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: Thinking statistically

Topic title: Measurement and graphical representations of data

Learning Outcomes

- To understand why statistics works
- To understand when to use statistics
- To understand the overall goal of using statistics

○ The course does not require prior knowledge in statistics

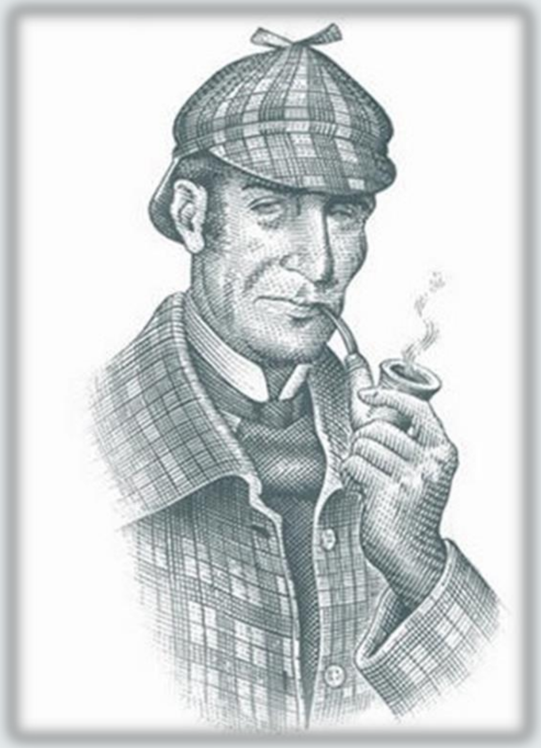
○ Invariably students will be at different levels depending on previous exposure to statistics

○ We will take it step by step and we will work together, helping each other with the ideas and the methods

A First Note on Statistics

Statistics is a valuable tool in modern research and is used extensively. It helps us unravel the mystery.

Sherlock Holmes quoting Winwood read:



"A famous statistician once stated that **while the individual man is an insoluble puzzle**, in the aggregate he becomes a mathematical certainty. You can, for example, **never foretell what any one man will do**. **But you can, with precision, say what an average man will do**. Individuals vary, percentages remain constant. So says the statistician. "

Introduction to Statistics

In this course we will see and understand why both parts of the quote, you have just heard, are true:

“you can **never foretell what a specific person will do**
but you **can, with precision, say what people on average** will do”

- **why** can we actually do so?
  (do statistics really work?)

- **when** does it make sense to use statistics?
  (we don’t always use it)

- **how** do we apply statistics?
  (you can learn to do so, give it a chance)

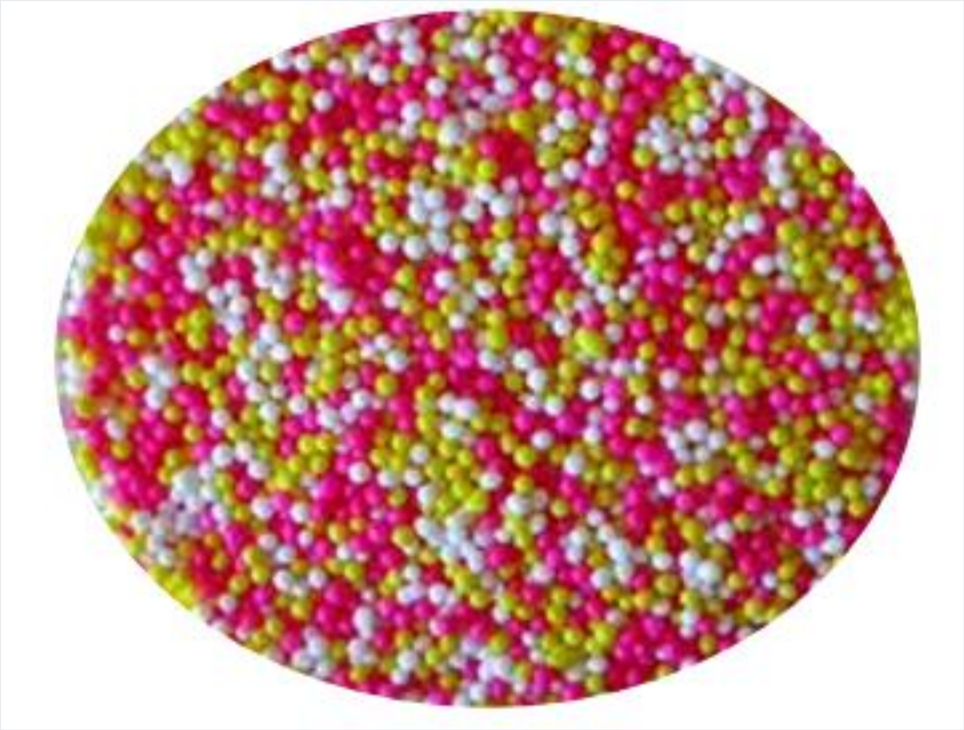
Introduction to Statistics: Why it Works?

