
**Dr Silia Vitoratou**

Department: Biostatistics and Health Informatics



**Topic materials:**
Silia Vitoratou

**Contributions:**
Zahra Abdula

**Improvements:**
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau



**Institute of Psychiatry, Psychology and Neuroscience**


KING'S College LONDON

**Module Title:** Introduction to Statistics

**Session Title:** Summarising numerical data

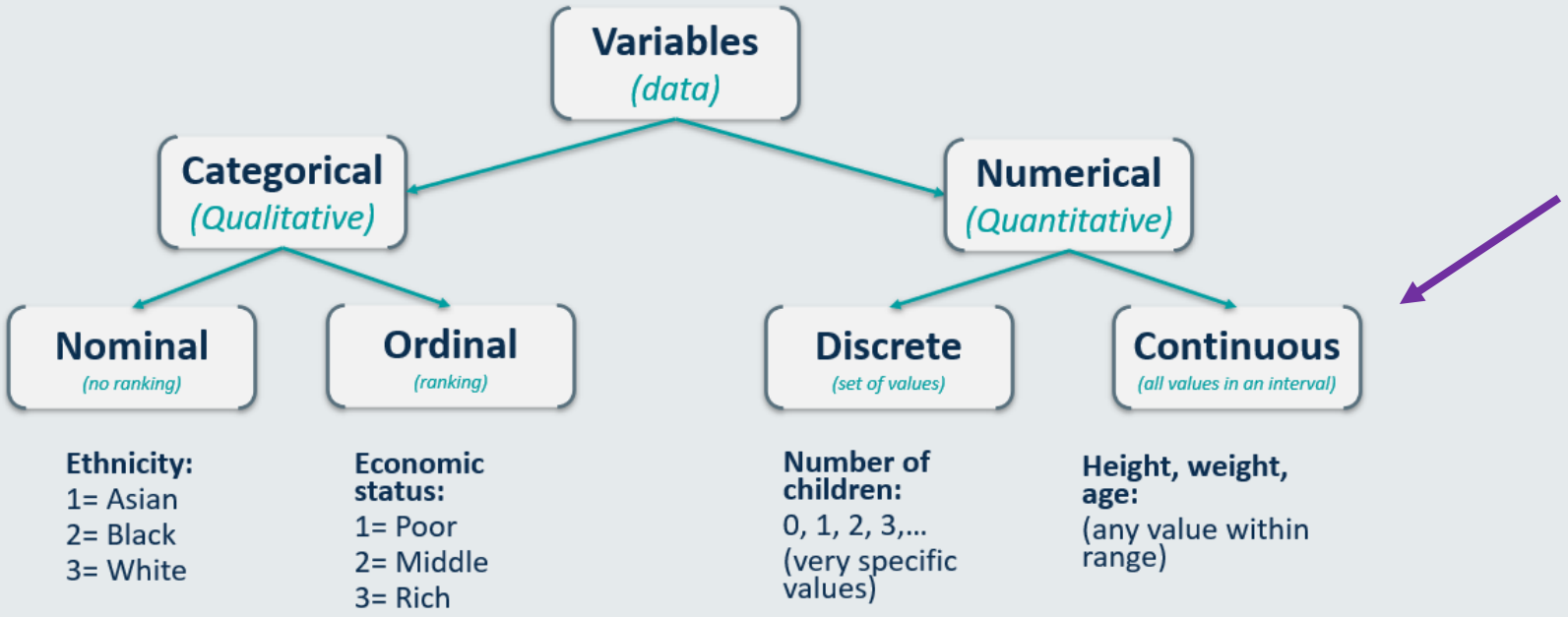
# Topic title: Measurement and graphical representations of data

# Learning Outcomes

- To understand the descriptive indices suitable for numerical data
- To understand the descriptive charts suitable for numerical data
- To be able to use SPSS to create descriptive indices and charts

# Types of Variables

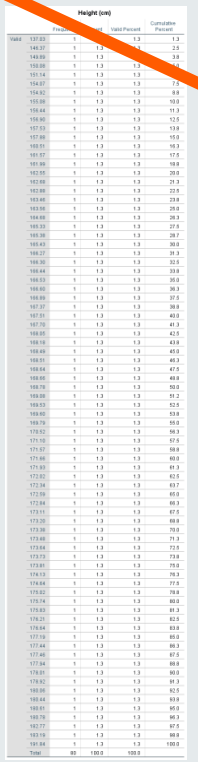
Based on the type of each variable, we use different ways to describe the data.

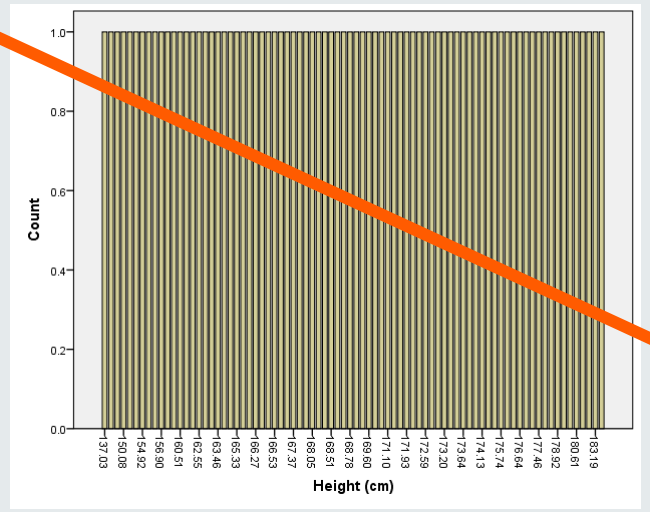


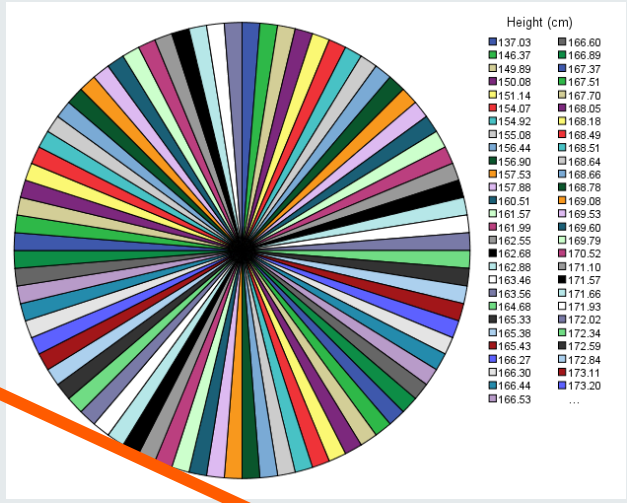

- Descriptive indices — Frequencies (Percentages %) — ?
- Charts/plots — Bar Chart — ?

# Types of Data

In numerical data, one would NOT be interested in how many people are in each category (here, value). For instance, let us see the frequencies, the bar and the pie chart for height:



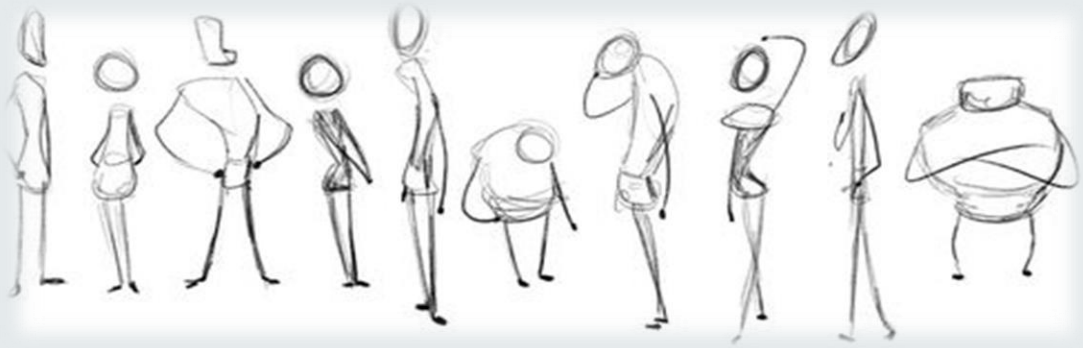



To describe a numerical variable, we need to properly summarise it properly.

# Quantitative (Numerical) Data

Let us start with the mean as a summary measure. Let us imagine that there are ten people in a room, with different ages.



$$\frac{37+41+18+21+17+86+31+33+21+55}{10}$$

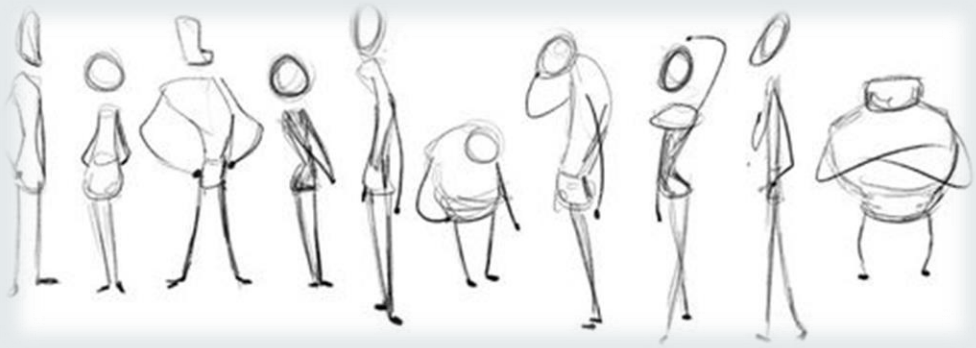
**= 36 years old on average**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean age (value) $\bar{x}$
is the sum **Σ** of the ages from the first person (i=1)
to the last person (**n-th)**,
divided by the number of people in the room **n**

# Quantitative (Numerical) Data

Is the mean enough for us to describe the data?



**37 41 18 21 17 86 31 33 21 55** **mean= 36 years**

Consider another set of values

**34 32 32 31 30 60 35 33 36 37** **mean= 36 years**

Even though the two sets of values have the same mean, it is clear that the values in the second are much closer to the mean (36yo).

# Quantitative (Numerical) Data

To understand how far from the mean value the values are we need to calculate a measure called **Variance**.

