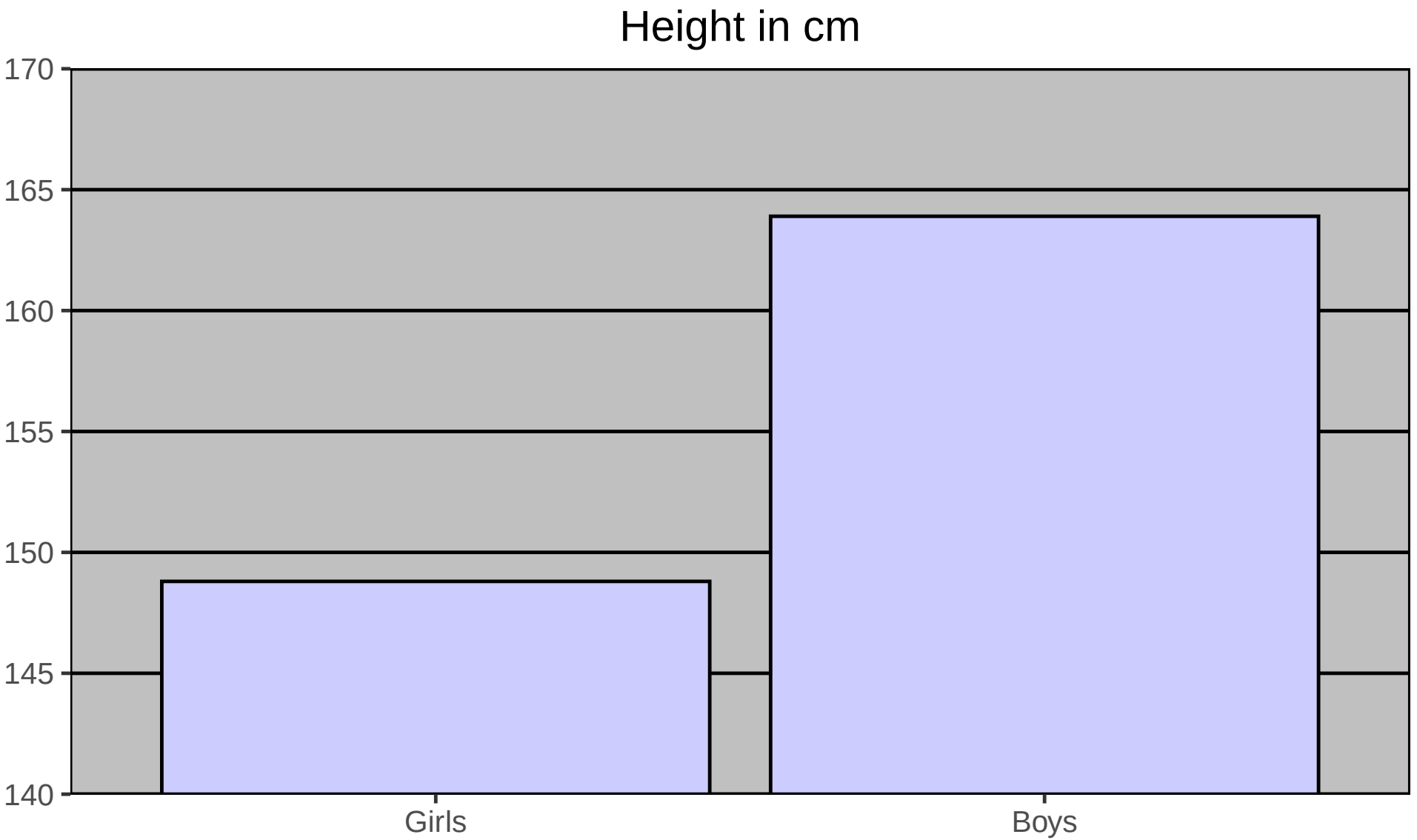
[1] Of course, it's possible to switch axes and have a horizontal bar plot instead.

Bar charts



Bad example

Same information, but misleading chart due to the axes!



Histograms

Definition 2 (Histogram)

A histogram is an approximate representation of the *distribution* of numerical data. It depicts frequencies of observations occurring in certain ranges of values.

Height (in cm) of respondents sorted in ascending order.

