Universiti Brunei Darussalam

SM-1402 Basic Statistics

Chapter 2: Descriptive statistics *[handout version]*

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

Semester II 2023/24

Learning outcomes

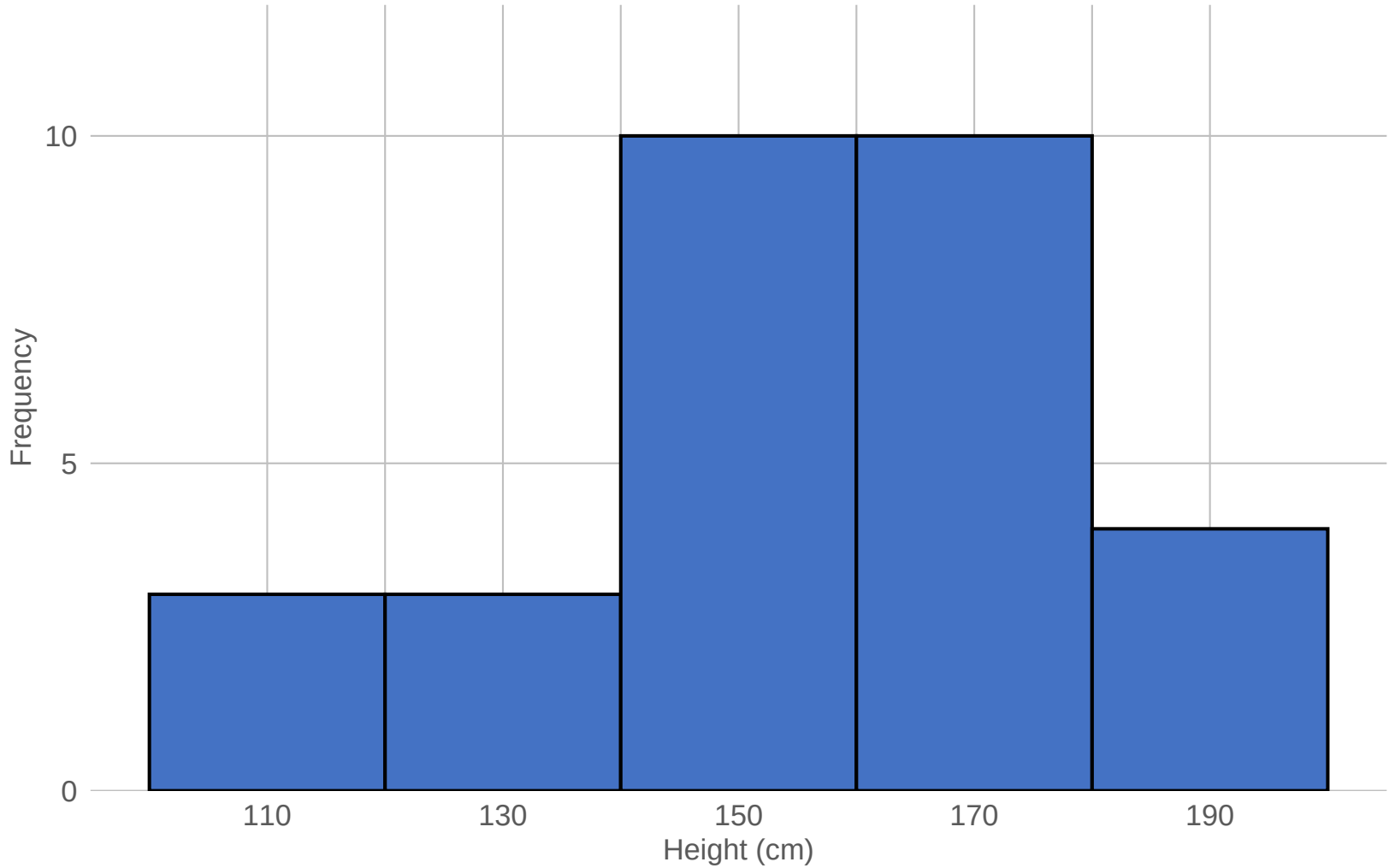
- Calculate the various measures of location (average, median, mode) and know the differences between them.
- Calculate the various measures of spread (range, IQR, SD) and know the differences between them.
- Understand that different data types require different measures of location and spread.