$\bar{x}$=mean= 36

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **37** | **41** | **18** | **21** | **17** | **86** | **31** | **33** | **21** | **55** |
| **+1** | **+5** | **-18** | **-15** | **-19** | **+50** | **-5** | **-3** | **-15** | **+19** |
| **1** | **25** | **324** | **225** | **361** | **2500** | **25** | **9** | **225** | **361** |

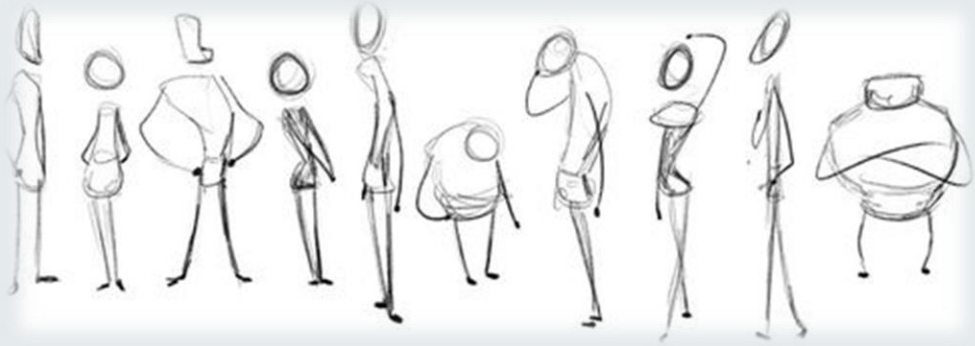
**Observations ($x_i$)**

**Distance ($x_i - \bar{x}$)**

**Squared Distances ($x_i - \bar{x})^2$**

The mean (squared) distance: $$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$ **Average Squared Distance**

# Quantitative (Numerical) Data



The mean is the average of the values…

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The variance measures the average of the values' distance from the mean…

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

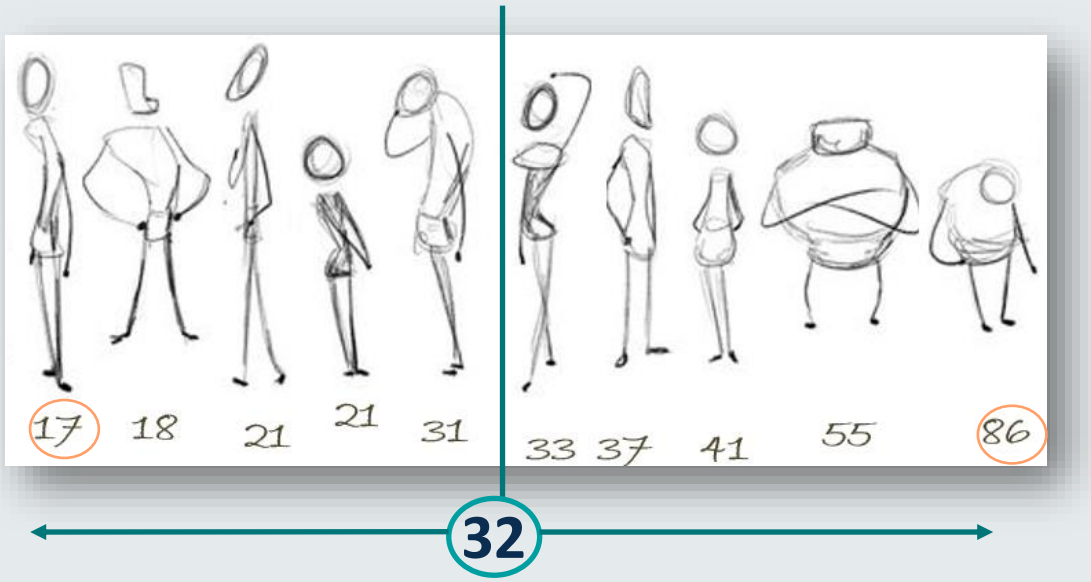
The standard deviation (SD) is how spread out a group of numbers is from the mean, by looking at the square root of the variance…

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

*Note: if this was a sample from a population then instead of dividing with n, we would divide by n-1, to obtain an 'unbiased' estimate for the population variance. We do not go into details about biased ad unbiased estimates in this module.*

# Quantitative (Numerical) Data

The mean and the SD are not the only summary measures. Let us put the values in ascending order.




Median:

- for an even number of values, it is the average of the two middle values
- for an odd number of values, it is simply the middle value (after ordering)

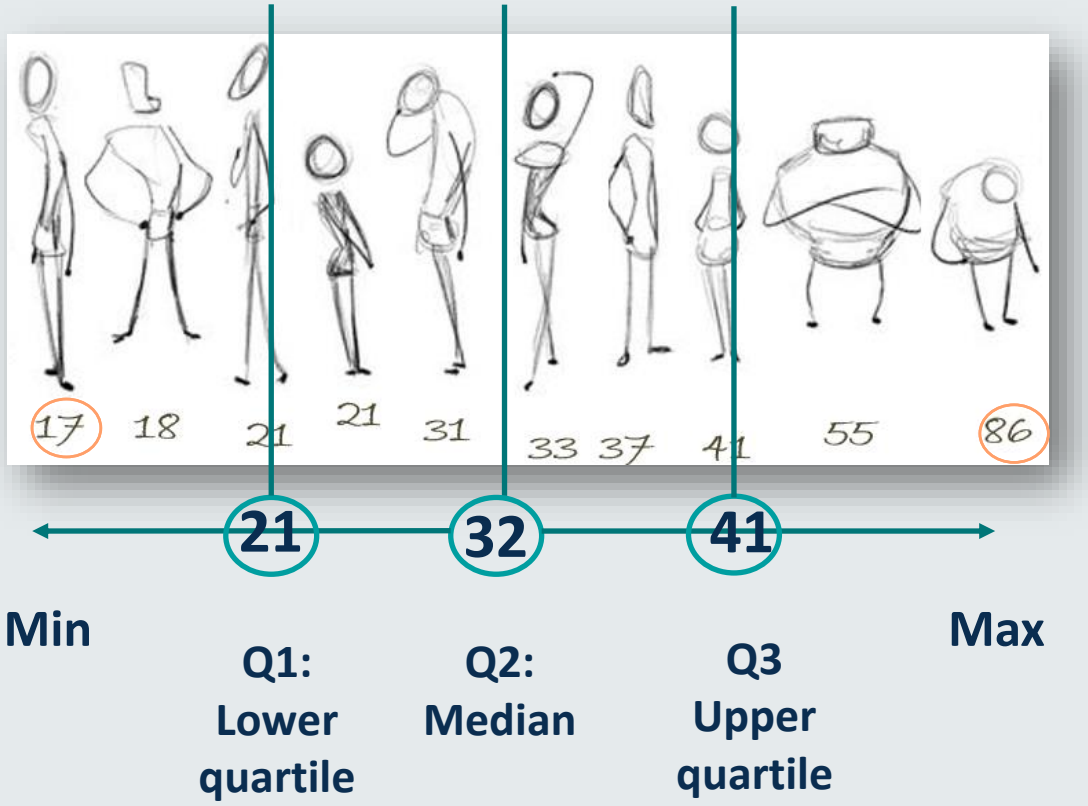
Median = 32

Minimum = 17, Maximum = 86

Range = 86 – 17= 69

# Quantitative (Numerical) Data

Other measures that are useful to describe numerical data are called **Quartiles.**




Lower quartile = 21

Median= 32

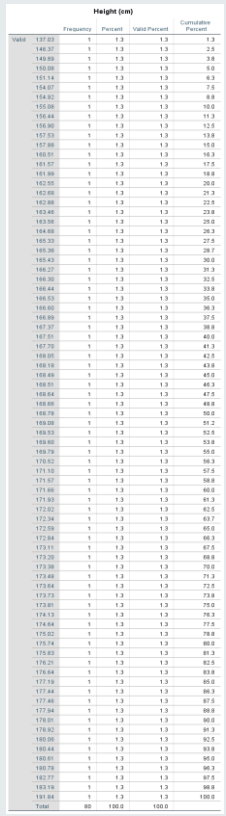
Upper quartile = 41

Interquartile Range = 41 – 21
= 20

# Quantitative (Numerical) Data

To describe the metrical or numerical variable, we need to properly summarise it.

**Instead of reporting this**



Height (cm)

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Valid 137.03 | 1 | 1.3 | 1.3 | 1.3 |
| 146.37 | 1 | 1.3 | 1.3 | 2.5 |
| 149.89 | 1 | 1.3 | 1.3 | 3.8 |
| 150.09 | 1 | 1.3 | 1.3 | 5.0 |
| 151.14 | 1 | 1.3 | 1.3 | 6.3 |
| 154.07 | 1 | 1.3 | 1.3 | 7.5 |
| 154.92 | 1 | 1.3 | 1.3 | 8.8 |
| 155.09 | 1 | 1.3 | 1.3 | 10.0 |
| 156.44 | 1 | 1.3 | 1.3 | 11.3 |
| 156.90 | 1 | 1.3 | 1.3 | 12.5 |
| 157.53 | 1 | 1.3 | 1.3 | 13.8 |
| 157.88 | 1 | 1.3 | 1.3 | 15.0 |
| 160.51 | 1 | 1.3 | 1.3 | 16.3 |
| 161.57 | 1 | 1.3 | 1.3 | 17.5 |
| 161.99 | 1 | 1.3 | 1.3 | 18.8 |
| 162.55 | 1 | 1.3 | 1.3 | 20.0 |
| 162.68 | 1 | 1.3 | 1.3 | 21.3 |
| 163.08 | 1 | 1.3 | 1.3 | 22.5 |
| 163.46 | 1 | 1.3 | 1.3 | 23.8 |
| 163.56 | 1 | 1.3 | 1.3 | 25.0 |
| 164.68 | 1 | 1.3 | 1.3 | 26.3 |
| 165.33 | 1 | 1.3 | 1.3 | 27.5 |
| 165.38 | 1 | 1.3 | 1.3 | 28.7 |
| 165.43 | 1 | 1.3 | 1.3 | 30.0 |
| 166.27 | 1 | 1.3 | 1.3 | 31.3 |
| 166.30 | 1 | 1.3 | 1.3 | 32.5 |
| 166.44 | 1 | 1.3 | 1.3 | 33.8 |
| 166.53 | 1 | 1.3 | 1.3 | 35.0 |
| 166.60 | 1 | 1.3 | 1.3 | 36.3 |
| 166.89 | 1 | 1.3 | 1.3 | 37.5 |
| 167.37 | 1 | 1.3 | 1.3 | 38.8 |
| 167.51 | 1 | 1.3 | 1.3 | 40.0 |
| 167.70 | 1 | 1.3 | 1.3 | 41.3 |
| 168.05 | 1 | 1.3 | 1.3 | 42.5 |
| 168.10 | 1 | 1.3 | 1.3 | 43.8 |
| 168.49 | 1 | 1.3 | 1.3 | 45.0 |
| 168.51 | 1 | 1.3 | 1.3 | 46.3 |
| 168.64 | 1 | 1.3 | 1.3 | 47.5 |
| 168.86 | 1 | 1.3 | 1.3 | 48.8 |
| 168.78 | 1 | 1.3 | 1.3 | 50.0 |
| 169.08 | 1 | 1.3 | 1.3 | 51.2 |
| 169.53 | 1 | 1.3 | 1.3 | 52.5 |
| 169.60 | 1 | 1.3 | 1.3 | 53.8 |
| 169.79 | 1 | 1.3 | 1.3 | 55.0 |
| 170.52 | 1 | 1.3 | 1.3 | 56.3 |
| 171.10 | 1 | 1.3 | 1.3 | 57.5 |
| 171.57 | 1 | 1.3 | 1.3 | 58.8 |
| 171.66 | 1 | 1.3 | 1.3 | 60.0 |
| 171.93 | 1 | 1.3 | 1.3 | 61.3 |
| 172.02 | 1 | 1.3 | 1.3 | 62.5 |
| 172.34 | 1 | 1.3 | 1.3 | 63.7 |
| 172.59 | 1 | 1.3 | 1.3 | 65.0 |
| 172.84 | 1 | 1.3 | 1.3 | 66.3 |
| 173.11 | 1 | 1.3 | 1.3 | 67.5 |
| 173.20 | 1 | 1.3 | 1.3 | 68.8 |
| 173.39 | 1 | 1.3 | 1.3 | 70.0 |
| 173.46 | 1 | 1.3 | 1.3 | 71.3 |
| 173.64 | 1 | 1.3 | 1.3 | 72.5 |
| 173.73 | 1 | 1.3 | 1.3 | 73.8 |
| 173.91 | 1 | 1.3 | 1.3 | 75.0 |
| 174.13 | 1 | 1.3 | 1.3 | 76.3 |
| 174.64 | 1 | 1.3 | 1.3 | 77.5 |
| 175.82 | 1 | 1.3 | 1.3 | 78.8 |
| 175.74 | 1 | 1.3 | 1.3 | 80.0 |
| 175.83 | 1 | 1.3 | 1.3 | 81.3 |
| 176.21 | 1 | 1.3 | 1.3 | 82.5 |
| 176.64 | 1 | 1.3 | 1.3 | 83.8 |
| 177.19 | 1 | 1.3 | 1.3 | 85.0 |
| 177.44 | 1 | 1.3 | 1.3 | 86.3 |
| 177.46 | 1 | 1.3 | 1.3 | 87.5 |
| 177.94 | 1 | 1.3 | 1.3 | 88.8 |
| 178.01 | 1 | 1.3 | 1.3 | 90.0 |
| 178.52 | 1 | 1.3 | 1.3 | 91.3 |
| 180.06 | 1 | 1.3 | 1.3 | 92.5 |
| 180.44 | 1 | 1.3 | 1.3 | 93.8 |
| 180.61 | 1 | 1.3 | 1.3 | 95.0 |
| 180.78 | 1 | 1.3 | 1.3 | 96.3 |
| 182.77 | 1 | 1.3 | 1.3 | 97.5 |
| 183.19 | 1 | 1.3 | 1.3 | 98.8 |
| 191.84 | 1 | 1.3 | 1.3 | 100.0 |
| Total | 80 | 100.0 | 100.0 | |