112 115 118 125 127 139 141 145 151
151 152 154 157 157 159 160 162 166
166 166 170 170 170 171 174 176 184
185 192 198

Interested in

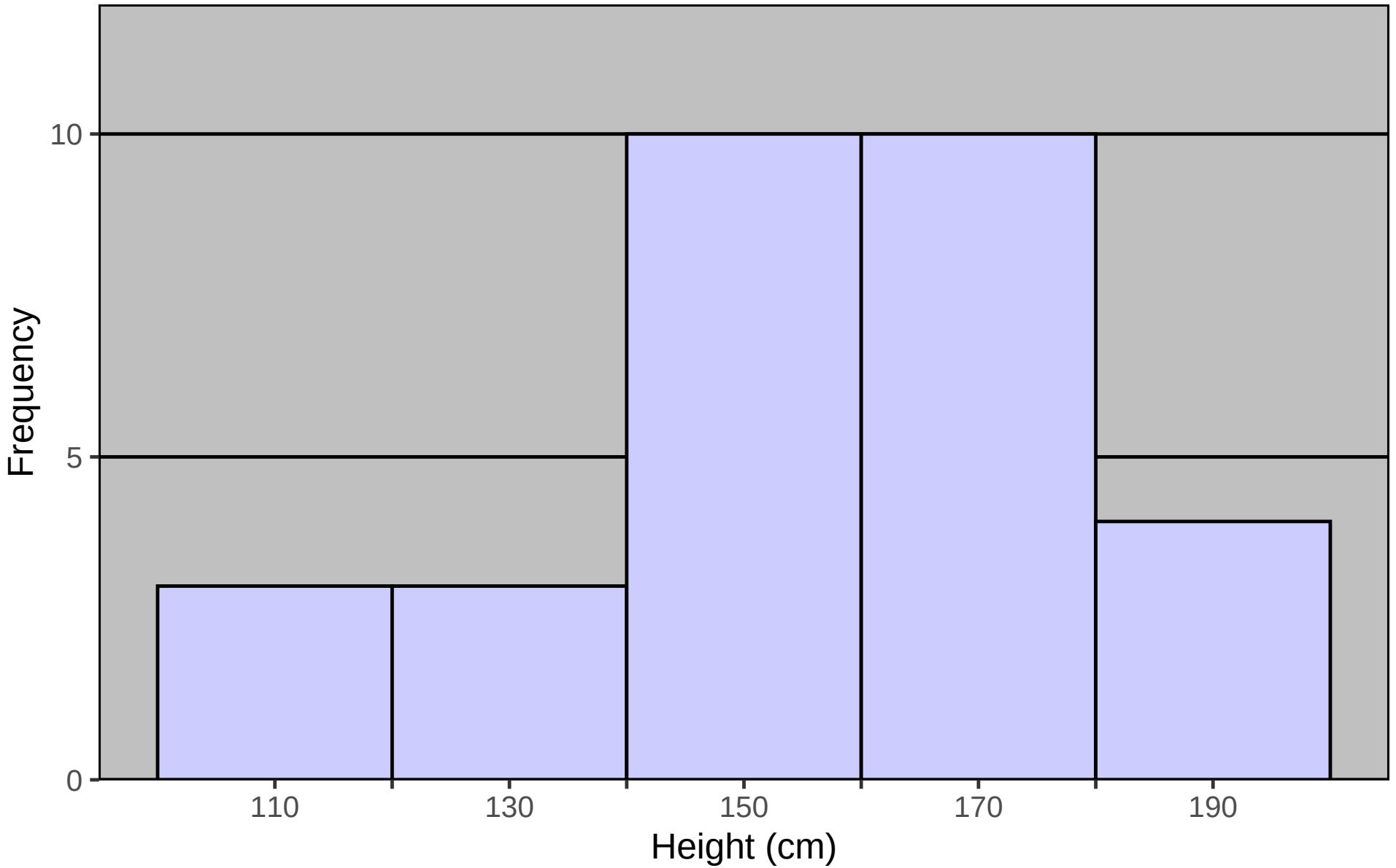
- Min & max
- Middle value
- Spread and shape

Group the data[a] in bins:

From	To	Freq	Centre
100	120	3	110
120	140	3	130
140	160	10	150
160	180	10	170
180	200	4	190

