Presentation of Data

Chapter 3

Methods of Presenting Data

Textual Presentation of Data – method of presenting the collected data in organized and narrative (paragraph) form.

Ex: Motorcycle Helmets – Can You See Those Ears?

In 2003, the US Department of transportation established standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report "Motorcycle Helmet Use in 2005-Overall Results" (National Highway Traffic Safety Administration, Aug. 2005) summarized data collected in June 2005 by observing 1700 motorcyclists nationwide. In total, there were 731 riders who wore no helmet, 153 who wore a noncompliant helmet, and 816 who wore a compliant helmet.

Tabular Presentation of Data – method of presenting the collected data by means of tables.

Components:

A. Heading – contains table number and table name

B. Box Head – categories contained in the column

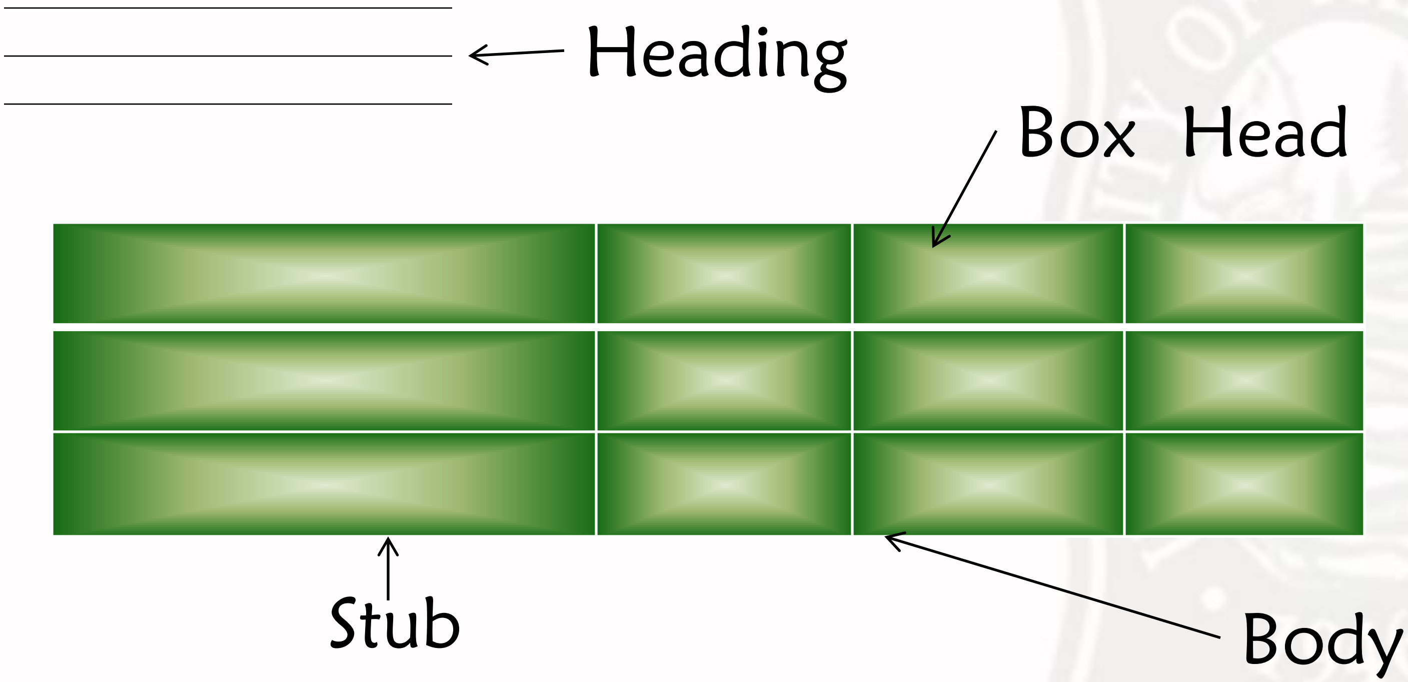
C. Stub – labels

D. Body – quantitative data

E. Foot-notes – information about data

F. Source of data

Illustration:



Source Note

Footnote

Frequency distribution – summary of the categories with their corresponding frequency

Frequency – actual presentation

- actual count

Relative Frequency (rf) – use to compare different distribution of the same situation

$$rf = \frac{\text{frequency of the categories}}{\text{total frequency}}$$

Nominal Data – categories are alphabetize

- chronological order (1st , 2nd,….)

Ex: Course distribution of Math CS 4 students

course	frequency (f)	rf (%)
ACT	6	0.15 or 15%
BSCS	8	0.20 or 20%
BSIT	23	0.575 or 57.5%
BSIS	3	0.075 or 7.5%
total	**40**	**100%**

Ordinal Data – categories are in rank order

ex: The instructor comes to class prepared for the lesson.

	f	rf(%)
Strongly agree		
Agree		
Disagree		
Strongly disagree		
total		

Interval and Ratio Data

Kinds:

1. Simple Frequency – short gap between lowest and highest data.

   ex: Age distribution of Math CS 4 students

Age	f	rf(%)

2. Grouped Frequency – wide gap between the highest to lowest class intervals.

Ex: Frequency distribution table (FDT) (to be discussed in detail)

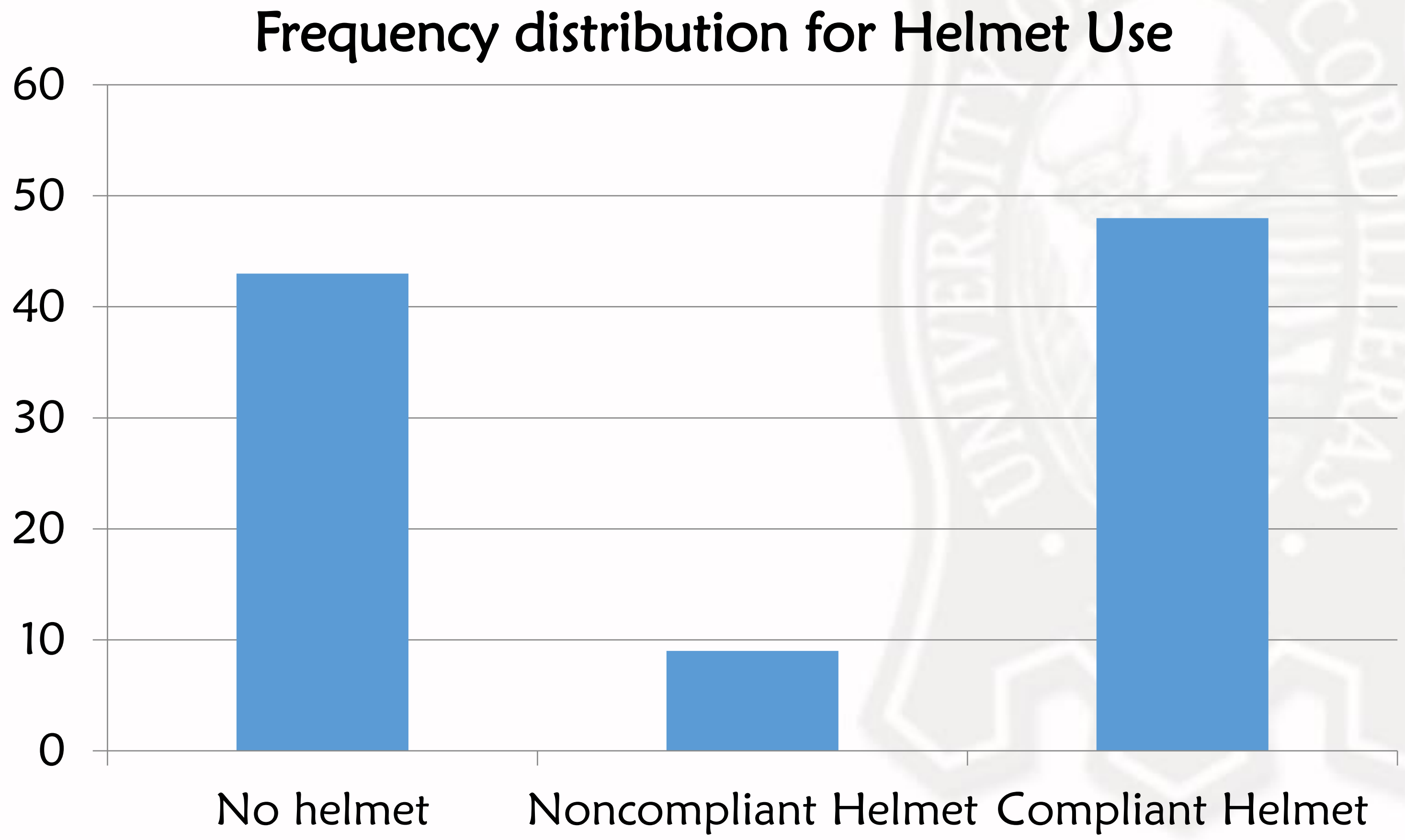
Graphical Presentation of Data – method of presenting data by means of graphs, charts or pictures.

Graphs – pictures of numerical data

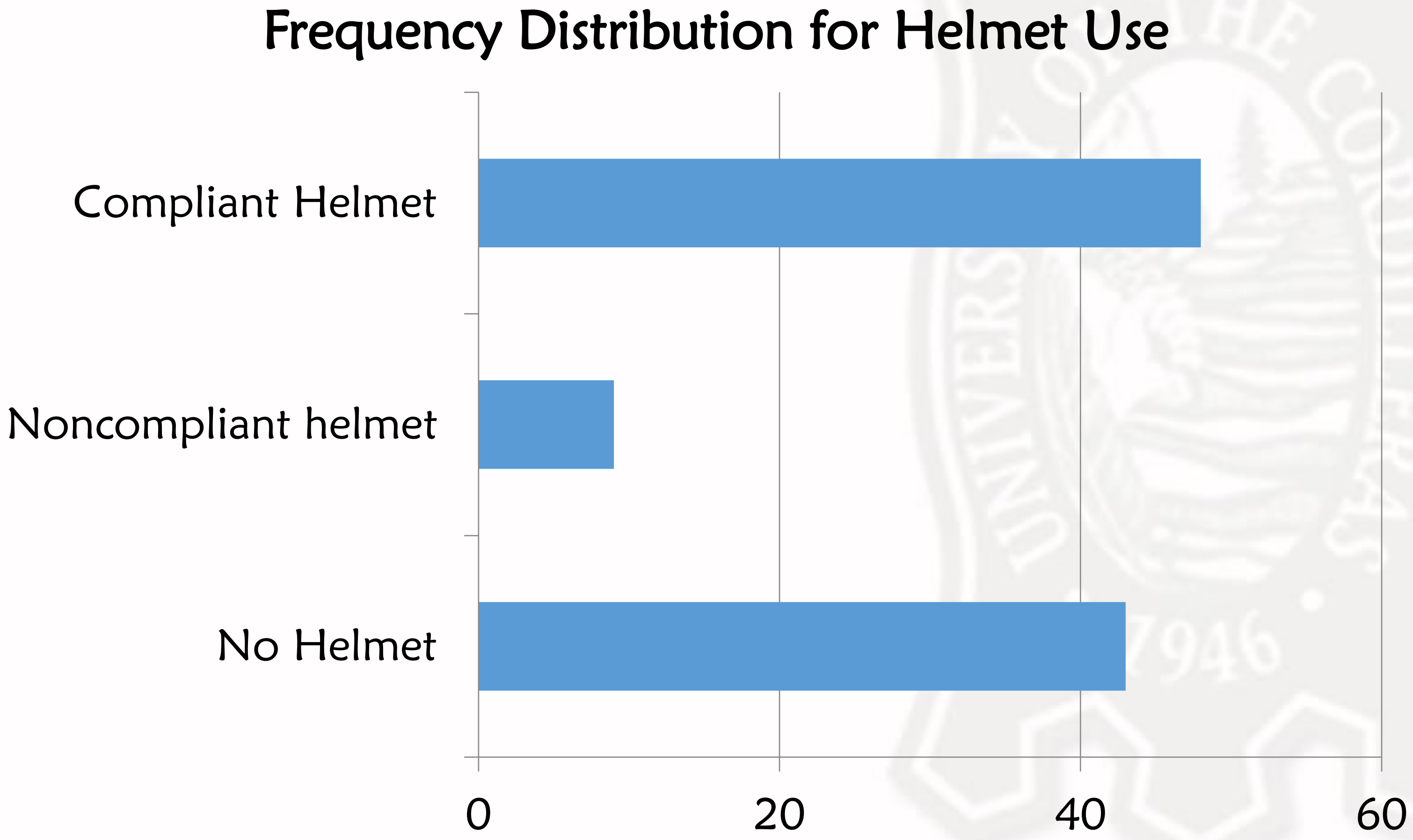
1. Bar Graphs – useful for making direct visual comparisons over a period of time.

Types:

a. Vertical bar graph



b. Horizontal Bar Graph



Quiz: Answer the ff. questions.

1. The Computer Assisted Assessment Center at the University of Luton published a report titled “Technical Review of Plagiarism Detection Software.” The authors of this report asked faculty at academic institutions about the extent to which they agreed with the statement “Plagiarism is a significant problem in academic institutions.” The responses are summarized in the accompanying table. Construct a **bar chart** of these data.

Response	frequency
Strongly disagree	5
Disagree	48
Not sure	90
Agree	140
Strongly agree	39

2. The article “So Close, Yet So Far: Predictors of Attrition in College Seniors” (*Journal of College Student Development[1998]*) examined the reasons college seniors leave their college programs before graduating. Forty-two college seniors at a large public university who dropped out before graduation were interviewed and asked the main reason for discontinuing enrolment at the university. Data consistent with that given in the article are summarized in the following frequency distribution:

Reason for leaving the University	frequency
Academic problems	7
Poor advising or teaching	3
Needed a break	2
Economic reasons	11
Family responsibilities	4
To attend another school	9
Personal problems	3
Others	3

Kinds:

1. Simple bar graphs – (as illustrated previously)
2. Compound (Multiple) bar graph – use to compare two or more variables

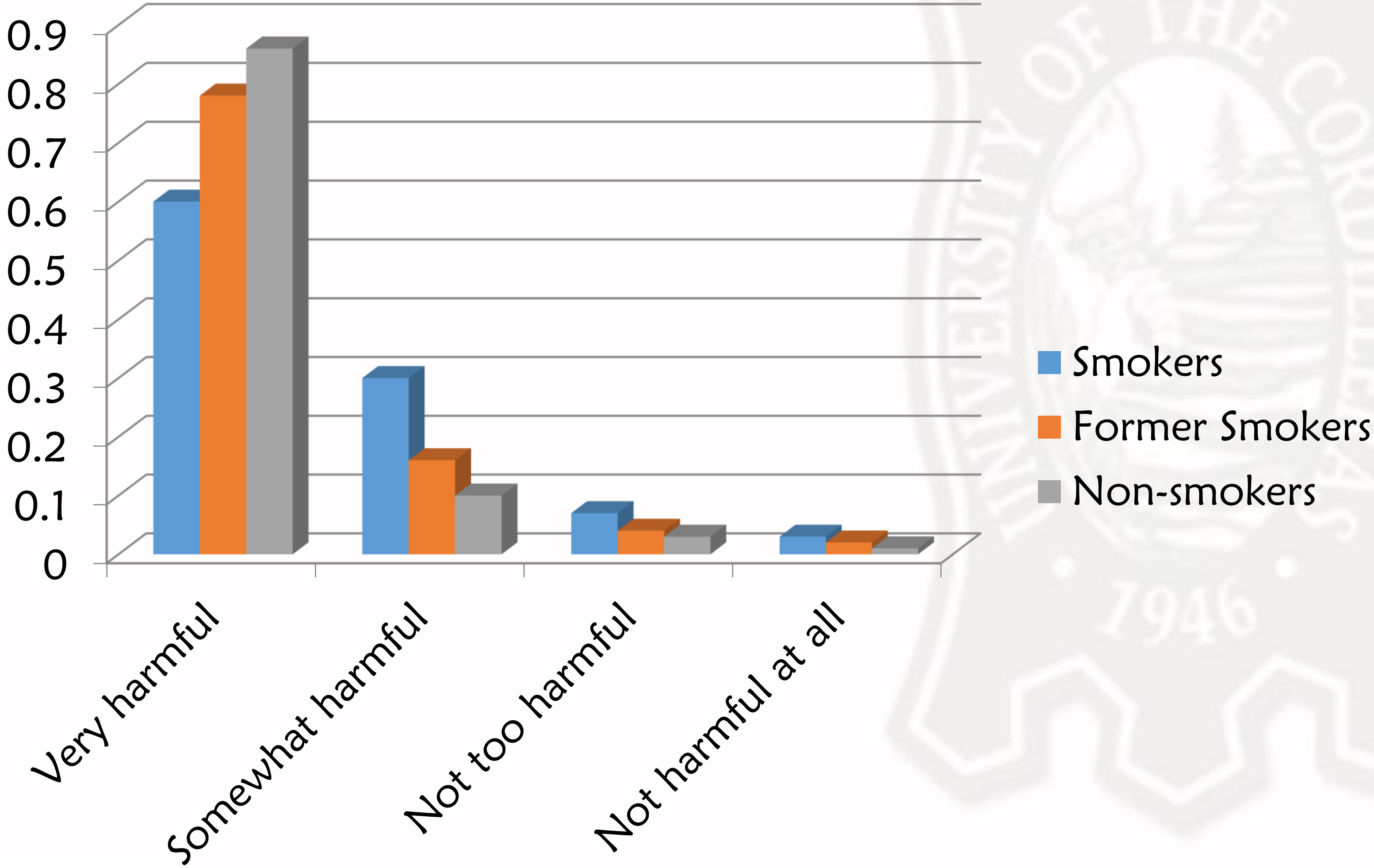
Ex: Perceived Risk of Smoking

The article "Most Smokers Wish They Could Quit" (*Gallup Poll Analyses*, Nov. 21, 2002) noted that smokers and non-smokers perceive the risks of smoking differently. The accompanying relative frequency table summarizes responses regarding the perceived harm of smoking for each three groups: a sample of 241 smokers, a sample of 261 former smokers, and a sample of 502 non-smokers.

frequency distribution:

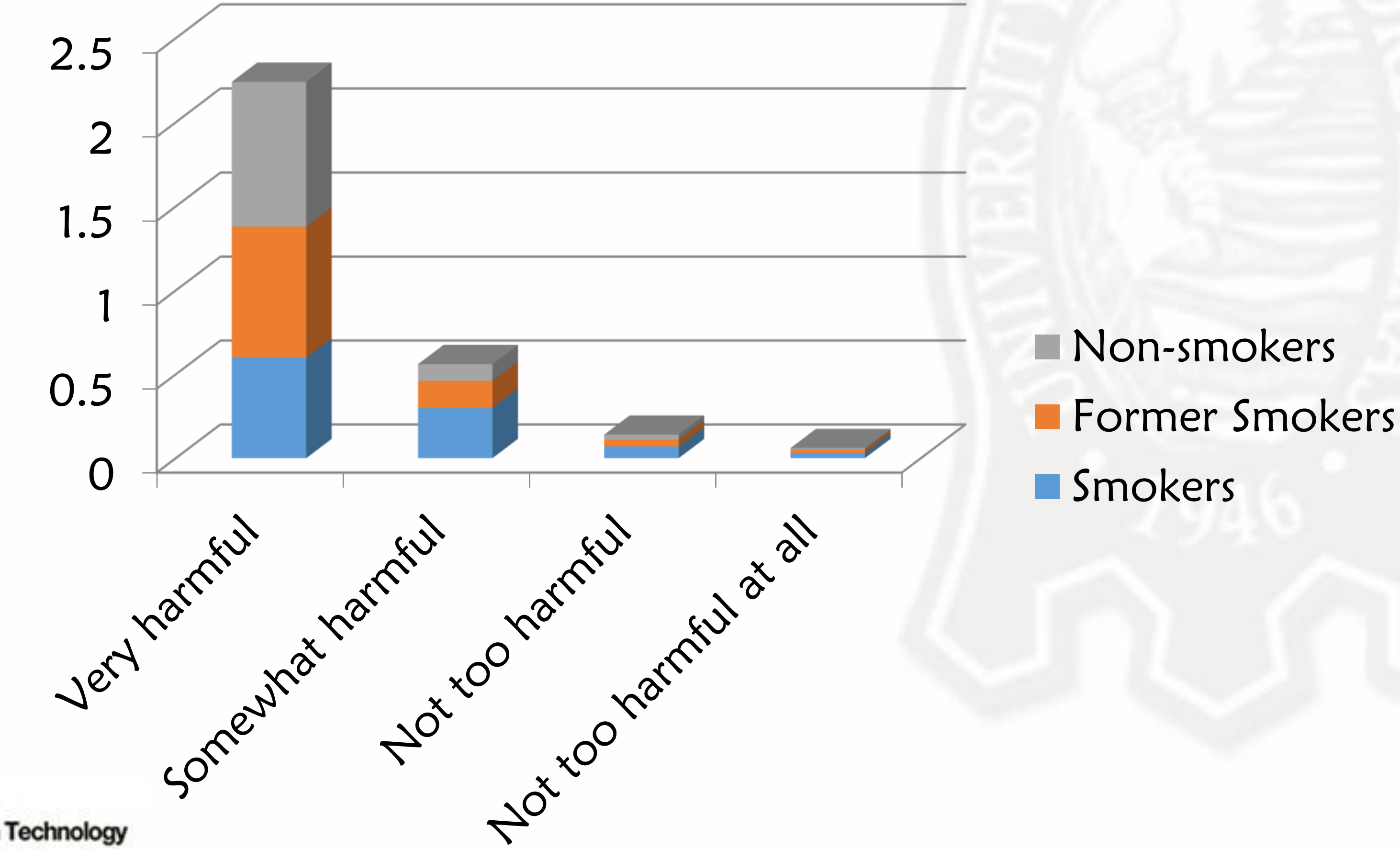
Perceived risk of smoking	frequency			Relative frequency		
	S	FS	NS	S	FS	NS
Very harmful	145	204	432	0.60	0.78	0.86
Somewhat harmful	72	42	50	0.30	0.16	0.10
Not too harmful	17	10	15	0.07	0.04	0.03
Not harmful at all	7	5	5	0.3	0.02	0.01
TOTAL	**241**	**261**	**502**	**1.00**	**1.00**	**1.00**

Graph:


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Very harmful
Somewhat harmful
Not too harmful
Not harmful at all
Smokers
Former Smokers
Non-smokers

3. Component bar graph – divides or breaks down the quantities into their components.

Ex: Perceived Risk of Smoking



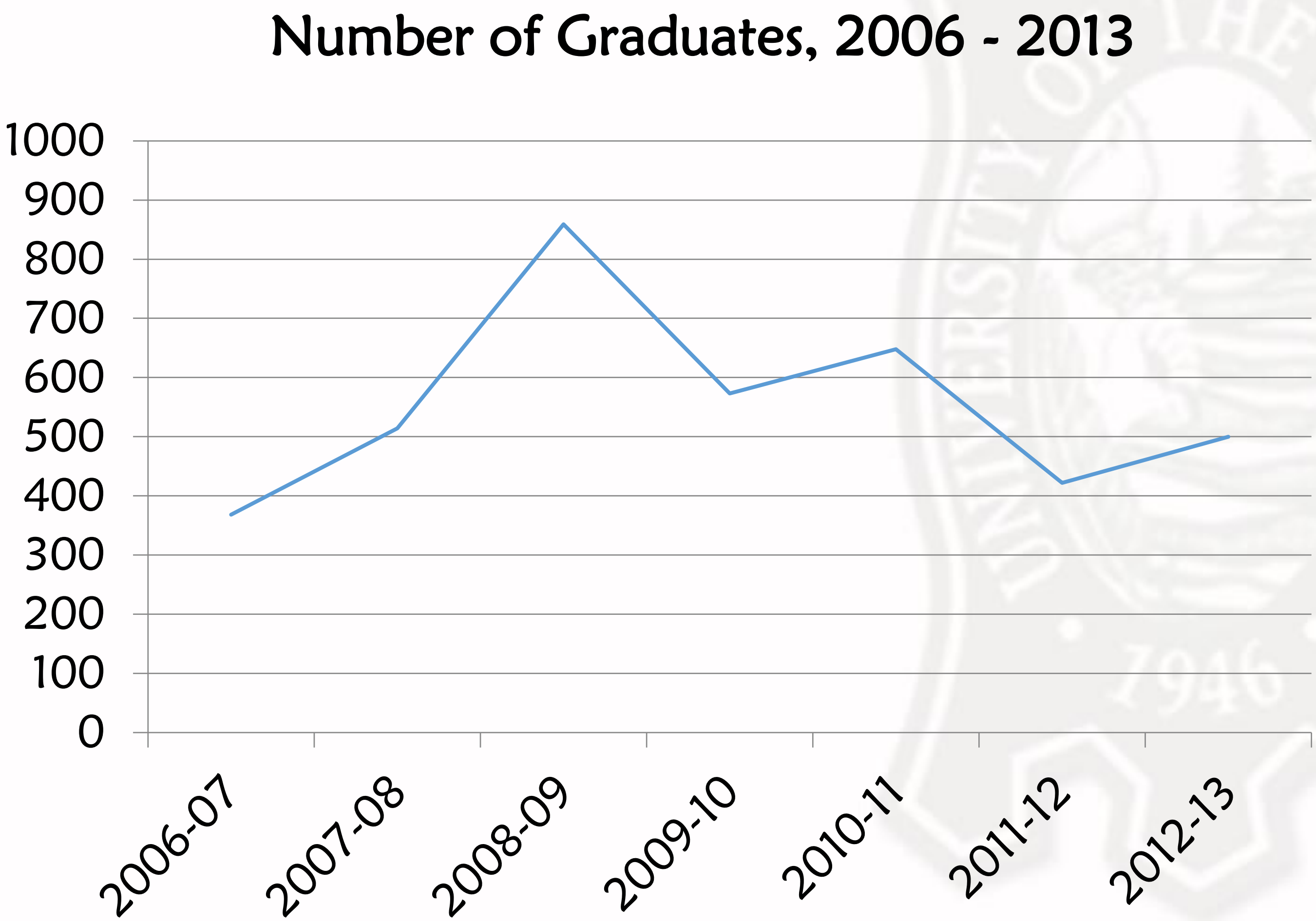
2. Line Graphs – used to picture related facts

- useful for plotting data over a period of time to indicate patterns or trends. The pattern is used to make predictions.