**We understand more by reporting on:**

## Measures of location (central tendency)

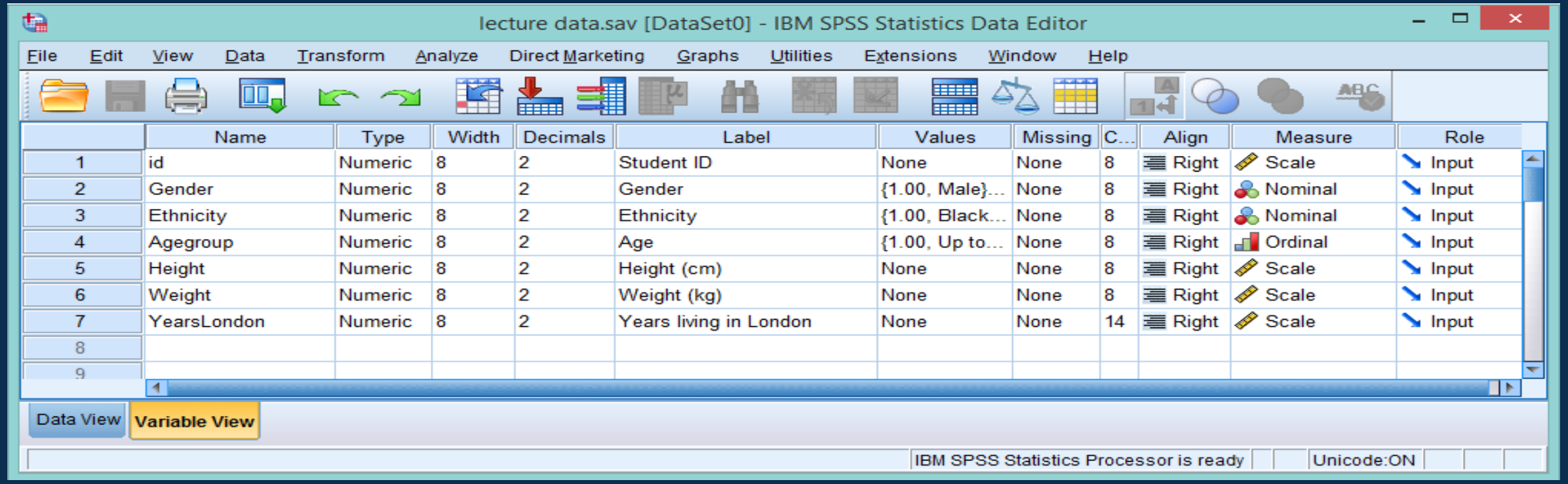
- Mean value: the **average** height of the students was 168.5cm (5.5ft)
- Median value: half of the students **were taller** than 169cm (5.5ft)
- Mode value: the height **most often** reported was 173cm (5.7ft)

## Measures of dispersion (spread, variability)

- Standard deviation: SD was 9cm (0.3ft): the heights were on average 9cm away from the mean height of 168.5cm
- Min and max values: **min** height =137cm (4.5ft), **max** height=192cm (6.3ft)
- Range or IQR: The difference between the tallest and the shortest student was 10 cm (0.3ft)

# SPSS Slide

To illustrate the how we can describe the different types of data we are going to use the below SPSS dataset "**lecture_1_data.sav".** Download the dataset to follow along




| | Name | Type | Width | Decimals | Label | Values | Missing | C... | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | Numeric | 8 | 2 | Student ID | None | None | 8 | Right | Scale | Input |
| 2 | Gender | Numeric | 8 | 2 | Gender | {1.00, Male}... | None | 8 | Right | Nominal | Input |
| 3 | Ethnicity | Numeric | 8 | 2 | Ethnicity | {1.00, Black... | None | 8 | Right | Nominal | Input |
| 4 | Agegroup | Numeric | 8 | 2 | Age | {1.00, Up to... | None | 8 | Right | Ordinal | Input |
| 5 | Height | Numeric | 8 | 2 | Height (cm) | None | None | 8 | Right | Scale | Input |
| 6 | Weight | Numeric | 8 | 2 | Weight (kg) | None | None | 8 | Right | Scale | Input |
| 7 | YearsLondon | Numeric | 8 | 2 | Years living in London | None | None | 14 | Right | Scale | Input |
| 8 | | | | | | | | | | | |
| 9 | | | | | | | | | | | |

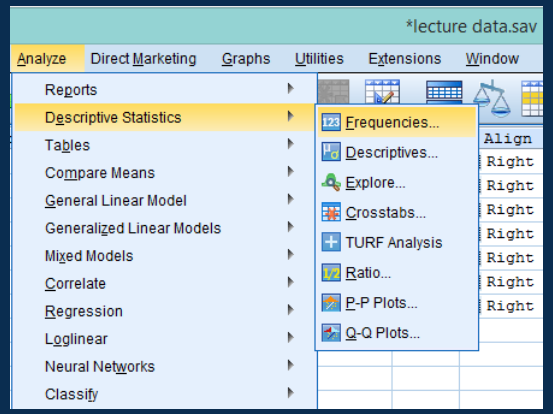
# SPSS Slide: ‘How to’ Steps

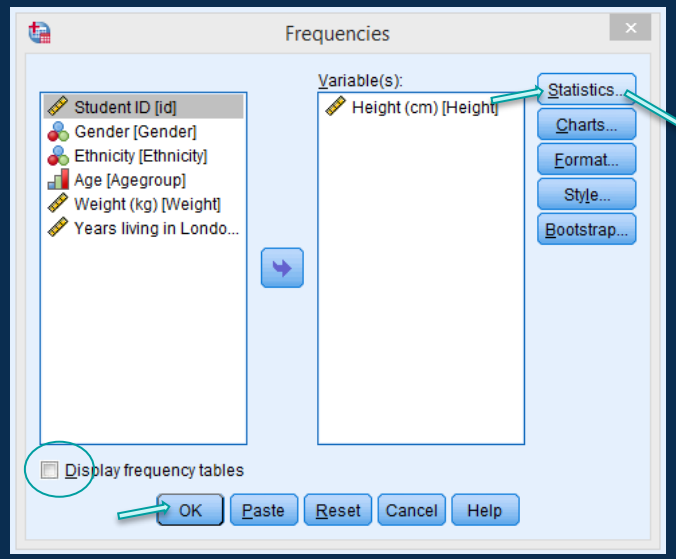
You can create the descriptive indices for height using the following steps:

Click on the ‘Analyse Tab’→ ‘Descriptive Statistics’ → ‘Frequencies’

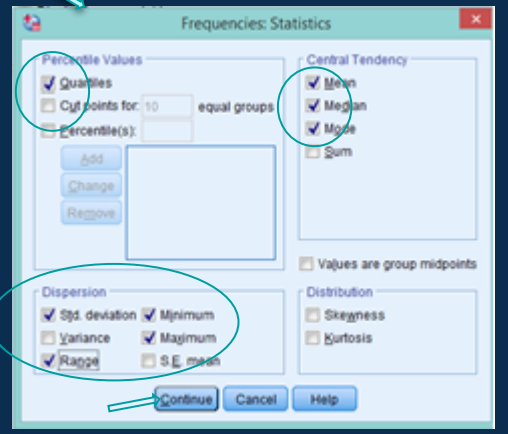
Add the variable of interest (height) into the ‘Variable(s)’ box

Make sure the ‘Display frequency tables’ box is unchecked




Click on ‘Statistics’ and choose the indices you want to report.




Instead of frequencies we now want measures of central tendency (location) and measures of dispersion (spread).

Click on ‘Continue’

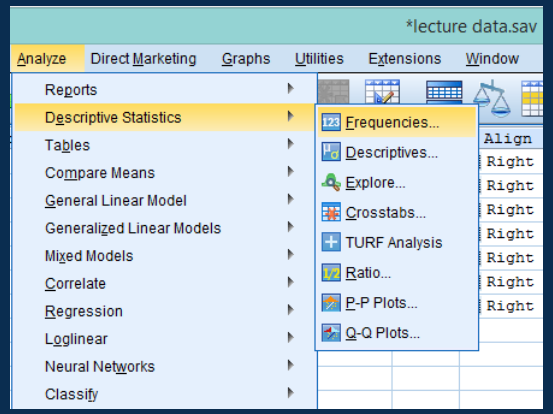
Click on ‘OK

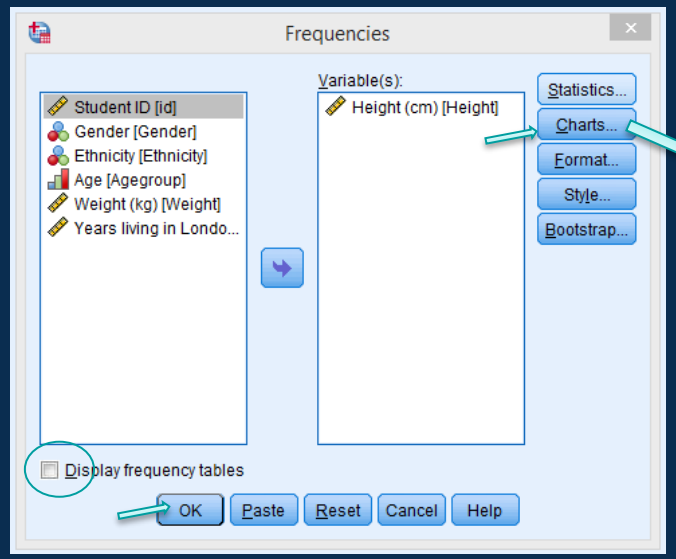
# SPSS Slide: ‘How to’ Steps

You can create a chart using the following steps:

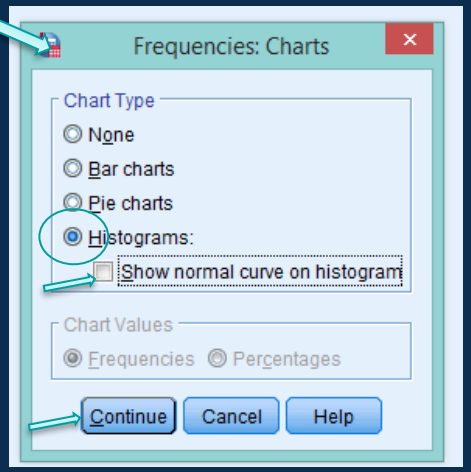
Click on the ‘Analyse Tab’→ ‘Descriptive Statistics’ → ‘Frequencies’

Add the variable of interest (height) into the ‘Variable(s)’ box
Make sure the ‘Display frequency tables’ box is unchecked




Click on ‘charts’ and choose the chart you want to report.