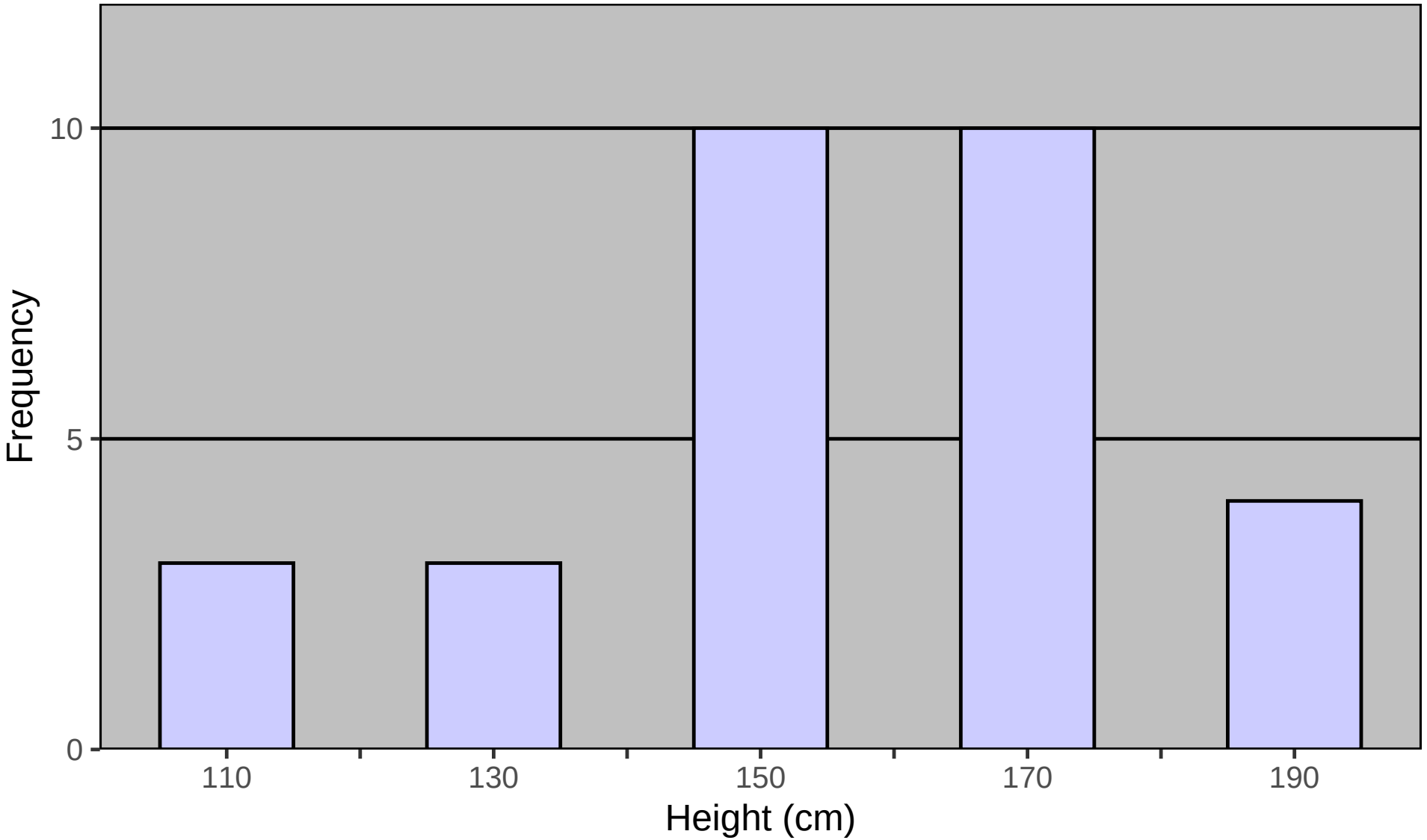
[a] `Histograms` sheet

Histograms (cont.)



Spaces between bars?

Best not to!