Required reading

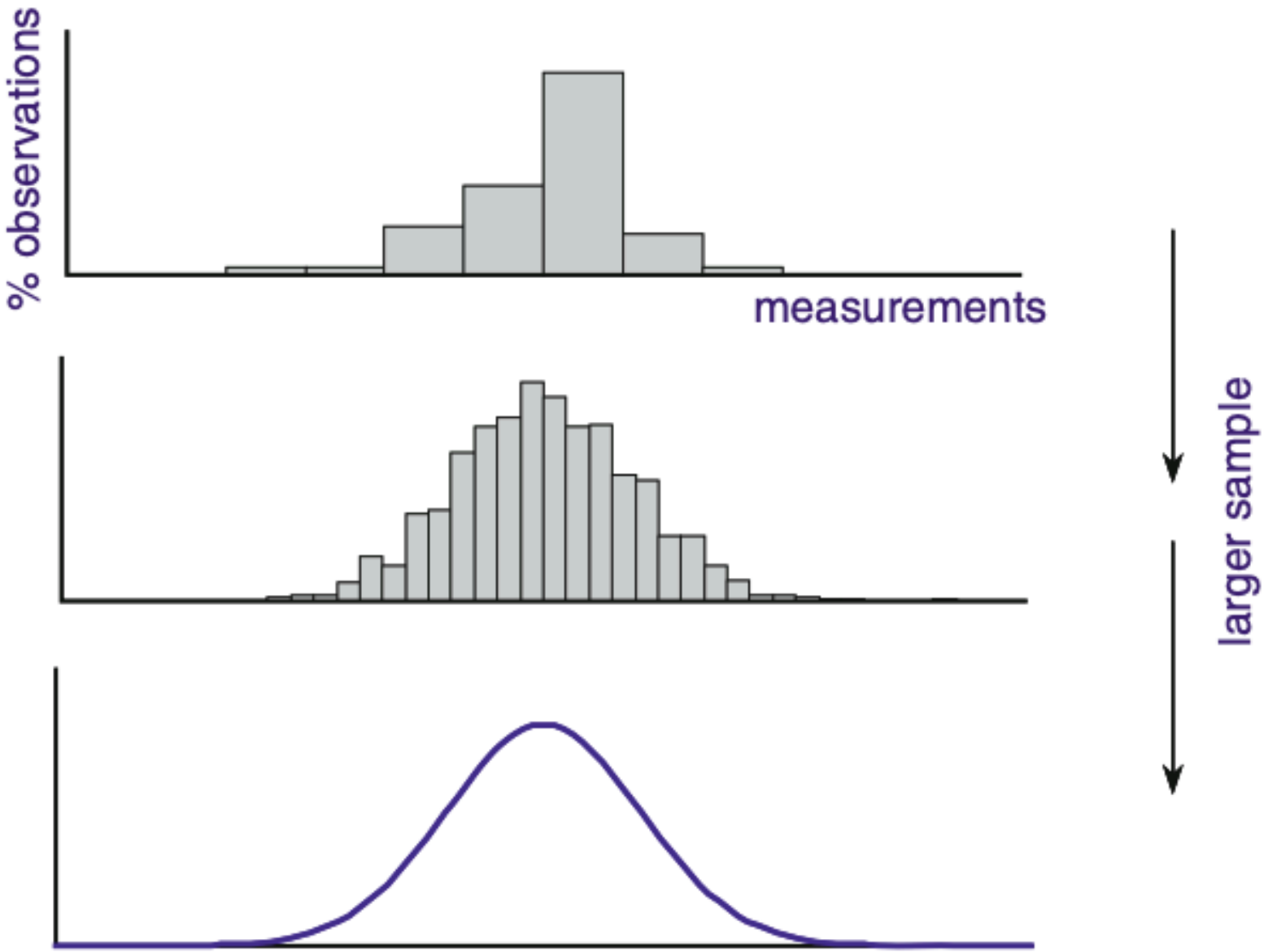
- Madsen (2016) Chapter 3.

Introduction

Statistics is all about describing the variation in data. Previously, we looked at histograms to depict the (approximate) distribution of data.