For the numerical variable height we would prefer the histogram

Tick ‘show the normal curve’

Click on ‘Continue’

Click on ‘OK

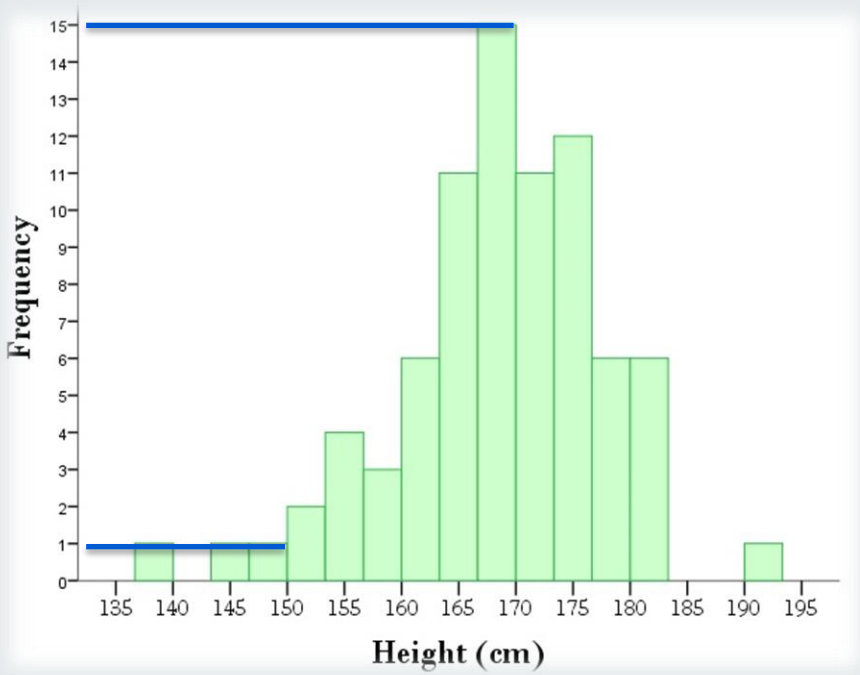
# Describing Quantitative (Numerical) Data using Charts

Descriptive indices depicted on the ***histogram:***

Bins represent intervals, not values (categories) as in the case of bar chart

1 person had height between 146 and 147cm

15 people had height between 166 and 170cm



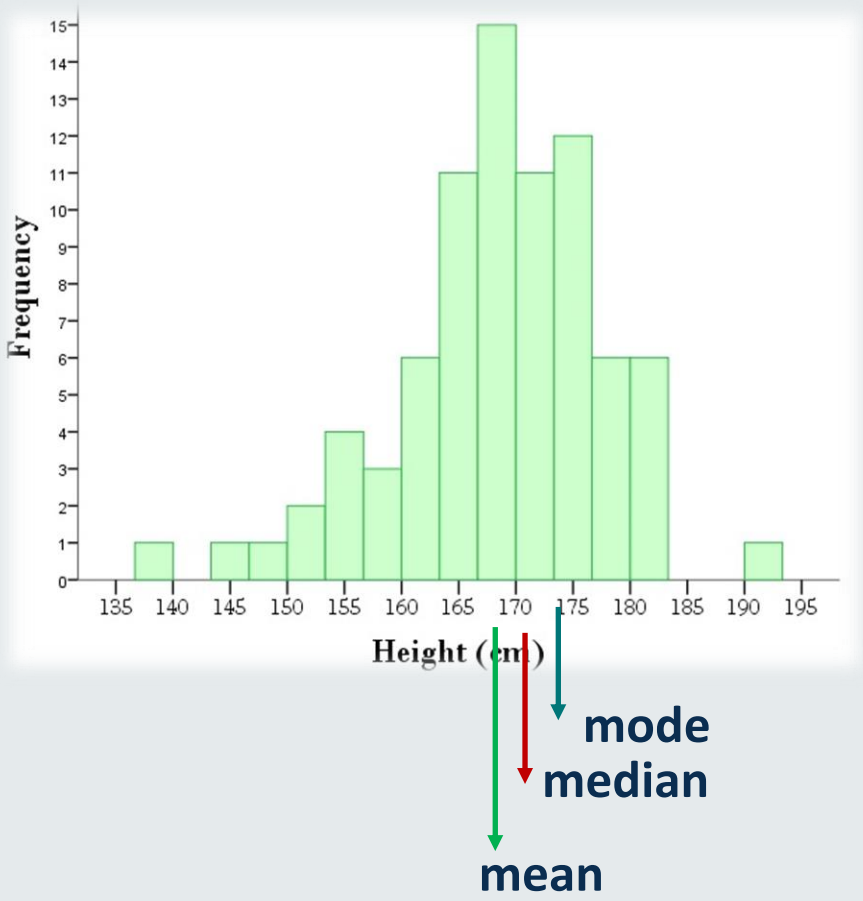
# Describing Quantitative (Numerical) Data using Charts

## Measures of location (central tendency): they show where about most of the values are

**Statistics**

Height (cm)

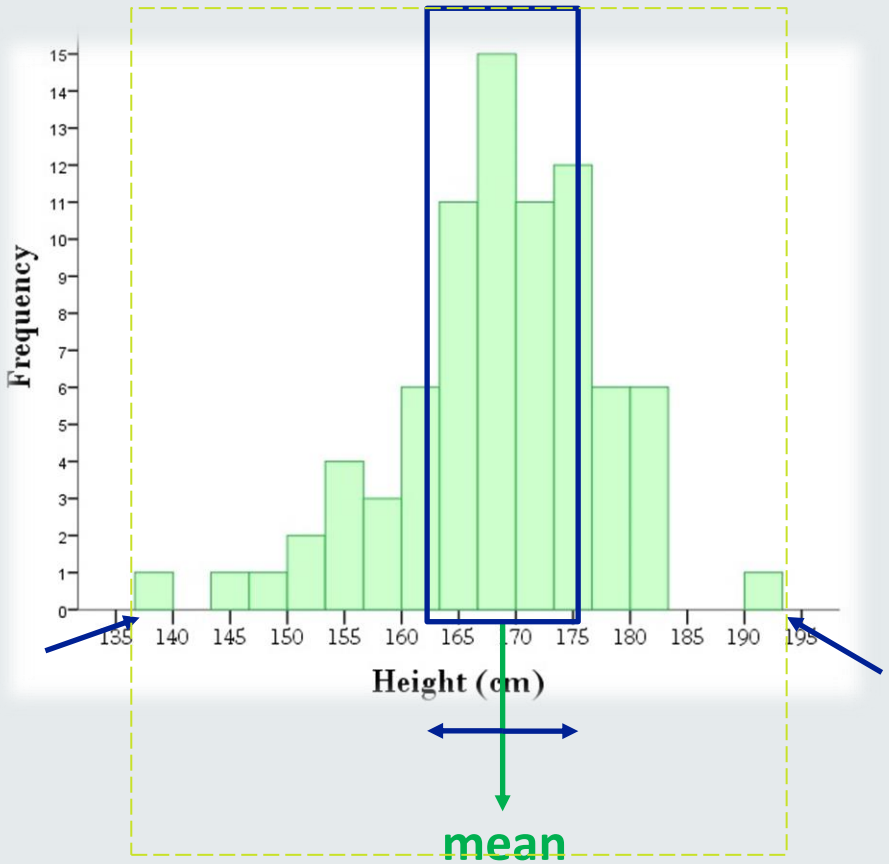
| | | |
|---|---|---|
| N | Valid | 80 |
| | Missing | 0 |
| Mean | | 168.5750 |
| Median | | 169.0000 |
| Mode | | 173.00 |
| Std. Deviation | | 9.16760 |
| Range | | 55.00 |
| Minimum | | 137.00 |
| Maximum | | 192.00 |



# Describing Quantitative (Numerical) Data using Charts

## Measures of dispersion (spread): they show how variable the values are

Each person obviously has a value within the interval [min, max]




**Statistics**

Height (cm)

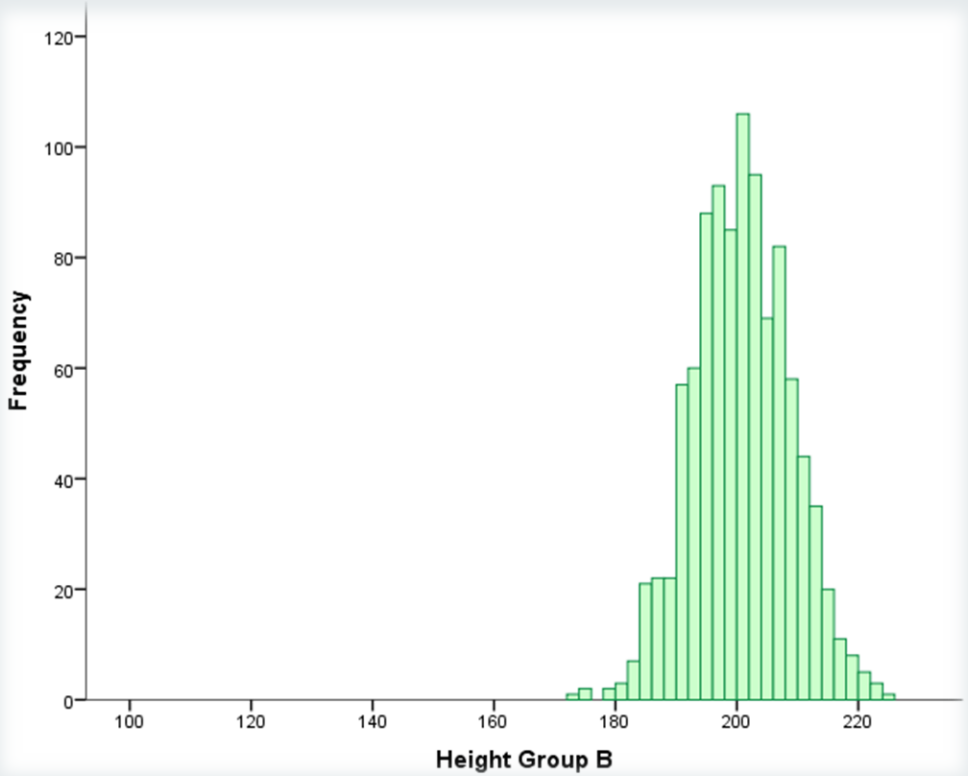
| | | |
|---|---|---|
| N | Valid | 80 |
| | Missing | 0 |
| Mean | | 168.5750 |
| Median | | 169.0000 |
| Mode | | 173.00 |
| Std. Deviation | | 9.16760 |
| Range | | 55.00 |
| Minimum | | 137.00 |
| Maximum | | 192.00 |

People with values within the interval [mean-sd, mean+sd].

# Describing Quantitative (Numerical) Data using Charts

How things change when the measures of **location** change?