Histogram considerations

1. Accommodate all observations–so must be aware of minimum and maximum values.
2. Appropriately choosing the number of bins/intervals.
   - Too few and you'll "hide" the true shape
   - Too many and you'll get an overly complex shape

   Aim for 3–13 bars. The more the observations, the more the bars. However, most software chooses this for you automatically.
3. Define the *intervals* (`From--To`) clearly. Typically,
   - `From` means 'in the interval from **but not including**' (greater than)
   - `To` means 'up to **and including**' (less than or equal to)

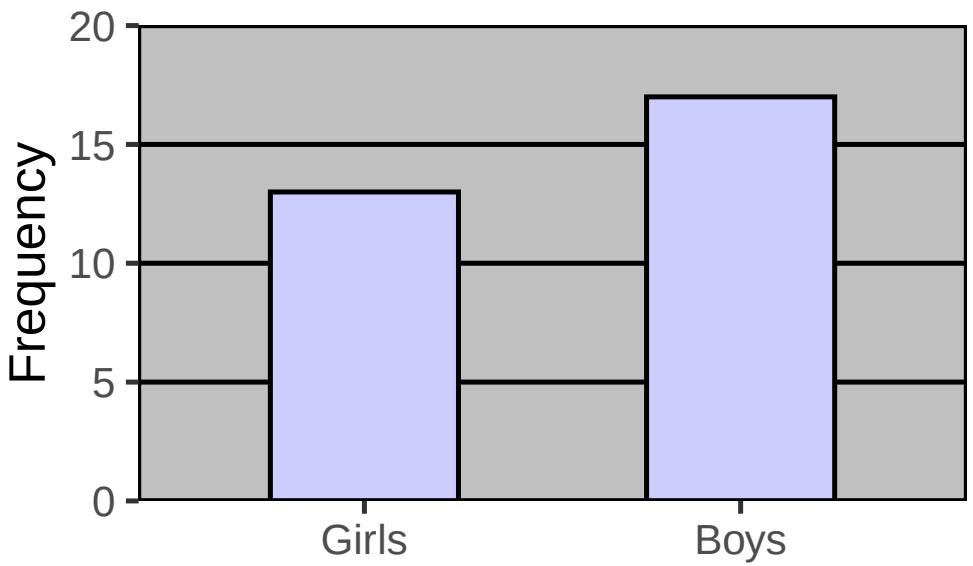
   so that each observation belongs to one and only one interval. Note that some histograms have uneven interval widths. These are much more difficult to read–avoid!

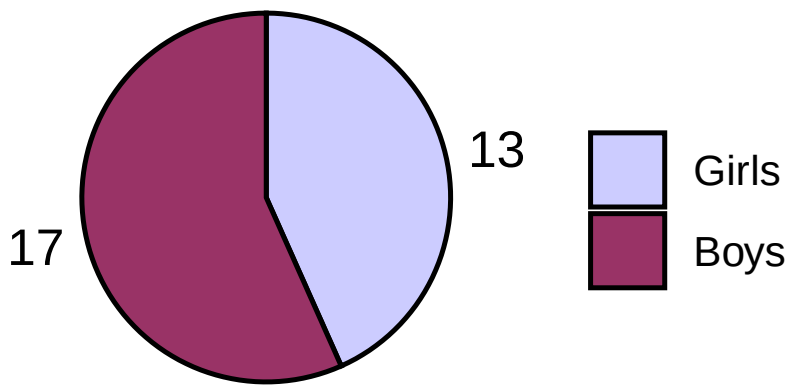
Pie charts

Definition 3 (Pie charts)

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

Sex	Frequency
Girls	13
Boys	17





Scatter plots

Definition 4 (Scatter plot)

A scatter plot is a graphic to display the values of two variables, often to show the relationship between the two variables.

Conjecture

There is a relationship between height and weight.