Ex: Based on the list below of college students who took Bachelor of Science in Information Technology, draw a line graph.

School Year	No. of students
06-07	368
07-08	514
08-09	859
09-10	573
10-11	648
11-12	422
12-13	500

Graph:

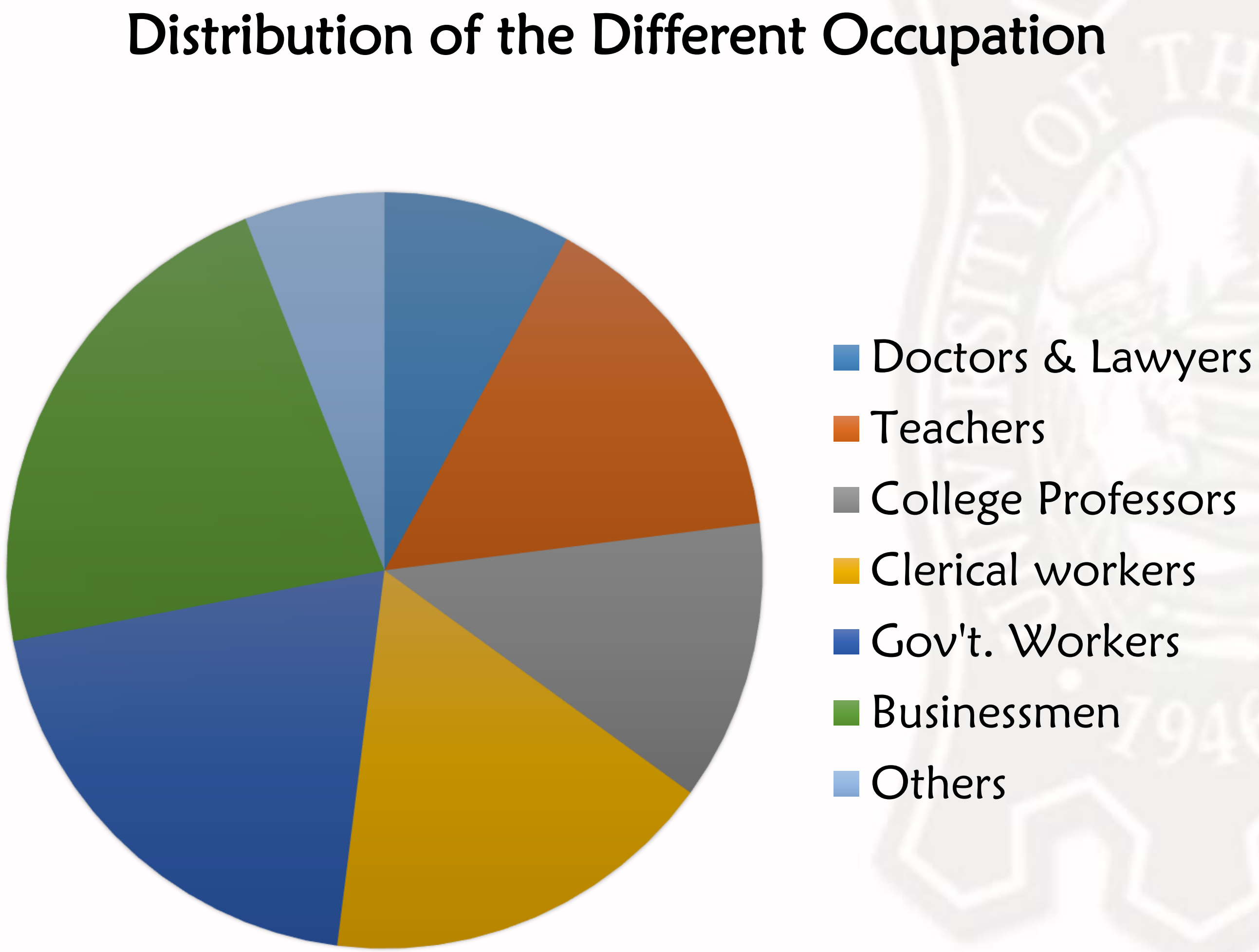

Number of Graduates, 2006 - 2013
1000
900
800
700
600
500
400
300
200
100
0
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13

3. Pie or Circle Graphs – shows the relative sizes of parts of a whole.

Ex: Distribution of 10,000 working people among the different occupation in a certain city:

Occupation	Number	%tage	"Slice"
Doctors & lawyers	800	8%	28.8°
Teachers	1,500	15%	54°
College professors	1,200	12%	43.2°
Clerical/Office Workers	1,700	17%	61.2°
Gov't Workers	2,000	20%	72°
Businessmen	2, 200	22%	79.2°
others	600	6%	21.6°
TOTAL	10,000	100%	360°

Graph:

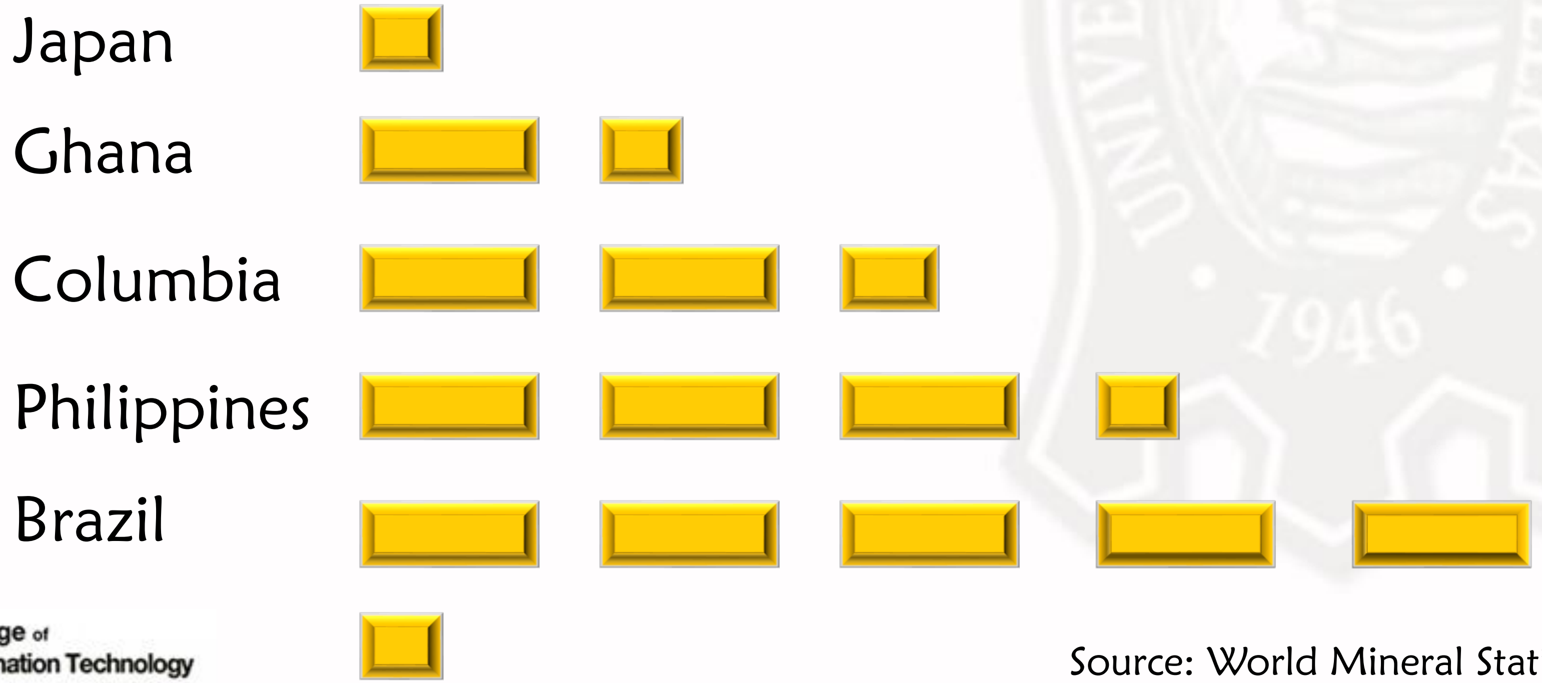

Distribution of the Different Occupation
Doctors & Lawyers
Teachers
College Professors
Clerical workers
Gov't. Workers
Businessmen
Others

4. Pictograph – use icons/pictures to symbolize the quantities being presented.

- show changes in quantity

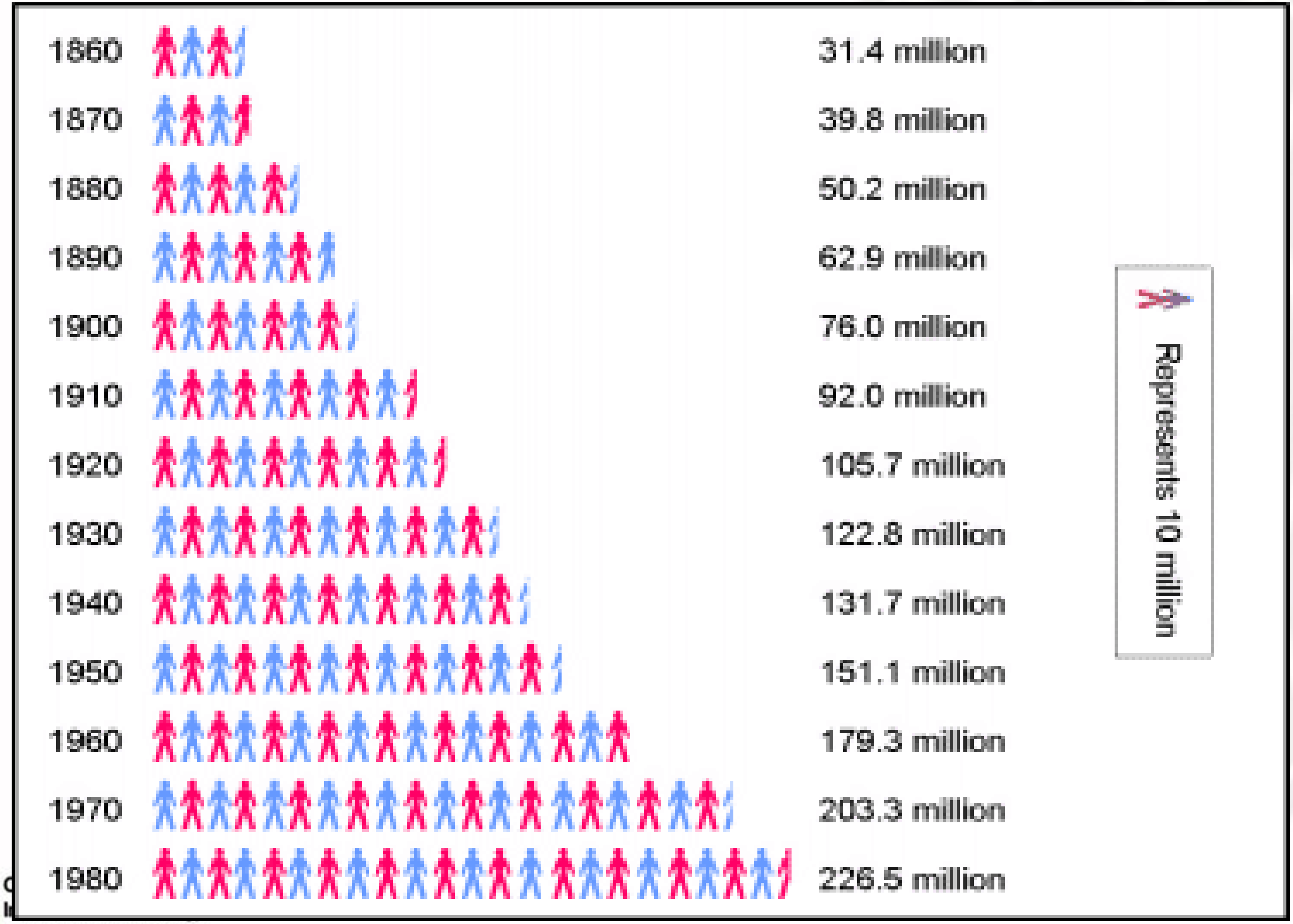
- make comparison between similar situations.

Ex: Gold Production in Selected Countries (in 10,000 kilograms) 1988



Source: World Mineral Statistics

Ex 2:


1860
31.4 million
1870
39.8 million
1880
50.2 million
1890
62.9 million
1900
76.0 million
1910
92.0 million
1920
105.7 million
1930
122.8 million
1940
131.7 million
1950
151.1 million
1960
179.3 million
1970
203.3 million
1980
226.5 million
Represents 10 million

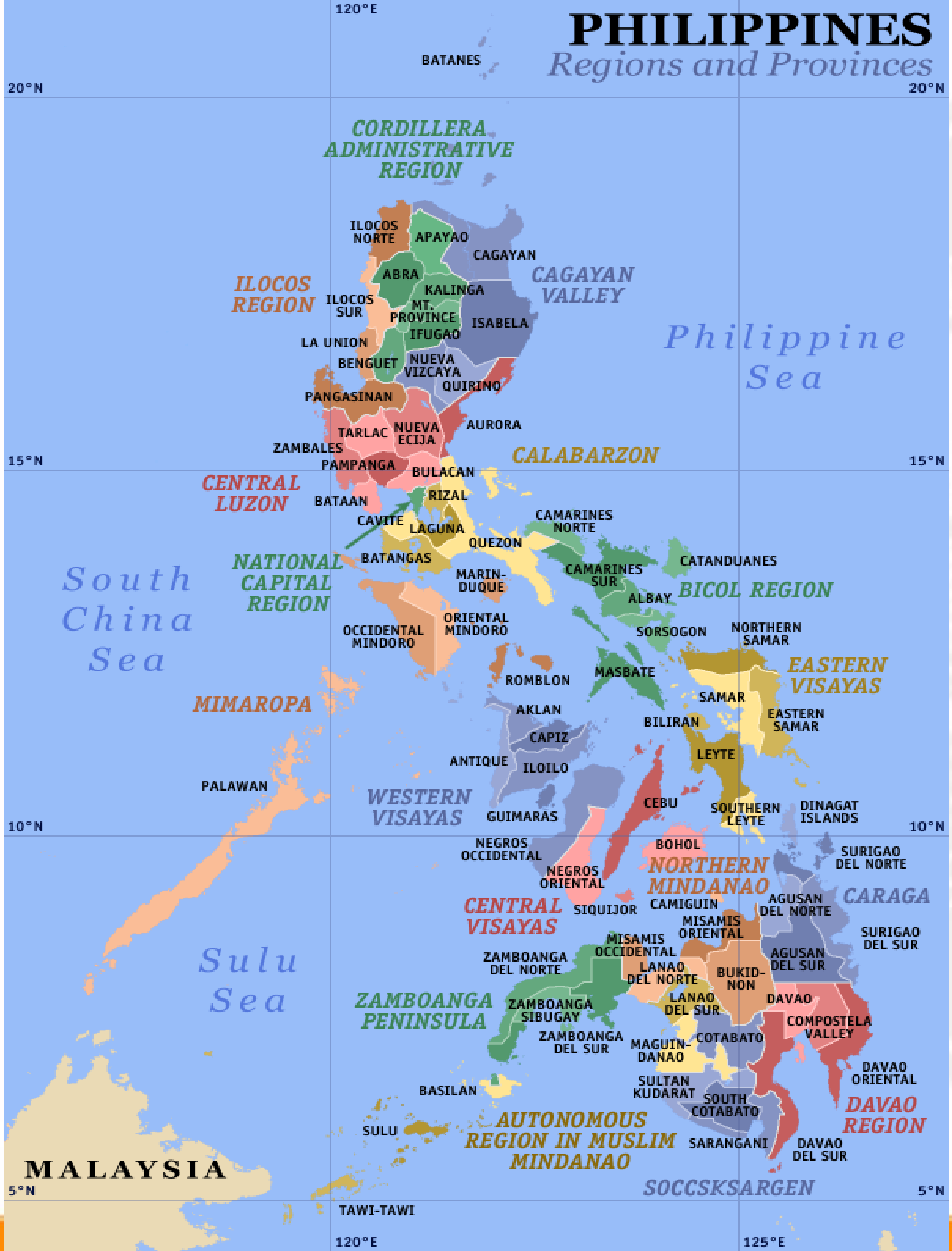
5. Map Graph or Cartogram – best way to represent geographical data.

- a map graph is drawn and divided into desired regions.

- always accompanied by a legend which tells the meaning of the lines, colours, or symbols used.

ex:

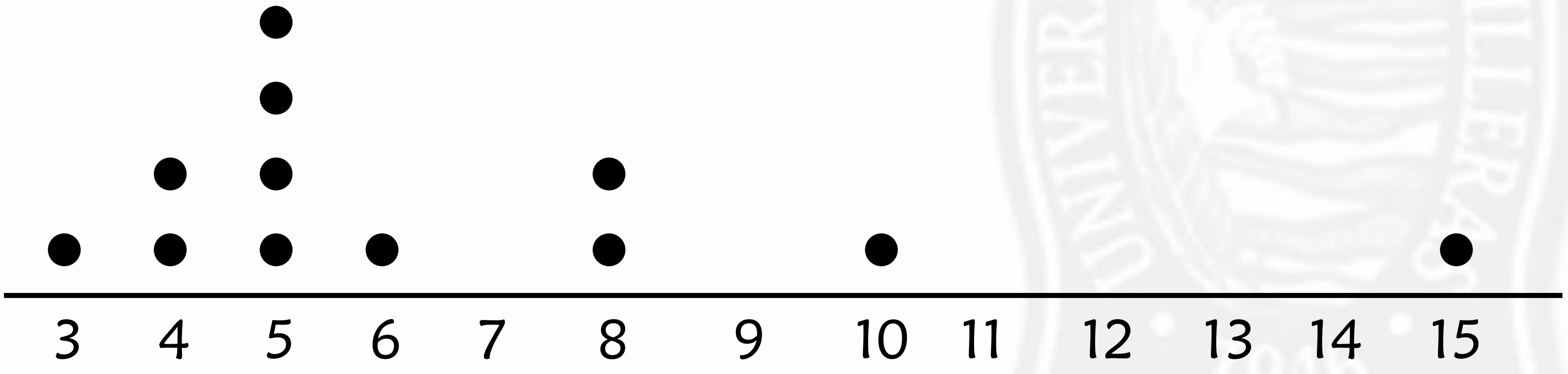
Ex:


PHILIPPINES
Regions and Provinces
120°E
20°N
20°N
BATANES
CORDILLERA ADMINISTRATIVE REGION
ILOCOS NORTE
APAYAO
CAGAYAN
CAGAYAN VALLEY
ABRA
KALINGA
ILOCOS REGION
ILOCOS SUR
MT. PROVINCE
ISABELA
IFUGAO
Philippine Sea
LA UNION
BENGUET
NUEVA VIZCAYA
QUIRINO
PANGASINAN
AURORA
TARLAC
NUEVA ECIJA
ZAMBALES
CALABARZON
15°N
15°N
PAMPANGA
BULACAN
CENTRAL LUZON
BATAAN
RIZAL
CAVITE
LAGUNA
CAMARINES NORTE
QUEZON
South China Sea
NATIONAL CAPITAL REGION
BATANGAS
CAMARINES SUR
CATANDUANES
MARIN-DUQUE
ALBAY
BICOL REGION
ORIENTAL MINDORO
OCCIDENTAL MINDORO
SORSOGON
NORTHERN SAMAR
MASBATE
EASTERN VISAYAS
ROMBLON
MIMAROPA
SAMAR
AKLAN
EASTERN SAMAR
BILIRAN
CAPIZ
LEYTE
ANTIQUE
ILOILO
PALAWAN
WESTERN VISAYAS
CEBU
SOUTHERN LEYTE
DINAGAT ISLANDS
GUIMARAS
10°N
10°N
NEGROS OCCIDENTAL
BOHOL
SURIGAO DEL NORTE
NEGROS ORIENTAL
NORTHERN MINDANAO
CARAGA
CENTRAL VISAYAS
SIQUIJOR
CAMIGUIN
AGUSAN DEL NORTE
MISAMIS ORIENTAL
SURIGAO DEL SUR
MISAMIS OCCIDENTAL
ZAMBOANGA DEL NORTE
LANAO DEL NORTE
BUKID-NON
AGUSAN DEL SUR
Sulu Sea
ZAMBOANGA PENINSULA
ZAMBOANGA SIBUGAY
LANAO DEL SUR
DAVAO
COMPOSTELA VALLEY
ZAMBOANGA DEL SUR
MAGUIN-DANAO
COTABATO
DAVAO ORIENTAL
BASILAN
SULTAN KUDARAT
SOUTH COTABATO
DAVAO REGION
AUTONOMOUS REGION IN MUSLIM MINDANAO
SULU
SARANGANI
DAVAO DEL SUR
MALAYSIA
5°N
SOCCSKSARGEN
5°N
TAWI-TAWI
120°E
125°E

6. Dot Plots – it is a simple way to display numerical data when the data set is reasonably small.

- it is use in a small numerical data sets.

ex: 3, 4, 4, 5, 5, 5, 5, 6, 8, 8, 10, 15



Outlier: 15

– unusual observation

7. Stem and Leaf Distribution

Steps:

1. Divide each measurement into two parts: the stem and the leaf.
2. List the stems in a column with vertical line to the right.
3. For each measurement, record the leaf portion in the same row as its corresponding stem.
4. Order the leaves from lowest to highest in each stem.
5. Provide a legend to your own stem and leaf coding.

Ex: The following are prices of 20 brands of candies (in peso). Make a stem and leaf distribution.