**Location**

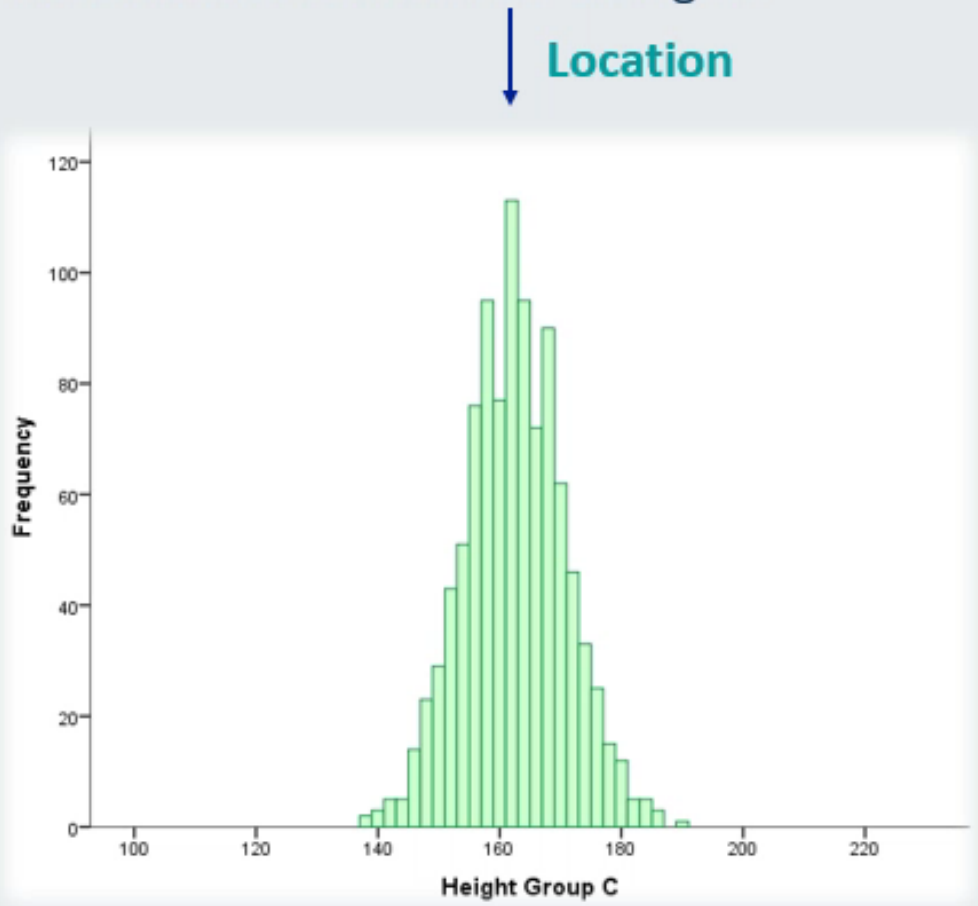



**Mean = 200cm**
SD = 8cm

**dispersion**

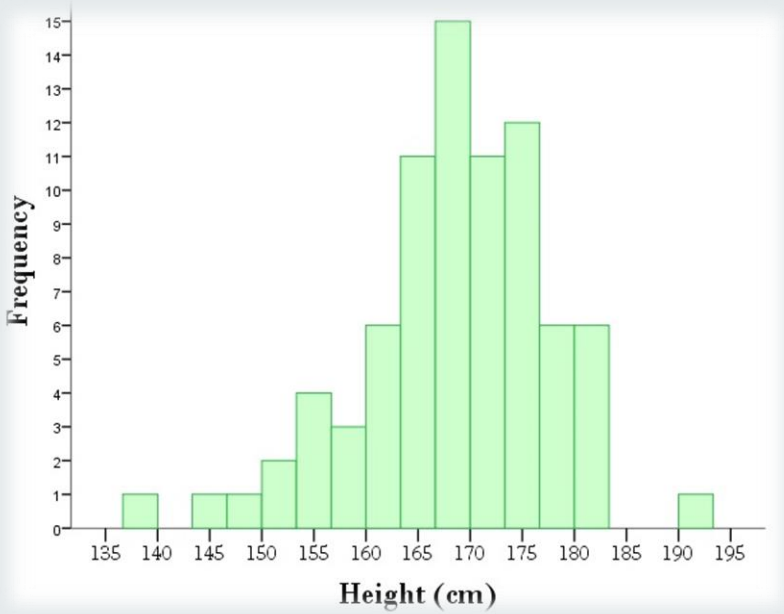
# Describing Quantitative (Numerical) Data using Charts

How things change when the measures of **location** change?




**Mean = 162cm**
SD = 8cm

# Output and Interpretation




**Statistics**

Height (cm)

| | | |
|---|---|---|
| N | Valid | 80 |
| | Missing | 0 |
| Mean | | 168.5253 |
| Median | | 168.9280 |
| Mode | | 137.03[a] |
| Std. Deviation | | 9.15218 |
| Range | | 54.81 |
| Minimum | | 137.03 |
| Maximum | | 191.84 |

a. Multiple modes exist. The smallest value is shown

The height of the individuals in our sample varied between 137.03cm and 191.84 cm.

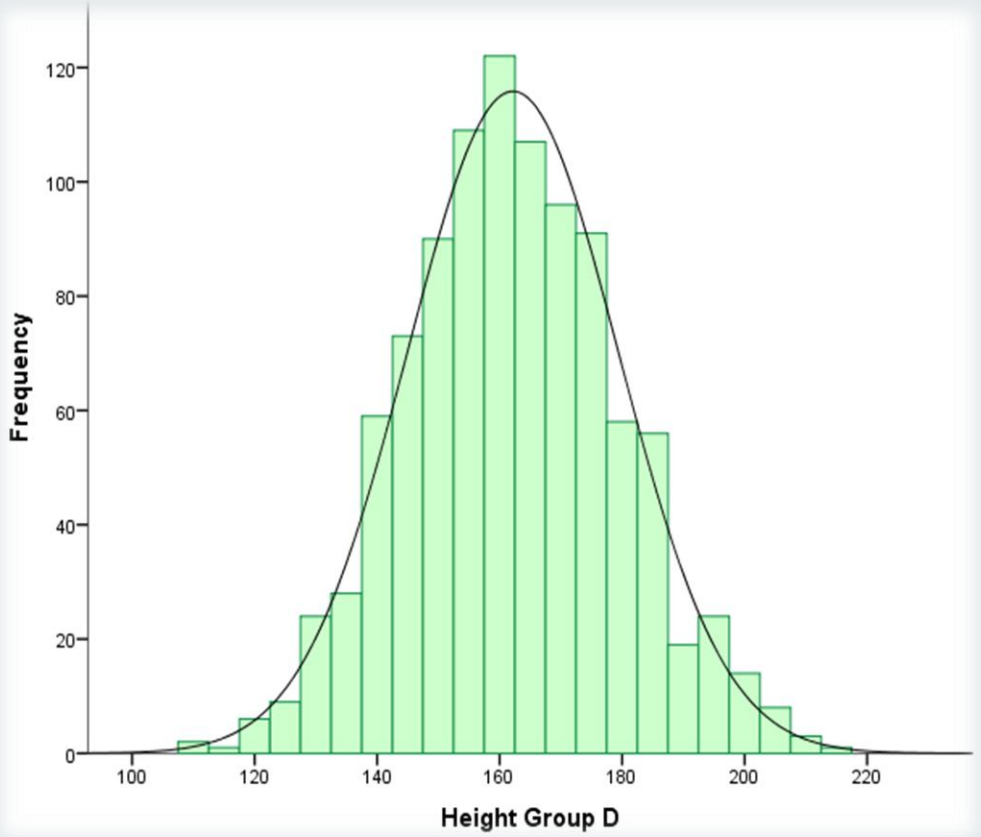
The average height was 168.53cm (SD=9.15cm).

Half of the people were taller than 168.93cm, while the height most often reported was 137.03cm.

The difference in the height between the shortest and the tallest person was 54.81cm.

# The Normal Curve

Usually, when we present the histogram, we also add the ***normal distribution*** curve




Mean = 162cm
**SD = 17cm**

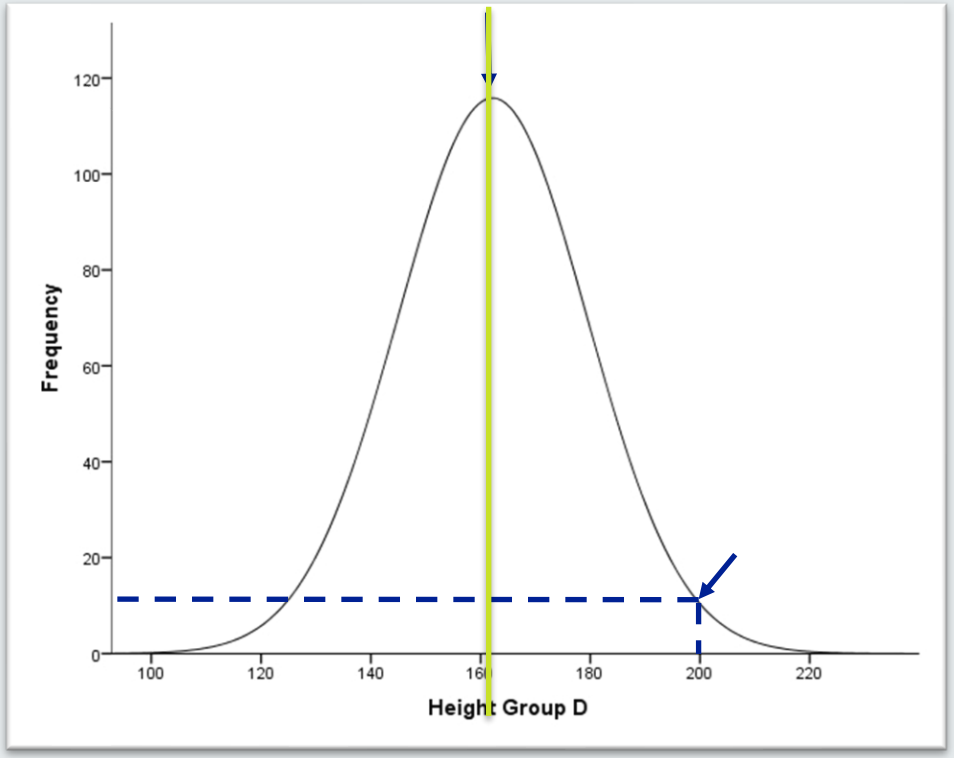
That is, the curve of a normal distribution with the same mean and standard deviation as our data…

# The Normal Curve

The **normal** distribution is a distribution which looks like a bell and where the data are **symmetrical** around the mean.

Mean = 162
Sd = 28

Mean (and median and mode)




The normal distribution looks like a bell and:

- half of the people (median) have values lower than the average (and half higher than the **average**)
- the most common value (mode) is the average
- the **majority of the people** are close to the average
- as we move away from the average, we have **fewer** observations.

We will study the normal distribution in detail in Topic 2.

# Which Statistical Measure to Use

Let us see another example

**Annual income (£1,000)**

**18 24 21 20 26 23 17 18 16 20 25 23 17**

Mean = £21K,
SD = £3K

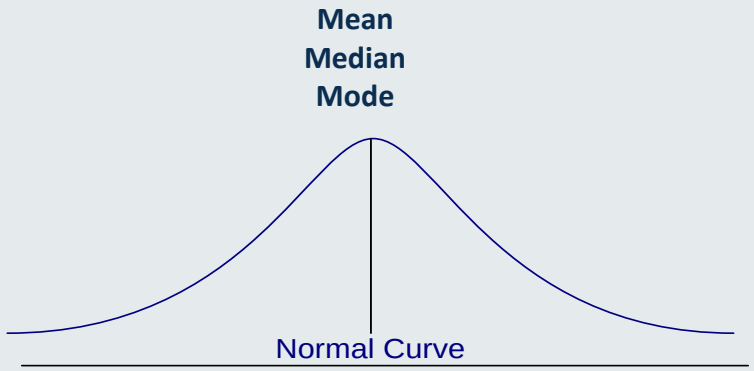
**18 294 21 285 26 23 17 18 16 20 25 23 17**

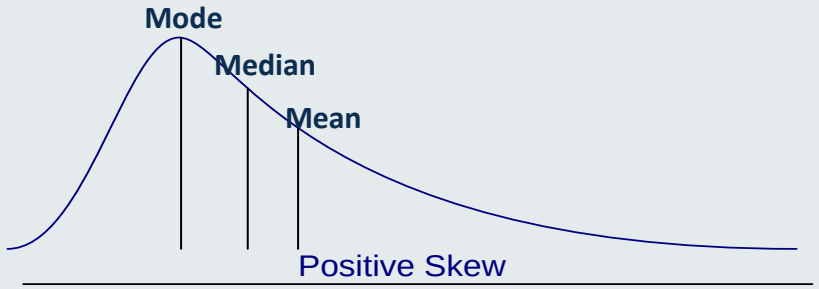
Mean £62K, SD = £101K
Median = £21K
min = £16K / max =294K

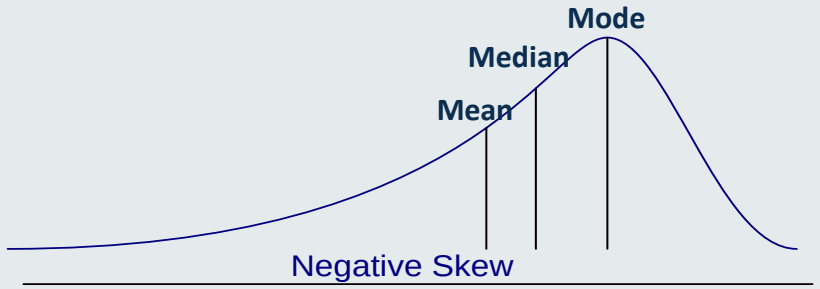
# Describing Quantitative (Numerical) Data using Charts

Is our data **Normal** (symmetrical about the mean) or **Skewed** (non symmetrical data)?