If we plot a graph such that

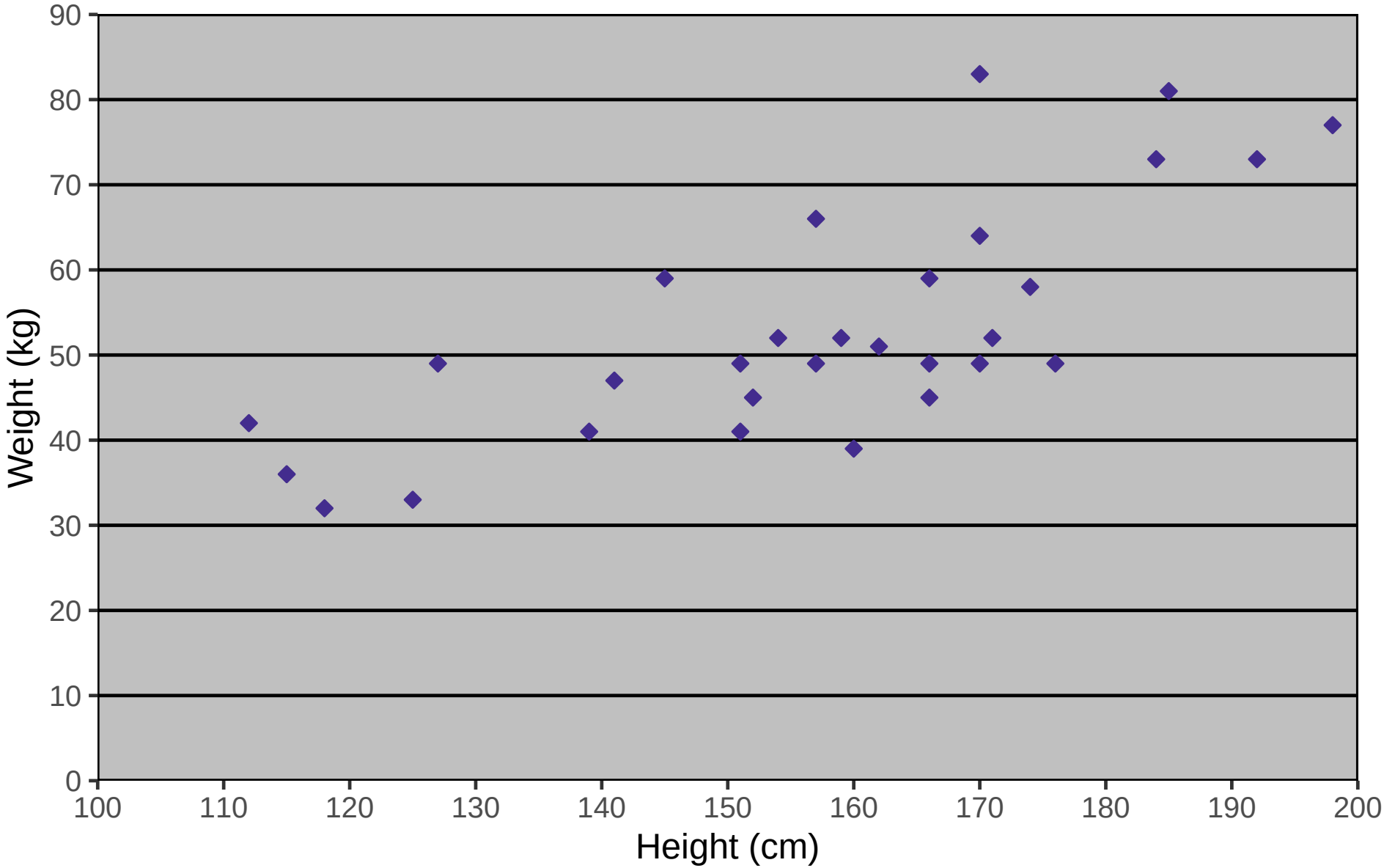
- Height (cm) is on the X-axis [independent variable, aka "cause"]
- Weight (cm) is on the Y-axis [dependent variable, aka "effect"]

then the scatter plot may support our conjecture (or not!).

Caution

One cannot, by statistical methods or by studying graphs, determine whether there is a "cause" and an "effect". Correlation does not imply causation–more on this in coming chapters.

Scatter plots (cont.)



Line charts

Definition 5 (Line charts)