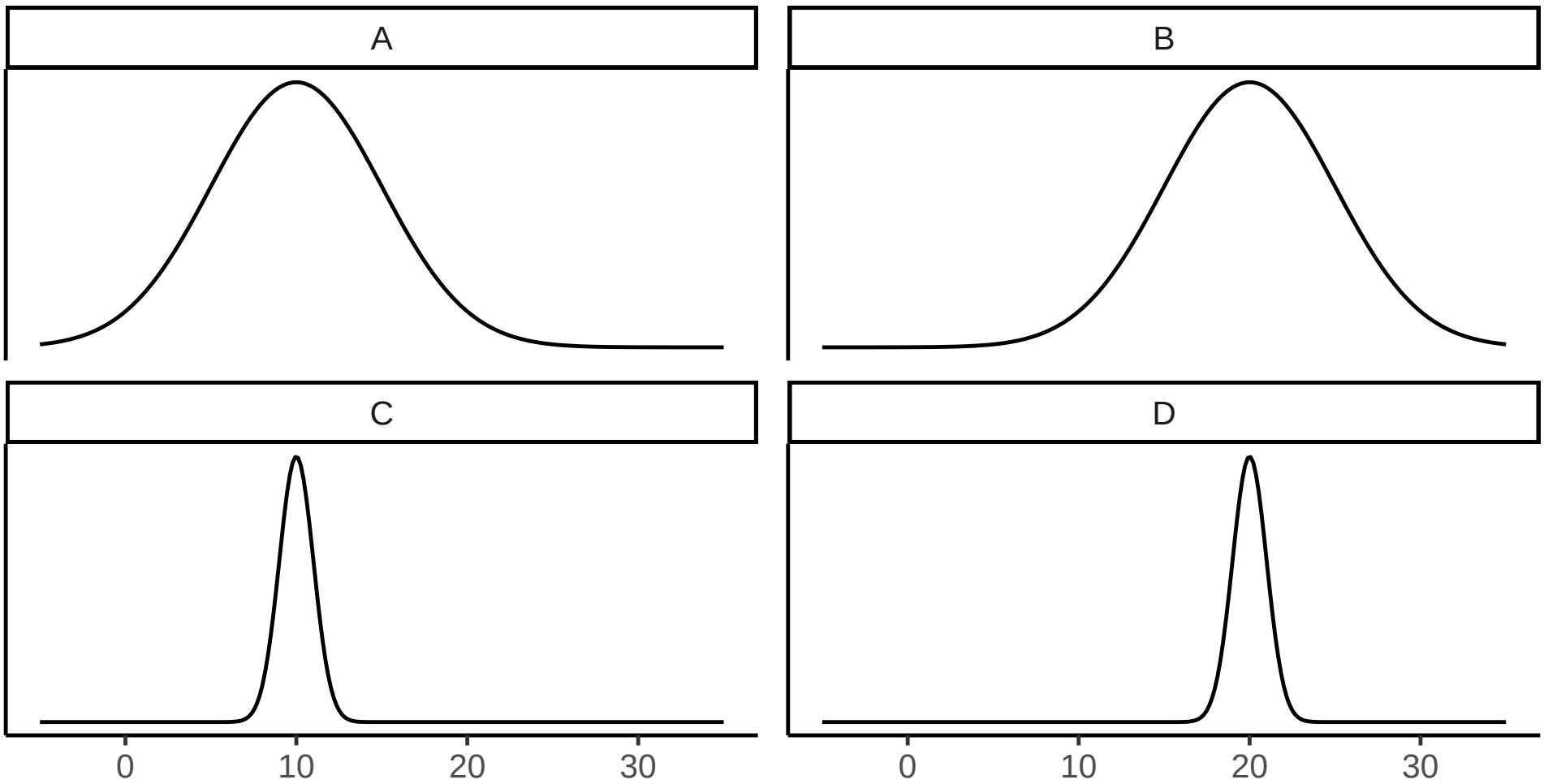
Distribution



Centre and spread

There are two key concepts we are interested in:

1. The centre (or location)–[Systematic variation]
2. The spread (or variability)–[Random variation]



Universiti Brunei Darussalam

Data types

Knowing the data type allows us to correctly apply **meaningful** statistics.

Sample statistic	Nominal	Ordinal	Integer	Interval	Ratio
Measures of location					
Mode	✔	✔	✔	✔	✔
Median	✗	✔	✔	✔	✔
Average	✗	(✗)	(✔)	✔	✔
Quartiles	✗	✔	✔	✔	✔
Measures of spread					
Interquartile range	✗	(✗)	✔	✔	✔
Range	✗	(✗)	✔	✔	✔
Standard deviation	✗	✗	(✔)	✔	✔
Coef. of variation	✗	✗	(✔)	✗	✔

Average

Definition 1 (Average)

The *average* (aka *mean*) is a measure of the centre in the distribution of data values.

To calculate the average, take the sum of all the data values, and divide by the count.

- The symbol for average is $\bar{x}$ (read x bar).
- The average is highly influenced by "extreme values", which can give a poor reflection of the centre of the data.

Example 2

Take the average of 3, 5, 6, 4 (4 numbers).

- The sum is $3 + 5 + 6 + 4 = 18$.
- The average is $\bar{x} = \frac{18}{4} = 4.5$.