If the data are symmetrical, typically report on: **mean and sd**



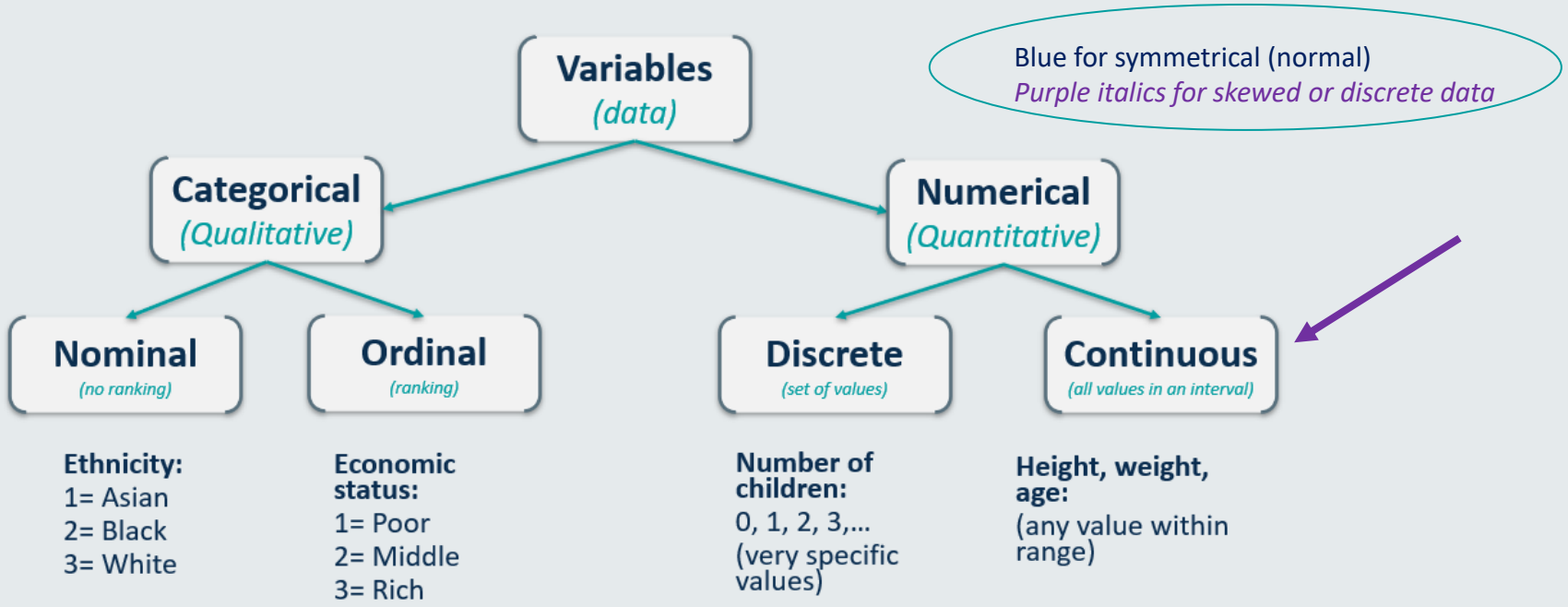




If the data are skewed, typically report on: **median and min-max and IQR**

# Types of Variables

Based on the type of each variable, we use different ways to describe the data.




- Descriptive indices — Frequencies (Percentages %)
  Location: mean, *median*, mode
  Dispersion: SD, *range*, IQR
- Charts/plots — Pie Chart (only for nominal)
  Bar Chart

# Describing Quantitative (Numerical) Data using Charts

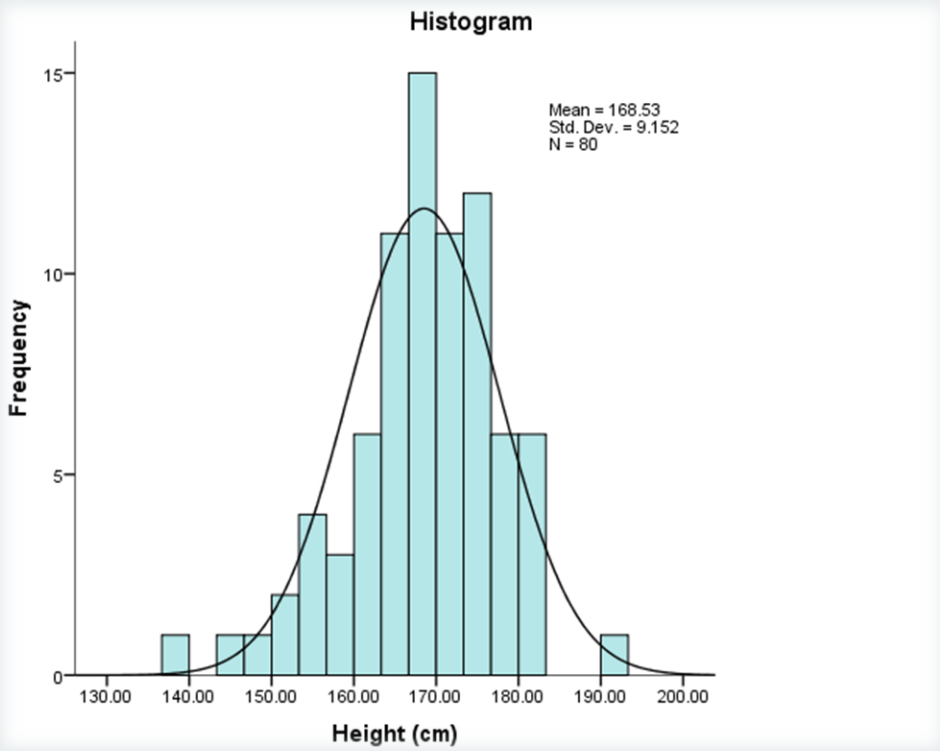
A chart has all the information we need and is easier to understand

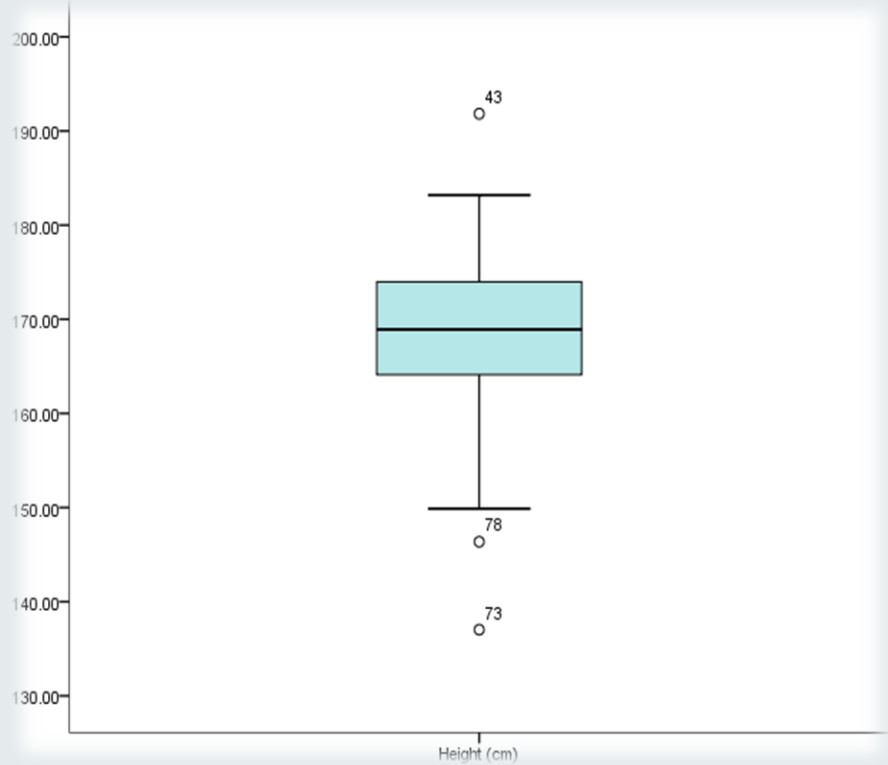
**Statistics**

Height (cm)

| | | |
|---|---|---|
| N | Valid | 80 |
| | Missing | 0 |
| Mean | | 168.5253 |
| Median | | 168.9280 |
| Mode | | 137.03[a] |
| Std. Deviation | | 9.15218 |
| Minimum | | 137.03 |
| Maximum | | 191.84 |
| Percentiles | 25 | 163.8393 |
| | 50 | 168.9280 |
| | 75 | 174.0473 |

a. Multiple modes exist. The smallest value is shown



# Describing Quantitative (Numerical) Data using Charts

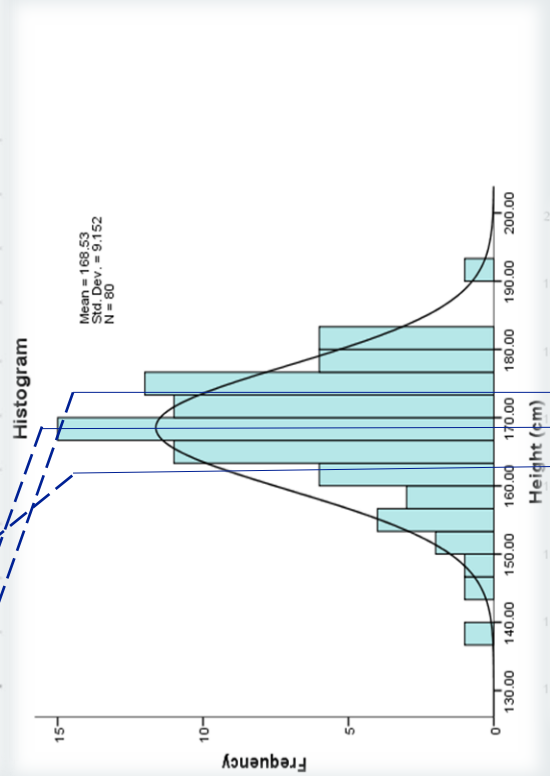
To describe a numerical variable, we need to properly **summarise** it.