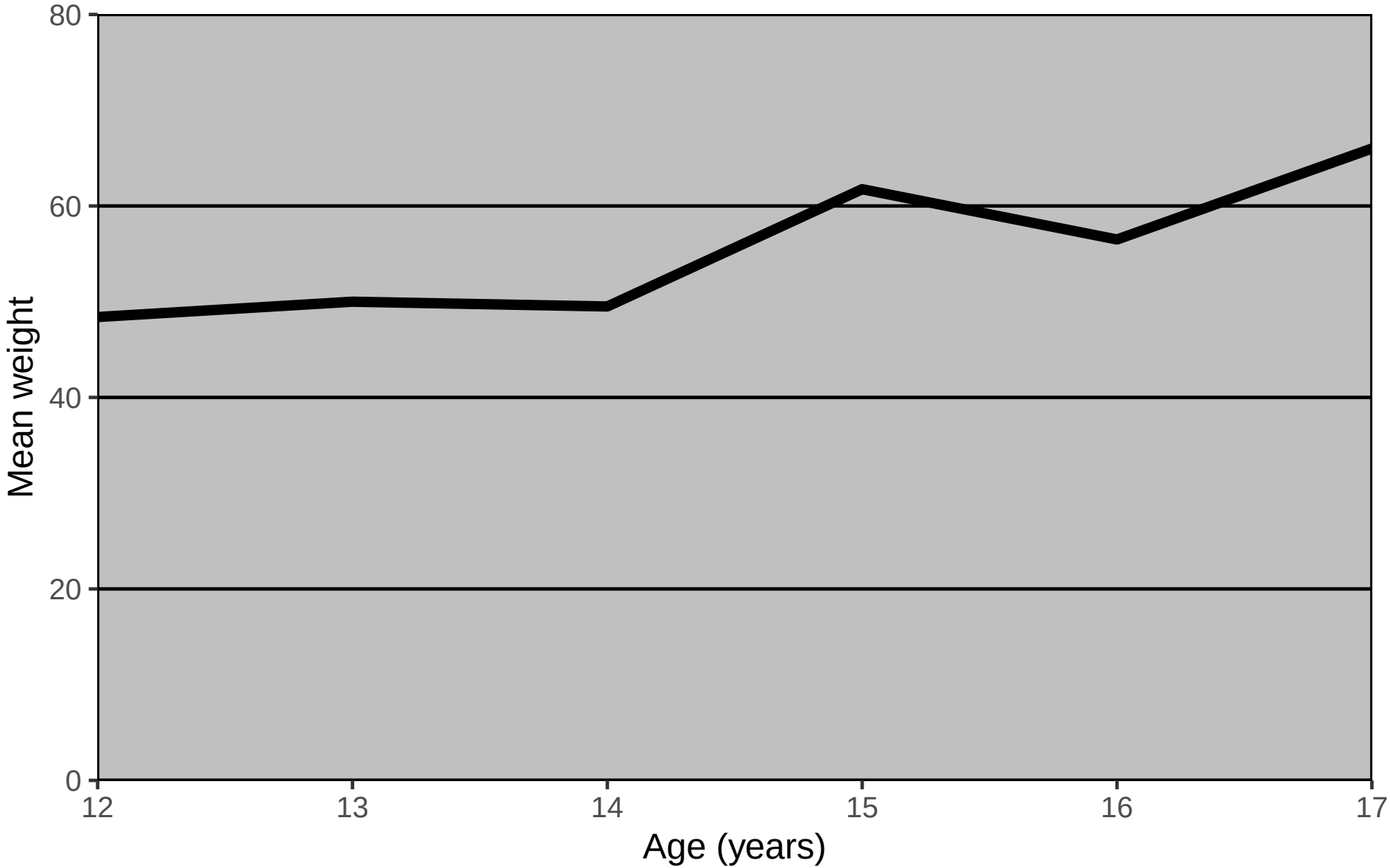
Line charts connect two-dimensional observations together with a straight line, and are used to illustrate a *trend* (typically the X variable is time).

For the Fitness Club example, let's show the *average* weight (Y) of the kids by age (X).