Calculation formula

Suppose we have n data values, labelled as $x_1, x_2, \ldots, x_n$. The average can be calculated using this formula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Example 3

For the previous example,

$$\overbrace{3}^{x_1}, \overbrace{5}^{x_2}, \overbrace{6}^{x_3}, \overbrace{4}^{x_4},$$

with $n = 4$.

- The sum is $\sum_{i=1}^{4} x_i = 18$.
- The average is $\bar{x} = \frac{18}{4} = 4.5$.

Median

Definition 4 (Median)

The median is the data value that is "in the middle", i.e. it is a number that divides the ordered data into two parts with an equal number of values.

Steps to find the median:

1. Arrange the data values $x_1, \ldots, x_n$ in ascending order.
2. Depending on whether n is even or odd:
   - If n is odd, then the median is the middle value. [The $\frac{n-1}{2}$th value]
   - If n is even, then the median is the *average* of the two middle values. [The average of the $\frac{n}{2}$th and $(\frac{n}{2}+1)$th value]

The median is not as sensitive to extreme values as the average!

Median (cont.)

Example 5

Find the median of 3, 5, 6, 4. First, we sort the data in ascending order

$$3, 4, 5, 6$$

Since there are an even number of data values, we take the average of the two middle values 4 and 5, so $M = (4 + 5)/2 = 4.5$.

Some remarks

- The median and the average are the same for this example, but they need not be.
- For skewed distributions, these will definitely be different.

Median (cont.)

Example 6

Consider the Fitness Club example; data values are the age of the $n = 17$ boys. First arrange the data in ascending order:

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Age	12	12	13	13	13	13	14	14	14	14	14	15	15	15	16	17	17

As there are 17 data values, the middle value is no. 9, i.e., the data value of 14. Out of the 17 data values, there are 8 data values on both sides of data value no. 9.

Universiti Brunei Darussalam

Mode

Definition 7 (Mode)

The *mode* is simply the most frequent data value.

All that is needed to compute the mode is a frequency count.

Example 8

The tabulated frequency of the age of boys in the Fitness Club example:

x	12	13	14	15	16	17
Freq	2	4	5	3	1	2

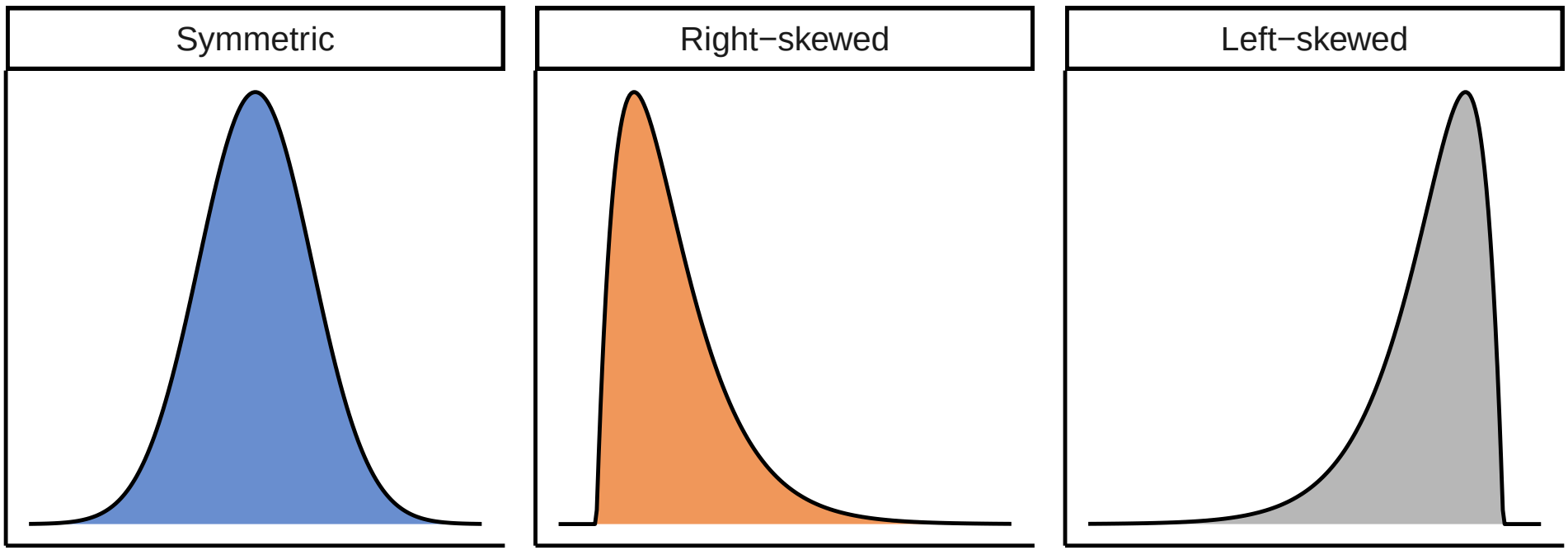
The mode has one very big disadvantage: There can be multiple modes!