**Statistics**

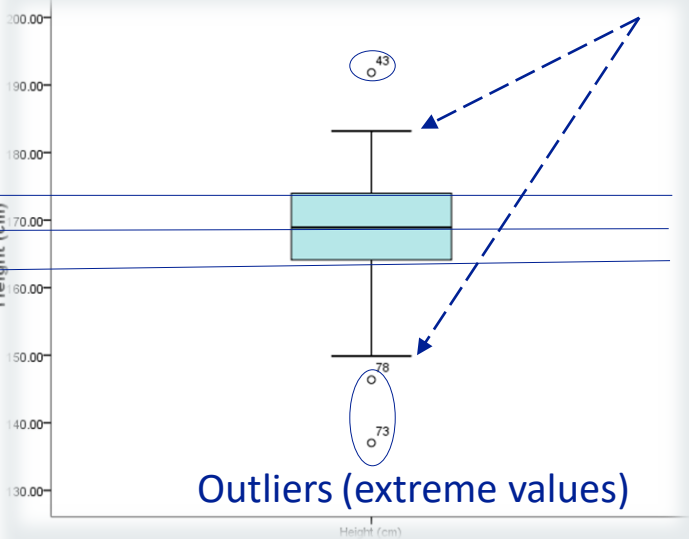
Height (cm)

| | | | |
|---|---|---|---|
| N | Valid | | 80 |
| | Missing | | 0 |
| Mean | | | 168.5253 |
| Median | | | 168.9280 |
| Mode | | | 137.03[a] |
| Std. Deviation | | | 9.15218 |
| Minimum | | | 137.03 |
| Maximum | | | 191.84 |
| Percentiles | 25 | Q1 | 163.8393 |
| **Quartiles** | 50 | Q2 | 168.9280 |
| | 75 | Q3 | 174.0473 |

a. Multiple modes exist. The smallest value is shown




Min and max if outliers were omitted

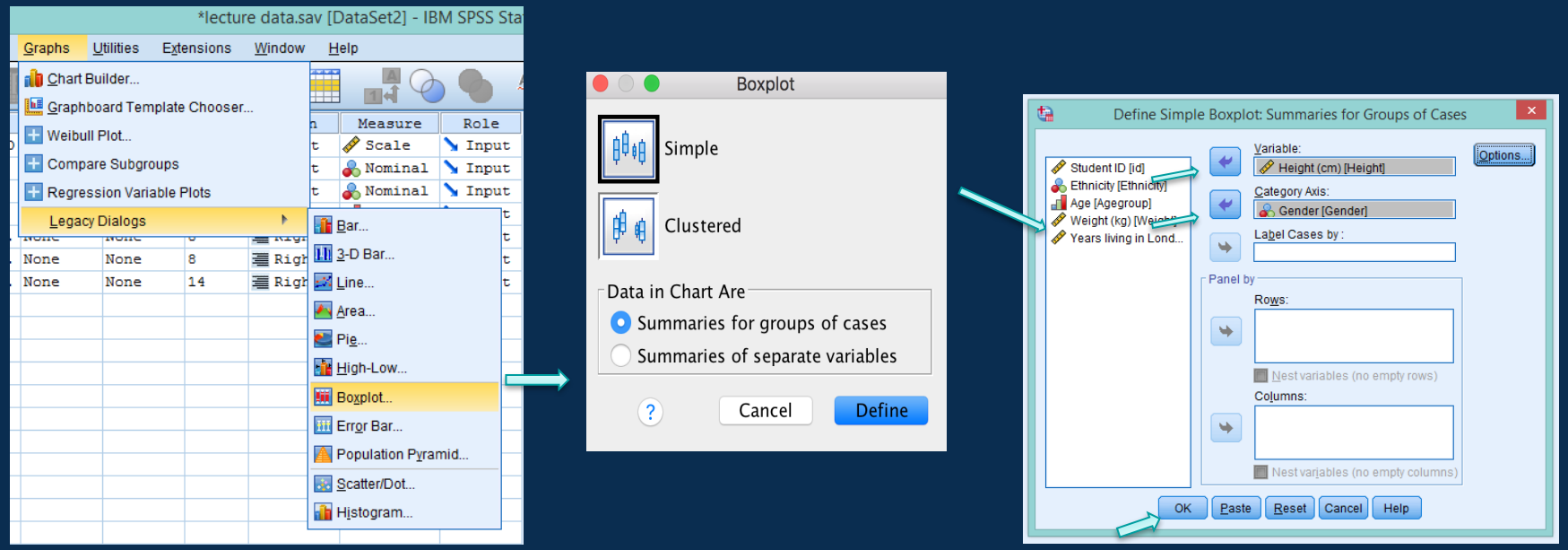



Outliers (extreme values)

# SPSS Slide: ‘How to’ Steps

You can create the boxplot for height **over gender,** using the following steps:

Click on ‘Graphs’→ ‘Legacy Dialogues’→ ‘Boxplot’



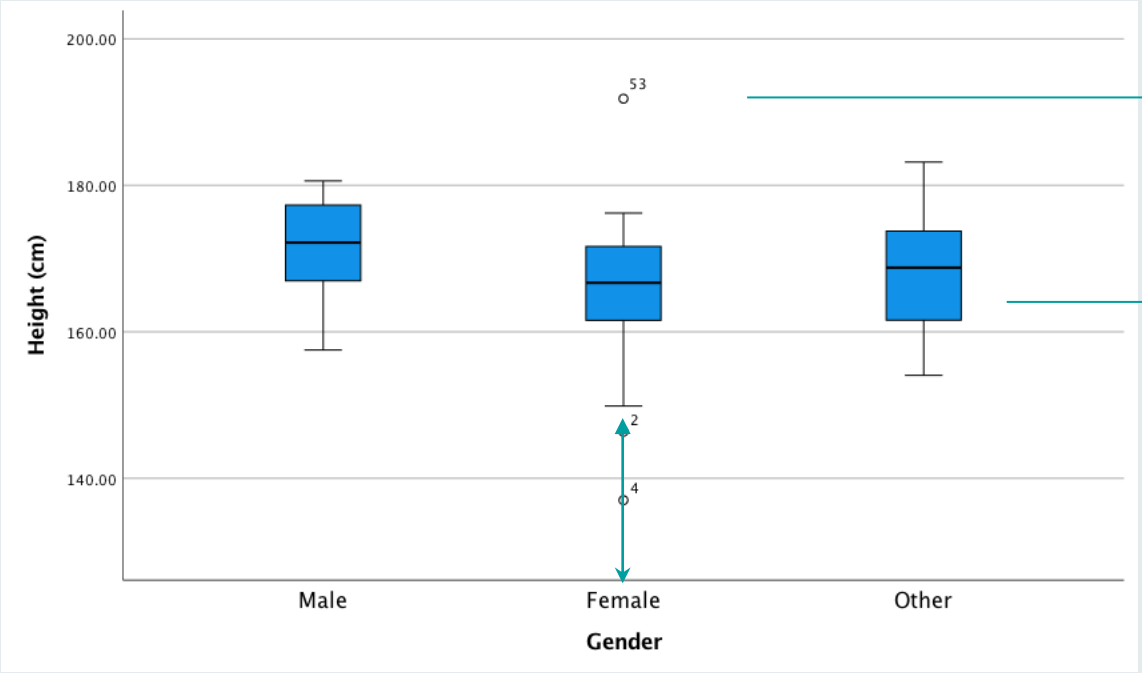

Choose a ‘simple’ layout and click ‘Define’

Add the variable of interest (height) into the ‘Variable(s)’ box
Add the grouping variable (gender) into the ‘Category axis’ box

Click on ‘OK’

# Describing Quantitative (Numerical) Data using Charts

The box plot is very useful in comparing groups visually.



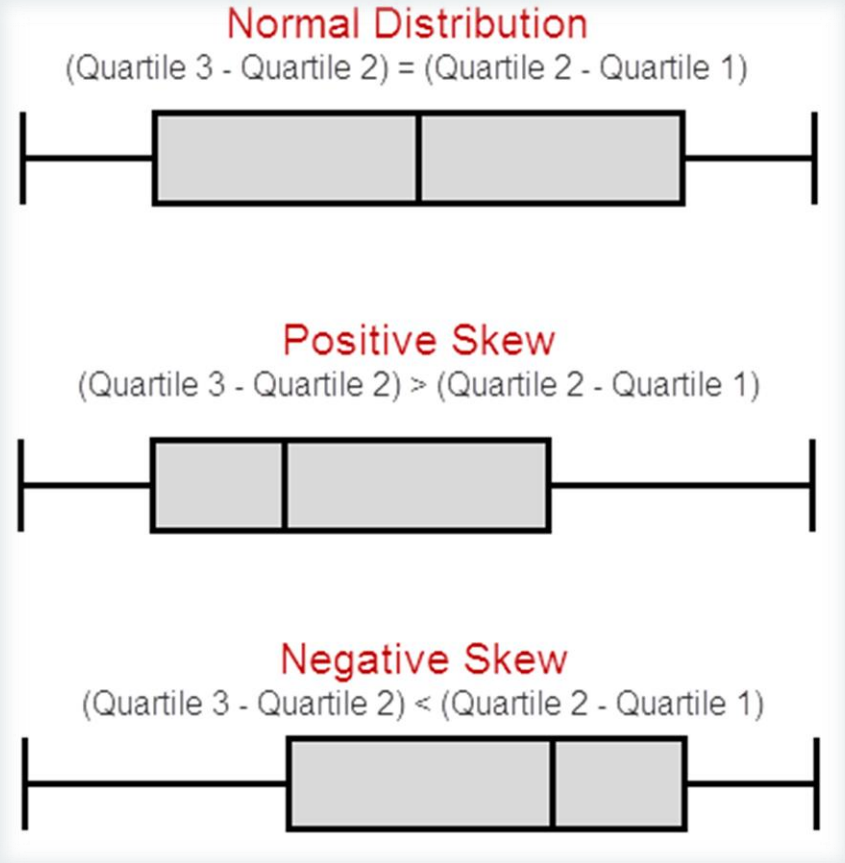

Outliers were ‘females’

The values of ‘other’ completely cover the values of ‘males’.

The low values of ‘females’ were lower than all the values of ‘males’ and ‘other’.

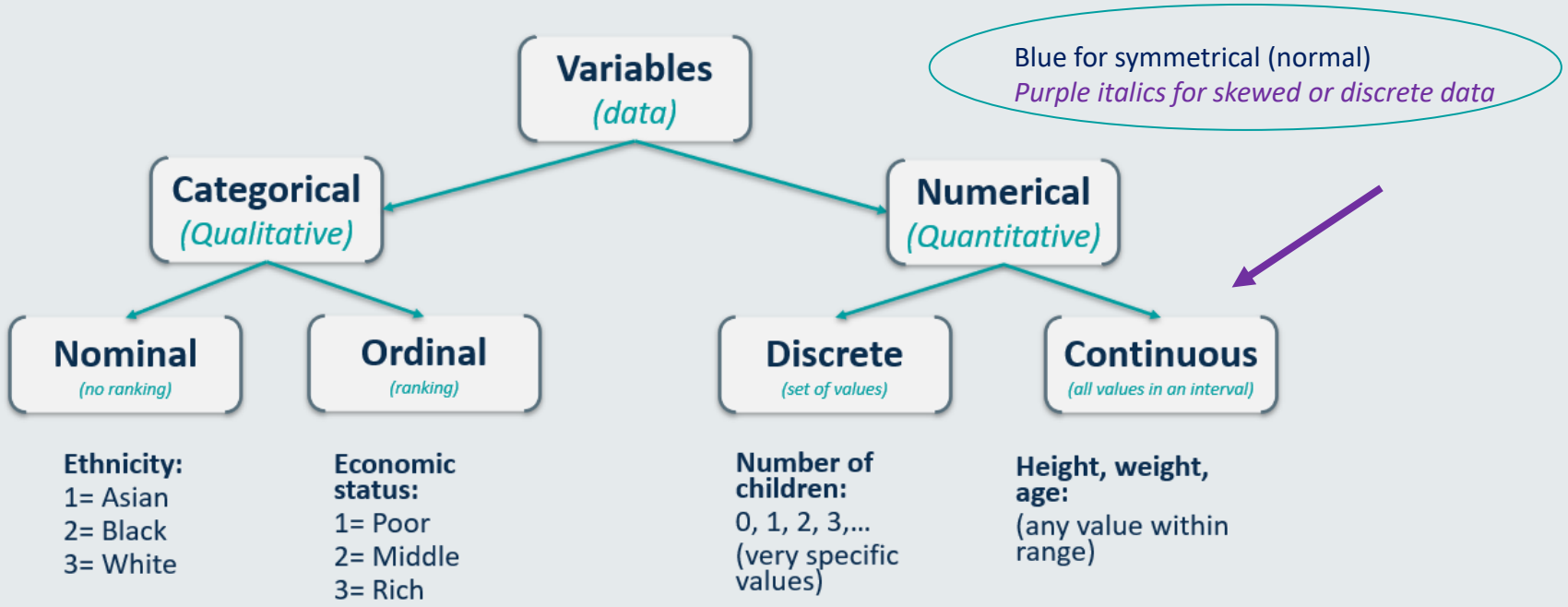
# Describing Quantitative (Numerical) Data using Charts