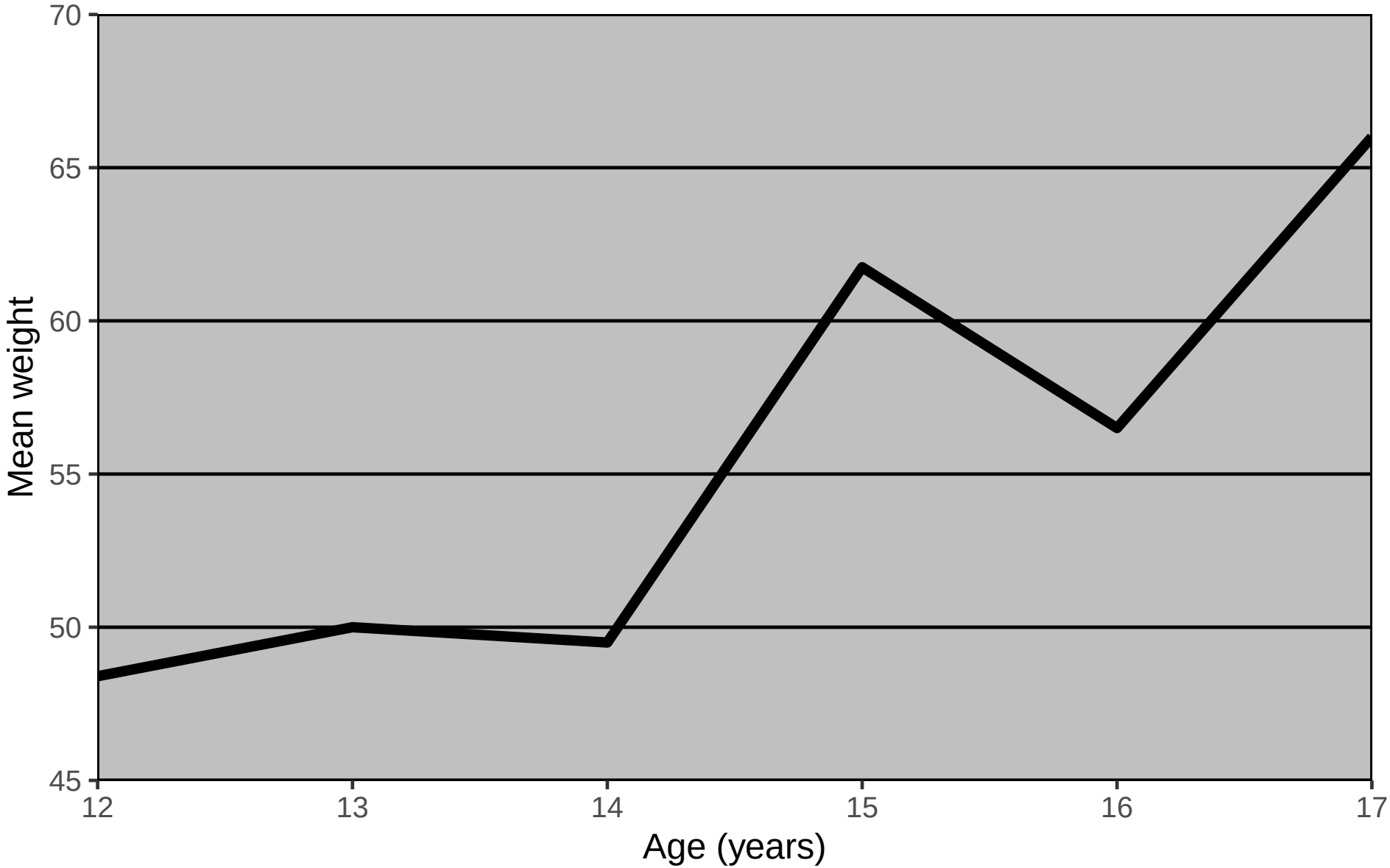
We can then visualise these data points in the form of a line chart.

Age (years)	Mean weight
12	48.40
13	50.00
14	49.50
15	61.75
16	56.50
17	66.00

Line charts (cont.)



