Outcome measures based on taking measurements on people or things (i.e. continuous data)

Understanding measurement data

Example:

The following data are cholesterol levels (mmol/L) of 40 healthy men all aged 45 years. How can we describe the distribution of cholesterol in this SAMPLE of men?

3.6	3.8	3.9	4.1	4.2	4.5	4.5	4.8
5.1	5.3	5.4	5.4	5.6	5.8	5.9	6.0
6.1	6.1	6.2	6.3	6.4	6.5	6.6	6.8
6.9	7.1	7.2	7.2	7.3	7.4	7.5	7.7
8.0	8.1	8.1	8.2	8.3	9.0	9.1	10.0

For this type of data, we need:

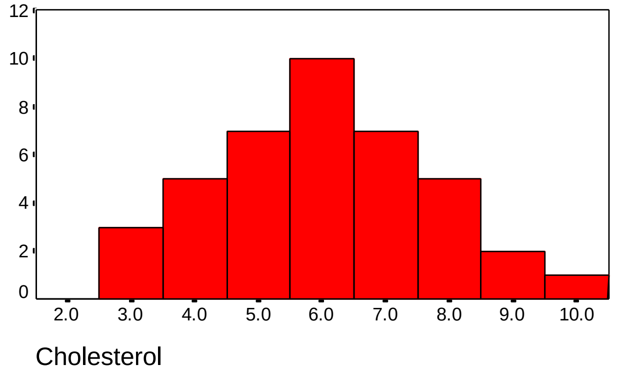
- Some measure of what a typical value is
- Some measure of spread

- What do you think they might be?

Divide the cholesterol values into groups, and count how many men are in each group (**frequency table**)

Cholesterol (mmol/L)	Number	Percentage (%)
3.0 – 3.9	3	7.5
4.0 – 4.9	5	12.5
5.0 – 5.9	7	17.5
6.0 – 6.9	10	25.0
7.0 – 7.9	7	17.5
8.0 – 8.9	5	12.5
9.0 – 9.9	2	5.0
10.0 – 10.9	1	2.5
Total	40	100.0

Histogram of cholesterol (mmol/L)



Measure of typical value

Mean = $\frac{\text{Sum of all the data values}}{\text{Number of data values}}$

Median = the value that has half the data points below it and half above

Symmetrical data

Mean cholesterol:

Mean = $\frac{256}{40}$ = 6.4 mmol/L

(the mean for a sample is often notated by $\bar{x}$)

Median cholesterol:

For the cholesterol data there are 40 data points, so median has 20 data points above it and 20 data points below.

It is halfway between the 20th and 21st data points, i.e. between 6.3 and 6.4

Median = 6.35 mmol/L

Symmetric data

Mean cholesterol = 6.4 mmol/L

Median cholesterol = 6.35 mmol/L

Mean and median here are very similar.
This is because these data are **symmetrical** about their middle.

Measure of centre - Skewed data

Skewed data is not symmetric about the middle:

e.g. Data values: 1 2 3 4 100

Mean =

Median =

Measure of centre - Skewed data

Skewed data is not symmetric about the middle:

e.g. Data values: 1 2 3 4 100

Mean = $\frac{110}{5}$ = 22

Median = 3

The mean and median are very different

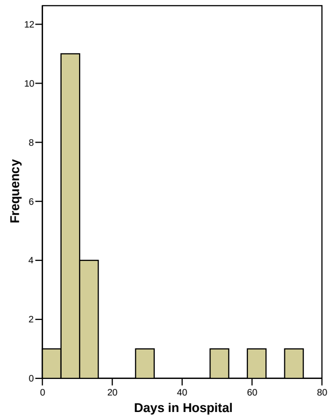
Mean takes the **size** of each value into account, so will be influenced by a few very large (or small) values.

Median uses only the **ranking** of the values, so is not influenced by a few very large (or small) values.

Example:

Length of stay in hospital for 20 patients

Days in Hospital	Frequency	Cumulative Percent
5	1	5
6	1	10
7	2	20
8	3	35
9	4	55
10	1	60
12	2	70
14	2	80
28	1	85
52	1	90
61	1	95
71	1	100
Total	20	



Mean = 17.95 - influenced by the extreme values

Median = 9

For the hospital data we use the median as the measure of centre of the data, because it better represents a ‘typical’ length of stay than the mean (ie it better represents <u>most</u> of the patients)

Percentile (or centile)

The kth centile is the point below which k% of the data values lie.