## Box Plots and skewness



# Types of Variables

Based on the type of each variable, we use different ways to describe the data.




| | | |
|---|---|---|
| • Descriptive indices | Frequencies (Percentages %) | Location: mean, *median*, mode<br>Dispersion: SD, *range*, IQR |
| • Charts/plots | Pie Chart (only for nominal)<br>Bar Chart | Histogram, Box plot |

# Knowledge Check

Q1. Below are the descriptive statistics for Height and LDL. Please give an interpretation of this information.

| ID | Age | Gender | Height | Blood group | LDL† | Feeling happy? | Number of children | Smoke? | Social class |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | F | 1.62 | B | 150 | Agree | 0 | No | I |
| 2 | 35 | F | 1.58 | O | 123 | Strongly agree | 1 | Yes | II |
| 3 | 44 | M | 1.35 | A | 178 | Disagree | 3 | Yes | I |
| 4 | 28 | F | 1.54 | AB | 205 | Disagree | 0 | No | III |
| 5 | 35 | M | 1.35 | O | 229 | Indifferent | 2 | Yes | I |
| 6 | 42 | M | 1.21 | B | 215 | Agree | 2 | Yes | IV |
| 7 | 36 | F | 1.76 | A | 130 | Strongly disagree | 1 | No | IV |
| 8 | 38 | M | 1.57 | A | 175 | Disagree | 1 | Yes | V |
| 9 | 30 | M | 1.47 | AB | 240 | Indifferent | 0 | No | III |
| 10 | 40 | F | 1.18 | B | 167 | Strongly agree | 6 | No | I |
| : | : | : | : | : | : | : | : | : | : |

† LDL =Low Density Lipoprotein

**Statistics**

| | | Height | LDL† |
|---|---|---|---|
| N | Valid | 10 | 10 |
| | Missing | 0 | 0 |
| Mean | | 1.4630 | 181.20 |
| Median | | 1.5050 | 176.50 |
| Mode | | 1.35 | 123[a] |
| Std. Deviation | | .18667 | 40.392 |
| Range | | .58 | 117 |
| Minimum | | 1.18 | 123 |
| Maximum | | 1.76 | 240 |
| Percentiles | 25 | 1.3150 | 145.00 |
| | 50 | 1.5050 | 176.50 |
| | 75 | 1.5900 | 218.50 |

a. Multiple modes exist. The smallest value is shown

# Knowledge Check Solutions

**Q1. Below are the descriptive statistics for Height and LDL give an interpretation of this information.**

**Height**: The height of the individuals in our sample varied between 1.18m and 1.76m. The average height was 1.46m (sd=0.187). Half of the people were taller than 1.51m, while the height most often reported was 1.35m. The difference in the height between the shortest and the tallest person was 0.58m.

***Note:*** *There appears to be some difference between the mean, median and mode and may be indicative that the distribution for height may not be entirely normally distributed. You would need to conduct a histogram or further tests for normality to check if this is the case.*

**LDL**: The LDL of the individuals in our sample varied between 123 and 240. The average LDL measure was 181.2 (sd=40.39). Half of the people had a LDL higher than 176.5, while the LDL value most often reported was 123. The difference in the LDL values between the lowest and the highest was 117.

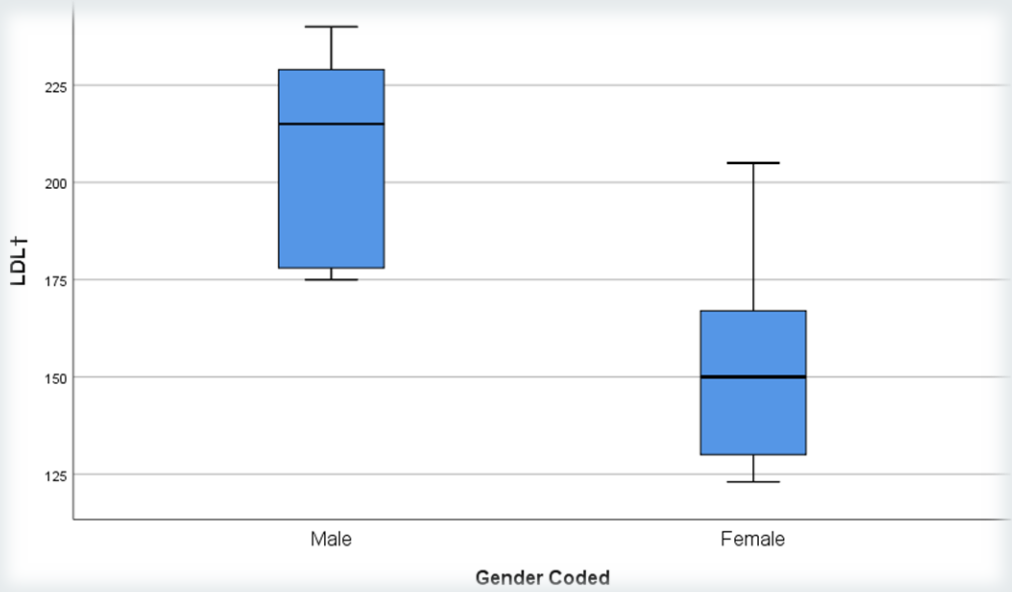
***Note:*** *There appears to be quite a bit of difference between the mean, median and mode and may be indicative that the distribution for LDL may not be normally distributed. You would need to conduct a histogram or further tests for normality to check if this is the case. We can also see that the standard deviation is high indicating a lot of variability in the data.*

# Knowledge Check

## Q2. Below is a box plot of the variable 'LDL Group' grouped by gender what does the chart show?

| ID | Age | Gender | Height | Blood group | LDL† | Feeling happy? | Number of children | Smoke? | Social class |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | F | 1.62 | B | 150 | Agree | 0 | No | I |
| 2 | 35 | F | 1.58 | O | 123 | Strongly agree | 1 | Yes | II |
| 3 | 44 | M | 1.35 | A | 178 | Disagree | 3 | Yes | I |
| 4 | 28 | F | 1.54 | AB | 205 | Disagree | 0 | No | III |
| 5 | 35 | M | 1.35 | O | 229 | Indifferent | 2 | Yes | I |
| 6 | 42 | M | 1.21 | B | 215 | Agree | 2 | Yes | IV |
| 7 | 36 | F | 1.76 | A | 130 | Strongly disagree | 1 | No | IV |
| 8 | 38 | M | 1.57 | A | 175 | Disagree | 1 | Yes | V |
| 9 | 30 | M | 1.47 | AB | 240 | Indifferent | 0 | No | III |
| 10 | 40 | F | 1.18 | B | 167 | Strongly agree | 6 | No | I |
| : | : | : | : | : | : | : | : | : | : |

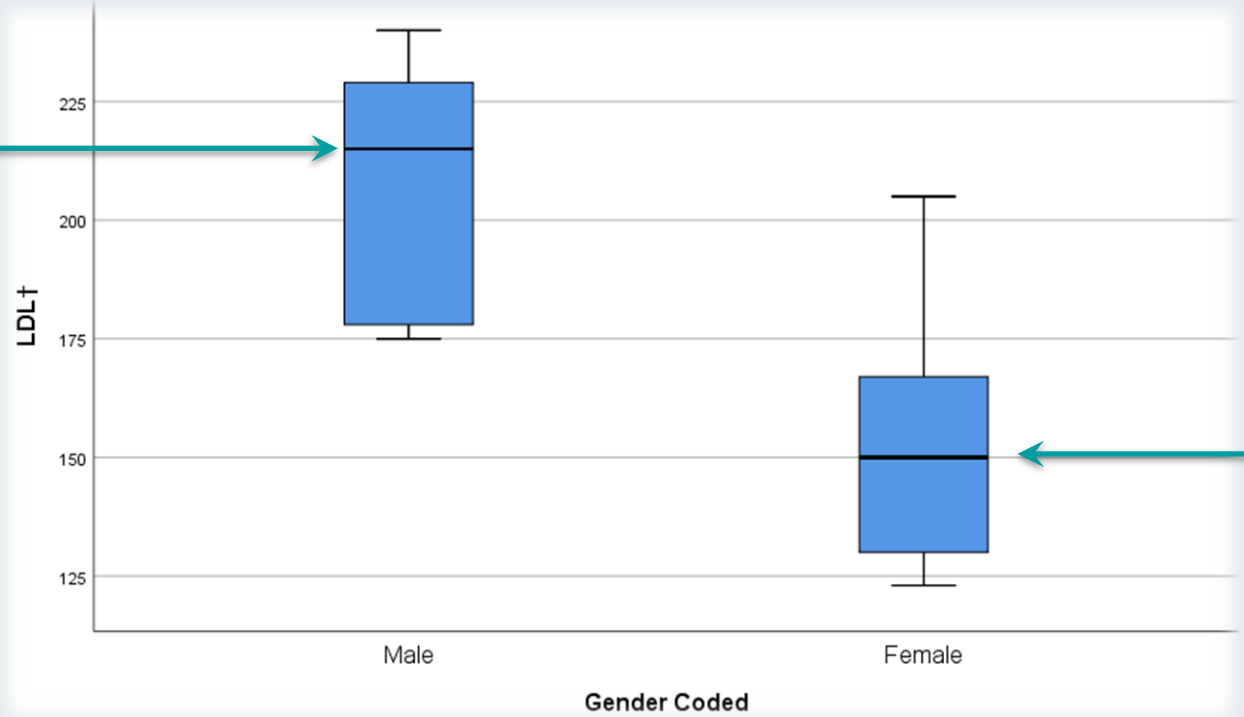
**† LDL =Low Density Lipoprotein**



# Knowledge Check Solutions

**Q2. Below is a box plot of the variable 'LDL Group' grouped by gender what does the chart show?**

Median LDL Value for Males




Median LDL Value for Females

Males have a higher median LDL value compared to females (215 vs 150 respectively). No outliers have been identified in the two groups. Males LDL values ranged from 175 to 240 and Females LDL values 123 to 205) with males having a much smaller range of values. Both the male and female LDL distributions are skewed, females Positively skewed (Q3-Q2>Q2-Q1) and males negatively skewed (Q3-Q2<Q2-Q1)

# Reference List


**For more details of the concepts covered in Topic 1, see Chapters 1- 3 of the book:**

Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences (4th Edition), Prentice Hall In Chapters 1-3.

**For more details on SPSS implementation see:**
Field (2005) Discovering Statistics using SPSS 2nd Edition, Sage, London.
The SPSS Environment, Chapter 2.

**For more details on measurement issues see:**
Streiner & Norman (2003) Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford University Press.

**Cleaning Data References**

https://www.betterevaluation.org/en/evaluation-options/data_cleaning

Google Refine: Tool of the Year for Evaluators: provides an overview of Google Refine which is a desktop application (downloadable) that can be used to calculate frequencies and multi-tabulate data from large datasets and also clean up your data. (AEA)

Data Cleaning: Problems and Current Approaches: explains the main problems that data cleaning is able to correct and then provides an overview of the solutions that are available to implement the cleansing of data. (University of Leipzig)
Guides

Data Cleaning 101: outlines a step-by-step process for verifying that data values are correct or, at the very least, conform to some a set of rules through the use of a data cleaning process.

Rahm, E., & Hai Do, H. University of Leipzig, Germany, (n.d.). Data cleaning: Problems and current approaches. Retrieved from website: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf

Wikipedia (2012). Data cleansing. Retrieved from http://en.wikipedia.org/wiki/Data_cleansing

# Thank you

**Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.**

**If you have comments on the materials (spotted typos or missing points) please contact Dr Vitoratou:**

Silia Vitoratou, PhD
Psychometrics & Measurement Lab,
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
**silia.vitoratou@kcl.ac.uk**

**For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:**

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**