90	68	68	75	95
65	70	70	70	65
70	70	74	40	75
70	60	65	70	84

9	0 5
8	4
7	0 0 0 0 0 4 5 0 0 5
6	5 8 0 8 5 5
5	
4	0

Arrange:

9	0 5
8	4
7	0 0 0 0 0 0 0 4 5 5
6	0 5 5 5 8 8
5	
4	0

Legend:

stem: ten's digit
leaf: unit's digit

Interpretation: Most of the 20 brands of candies cost 70 pesos.

Quiz:

1. pie graph (household monthly budget)- enumerate atleast 5 items your family spend in a month

2. The following are lists of a science test scores. Construct its a) dot plots and b) stem and leaf plot to display the distribution. Interpret the result.

22	23	24	45	39	11
29	26	22	25	16	28
33	36	16	39	19	17
22	32	34	22	18	21
27	34	26	41	28	25

Frequency Distribution Table

Frequency Distribution Table (FDT) – statistical table showing the frequency or number of observation contained in each defined classes or categories.

Class Frequency – refers to the number of observations belonging to a class interval.

Class Interval – is a grouping or category defined by a lower limit and upper limit.

Class Size/ Class Width (i) – difference between two successive lower class limits.

Class Boundaries – are more precise expressions of the class limits by at least 0.5 of their values.

- used to make scale continuous.

Class Marks / Midpoints – single values that represent every class intervals.

$$= \frac{LL + UL}{2}$$

Steps in Constructing FDT

Steps in constructing frequency distribution

1. Determine the range – highest value minus the lowest value: $R = HV - LV$
2. Find the number of class intervals (ci), using the Sturge's formula. (ideal is 5 to 15).
3. Determine class size. (Round up)

$$\text{Class size} = \frac{range}{ci}$$

4. Determine the lowest class interval of the table.

   *Lowest value must be divisible by i.
5. Enumerate the classes / categories.
6. Tally frequencies

Sturge’s Formula – formula used to determine the number of classes

$$ci = 1 + 3.3 \log N$$

Where: ci = number of class intervals

N = total number of observations

log N = logarithm of N to the base 10

Ex: Class scores of students in Math CS 4. Construct fd.

Scores:

174 171 185 164 166 176 162 172
145 142 178 180 147 167 178 176
162 172 166 170 184 173 173 169
148 168 165 172 187 181 183 173
175 193 156 168 171 191 161 156
179 188 197 179 181 151 177 158
187 152

Construct FDT:

1. Determine: range = hv – lv

   = 197 – 142 = 55

2. Determine number of classes:

   Ci = 55 ÷ 5 = 11; if i = 5

   *Number of class interval ranges from 5 – 15

   *Class width (i) – depends on the researcher

3. Determine the lowest class interval of the table.

   *Lowest class limit must be divisible by the class width

4. Enumerate the classes/ categories
5. Tally frequencies.

FDT:

score	tally	f	rf	xm/cm	cb	<cf	>cf
195-199		1	2%	197	194.5-199.5	50	1
190-194		2	4%	192	189.5-194.5	49	3
185-189		4	8%	187	184.5-189.5	47	7
180-184		5	10%	182	179.5-184.5	43	12
175-179		8	16%	177	174.5-179.5	38	20
170-174		10	20%	172	169.5-174.5	30	30
165-169		6	12%	167	164.5-169.5	20	36
160-164		4	8%	162	159.5-164.5	14	40
155-159		4	8%	157	154.5-159.5	10	44
150-154		2	4%	152	149.5-154.5	6	46
145-149		3	6%	147	144.5-149.5	4	49
140-144		1	2%c	142	139.5-144.5	1	50
Class limits	total	50	100%				

*class boundaries – exact limits/true limits

- used to make scale continuous

Interpretation of Cumulative Frequency(cf)

Cumulative frequency – tabular summary of a set of data showing the total number of data items less than or equal to the upper limits of each class.

a. <cf – number of cases falling below a particular score.

ex: <168 – 175 is 31

*there are 31 students who got scores less than/below 175.5 (upper class boundary)

b. >cf – number of cases falling above a particular score.

ex: >168 – 175 is 33

* there are 33 students who got scores above 167.5 (lower class boundary)

Graphical Presentation for a Frequency Distribution

Histogram

- bar graph

- consists of a set of rectangles having bases on a horizontal axis which centers on the classmarks.

- base – class size

- height – frequency

*x-axis – class boundaries

*y-axis – frequencies

Frequency Polygon

- line graph

- x-axis – midpoints/class marks

- y-axis - frequency

Graph of Cumulative Frequency Distribution

Cumulative frequency distribution is a tabular arrangements of data by class intervals whose frequencies are cumulated.

*Less than cumulative frequency (<ogive)

P(upper limit, <cf)

*Greater than cumulative frequency (> ogive)

P(lower limit, >cf)

Quiz: Answer the ff problems:

1. A sample of fifty customers at a new-open supermarket has been selected at random. The following data show the customer's ages. Construct a frequency distribution table, histogram and frequency polygon.

12	23	19	21	35	37	28	29
19	29	10	21	27	23	21	33
20	27	18	53	23	16	23	11
13	23	18	19	50	47	41	32
21	21	39	25	17	34	46	21
59	42	27	28	60	48	26	47
14	52						

2. The following are scores of Math CS 1 students. Construct the frequency distribution table, histogram, and frequency polygon.

59	31	37	33	35	47	30	19
22	47	40	44	37	26	47	27
25	44	43	42	30	29	43	49
48	30	36	31	52	56	20	29
52	45	54	51	62	50	55	34
36	33	39	28	42	45	38	51
21	37	21	47	58	60	21	34
23	52	61	31				

Measures of Central Tendency

By: BRETZ HARLLYNNE M. MOLTIO

What is a Central Tendency?

- The **central tendency** of your data set tells you where most of your values lie.
- Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

What is a Central Tendency?

- **Mode**: the most frequent value.
- **Median**: the middle number in an ordered dataset.
- **Mean**: the sum of all values divided by the total number of values.

What is a Central Tendency?

- The mode, mean, and median are three most commonly used measures of central tendency. However, only the mode can be used with nominal data.

What is a Central Tendency?

- To get the **median** of a data set, you have to be able to **order values from low to high**.
- For the **mean**, you need to be able to perform arithmetic operations like addition and division on the values in the data set.

What is a Central Tendency?

- While nominal data can be grouped by category, it cannot be ordered nor summed up.
- Therefore, the central tendency of nominal data can only be expressed by the **mode** – the **most frequently recurring value**.

The Average/Mean Vs The Median

What is Mean?

What is Mean?

- The most common and effective **numeric measure of the "center"** of a set of data is the **(arithmetic) mean.**
- The **mean** of a dataset is the sum of all values divided by the total number of values.
- It's the most commonly used measure of central tendency and is often referred to as the "**average**."

Measuring the Central Tendency (Mean)

- Let x_1,x_2, : : : ,x_N be a set of N values or observations, such as for some numeric attribute X, like salary. For calculating the mean of a sample, use this formula:

$$\boxed{\bar{x}} = \frac{\boxed{\sum_{i=1}^{N} x_i}}{\boxed{N}} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Measuring the Central Tendency (Mean)

x̄: sample mean

$\sum x$:sum of all values in the sample dataset

N: number of values in the sample dataset

Mean: Example

- Suppose we have the following values for salary (in thousands of dollars), shown in increasing order:

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$\bar{x} = \frac{696}{12}$$

$$\boxed{\bar{x} = 58}$$

Thus, the mean salary is $58,000.

What is Median?

What is Median?

- For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values.
- It is the value that separates the higher half of a data set from the lower half.

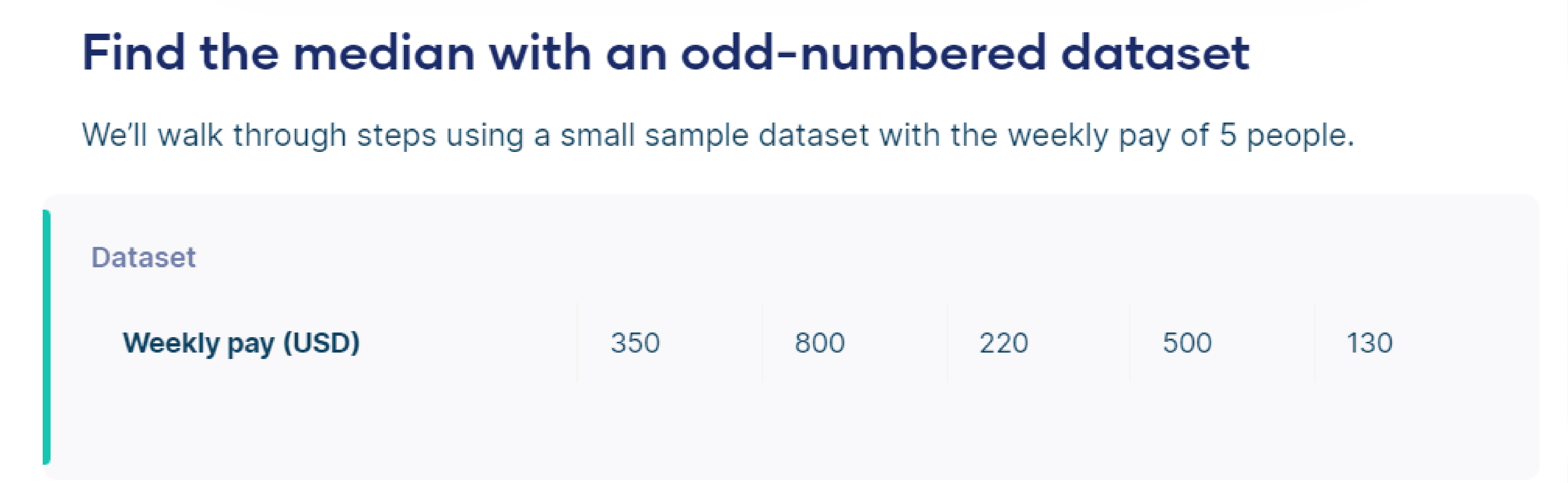
What is Median?

- In probability and statistics, the median generally applies to numeric data; however, we may extend the concept to ordinal data.
- Suppose that a given data set of N values for an attribute X is sorted in increasing order.

What is Median?

- **If N is odd**, then the **median is the middle value** of the ordered set.
- **If N is even**, then the **median is not unique**; it is the **two middlemost values** and any value in between.

Median with an Odd-Numbered Dataset



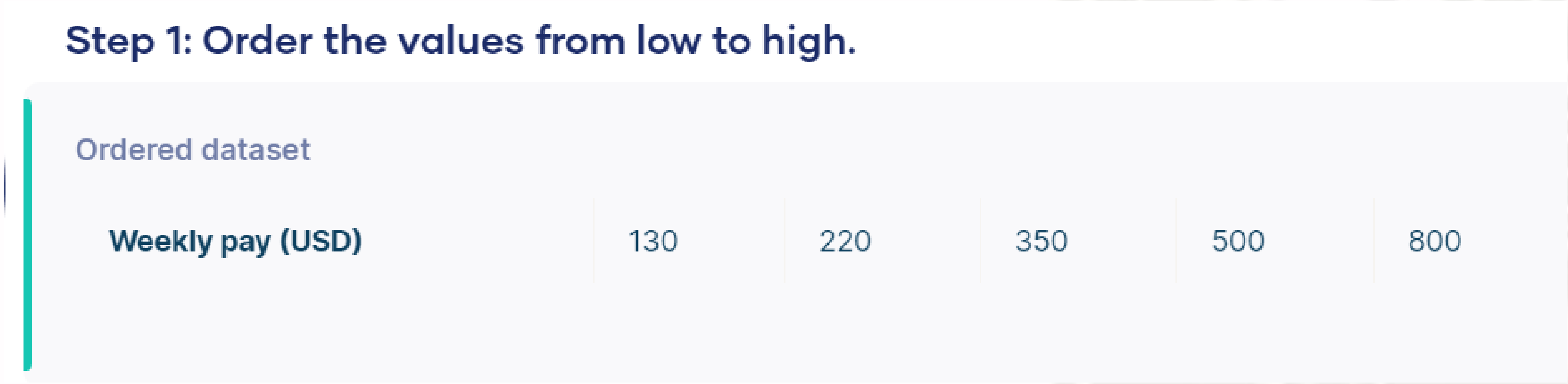
Find the median with an odd-numbered dataset

We'll walk through steps using a small sample dataset with the weekly pay of 5 people.

Dataset

Weekly pay (USD)	350	800	220	500	130

Median with an Odd-Numbered Dataset

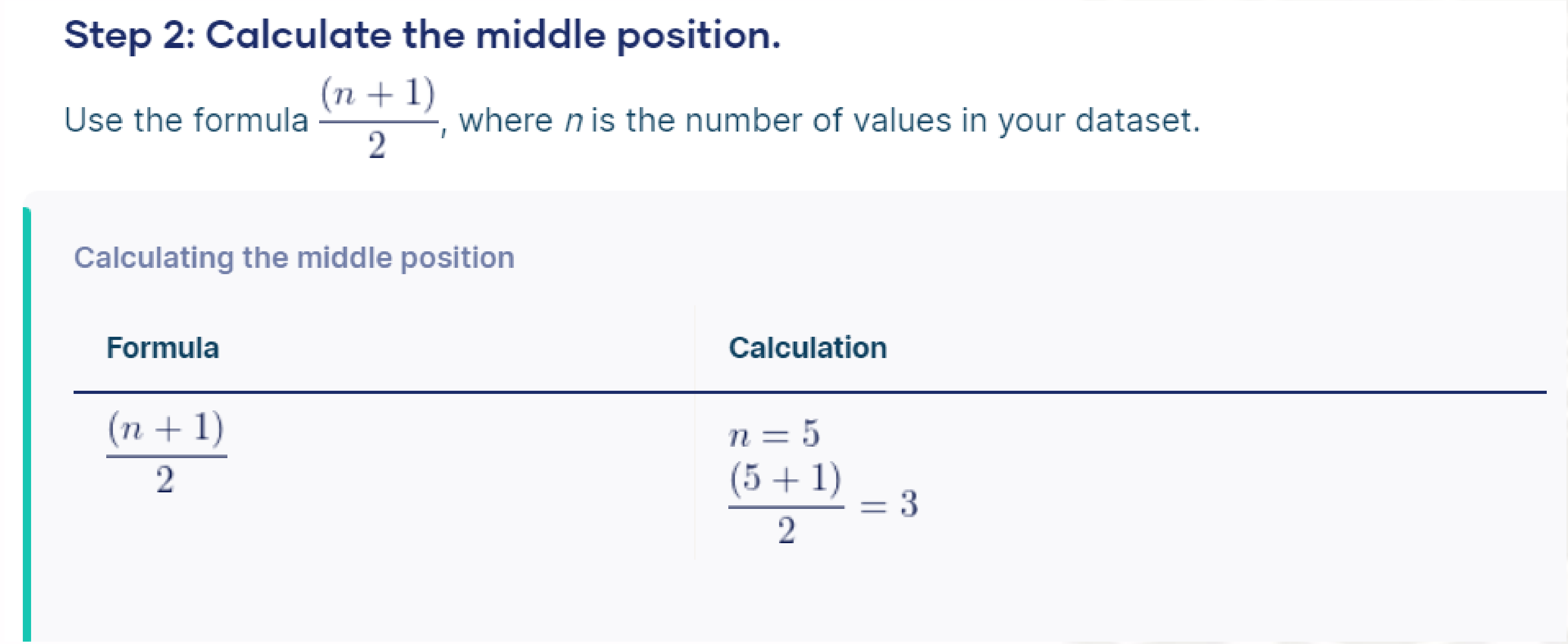


Step 1: Order the values from low to high.

Ordered dataset

Weekly pay (USD)	130	220	350	500	800

Median with an Odd-Numbered Dataset



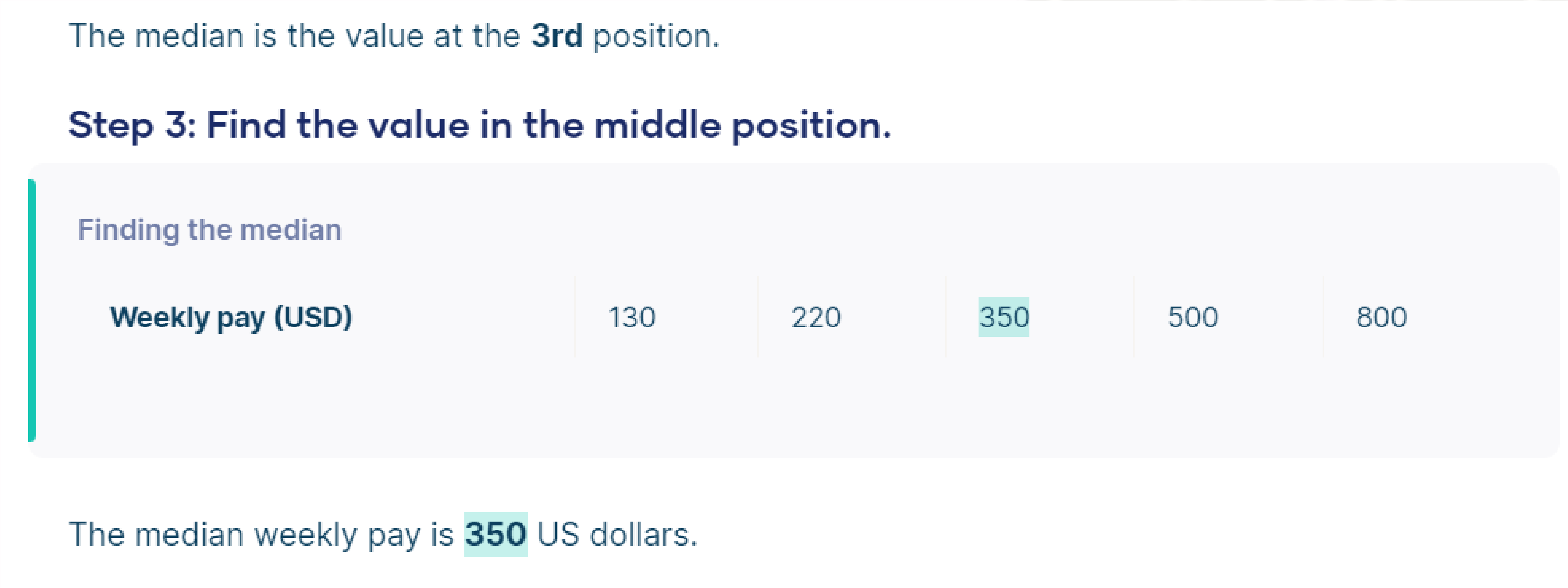
Step 2: Calculate the middle position.

Use the formula $\frac{(n+1)}{2}$, where n is the number of values in your dataset.

Calculating the middle position

Formula	Calculation
$\frac{(n+1)}{2}$	$n = 5$ $\frac{(5+1)}{2} = 3$

Median with an Odd-Numbered Dataset



The median is the value at the **3rd** position.

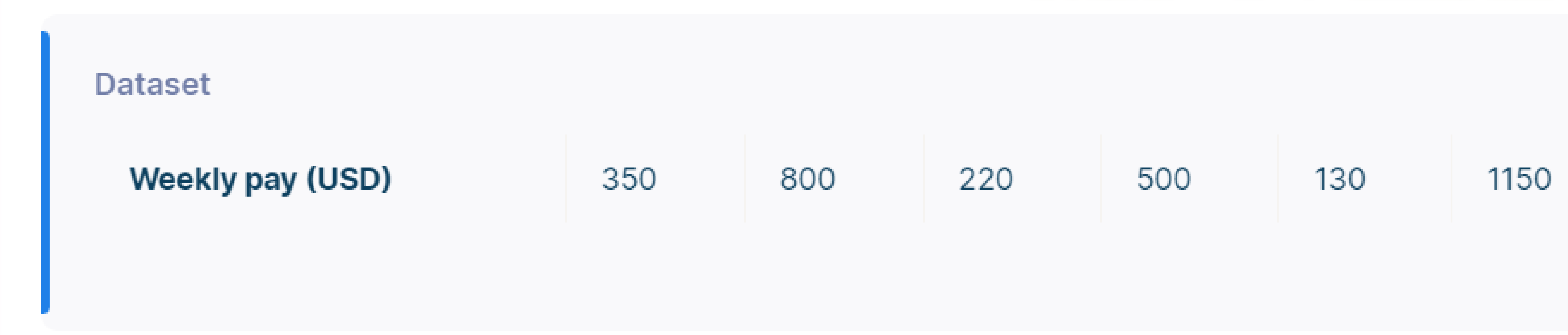
Step 3: Find the value in the middle position.

Finding the median

Weekly pay (USD)	130	220	350	500	800

The median weekly pay is **350** US dollars.

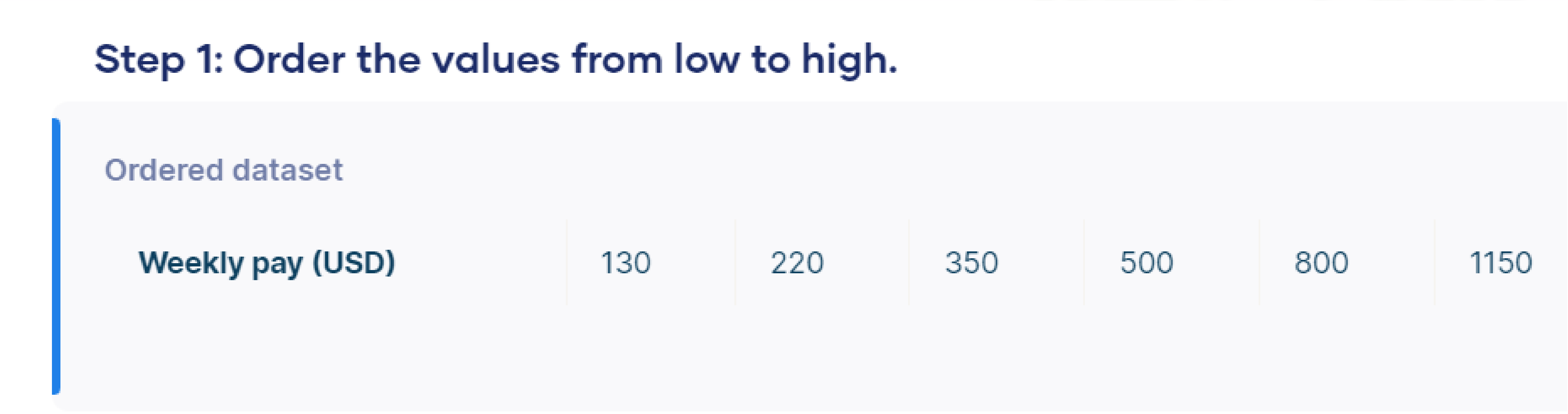
Median with an Even-Numbered Dataset



Dataset

Weekly pay (USD)	350	800	220	500	130	1150

Median with an Even-Numbered Dataset

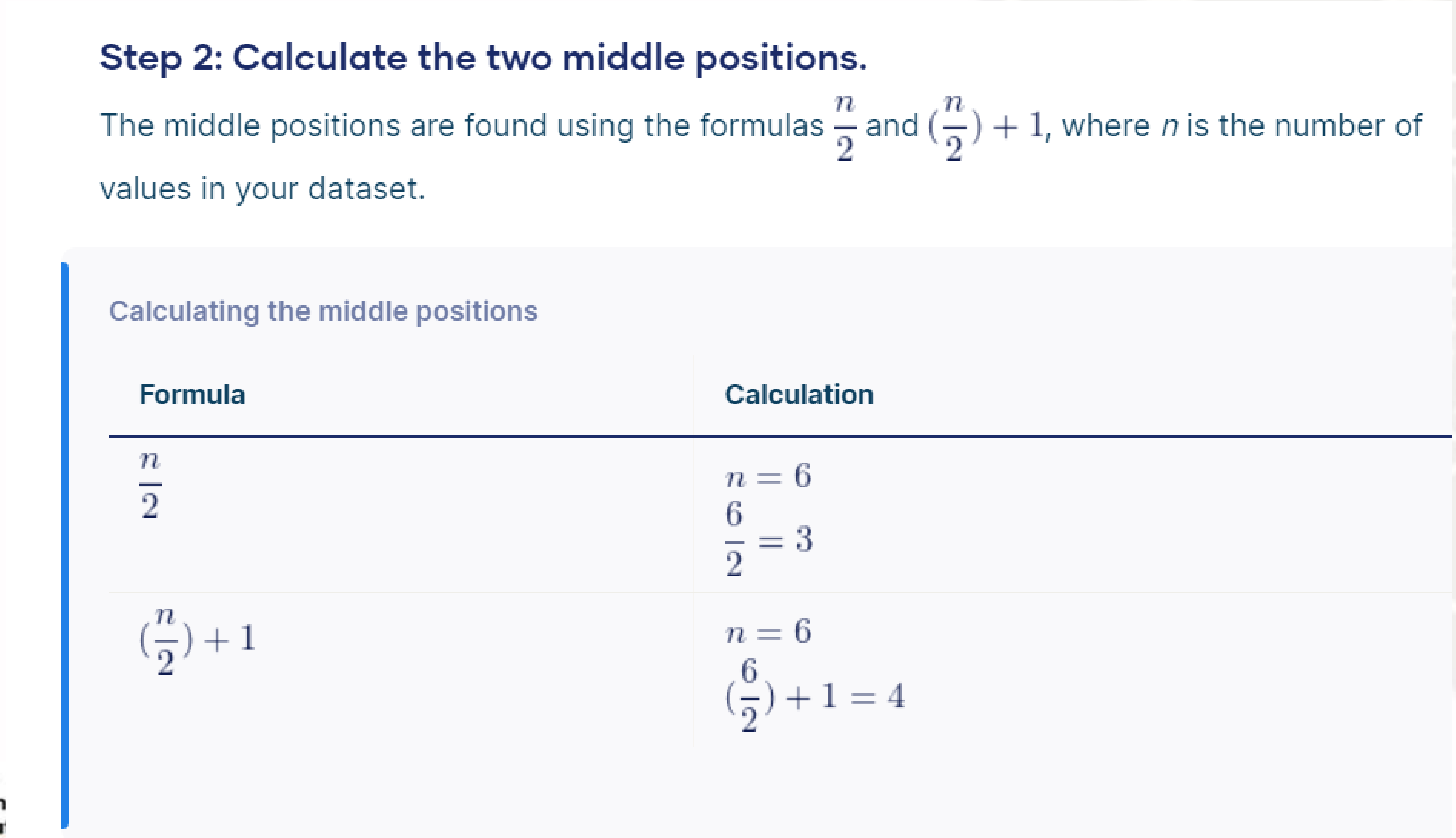


Step 1: Order the values from low to high.