The 50th centile is the median.

Measures of spread

1. Standard deviation

2. Inter-Quartile Range (IQR)

Example

Calculating standard deviation of 5 cholesterol values

Cholesterol value	6.2	6.3	6.4	6.5	6.6
Mean value =	6.4 mmol/L				
Difference from the mean:	-0.2	-0.1	0	+0.1	+0.2
Sum the differences:	0				

Cholesterol value	6.2	6.3	6.4	6.5	6.6
Difference from mean:	-0.2	-0.1	0	+0.1	+0.2
Square the differences:	0.04	0.01	0	0.01	0.04
Sum of squares:	0.10				

Standard deviation (SD) =

$$\sqrt{\frac{\text{Sum of (the distances of each data point from the mean)}^2}{(\text{Number of data values} - 1)}}$$

[Divide sum of squares by (number of observations –1): 0.10/4 = 0.025

Take square root to get back to original scale, standard deviation = √0.025 = 0.158mmol/L]

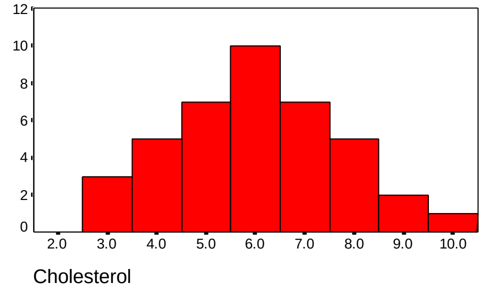
SD = measure of the average spread of the data about their mean

[the standard deviation squared is called ‘variance’]

NB: if you only have very few data values (eg <5) , showing them all could be just as informative as giving the SD

Standard deviation for symmetric data

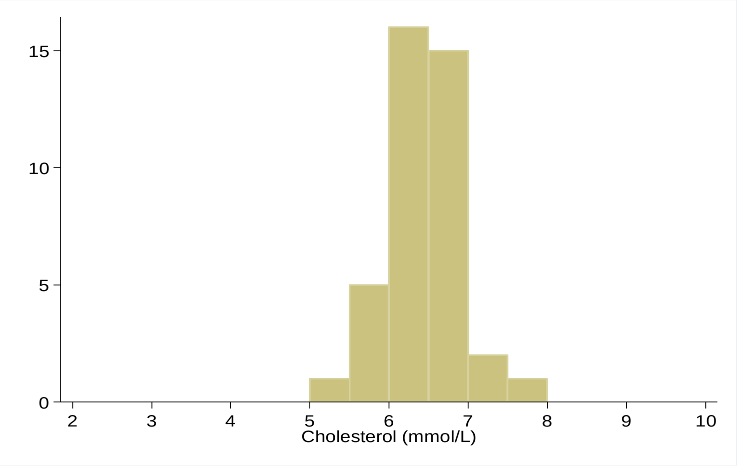
Example1: Cholesterol levels for 40 men aged 45 years



Standard deviation = 1.57 mmol/L

The cholesterol values differ from the mean of 6.4mmol/L by, on average, 1.57 mmol/L

Here, you can see data with same mean, but much less variability

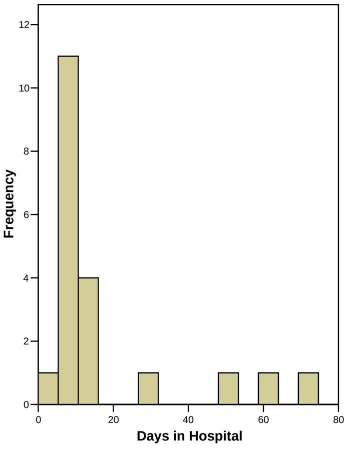


Mean = 6.4 mmol/L

Standard deviation = 0.48 mmol/L

Standard deviation for skewed data

Example 2: Length of stay in hospital for 20 patients

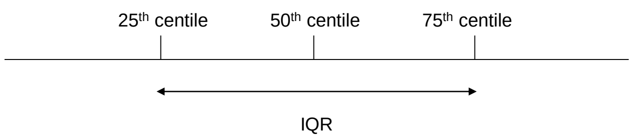


Standard deviation = 19.55 days
(which doesn't look right from the histogram – too big)