Universiti Brunei Darussalam

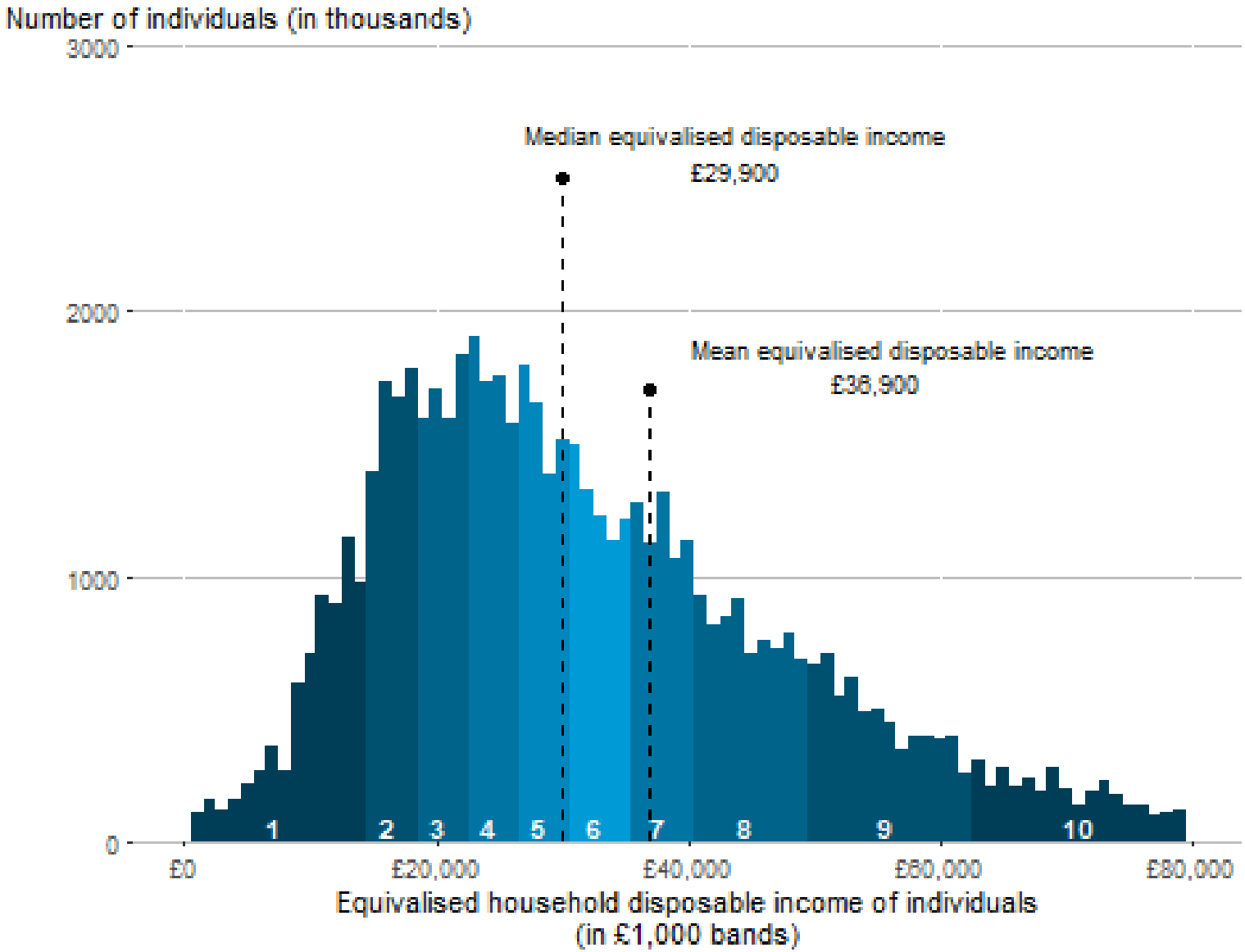
Choosing measure of location

Appropriate measures of location depend on the *skewness* of the distribution.

- For symmetrical distributions, then there is little difference between the average, median and mode.
- For skewed (i.e. non-symmetrical) distributions, then the average, median and mode are not the same.