Ordered dataset

Weekly pay (USD)	130	220	350	500	800	1150

Median with an Even-Numbered Dataset



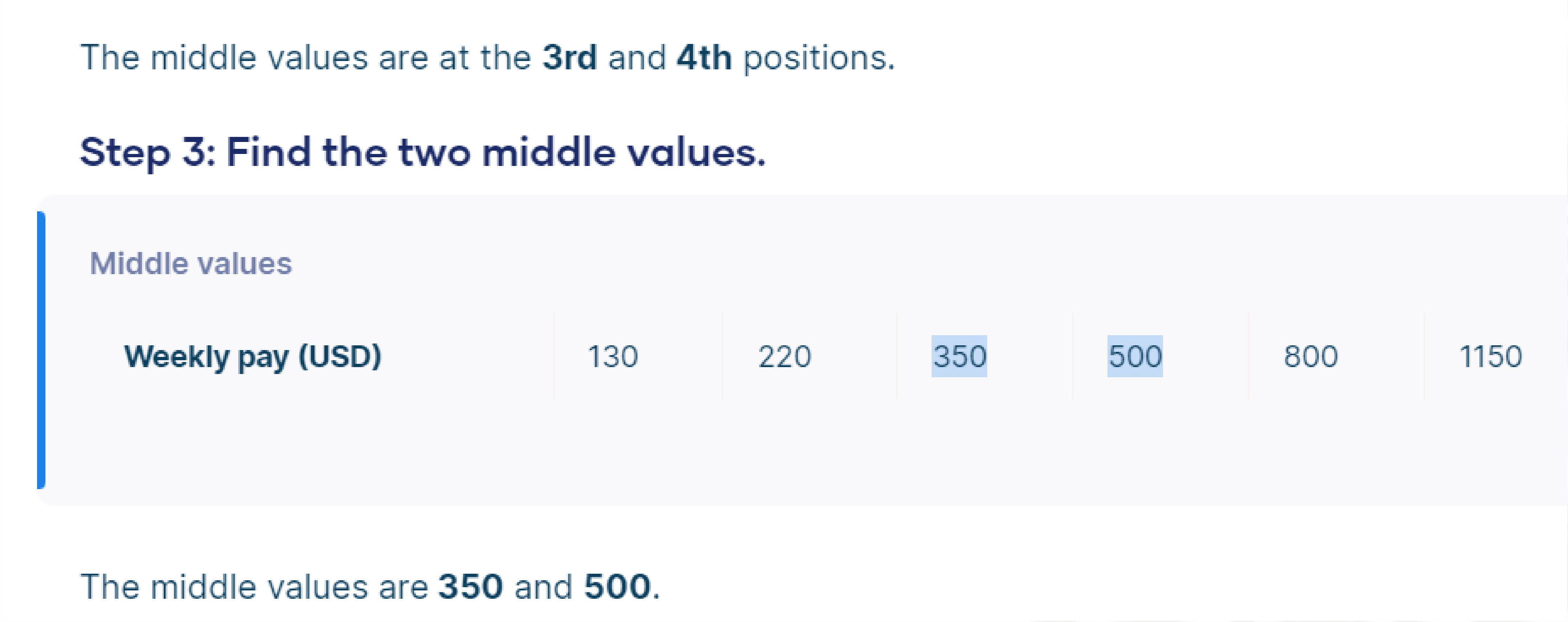
Step 2: Calculate the two middle positions.

The middle positions are found using the formulas $\frac{n}{2}$ and $(\frac{n}{2}) + 1$, where *n* is the number of values in your dataset.

Calculating the middle positions

Formula	Calculation
$\frac{n}{2}$	$n = 6$ $\frac{6}{2} = 3$
$(\frac{n}{2}) + 1$	$n = 6$ $(\frac{6}{2}) + 1 = 4$

Median with an Even-Numbered Dataset



The middle values are at the **3rd** and **4th** positions.

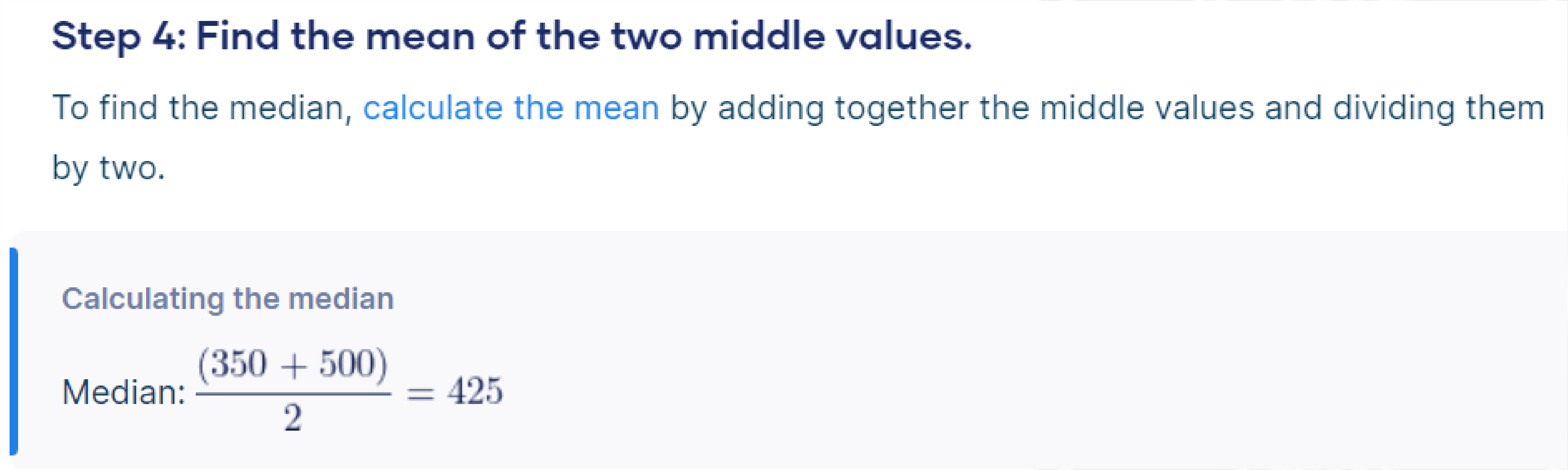
Step 3: Find the two middle values.

Middle values

Weekly pay (USD)	130	220	350	500	800	1150

The middle values are **350** and **500**.

Median with an Even-Numbered Dataset

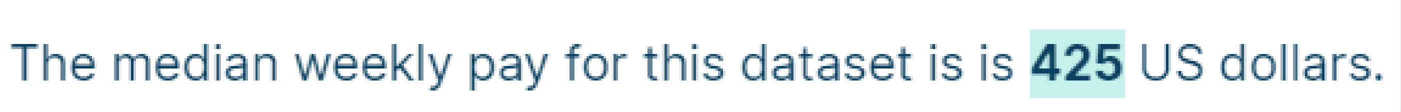


Step 4: Find the mean of the two middle values.

To find the median, calculate the mean by adding together the middle values and dividing them by two.

Calculating the median

Median: $\frac{(350 + 500)}{2} = 425$



The median weekly pay for this dataset is is **425** US dollars.

What is Mode?

What is Mode?

- The **mode** is another measure of central tendency.
- The **mode** for a set of data is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative and quantitative attributes.

What is Mode?

- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
- Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal.

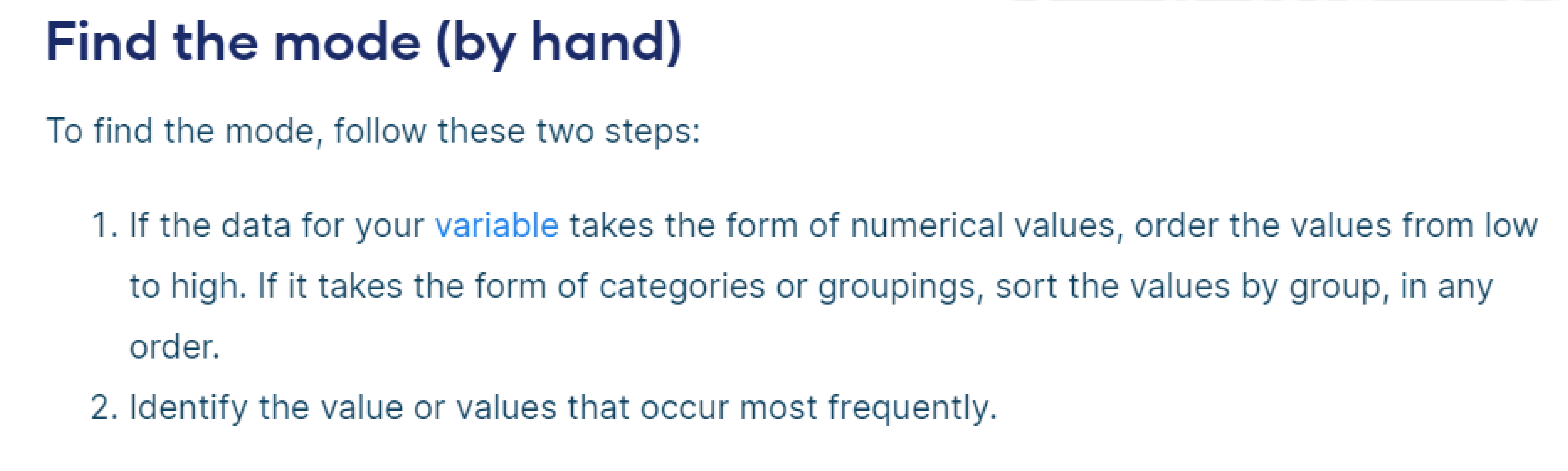
What is Mode?

- A data set can often have no mode, one mode or more than one mode – it all depends on how many different values repeat most frequently.

What is Mode?

- Your data can be:
- without any mode
- **unimodal**, with one mode,
- **bimodal**, with two modes,
- **trimodal**, with three modes, or
- **multimodal**, with four or more modes.

Finding the Mode

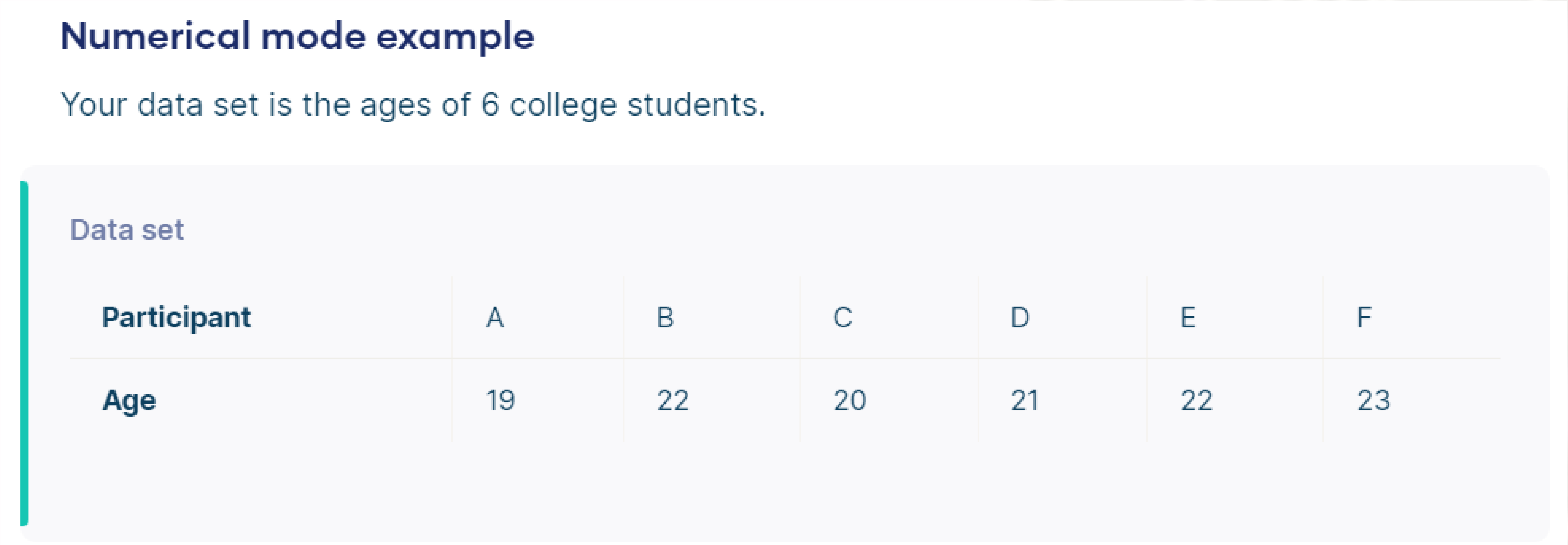


Find the mode (by hand)

To find the mode, follow these two steps:

1. If the data for your variable takes the form of numerical values, order the values from low to high. If it takes the form of categories or groupings, sort the values by group, in any order.
2. Identify the value or values that occur most frequently.

Finding the Mode



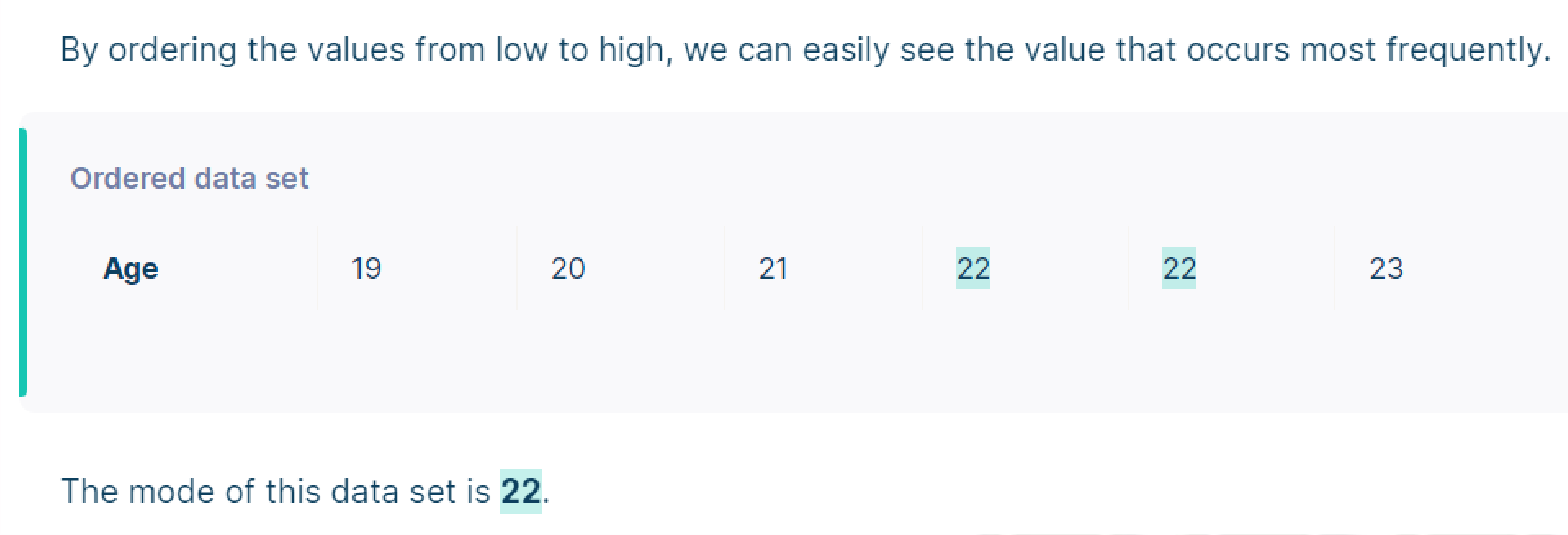
Numerical mode example

Your data set is the ages of 6 college students.

Data set

Participant	A	B	C	D	E	F
Age	19	22	20	21	22	23

Finding the Mode



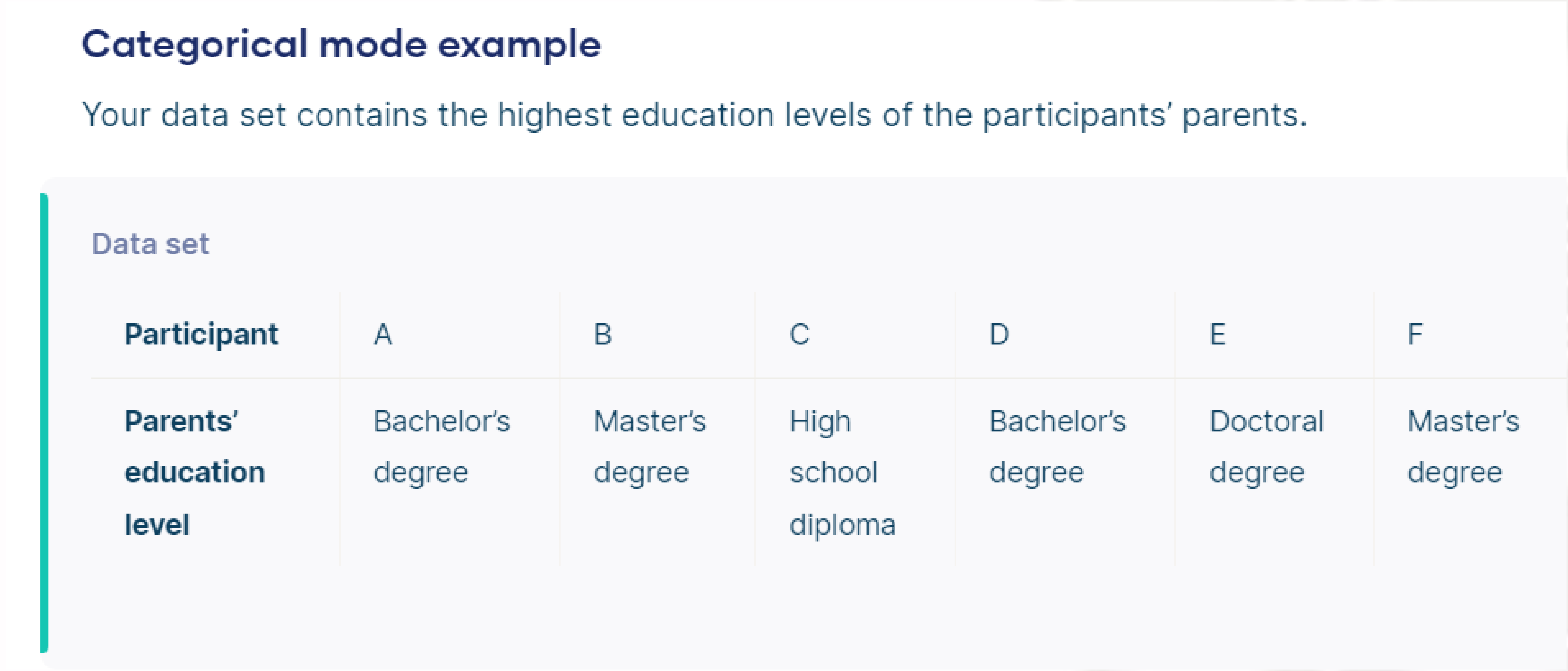
By ordering the values from low to high, we can easily see the value that occurs most frequently.

Ordered data set

Age	19	20	21	22	22	23

The mode of this data set is **22**.

Finding the Mode



Categorical mode example

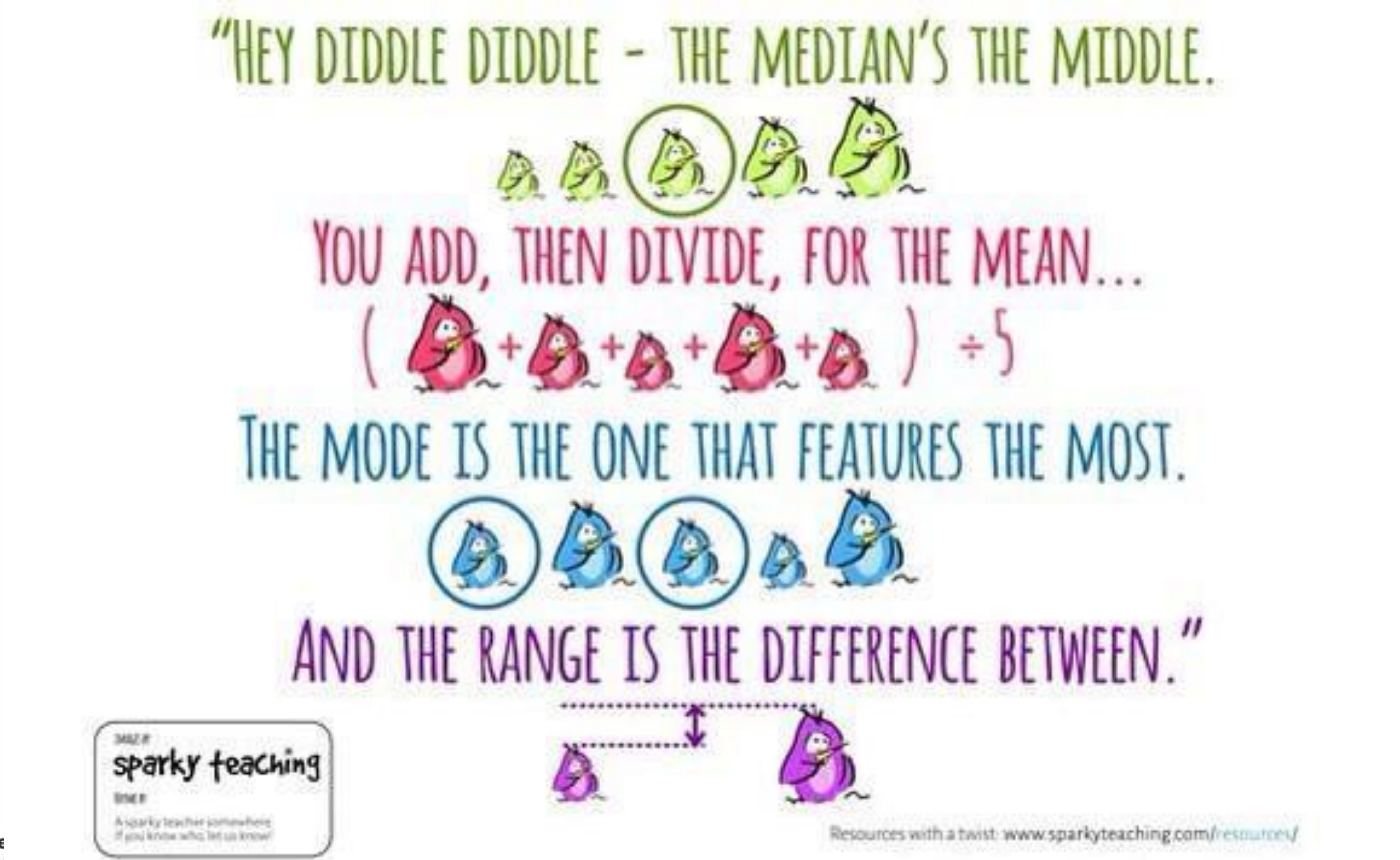
Your data set contains the highest education levels of the participants' parents.