Careful with axes

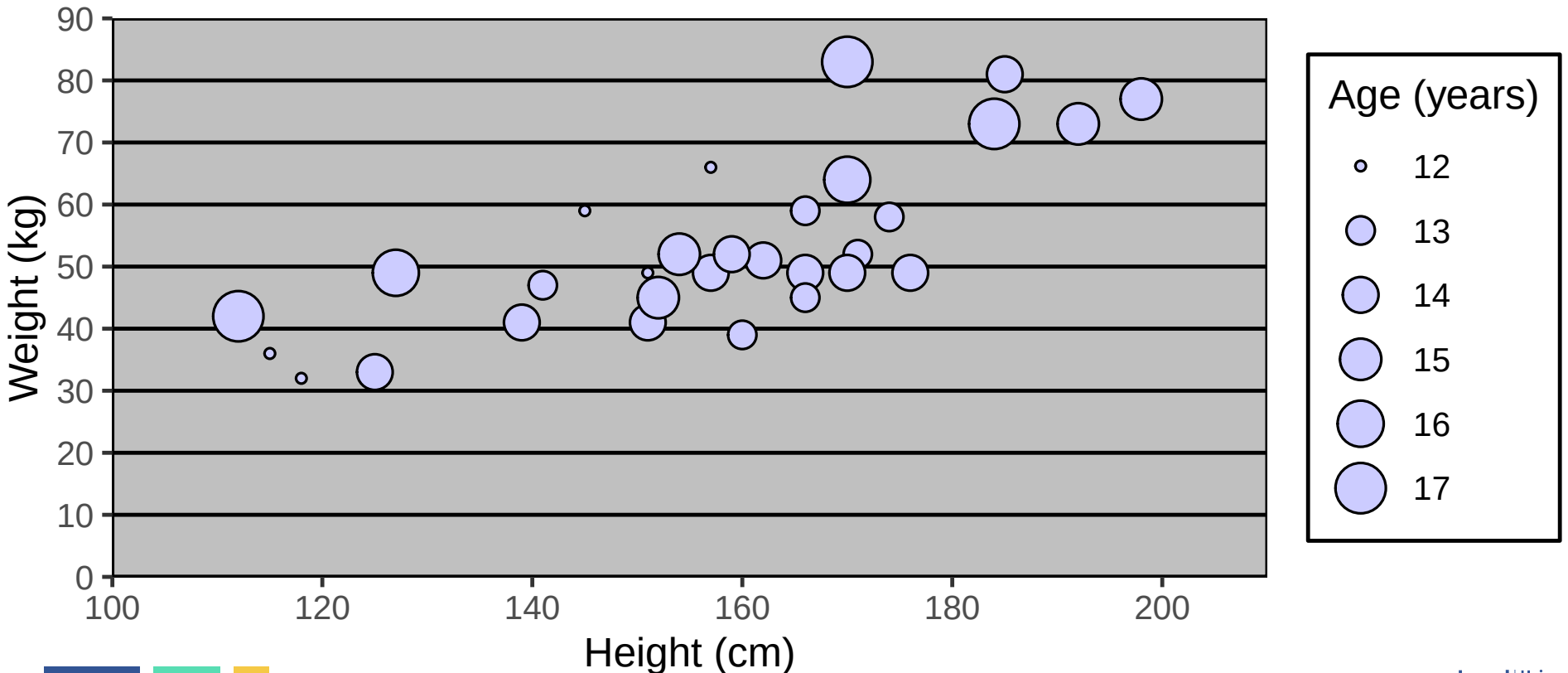


Bubble plots

Definition 6 (Bubble plots)

A variant of the scatter plot, bubble plots depict "bubbles" whose sizes represent the value of a third variable.

Continuing our scatter plot example, let the third variable (bubbles) be age.



Data

Visualisations

Tables

Tables

Visualisations might be more preferred than tables, but used sparingly, tables can also offer a way to concisely convey information. Think about what you want to convey:

- Frequencies (count) of grouping of data
- Average value of variables
- Other statistics such as minimum, maximum, median, etc.

Table 1: No. of kids by sex and age.

	Age			
Sex	12–13	14–15	16–17	Total
Girls	5	6	2	13
Boys	6	8	3	17
Total	11	14	5	30

Source: Sample survey, Fitness Club.
Other notes and comments.

Tables (cont.)

A common scenario: Summarise variables to compare and contrast two or more demographic groups.

Table 2: Demographic comparison between girls and boys.

	Girls			Boys		
	Min	Average	Max	Min	Average	Max
Height (cm)	112.0	148.8	185.0	115.0	163.9	198.0
Weight (kg)	32.0	48.2	81	36.0	56.9	83
Age (years)	12.0	13.8	17	12.0	14.2	17

Percentages

Sample frequencies themselves may be less interesting, than say, percentages, which allow for comparison to population percentages (if available).

Table 3: No. of kids by sex and age, percent

	Age			
Sex	12–13	14–15	16–17	Total
Girls	16.7%	20.0%	6.7%	43.3%
Boys	20.0%	26.7%	10.0%	56.7%
Total	36.7%	46.7%	16.7%	100%

This is just Table 1 divided by the total number of sample respondents (30).

Percentage of what?

Table 4: No. of kids by sex and age, row percent.

	Age			
Sex	12–13	14–15	16–17	Total
Girls	38.5%	46.2%	15.4%	100%
Boys	35.3%	47.1%	17.6%	100%

Table 5: No. of kids by sex and age, column percent.

	Age		
Sex	12–13	14–15	16–17
Girls	45.5%	42.9%	40.0%
Boys	54.5%	57.1%	60.0%
Total	100%	100%	100%

Reponses to survey questions

Questionnaire contained

1. Do you do cardiovascular workouts?
2. How would you assess your physical fitness?

Table 6: Cardiovascular workout and physical fitness.

	Physical fitness				
Cardiovascular workouts?	Bad	Medium	Good	Total	Freq
No	40%	40%	20%	100%	15
Yes	20%	40%	40%	100%	15

Possible to suggest a trend that cardio workouts have an impact on fitness?