Income distribution



Source: Office of National Statistics, UK

Range

Definition 9 (Range)

The *range* is simply the width of the interval of the data values.

It is calculated as the difference between the largest and the smallest data values:

$$R = x_{\text{max}} - x_{\text{min}}.$$

- The advantage is that it is easy to compute and easy to understand.
- If there are too many data values or extreme values in the data, then the range is affected.
- Range is mainly used for small samples.

Example 10

The range for the numbers 3, 5, 6, 4 is $R = 6 - 3 = 3$.

Variance and standard deviation

The *standard deviation* is the most common measure of dispersion. It can be interpreted as the average distance between the data values and the mean value.

Definition 11 (Variance)

The *variance* is the average of the squared distances between the data and the mean value.

Definition 12 (Standard deviation (SD))

The *standard deviation* (SD) is the square root of the variance.

Both variance and SD are dispersion measurements. However:

- The variance is measured in the square units of the original measure.
- The SD, on the other hand, is in the same units of measurement.

Steps to compute the SD

To calculate the standard deviation s of a sample data $x_1, \ldots, x_n$, do the following:

1. Calculate the sample average, $\bar{x}$.
2. For each reading x_i, calculate $x_i - \bar{x}$, its deviation from the mean.
3. Square this deviation to give $(x_i - \bar{x})^2$.
4. Sum all of these squared deviations to give $\sum_{i=1}^{n}(x_i - \bar{x})^2$.
5. Divide this sum by $n - 1$ to give the (unbiased) sample variance, s^2.
6. Finally, take the positive square root of the variance to obtain the standard deviation s.

More succintly, we have the formula

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Example

Example 13

Compute the SD of 3, 5, 6, 4. Firstly, compute the mean, which is $\bar{x} = 4.5$. Then, in tabular form, we perform the steps above.

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	3	-1.5	2.25
2	5	0.5	0.25
3	6	1.5	2.25
4	4	-0.5	0.25

Then, $s^2 = \frac{1}{3}(2.25 + 0.25 + 2.25 + 0.25) = 5/3$. The SD is therefore $s = \sqrt{5/3} = 1.29$.

Universiti Brunei Darussalam

Some remarks

Why divide by $n-1$ (aka Bessel's correction)?

- Short answer: It is customary!
- Long answer: If we were dealing with the population variance, then dividing by n is correct. But, we're dealing with a sample. As such, the true population mean is not known! Making the calculation using the sample mean leads to errors which consequently underestimates variance. So dividing by a smaller number $(n-1)$ corrects this fact.

Good news is, if you have a respectable sample size, using either n or $n-1$ doesn't make too much of a difference anyway!

On another note, there is an alternative formula which can be used:

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}.$$

Interquartile range

Recall when the median was calculated, we divided the data into two parts. We can further divide each part in half again, effectively getting four parts with (roughly) the same number of data values.

- The lower quartile (Q1) is the value under which 25%...
- The median (Q2) is the value under which 50%...
- The upper quartile (Q3) is the value under which 75%...

...of data points are found when they are arranged in increasing order.

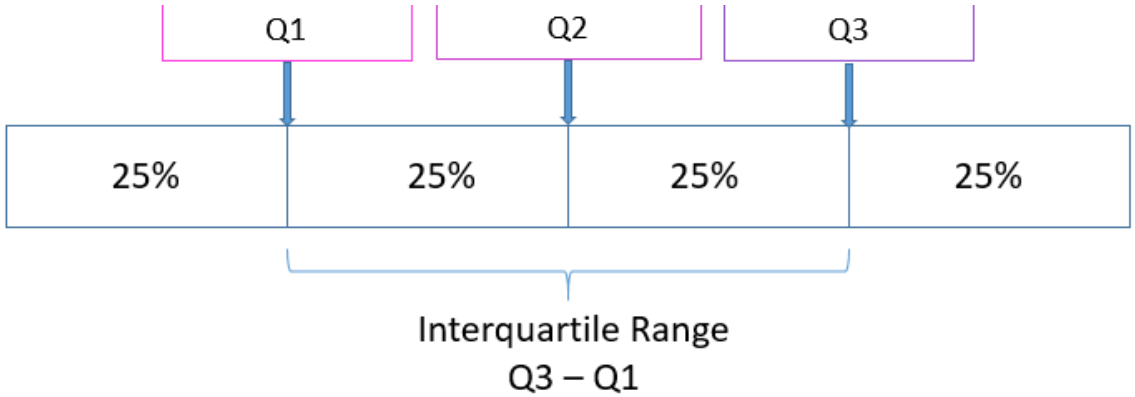
Example 14

Back to Example 6 earlier (age of boys in Fitness Club sample).