Data set

Participant	A	B	C	D	E	F
Parents' education level	Bachelor's degree	Master's degree	High school diploma	Bachelor's degree	Doctoral degree	Master's degree

Finding the Mode

- From the table, you can see that there are two modes. This means you have a **bimodal** data set.
- The modes are **Bachelor's degree** and **Master's degree**.



"Hey diddle diddle - the median's the middle.

You add, then divide, for the mean...

The mode is the one that features the most.

And the range is the difference between."

Resources with a twist: www.sparkyteaching.com/resources/

Rhyme to Remember the Measures of Central Tendency and Range

- Hey Diddle Diddle,
- **Median's** the **middle**,
- You **add** and **divide** for the **mean**,
- The **mode** is the one that you see the **most**,
- and **range** is the **difference between**.

Activity

- Suppose that the data for analysis includes the attribute age.

32, **6**, **21**, **10**, **8**, **11**, **12**, **36**, **17**, **16**, **15**, **18**, **40**, **24**, **21**, **23**, **24**, **24**, **29**, **16**, **32**, **31**, **10**, **30**, **35**, **32**, **18**, **39**, **12**, **20**

1. What is the mean of the data?
2. What is the median?
3. What is the mode of the data? Comment on the data's modality (i.e., bimodal,trimodal, etc.).
4. What is the range?

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

1. What is the mean of the data?
2. What is the median?
3. What is the mode of the data? Comment on the data’s modality (i.e., bimodal,trimodal, etc.).
4. What is the range?

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

1. What is the mean of the data?

$$\bar{x} = \frac{6+8+10+10+11+12+12+15+16+16+17+18+18+20+21+21+23+24+24+29+30+31+32+32+32+35+36+39+40}{30}$$

$$\bar{x} = \frac{662}{30}$$

$$\boxed{\bar{x} = 22.07}$$

Thus, the mean is 22.07.

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

What is the median?

21+21

=---------

2

42

=----

2

=21

Thus, the median is 21.

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

What is the mode of the data? Comment on the data's modality (i.e., bimodal,trimodal, etc.).

24, 32

bimodal

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

What is the range?

Range = 40-6

Range = 34

Thus, the range is 34.

Measures of Central Tendency

Measures of Central Tendency of Grouped Data

MEAN

❑It is best to compute the measure of central tendency for grouped data using frequency distribution when the range is too large.

Methods:

❑Long/Direct Method or Midpoint Method

$$\bar{x} = \frac{\sum fxm}{\sum f}$$

Where: f = frequency
xm = midpoint
Σf = summation of f

Steps:

1. Get the midpoint (xm).

2. Multiply the frequency (f) with every midpoint (xm) in every interval.

3. Get the summation of the products of the frequency (f) and the midpoint (xm).

4. Get the summation of the frequency(f).

5. Substitute the values in the formula and compute.

Ex: Scores of Math CS 4 students.

Class interval	f	xm	fxm
195 – 199	1	197	197
190 – 194	2	192	384
185 – 189	4	187	748
180 – 184	5	182	910
175 – 179	8	177	1416
170 – 174	10	172	1720
165 – 169	6	167	1002
160 – 164	4	162	648
155 – 159	4	157	628
150 – 154	2	152	304
145 – 149	3	147	441
140 – 144	1	142	142
	Σf = 50		Σfxm=8,540

Solution:

$$\bar{x} = \frac{\sum fxm}{\sum f}$$

$$= \frac{8,540}{50}$$

$$\boxed{= 170.80}$$

The average score of Math CS 4 students is 170.80.

Exercises:

Find the mean of the distribution using a) long method; b) deviation method

Class Interval	Frequency
172 – 180	2
163 – 171	4
154 – 162	5
145 – 153	12
136 – 144	9
127 – 135	5
118 – 126	3

MEDIAN

Properties	When to use
1. An ordinal statistic	1. An ordinal interpretation is needed.
2. A rank or position average	2. The middle score is desired.
3. Value is determined by scores near the middle value of the distribution.	3. We want to avoid influence of extreme values.
4. Not affected by extreme values.	
5. Can be subjected to only a few mathematical computations.	
6. Less widely used than the mean	
7. Represents typical score.	

Measuring the Dispersion of Data

CC19 – DATA MINING

Topics

- Range
- Quartiles
- Variance
- Standard Deviation
- Interquartile Range

Variability: Overview

- Variability describes how far apart data points lie from each other and from the center of a distribution. Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.

Variability: Overview

- Variability is also referred to as **spread, scatter or dispersion**. It is most commonly measured with the following:
- **Range**: the difference between the highest and lowest values
- **Interquartile range**: the range of the middle half of a distribution
- **Standard deviation**: average distance from the mean
- **Variance**: average of squared distances from the mean

Why does variability matter?

- While the **central tendency**, or average, tells you where most of your points lie, variability summarizes how far apart they are. This is important because the amount of variability determines how well you can **generalize** results from the sample to your population.

Why does variability matter?

- Low variability is ideal because it means that you can better predict information about the population based on sample data. High variability means that the values are less consistent, so it’s harder to make predictions.

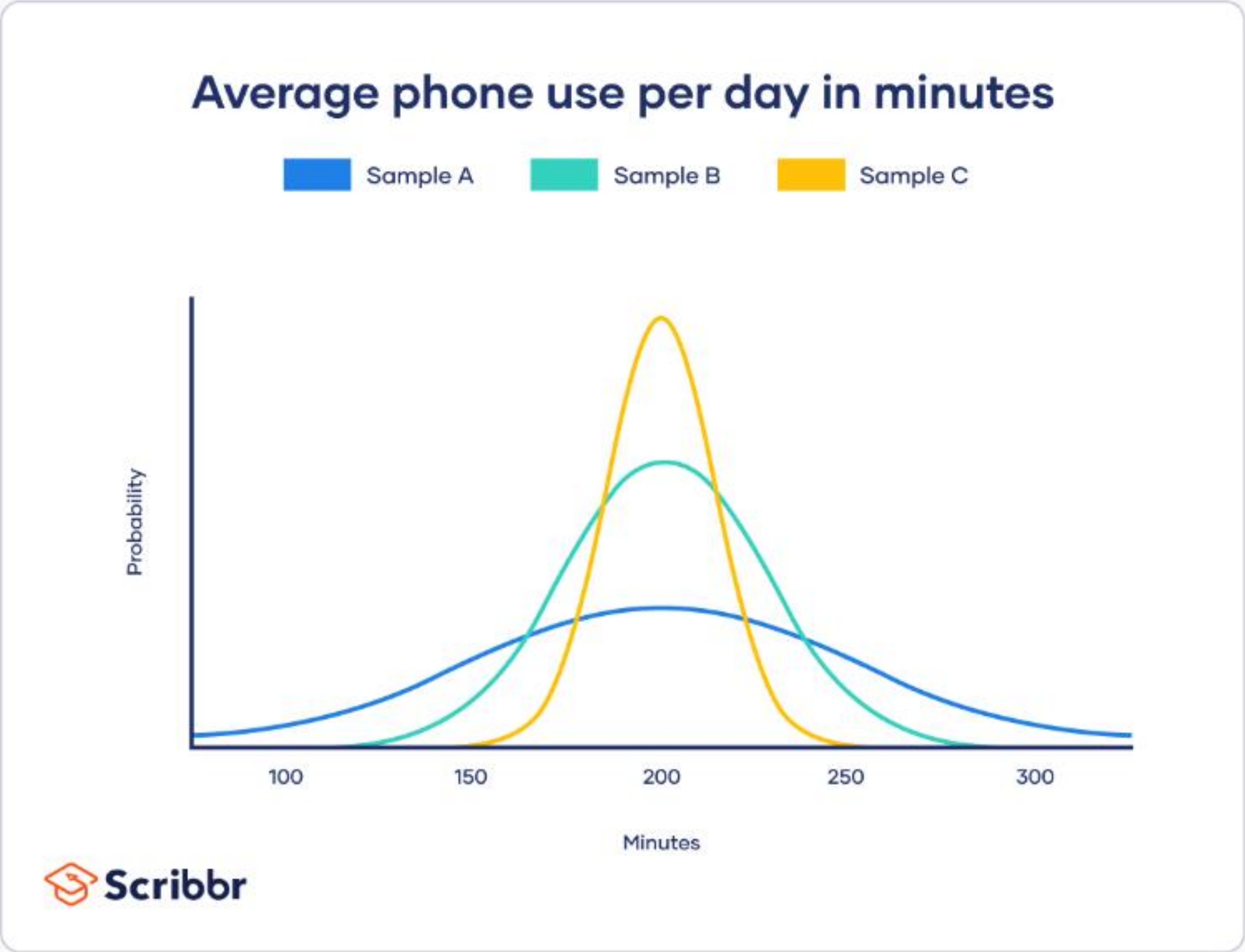
Why does variability matter?

- Data sets can have the same central tendency but different levels of variability or vice versa. If you know only the central tendency or the variability, you can't say anything about the other aspect. Both of them together give you a complete picture of your data.

Example: Variability in normal distributions

- You are investigating the amounts of time spent on phones daily by different groups of people.
- Using simple random samples, you collect data from 3 groups:

- Sample A: high school students,
- Sample B: college students,
- Sample C: adult full-time employees.

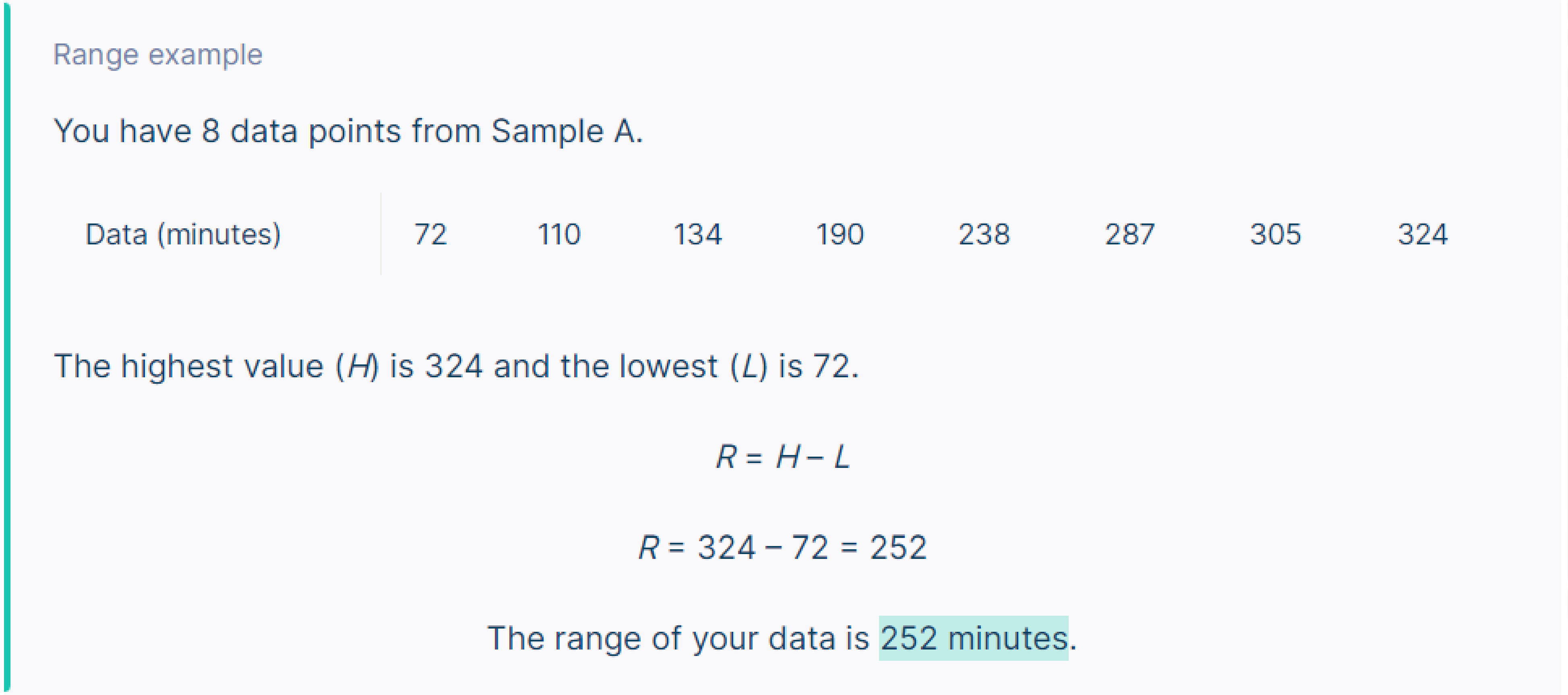

Average phone use per day in minutes
Sample A
Sample B
Sample C
Probability
100
150
200
250
300
Minutes
Scribbr

Range

- The **range** of the set is the difference between the largest (max()) and smallest (min()) values.
- The range tells you the spread of your data from the lowest to the highest value in the distribution. It’s the easiest measure of variability to calculate.

Range

- To find the range, simply subtract the lowest value from the highest value in the data set.



Range example

You have 8 data points from Sample A.

Data (minutes)	72	110	134	190	238	287	305	324

The highest value (*H*) is 324 and the lowest (*L*) is 72.

$$R = H - L$$

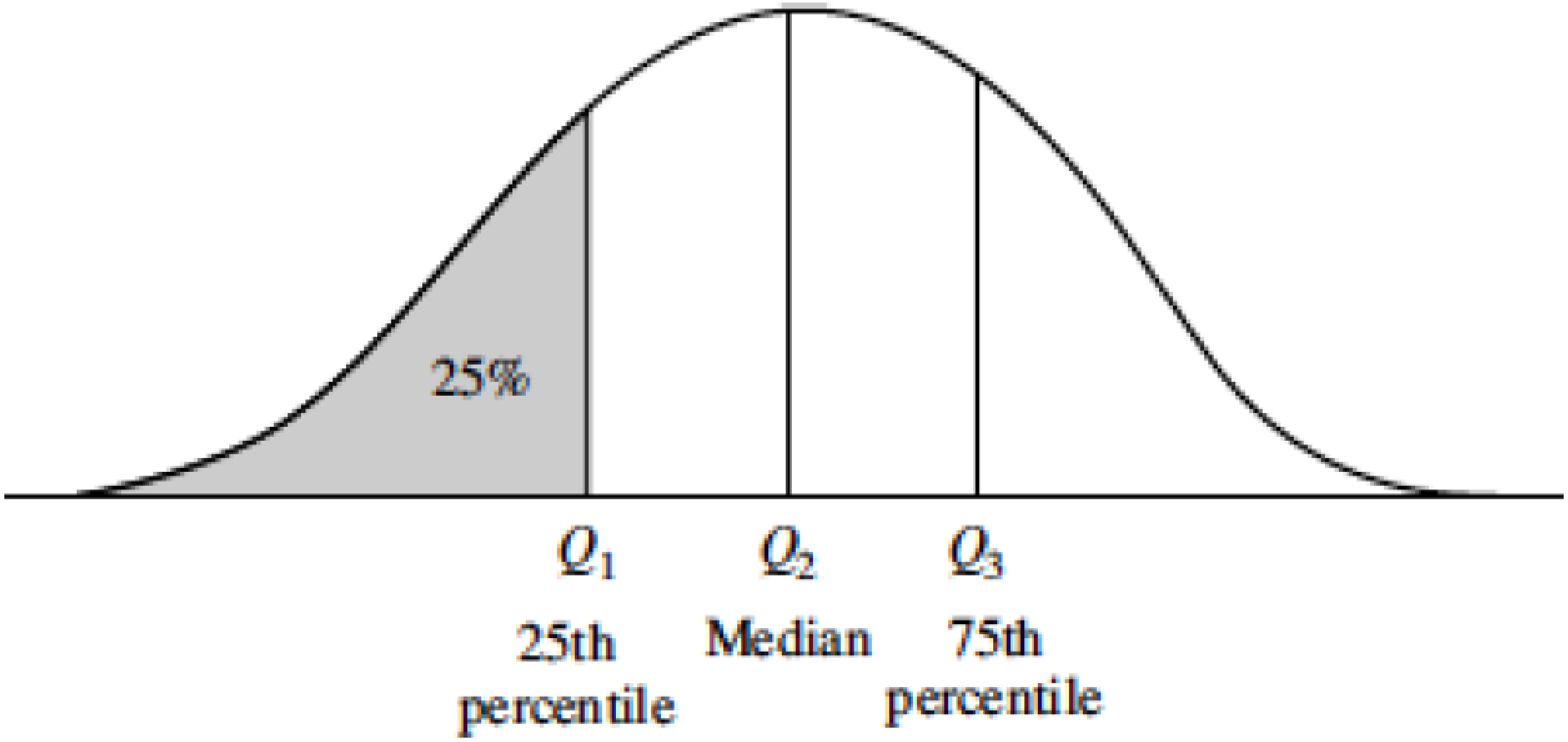
$$R = 324 - 72 = 252$$

The range of your data is 252 minutes.

Quartiles

- Q1 (25th percentile), Q3 (75th percentile)
- The quartiles give an indication of a distribution's center, spread, and shape. The first quartile, denoted by Q1, is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by Q3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

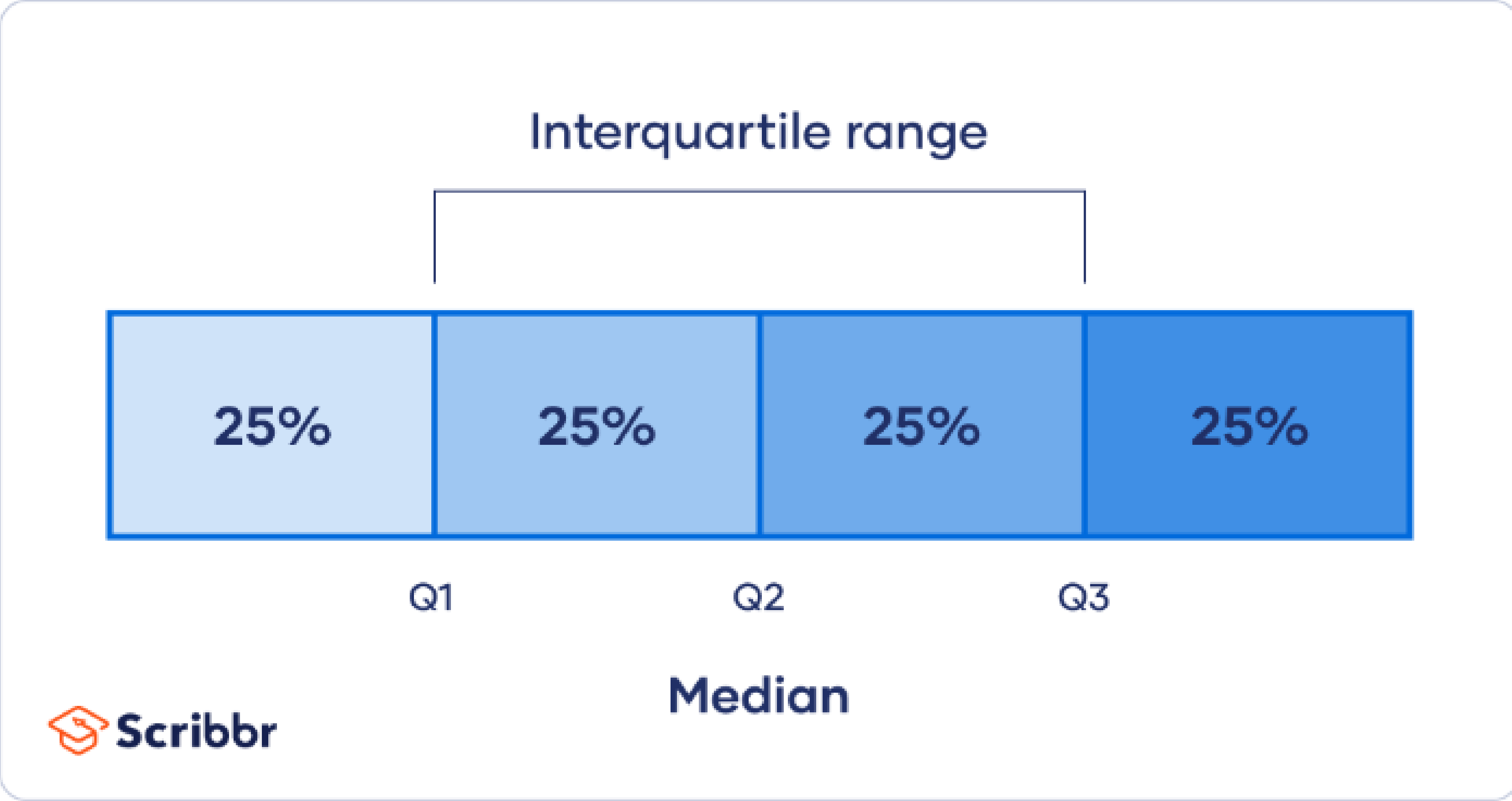
Quartiles



Interquartile Range

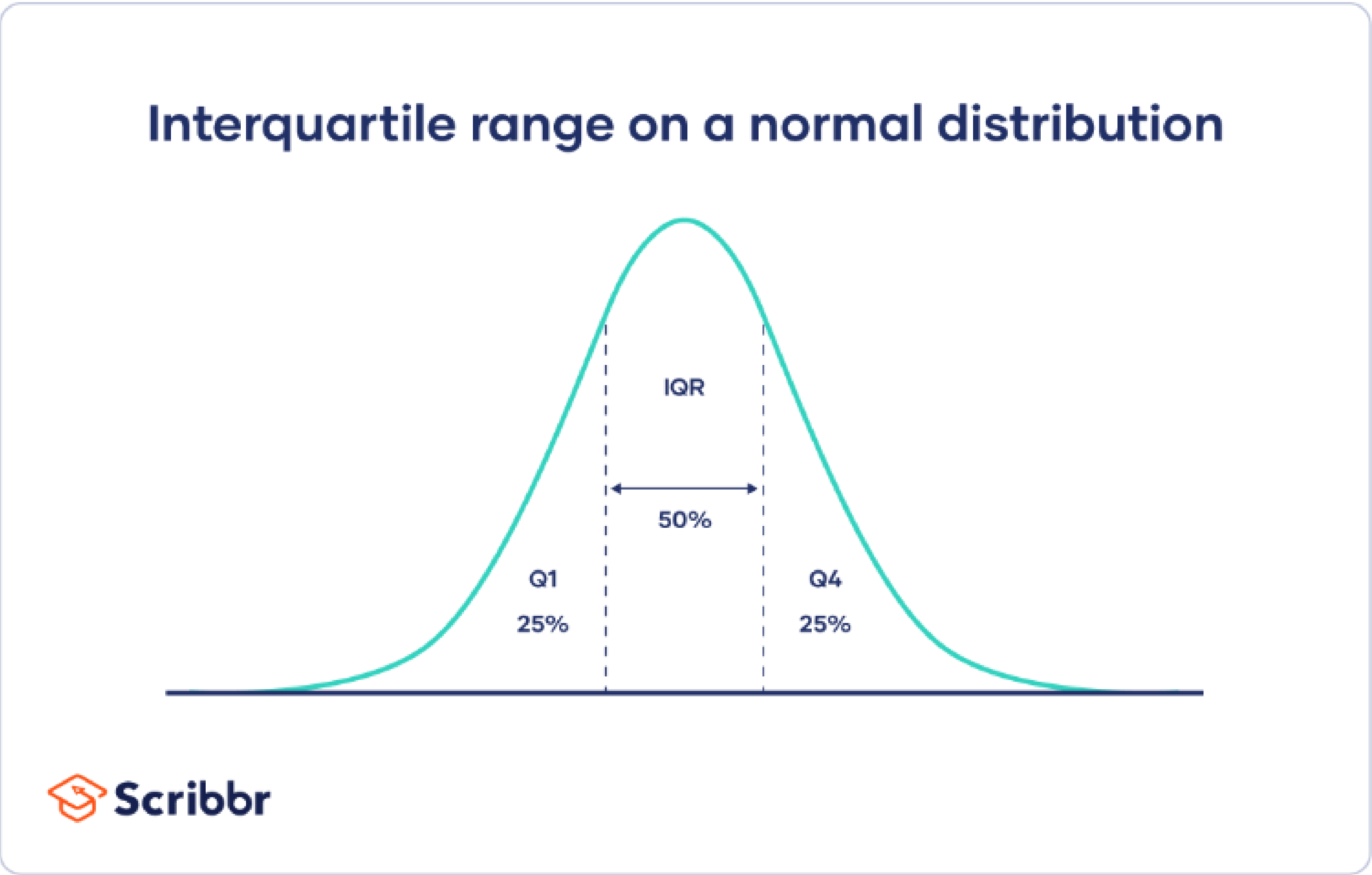
- The interquartile range gives you the spread of the middle of your distribution.
- For any distribution that’s ordered from low to high, the interquartile range contains half of the values. While the first quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.