Imagine that we have a bowl and we pour in the same amounts of pink, yellow and white candies.

1/3

1/3

1/3



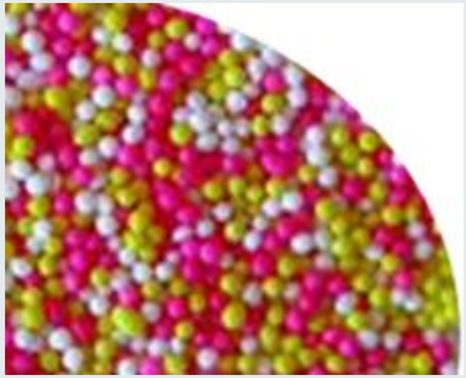
Introduction to Statistics: Why it Works?

Assume that we did not know the proportions of each colour and we can only see a random part of the bowl.

 ?

 ?

 ?



Can we tell, if we only saw part of the bowl, that there are equal proportions from each colour?

Actually, we could even count how many of each colour there are in a **random sample** of the bowl, and based on that **estimate** how many there are in the entire bowl.

Introduction to Statistics: <u>When to Use?</u>

That is how statistics works. We study a representative sample and find out (**infer**) what happens in the entire population.

<u>So when do we use statistics?</u>

When we want to study a **<u>characteristic</u>** and we can only access a **<u>sample</u>** of the population, not every one on the planet.

But that is not all: what kind of characteristics?

Introduction to Statistics: When to Use?

Example: we want to study how fast a runner's heart beats after 10 minutes of running.

Heart rate variable

170

150 values

195

sample

https://depositphotos.com/10476109/stock-illustration-sketch-of-female-marathon-runner.html

The values "vary", not everyone has the same rate.

A characteristic that varies (a variable) is a characteristic that we study using statistics.

Introduction to Statistics: When to Use?

Observing that the heartbeats vary is one thing, but why do the heart rates vary? To understand that, we add variables:

Heart rate	Age	BMI
170	22	22
150	21	35
195	32	18

In statistics the objective is to understand why things vary:

"to explain the variability" **"to reduce uncertainty"**

In this course we will learn ways to account for this variability.

sample

https://depositphotos.com/10476109/stock-illustration-sketch-of-female-marathon-runner.html

Introduction to Statistics: What is the Goal?

Statistics is the science whose objective is to understand why things are not the same for everyone:

Why don’t we all have the same weight? Why don’t we live the same amount of days? Why don’t we all feel equally happy?

Statistics helps us unravel the mystery of variability.

We study a sample and use probability theory to have an educated guess about the **population** based on our **sample**.

While out target is to understand the **population**, we work on the **sample**. We need to get to know our data first.

Reference List

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

Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences (4th Edition), Prentice Hall Inc. 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.

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**

Institute of Psychiatry, Psychology and Neuroscience

Module Title: Introduction to Statistics

Session Title: Types of data

Topic title: Measurement and graphical representations of data

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Learning Outcomes

- To understand the different types of data
- To classify variables according to their type of data
- To reflect on the data you are likely to come across in your own research

Descriptive Statistics

The very **first thing** to do is to familiarise with your sample data.

(Most people think this is all statistics is about, but it is not, it is just the first step!)

We do this using descriptive statistics

Descriptive Statistics

Descriptive statistics answer the questions:

- what **type** of variables you have?
- what are their **values** in your sample?

and allow you to:

- make sure you do not have errors or typos in your data (**data cleaning**)
  You need to make sure that your data are cleaned, otherwise