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Age	12	12	13	13	13	13	14	14	14	14	14	15	15	15	16	17	17

Q1 = (13 + 13)/2 = 13 (4.5th position); Q2 = 14 (9th position); Q3 = (15 + 15) / 2 = 15 (13.5th position).

Interquartile range (cont.)



Definition 15 (Interquartile range)

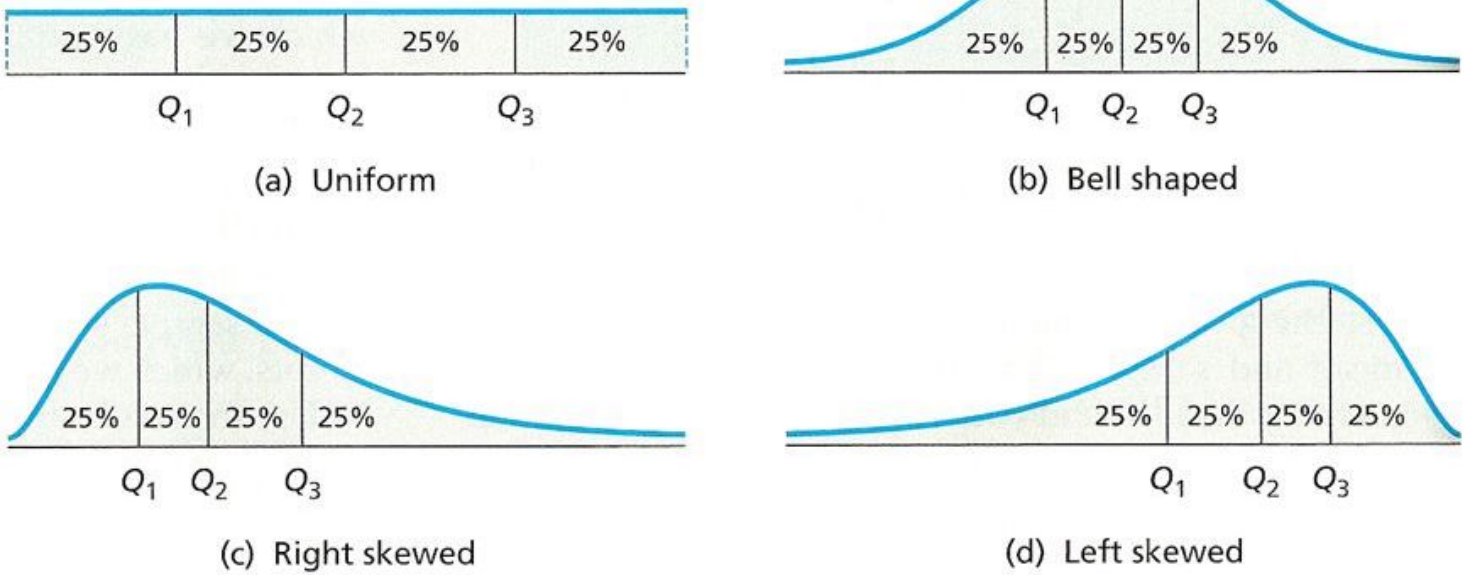
The interquartile range (IQR) is the difference between the upper and lower quartiles. Thus, IQR $= Q3 - Q1$.

Example 16

For the Fitness Club boys' ages, the IQR is $15 - 13 = 2$. Just for comparison,

- Range $= 5$
- SD $= 1.51$

Which measure to use?



For symmetric distributions

- Location: Mean, median or mode (they should be roughly the same)
- Spread: SD

For asymmetric (skewed) distributions

- Location: Median
- Spread: IQR

Relative spread

- When comparing samples from several time periods, the average often increases over time.
- At the same time, the spread increases with increasing average (usually). So to make a valid comparison, the relative spread is used.

Definition 17 (Coefficient of variation)

The *coefficient of variation* (CV) is used as a measure of relative spread, expressed as a percentage:

$$\text{CV} = \frac{s}{\bar{x}} \times 100\%$$

Example 18

The numbers 3, 5, 6, 4 has an average of $\bar{x} = 4.5$ and SD of $s = 1.29$. Thus, $\text{CV} = 1.29/4.5 \times 100\% = 29\%$.

Warning: Use only for data points with a lower limit of 0!

Relative spread (cont.)

Example 19

Imagine we have two classes, Class A and Class B, each taking a different test. Results from the test are:

- Class A: 80, 79, 80, 73, 84
- Class B: 45, 48, 43, 43, 43

CV compares the *consistency*[1] of the test scores between the two classes.