Interquartile Range



Interquartile Range

- The quartiles are the three values that split the sorted data set into four equal parts. Using the following data: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110, already sorted in increasing order. Thus, the quartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list.



Interquartile range on a normal distribution

Scribbr

Interquartile Range

- To find the interquartile range of your 8 data points, you first find the values at Q1 and Q3.
- Multiply the number of values in the data set (8) by 0.25 for the 25th percentile (Q1) and by 0.75 for the 75th percentile (Q3).

- Q1 position: 0.25 x 8 = 2

- Q3 position: 0.75 x 8 = 6

Interquartile Range

- Q1 is the value in the 2nd position, which is 110. Q3 is the value in the 6th position, which is 287.

- IQR = Q3 – Q1

- IQR = 287 – 110 = 177

- The interquartile range of your data is 177 minutes.

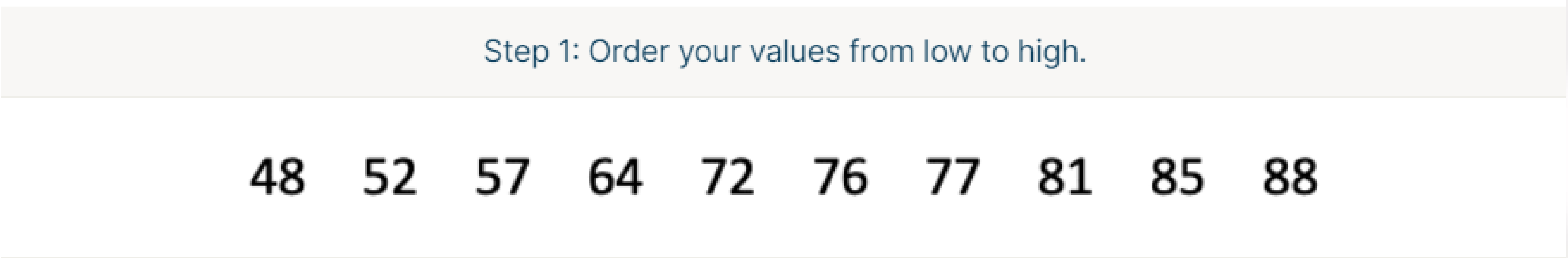
Interquartile Range

- Just like the range, the interquartile range uses only 2 values in its calculation. But the IQR is less affected by outliers: the 2 values come from the middle half of the data set, so they are unlikely to be extreme scores.

- The IQR gives a consistent measure of variability for skewed as well as normal distributions.

Even-numbered Data set

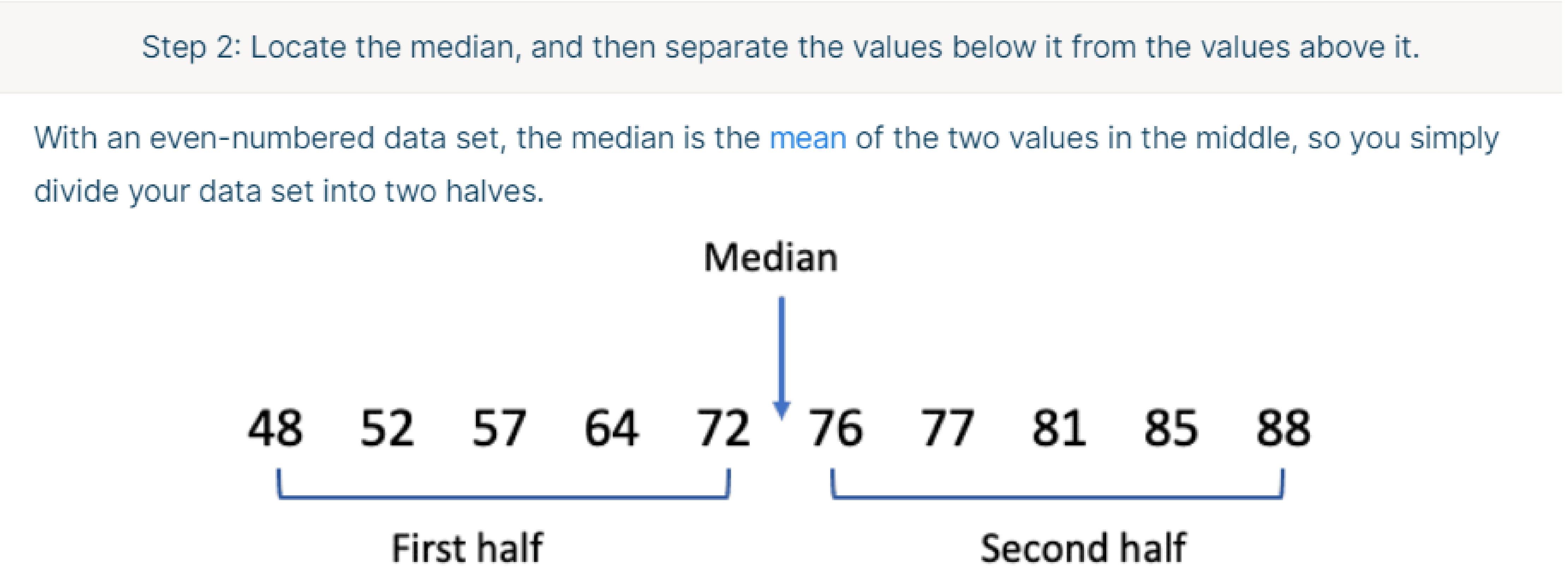
- We’ll walk through four steps using a sample data set with 10 values.



Step 1: Order your values from low to high.

48 52 57 64 72 76 77 81 85 88

Even-numbered Data set



Step 2: Locate the median, and then separate the values below it from the values above it.

With an even-numbered data set, the median is the mean of the two values in the middle, so you simply divide your data set into two halves.

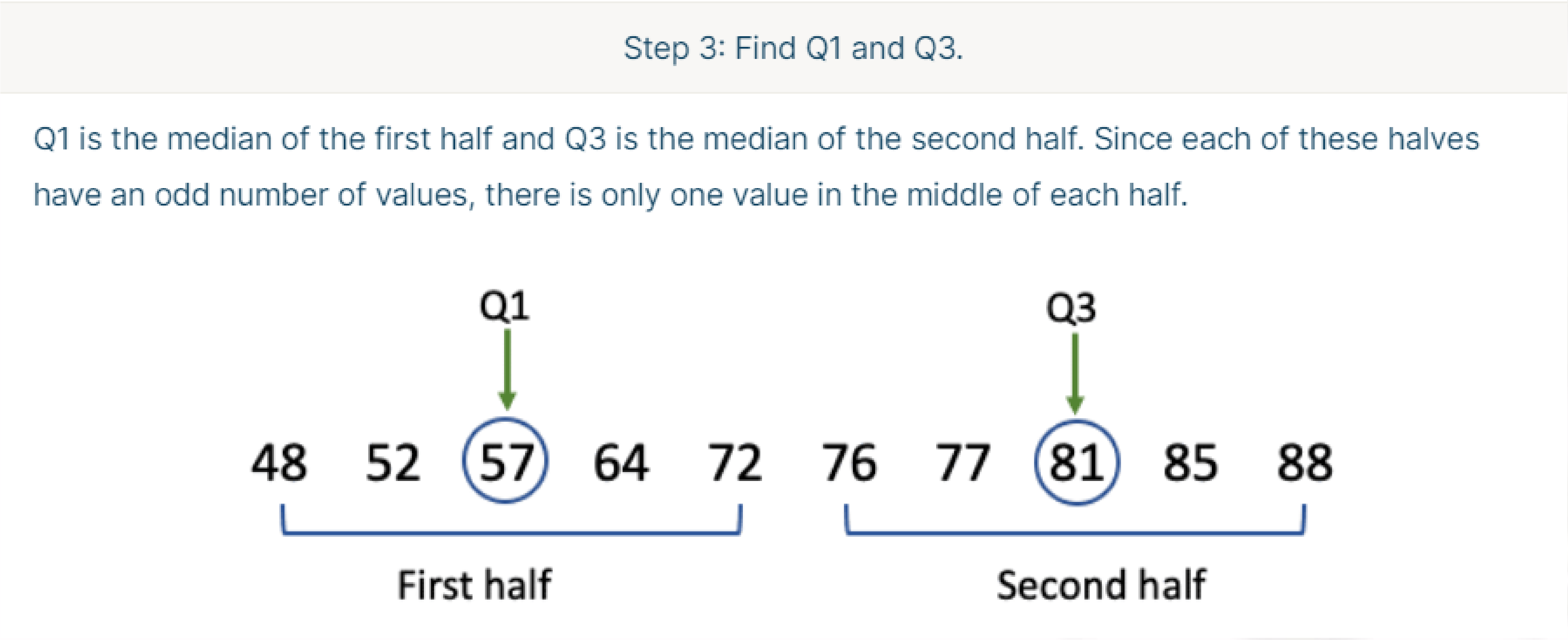
Median

48 52 57 64 72 76 77 81 85 88

First half

Second half

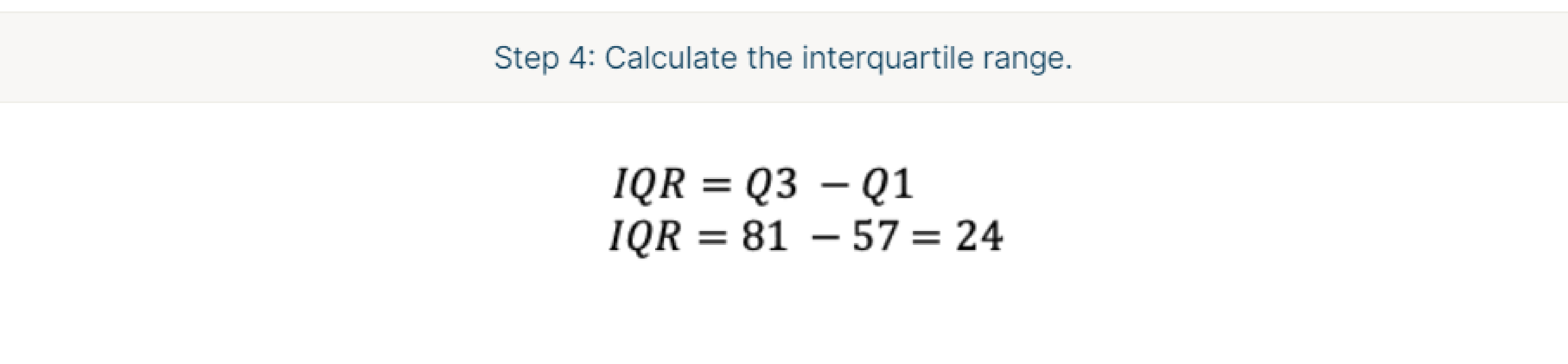
Even-numbered Data set



Step 3: Find Q1 and Q3.

Q1 is the median of the first half and Q3 is the median of the second half. Since each of these halves have an odd number of values, there is only one value in the middle of each half.

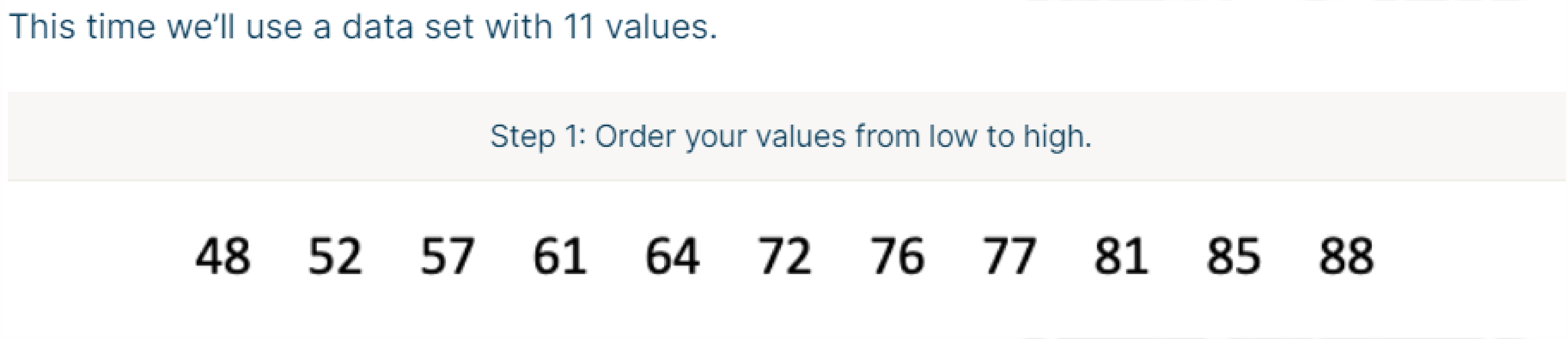
Even-numbered Data set



Step 4: Calculate the interquartile range.

$$IQR = Q3 \ - Q1$$
$$IQR = 81 \ - 57 = 24$$

Odd-numbered Data set

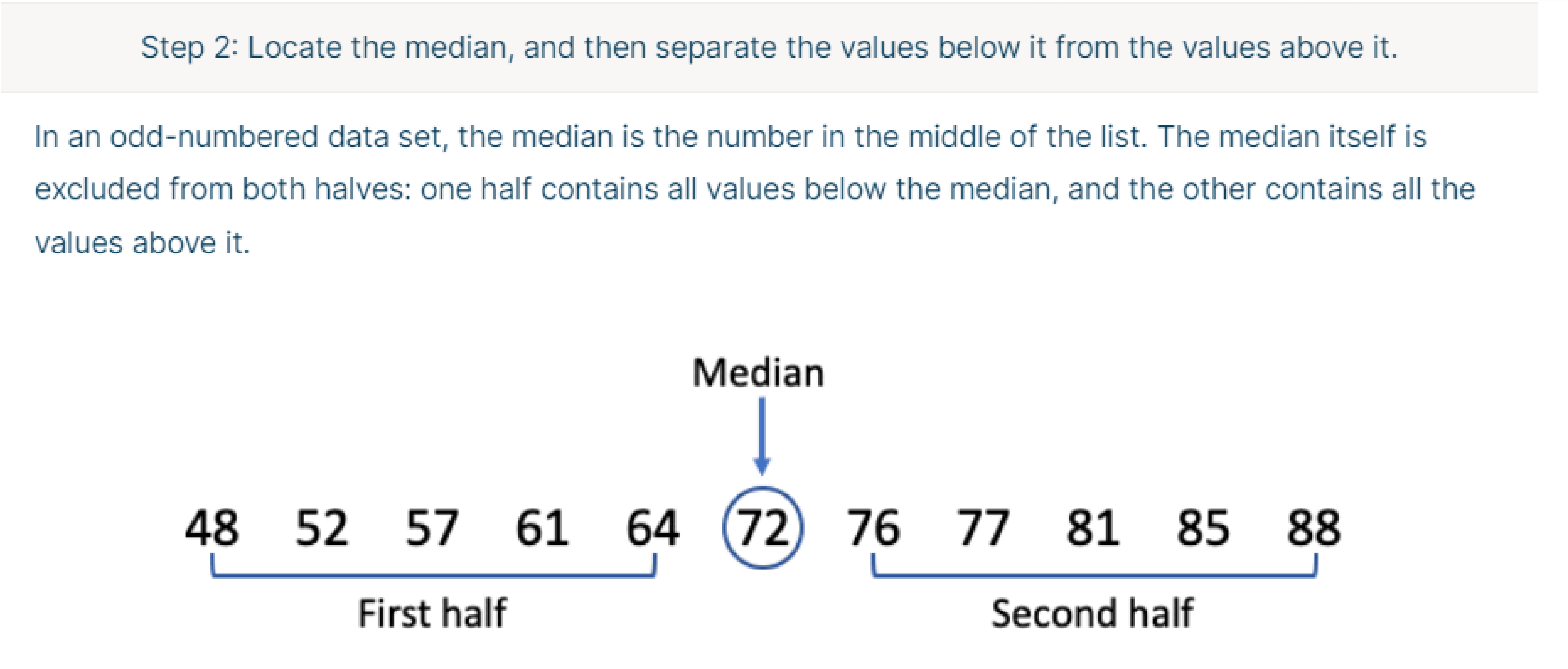


This time we'll use a data set with 11 values.

Step 1: Order your values from low to high.

48 52 57 61 64 72 76 77 81 85 88

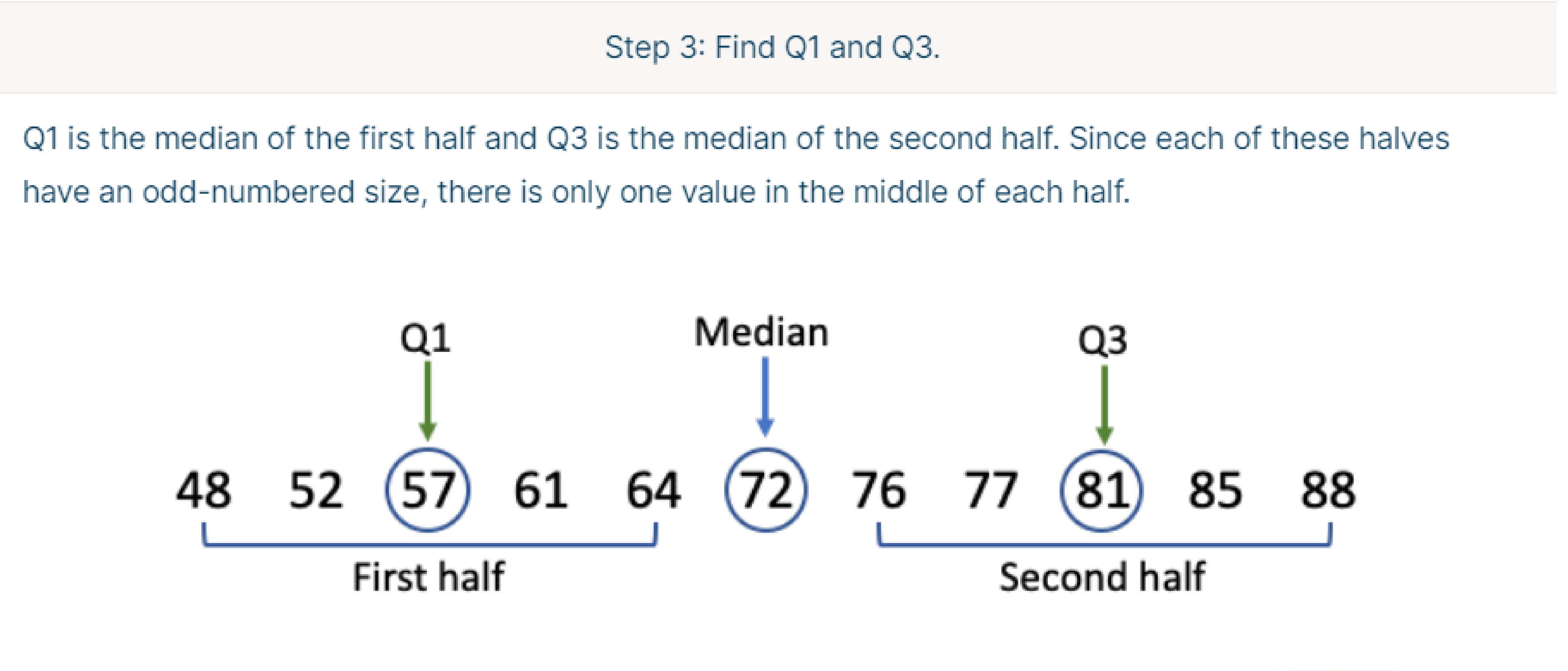
Odd-numbered Data set



Step 2: Locate the median, and then separate the values below it from the values above it.

In an odd-numbered data set, the median is the number in the middle of the list. The median itself is excluded from both halves: one half contains all values below the median, and the other contains all the values above it.

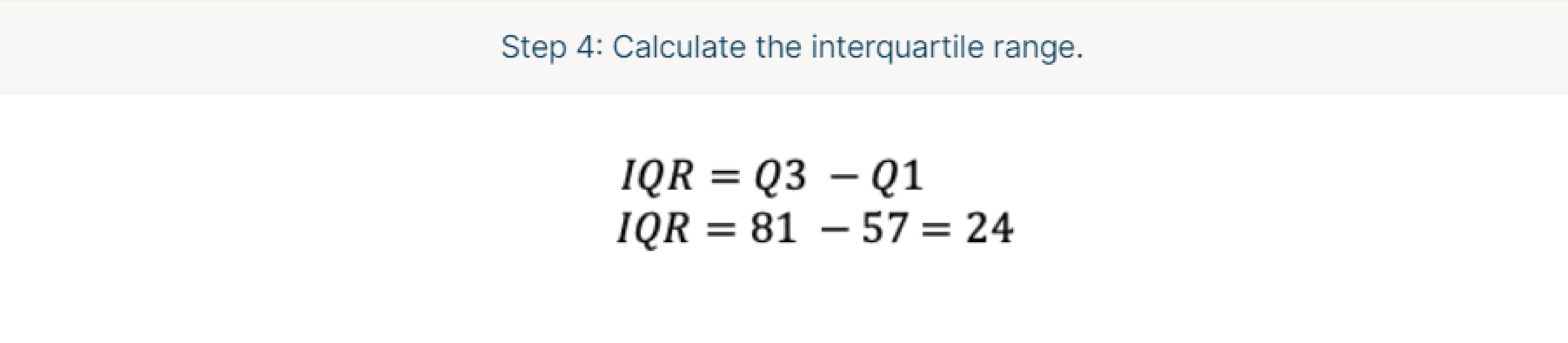
Odd-numbered Data set



Step 3: Find Q1 and Q3.