2. Inter-Quartile Range

The Inter-Quartile Range is the difference between the 25th and 75th centiles:

- Lower quartile (25th centile):
  the value below which 25% of the data lie
- Upper quartile (75th centile):
  the value below which 75% of the data lie



Example: Cholesterol data

3.6	3.8	3.9	4.1	4.2	4.5	4.5	4.8
5.1	5.3	5.4	5.4	5.6	5.8	5.9	6.0
6.1	6.1	6.2	6.3	6.4	6.5	6.6	6.8
6.9	7.1	7.2	7.2	7.3	7.4	7.5	7.7
8.0	8.1	8.1	8.2	8.3	9.0	9.1	10.0

Lower quartile has 10 data points below it (between the 10th and 11th data points) = 5.35

Upper quartile has 30 data points below it (between the 30th and 31st data points) = 7.45

Inter-Quartile Range = 7.45 – 5.35 = 2.1 mmol/L
(compares with SD of 1.57 mmol/L)

However, you can also quote (5.35 to 7.45 mmol/L), as this provides more information

Example 2: Hospital data

Days in Hospital	Frequency	Cumulative Percent
5	1	5
6	1	10
7	2	20
8	3	35
9	4	55
10	1	60
12	2	70
14	2	80
28	1	85
52	1	90
61	1	95
71	1	100
Total	20	

IQR?

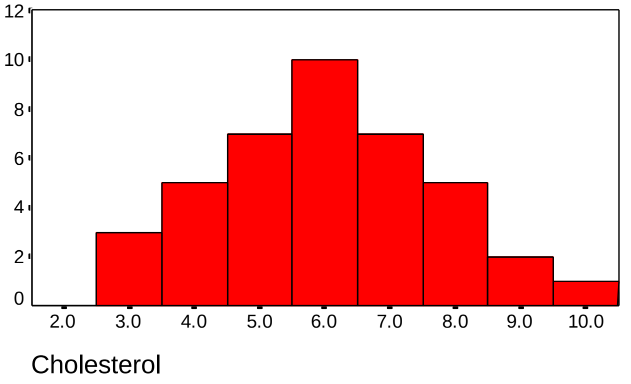
25^{th} centile = 8

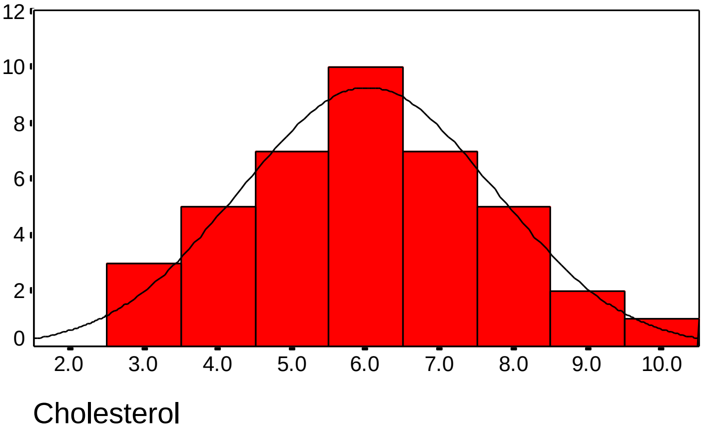
75^{th} centile = 14

IQR = 14-8 =6 days

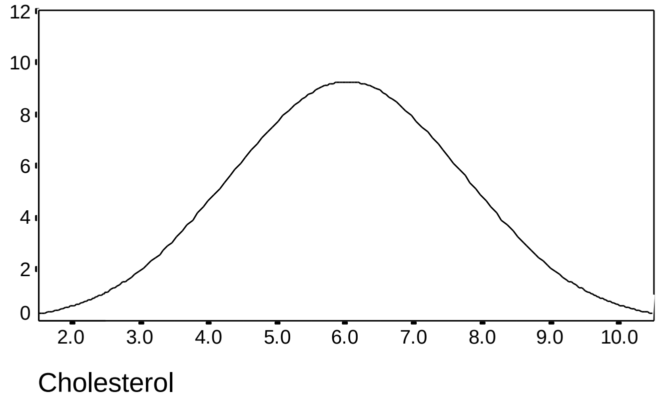
(SD = 19.6 days)