- Class A: $\bar{x} = 79.2$, $s = 4$, CV $= 5\%$.
- Class B: $\bar{x} = 44.4$, $s = 2.2$, CV $= 4.9\%$.

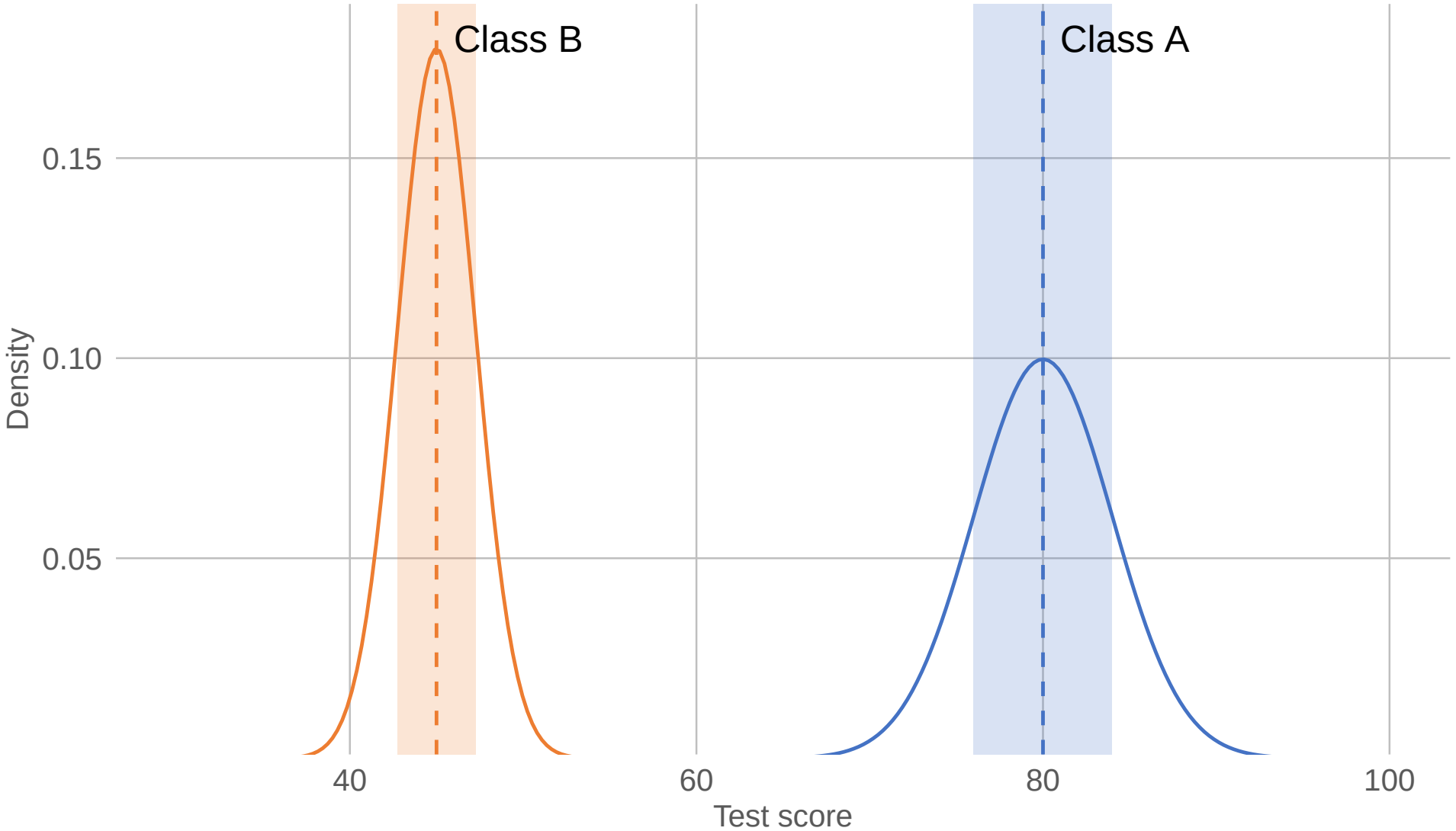
Class A's mean score is higher than Class B's, and so is the spread. However, the CV are similar, indicating that the two classes have similar consistency in their test scores.

Since scores start at 0, CV may be used. What if, for instance, everyone's scores were scaled up by 20 points in Class B. Would the CV be the same?

[1] The degree to which the data points are close to each other and the average value.

Relative spread (cont.)

Percentage of what, exactly?



Universiti Brunei Darussalam