Q1 is the median of the first half and Q3 is the median of the second half. Since each of these halves have an odd-numbered size, there is only one value in the middle of each half.

Odd-numbered Data set



Step 4: Calculate the interquartile range.

$$IQR = Q3 \ - Q1$$
$$IQR = 81 \ - 57 = 24$$

Interquartile Range

Five-number summary

- Every distribution can be organized using a five-number summary:

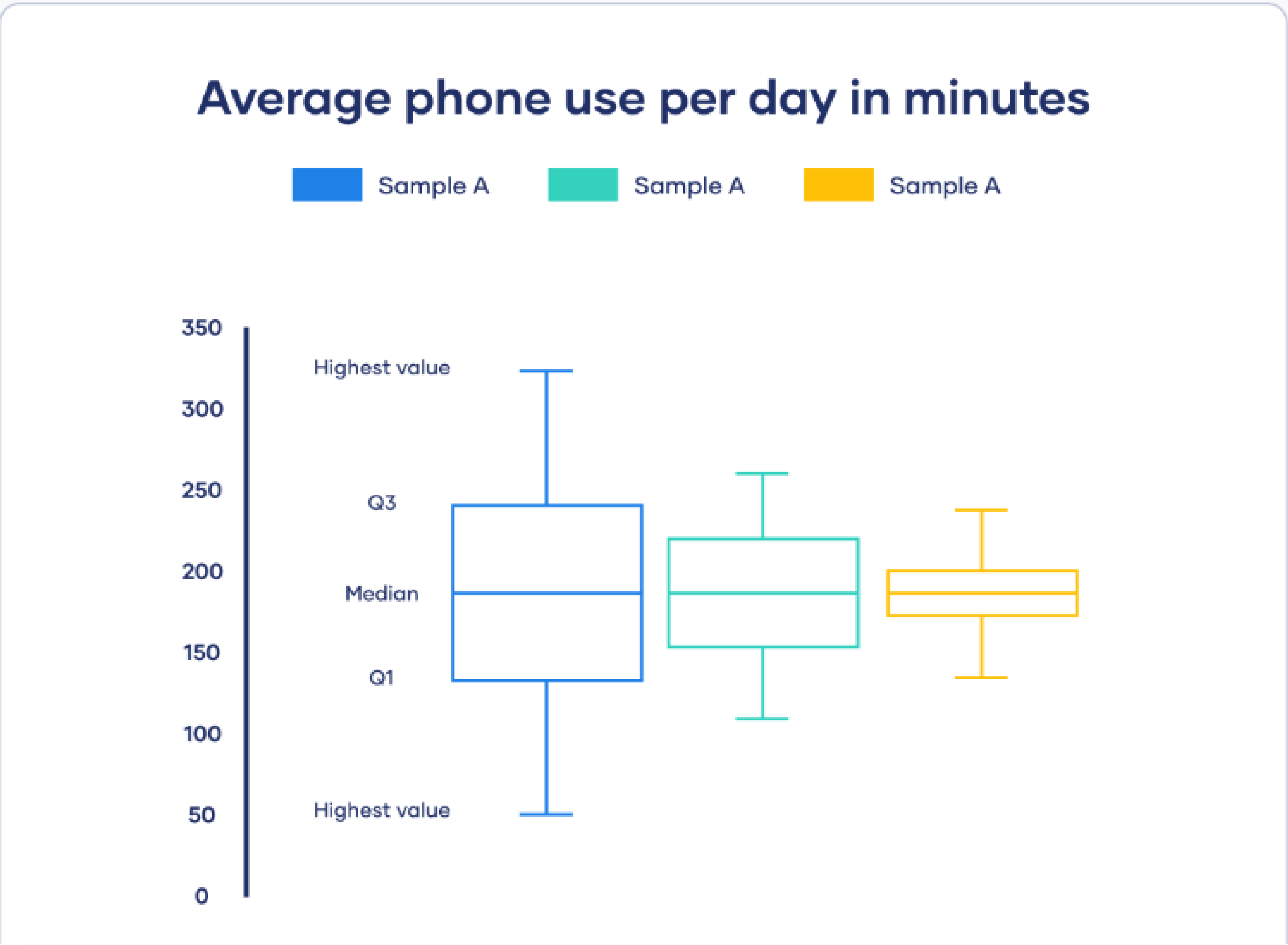
- Lowest value
- Q1: 25th percentile
- Q2: the median
- Q3: 75th percentile
- Highest value (Q4)
- These five-number summaries can be easily visualized using box and whisker plots.

Interquartile Range

Box and whisker plot example

For each of our samples, the horizontal lines in a box show Q1, the median and Q3, while the whiskers at the end show the highest and lowest values.

Box and Whisker Plot Example



Standard Deviation

- The standard deviation is the average amount of variability in your dataset.

- It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

Standard Deviation

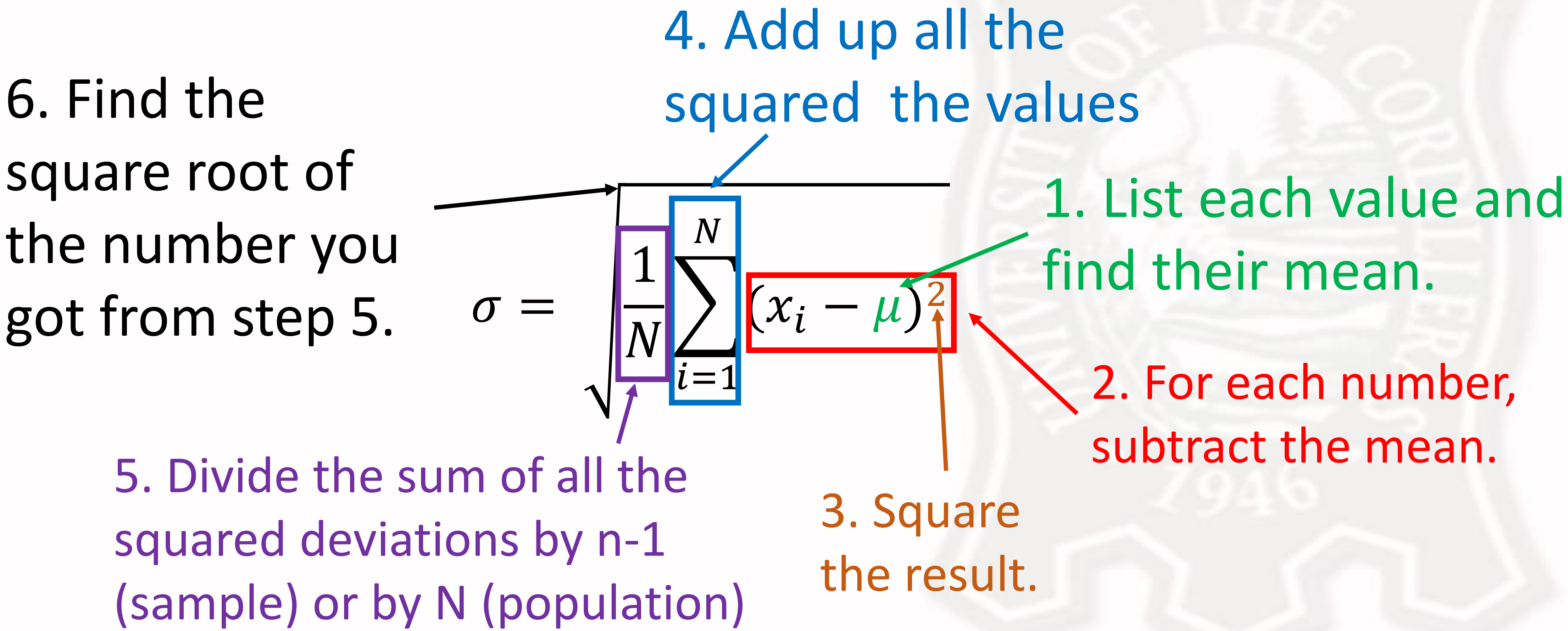
- There are six steps for finding the standard deviation by hand:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by n – 1 (for a sample) or N (for a population).
6. Find the square root of the number you found.

Standard Deviation Formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Standard Deviation Formula:



4. Add up all the squared the values

6. Find the square root of the number you got from step 5.

$$\sigma = \sqrt{\boxed{\frac{1}{N}}\boxed{\sum_{i=1}^{N}}\boxed{(x_i - \mu)^2}}$$

1. List each value and find their mean.

2. For each number, subtract the mean.

5. Divide the sum of all the squared deviations by n-1 (sample) or by N (population)

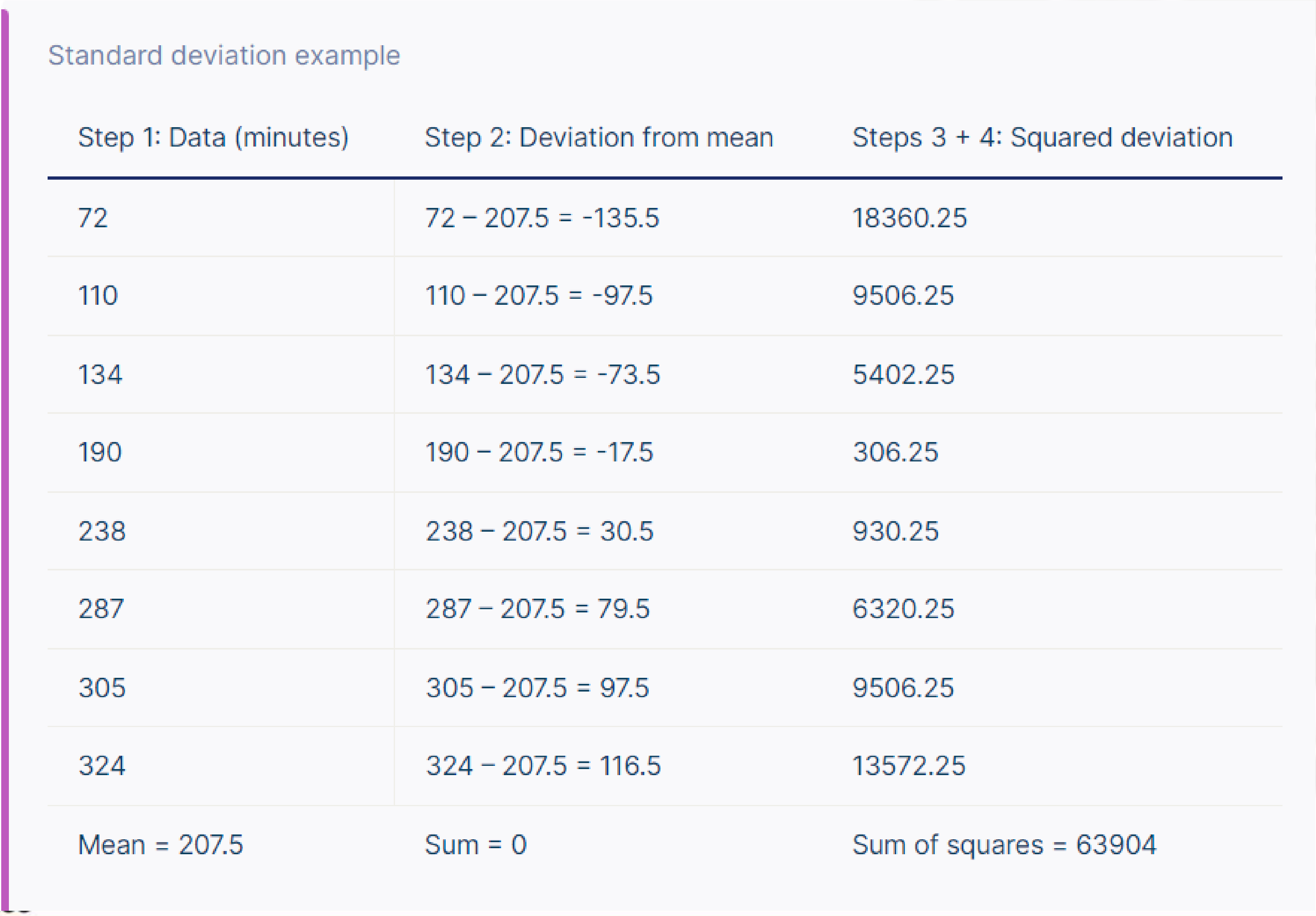
3. Square the result.

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

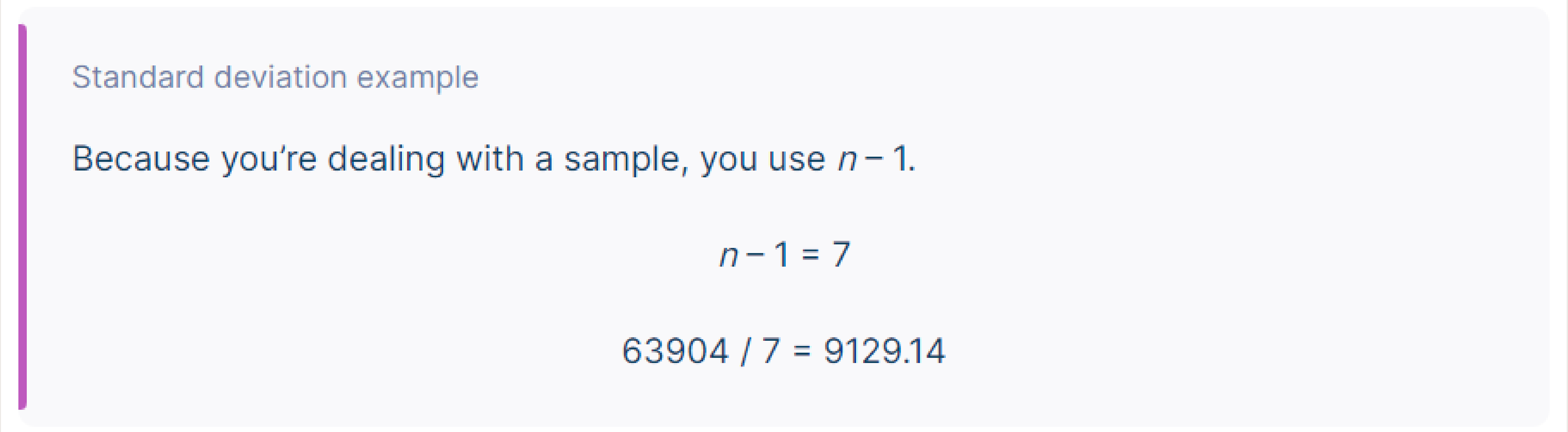
Step 1-4



Standard deviation example

Step 1: Data (minutes)	Step 2: Deviation from mean	Steps 3 + 4: Squared deviation
72	72 – 207.5 = -135.5	18360.25
110	110 – 207.5 = -97.5	9506.25
134	134 – 207.5 = -73.5	5402.25
190	190 – 207.5 = -17.5	306.25
238	238 – 207.5 = 30.5	930.25
287	287 – 207.5 = 79.5	6320.25
305	305 – 207.5 = 97.5	9506.25
324	324 – 207.5 = 116.5	13572.25
Mean = 207.5	Sum = 0	Sum of squares = 63904

Step 5



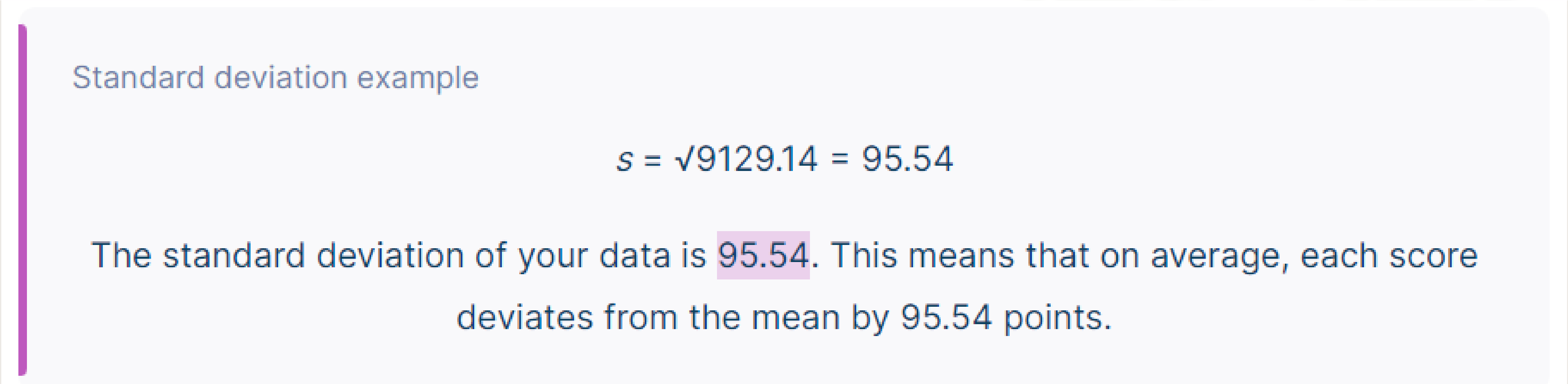
Standard deviation example

Because you're dealing with a sample, you use $n - 1$.

$$n - 1 = 7$$

$$63904 / 7 = 9129.14$$

Step 6



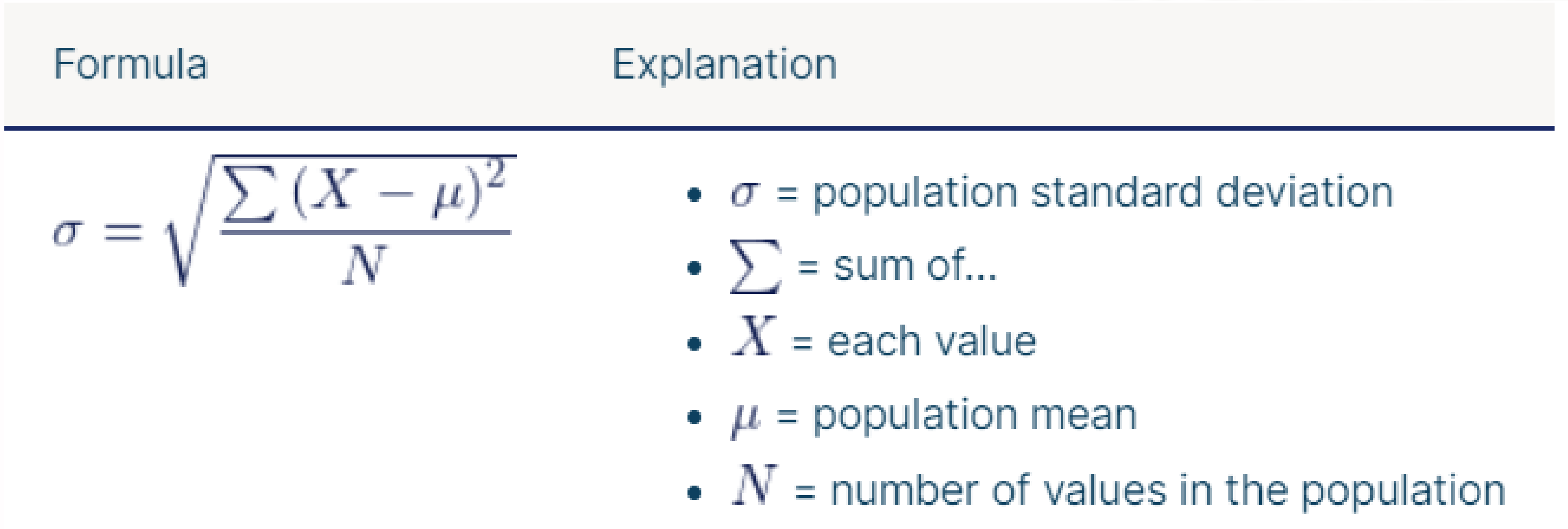
Standard deviation example

$$s = \sqrt{9129.14} = 95.54$$

The standard deviation of your data is 95.54. This means that on average, each score deviates from the mean by 95.54 points.

Standard deviation formula for populations

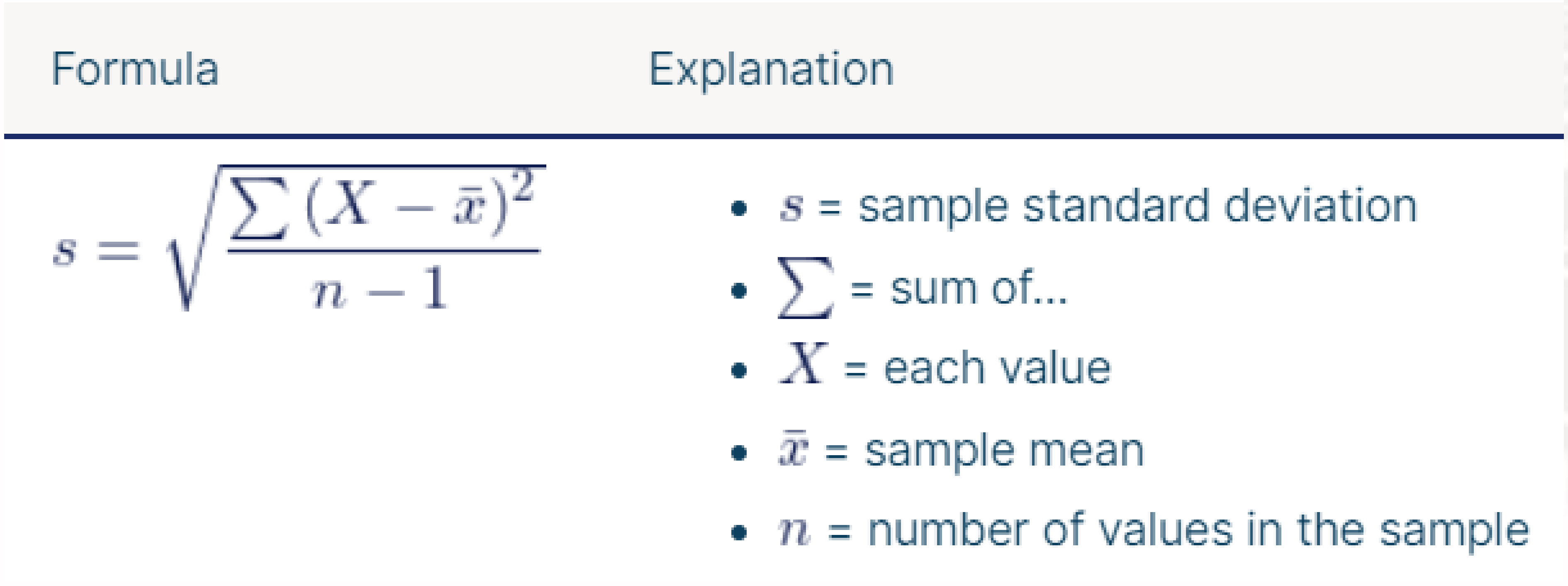
- If you have data from the entire population, use the population standard deviation formula:



Formula	Explanation
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	• σ = population standard deviation • $\sum$ = sum of... • X = each value • μ = population mean • N = number of values in the population

Standard deviation formula for samples

- If you have data from a sample, use the sample standard deviation formula:



Formula	Explanation
$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$	• s = sample standard deviation • $\sum$ = sum of... • X = each value • $\bar{x}$ = sample mean • n = number of values in the sample

Sources:

Bhandari, P.2020.*How to find interquartile range(IQR)|Calculator and examples*.Retrieved from: https://www.scribbr.com/statistics/interquartile-range/ on January 29, 2023.

Bhandari, P.2020.*Variability|Calculating range, IQR, variance, standard deviation*.Retrieved from: https://www.scribbr.com/statistics/variability/ on January 29, 2023.

Math is fun.nd.*Standard deviation formulas*.Retrieved from https://www.mathsisfun.com/data/standard-deviation-formulas.html on January 29, 2023.