	Best measure of a 'typical' value, i.e. centre of the data	Best measure of spread
Symmetric data	Mean	Standard deviation
Non-symmetric (skewed) data	Median	Interquartile range (25-75^{th} centile values)


12
10
8
6
4
2
0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
Cholesterol


12
10
8
6
4
2
0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
Cholesterol

Gaussian (or Normal) distribution



Normal ranges

We can use the Gaussian distribution to get a 'normal' range, ie where we expect most people to have values (often used in clinical practice, to see whether someone has a 'normal' or 'abnormally' high or low measurement of something (eg blood value)

95% central range = mean ± 1.96 x standard deviation

= 6.4 ± 1.96 x 1.57

= 3.3 to 9.5 mmol/L

Therefore, 95% of men aged 45 years are expected to have a cholesterol value between 3.3 and 9.5

This is not a confidence interval.

A central range deals with people (or things), a 95% CI deals with a summary measure (eg percentage or mean value)

- Most of the time, we wish to describe the distribution of a particular measurement in the whole population of interest, not just the sample in our study

- E.g. the mean cholesterol level is 6.4 in the 40 men, but can we be sure that this is the mean level in **all** men aged 45 years?

- What about natural variation?

- **What are the implications of conducting a study on a sample of people?**

Standard error of a mean

A measure of the precision of the observed mean

Standard error of the mean (SE) = $\frac{s}{\sqrt{n}}$
(s= standard deviation of sample)

Example: sample of 40 cholesterol values

Mean cholesterol = 6.4 mmol/L

SD cholesterol = 1.57 mmol/L

SE (mean) = $\frac{1.57}{\sqrt{40}} = 0.248$

- The standard error gives us an idea of how far our observed mean value (in the sample of 40 men) could be from the true mean

- As the sample size gets bigger, we should be getting closer to the true mean

- Therefore, the standard error should get smaller

Calculating the confidence interval (CI) for a mean

Lower limit of CI = observed mean – (1.96 x standard error of mean)

Upper limit of CI = observed mean + (1.96 x standard error of mean)

Example: 40 cholesterol values for men aged 45

Mean = 6.4 Standard error = 0.248

Lower limit = 6.4 - (1.96 x 0.248) = 6.4 – 0.486 = 5.914

Upper limit = 6.4 + (1.96 x 0.248) = 6.4 + 0.486 = 6.886

95% confidence interval is 5.9 to 6.9 mmol/L.

1.96 is used when there are about 30 or more observations; for smaller samples the multiplier used is larger and will depend on the sample size (the stats package will do this automatically)

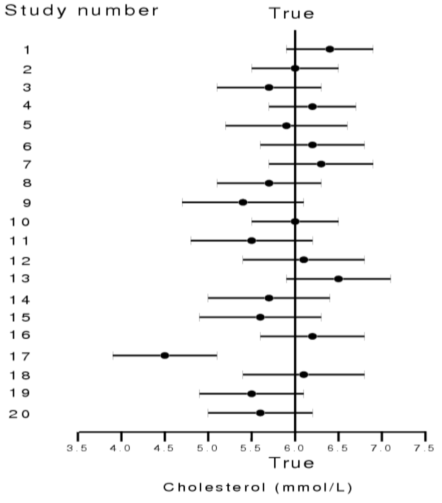
Interpretation

All men, aged 45 years
Sample of 40 men in study

True mean= ??
Observed mean=6.4

- Using the data from our study of 40 men, we think the true mean cholesterol is 6.4 mmol/L
- However, we cannot be completely sure of this
- But we are 95% certain, that the true mean is expected to lie somewhere between 5.9 and 6.9 mmol/L

20 different studies, each with 40 men
Each study is trying to estimate the true mean



- Because we specify a 95% confidence interval, it indicates that we are expected to get the wrong answer 5% of the time
- There may be nothing wrong with how the study was conducted
- It could just mean that we were unlucky to have a sample of men that had very different cholesterol levels (only because of natural variability)

STANDARD ERROR & STANDARD DEVIATION

Can be easy to confuse

Standard **D**eviation measures how far the **D**ata spreads out from the mean value
(it just describes how much the measurement varies between people/things)

Standard **E**rror measures the precision of our **E**stimate
(it is used when we are making inferences about the true mean value in a population)