Graphic Displays of Basic Statistical Descriptions of Data

By: Bretz Harllynne M. Moltio

Introduction

- There are graphic displays of basic statistical descriptions. These include **quantile plots**, **quantile–quantile plots, histograms,** and **scatter plots.** Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Outliers

- There are graphic displays of basic statistical descriptions. These include **quantile plots**, **quantile–quantile plots, histograms,** and **scatter plots.** Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Univariate descriptive statistics

- Univariate descriptive statistics focus on only one variable at a time. It’s important to examine data from each variable separately using multiple measures of distribution, central tendency and spread. Programs like SPSS and Excel can be used to easily calculate these.

Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute (allowing the user to assess both the overall behavior and unusual occurrences). Second, it plots quantile information.

Quantile Plot

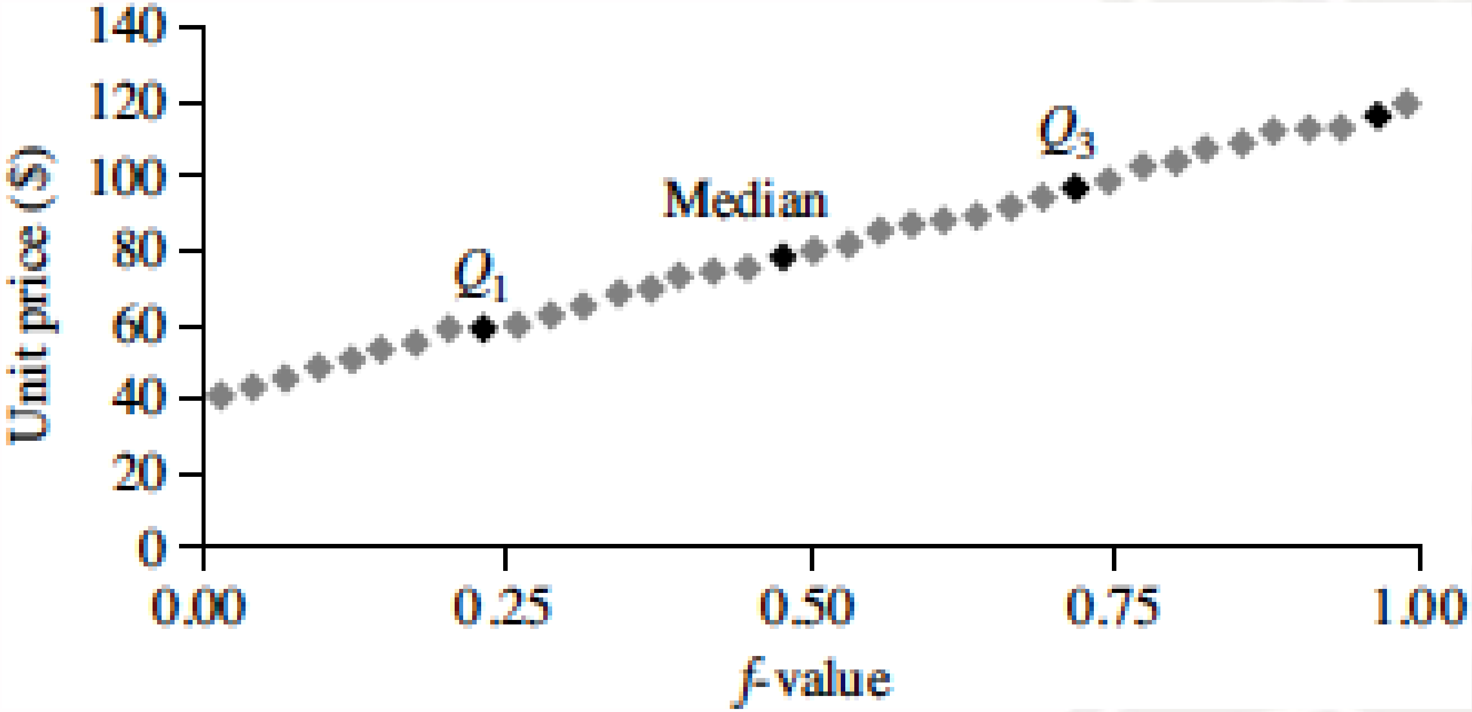
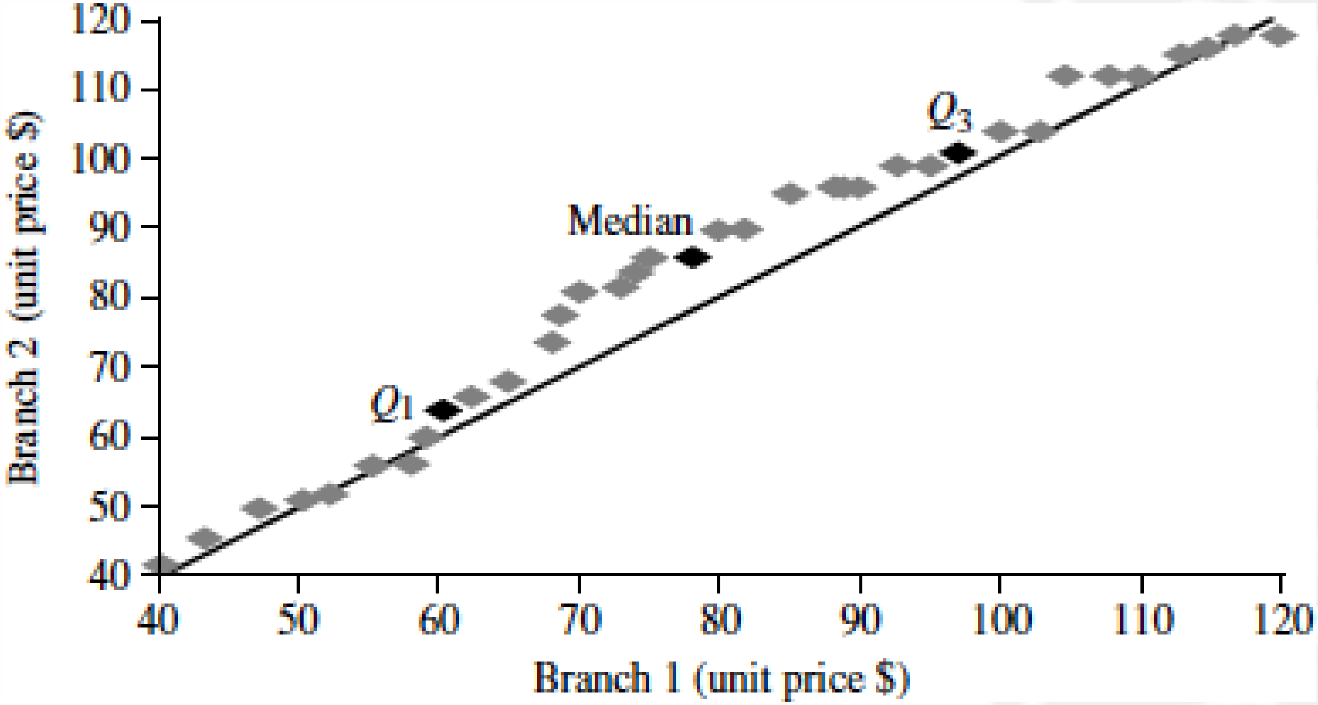


Figure 4. A quantile plot for the unit price data of Table 1.

Quantile-Quantile Plot

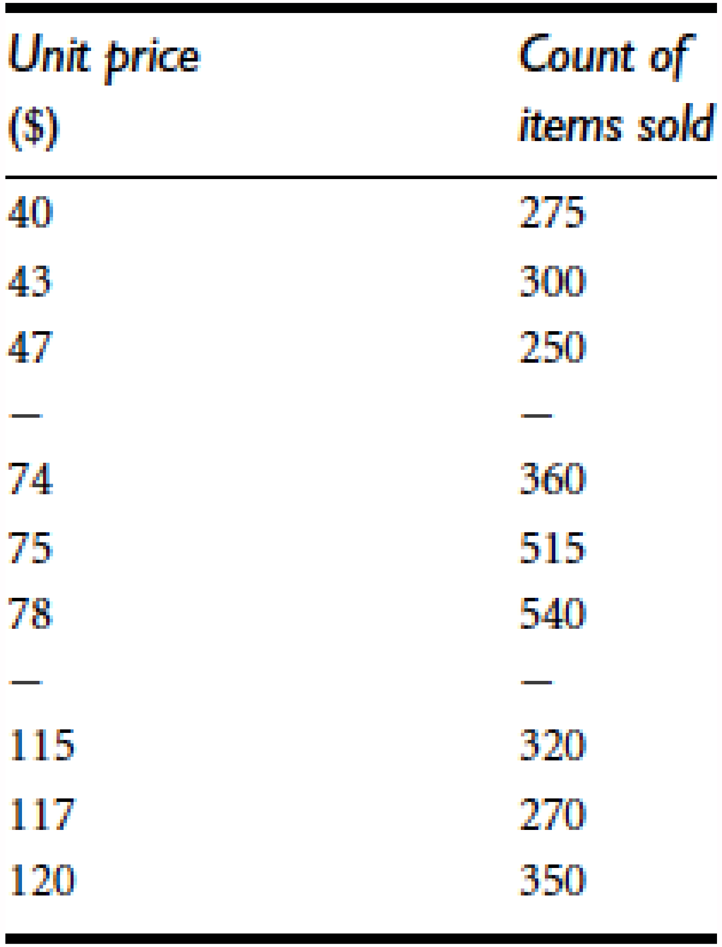
A quantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Quantile-Quantile Plot



- **Figure 5** A q-q plot for unit price data from two AllElectronics branches.

Quantile-Quantile Plot



Unit price ($)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

- **Table 1** A Set of Unit Price Data for Items Sold at a Branch of AllElectronics

Histograms

- Histograms (or frequency histograms) are at least a century old and are widely used. “Histos” means pole or mast, and “gram” means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X.

Histograms

- If X is nominal, such as automobile model or item type, then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a **bar chart**.

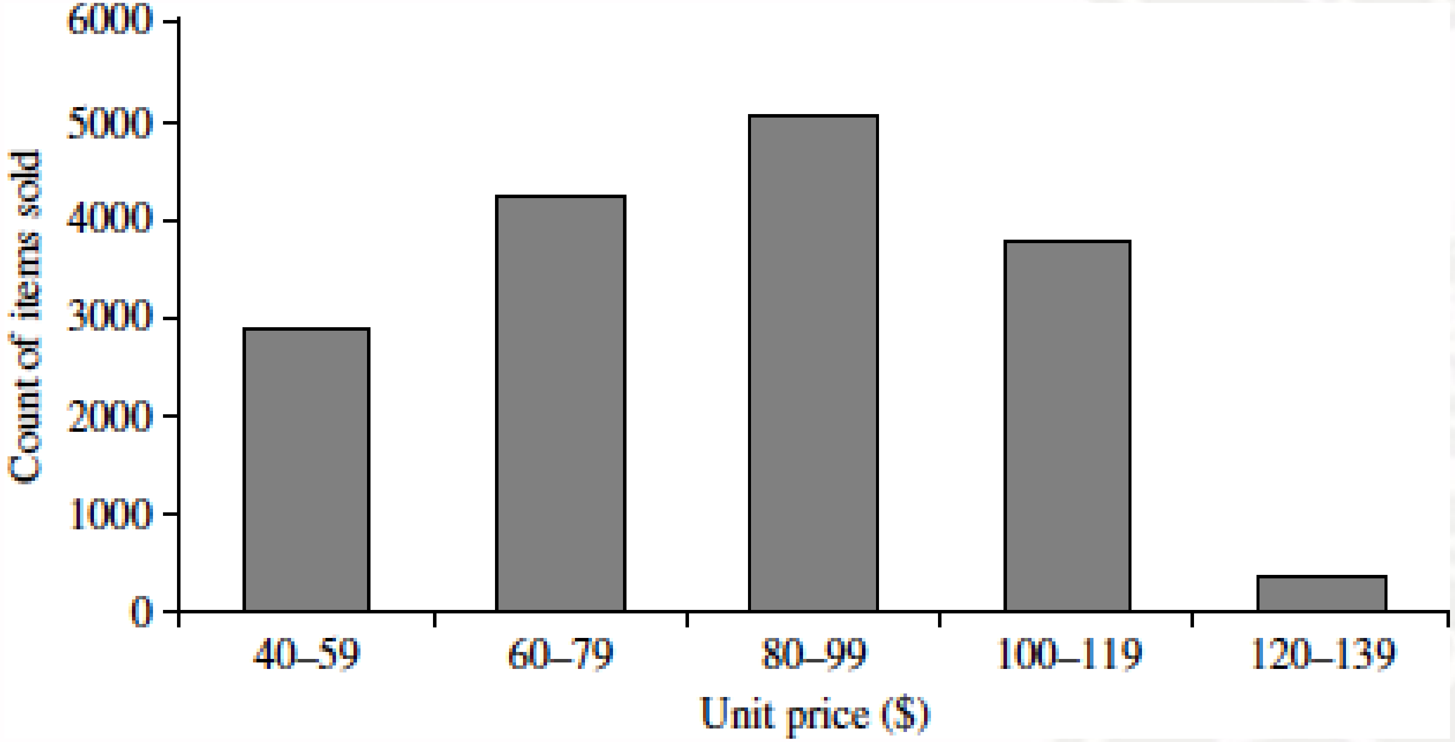
Histograms

- If X is numeric, the term histogram is preferred. The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as **buckets** or **bins**, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width.

Histograms

- Typically, the buckets are of equal width. For example, a price attribute with a value range of $1 to $200 (rounded up to the nearest dollar) can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on. For each subrange, a bar is drawn with a height that represents the total count of items observed within the subrange.

Histograms

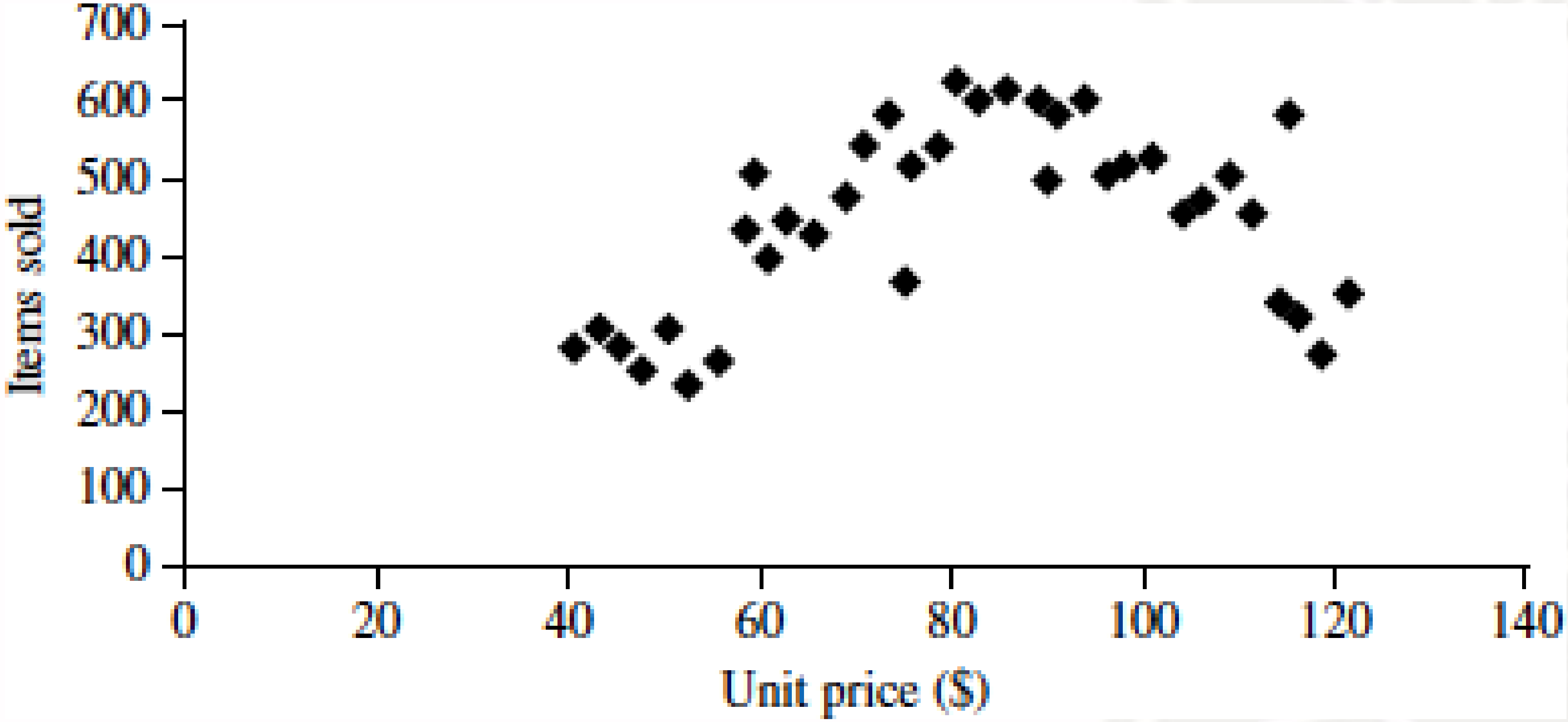


- **Figure 6** A histogram for the Table 1 data set.

Scatter Plot

- A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

Scatter Plot



- **Figure 7** A scatter plot for the Table 1 data set.

Scatter Plot

- Figure 7 shows a scatter plot for the set of data in Table 1. The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.
- Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated).

Measures of Position

Measures of Relative Standing

Measures of Position

- ❑These are values or measurements that divide the distribution into specified fraction or position.
- ❑Quantiles are measures of position that separates the distribution into equal parts.
- ❑Quantiles are extension of the median concept which divide a set of data into four equal parts (quartiles), ten equal parts (deciles), and one hundred equal parts (percentiles).
- ❑The median is equal to the second quartile, fifth decile, or 50th percentile.

Quartiles of Ungrouped Data

- ❑Quartiles are measurements or points that divide the distribution into four (4) equal parts.
  - ✓The values are denoted by Q_1 (lower than or equal to 25%), Q_2 (median), and Q_3 (lower than or equal to 75%) of the distribution.

Ex:

The following data give the average number of minutes required to do an assembly job by the 14 workers of a manufacturing plant:

78, 86, 64, 55, 63, 79, 81, 49, 64, 80, 70, 56, 64, 79

Find the three quartiles.

Solution: Arrange the data in ascending order

49, 55, 56, 63, 64, 64, 64, 70, 78, 79, 79, 80, 81, 86

Find median first: mdn = Q_2

a. mdn = Q_2

$$Q_2 = \frac{64+70}{2}$$

$\boxed{Q_2 = 67}$

b. Q_1= values less than the median

$\boxed{Q_1 = 63}$

c. Q_3= values greater than the median

$\boxed{Q_3 = 79}$

2. The average family income (in thousand pesos) in 1985 for each of the 13 regions in the Philippines was:

28, 23, 21, 20, 38, 31, 58, 27, 29, 24, 18, 27, 24

find the values of the three quartiles.

Quantiles of Grouped Data

- Quartiles are points or measurements that divide the distribution into 4 equal parts.

- Deciles are points or measurements that divide the distribution into 10 equal parts.

  $D_1 = P_{10}$

- Percentiles are points or measurements that divide the distribution into 100 equal parts.

  $D_5 = P_{50} = Q_2$

  $Q_3 = P_{75}$

Computations:

- The formula used in computing quartiles, deciles and percentiles is similar to the formula in computing the median.

Quartiles

$$Q_1 = Lcb + \left[\frac{\frac{1}{4}N - < cfb}{fQ_1c} \right] i$$

$$Q_3 = Lcb + \left[\frac{\frac{3}{4}N - < cfb}{fQ_3c} \right] i$$

Deciles

$$Dr = Lcb + \left[\frac{\frac{rN}{10} - < cfb}{fDc} \right] i$$

r = 1, 2, 3, …,9

Percentiles

$$\mathrm{Pr} = Lcb + \left[\frac{\frac{rN}{100} - < cfb}{fPc} \right] i$$

r = 1, 2, 3, 4, …,99

Ex: Scores of Math CS 4 students.

Class interval	f	Lcb	<cf	rank
195 – 199	1	194.5	50	50th
190 – 194	2	189.5	49	48th – 49th
185 – 189	4	184.5	47	44th – 47th
180 – 184	5	179.5	43	39th – 43rd
175 – 179	8	174.5	38	31st – 38th
170 – 174	10	169.5	30	21st – 30th
165 – 169	6	164.5	20	15th – 20th
160 – 164	4	159.5	14	11th – 14th
155 – 159	4	154.5	10	7th – 10th
150 – 154	2	149.5	6	5th – 6th
145 – 149	3	144.5	4	2nd – 4th
140 – 144	1	139.5	1	1st

Compute:

1. Q1
2. Q3
3. D7
4. D9
5. P55
6. P18

Solution: 1. Q1

Find rank:

r = $\dfrac{1(50)}{4}th$

r = 12.5th

*160 – 164 (Q1 class)

Find Q1:

$$Q_1 = Lcb + \left[\frac{\frac{1}{4}N - <cfb}{fQ_1c}\right]i$$

$$Q_1 = 159.5 + \left[\frac{12.5-10}{4}\right] 5$$

$$= 159.5 + \left[\frac{2.5}{4}\right] 5$$

$$= 159.5 + 3.125$$

$$\boxed{= 162.63}$$

Interpretation: 25% of Math CS 4 students got scores lower than or equal to 162.63.

3. D7

Solution:

Find r:

$$r = \frac{7(50)}{10} th$$

$r = 35^{th}$

*175 – 179 (D7 class)

Compute:

$$D_7 = Lcb + \left[\frac{\frac{7}{10} N - < cfb}{fD_7 c} \right] i$$

$$D7 = 174.5 + \left[\frac{35-30}{8} \right] 5$$

$$= 174.5 + \left[\frac{5}{8} \right] 5$$

$$= 174.5 + 3.125$$

$$\boxed{D7 = 177.625}$$

Interpretation: 70% of Math CS 4 students got lesser than or equal to 177.625

5. P55

Solution:

Find r:

$$r = \frac{55(50)}{100} th$$

r = 27.5th

* 170 – 174 (P55 class)

Compute:

$$P_{55} = Lcb + \left[\frac{\frac{55}{100} N - < cfb}{fP_{55}c} \right] i$$

$$P_{55} = 169.5 + \left[\frac{27.5 - 20}{10} \right] 5$$

$$= 169.5 + \left[\frac{7.5}{10} \right] 5$$

$$= 169.5 + 3.75$$

$$\boxed{P_{55} = 173.25}$$

Interpretation: 55% of the Math CS 4 students got lower than or equal 173.25.

Your Turn: Answer 2, 4, 6

Quiz: Show complete and orderly solution:

The distribution shows the CS scores of CCS.1107 students. Compute and interpret the ff.

1. Q3
2. D9
3. P8
4. P89

Class interval	f
118 – 125	3
110 – 117	6
102 – 109	11
94 – 101	13
86 – 93	14
78 – 85	18
70 - 77	12
62 – 69	10
54 – 61	9
46 – 53	4

2. *Psychology Today* reported on the characteristics of individuals caught in the act of shoplifting in a supermarket. The following frequency distribution shows the age of the shoplifters.

Age of shoplifter	frequency
6 – 11	9
12 – 17	23
18 – 23	40
24 – 29	38
30 – 35	24
36 – 41	17
42 – 47	13
48 – 53	6

Compute and interpret the ff:

1. Q_1
2. D_9
3. P_{35}
4. P_8

3. The accompanying frequency distribution summarizes a sample of speed of cars in kilometers per hour. Interpret each computed value.

Speed	No. of cars (f)
20.00 – 27.45	12
27.50 – 34.95	18
35.00 – 42.45	11
42.50 – 49.95	9
50.00 – 57.45	8
57.50 – 64.95	6
65.00 – 72.45	3
72.50 – 79.95	3

Compute the following and interpret each computed value

1. Q_3
2. D_7
3. P_{35}
4. P_8
5. D_3