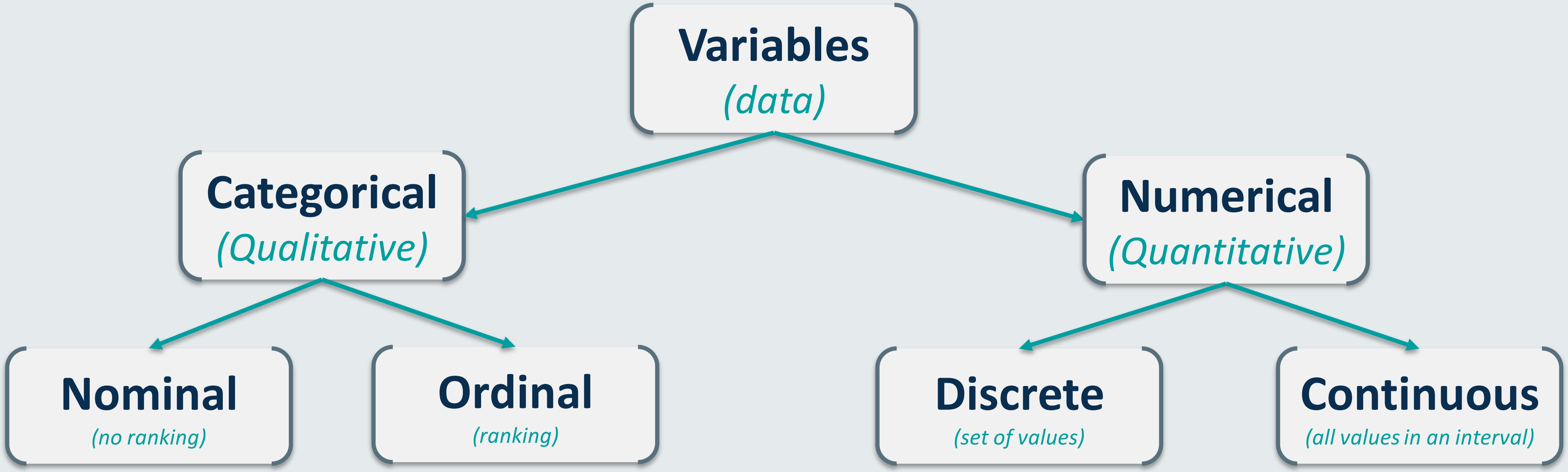
GIGO

(garbage in, garbage out)

- understand your information in your (clean) **sample**, so you can start thinking about the **population**

Types of Variables

Types of Variables



Ethnicity:
1= Asian
2= Black
3= White

Economic status:
1= Poor
2= Middle
3= Rich

Number of children:
0, 1, 2, 3,…
(very specific values)

Height, weight, age:
(any value within range)

Categorical Data: nominal or ordinal?

Nominal data can't be expressed as a number and can't be measured. They are **names** which represent qualities of the observations, characteristics, categories the observations belong to.

Nominal data can take on numerical values (example: 1 for male, 2 for female, 3 for other) but those numbers don't have mathematical meaning - are coded for ease of computation in most statistical software. Ordering has no meaning.

Ethnicity

i. Asian
ii. Black
iii. White
iv. Other

Gender

a) Cis man
b) Cis woman
c) Trans man
d) Trans woman
e) Other

Hair colour

1. Blonde
2. Brown
3. Brunette
4. Red

Marital Status

- Married
- Single
- Widowed
- Self-partnered

Housing Style

- ❑ Detached
- ❑ Semi-Detached
- ❑ Terraced
- ❑ Bungalow
- ❑ Flat

Religion

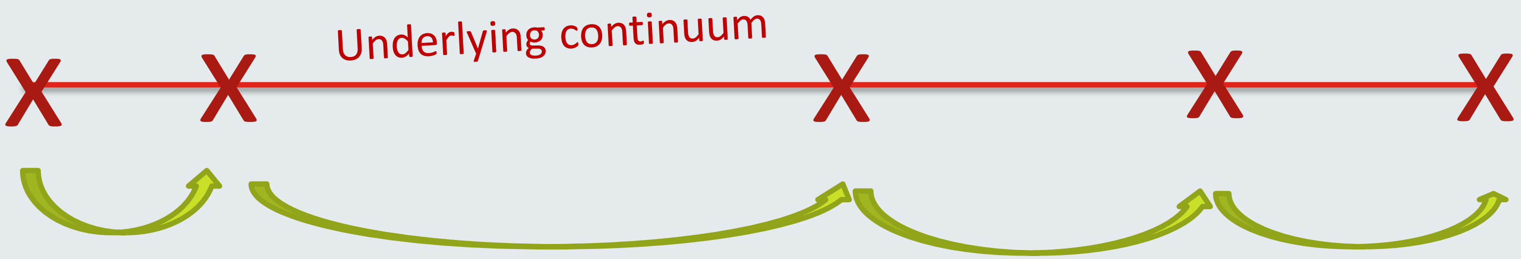
I. Buddhism
II. Christianity
III. Hinduism
IV. Islam
V. No religion

Categorical Data nominal or ordinal?

Ordinal data take on numerical values and those numbers represent the **order** of the categories. However they lack mathematical meaning as the spacing between categories is not necessarily equal.

How satisfied are you from this lecture?

very dissatisfied dissatisfied neutral satisfied very satisfied

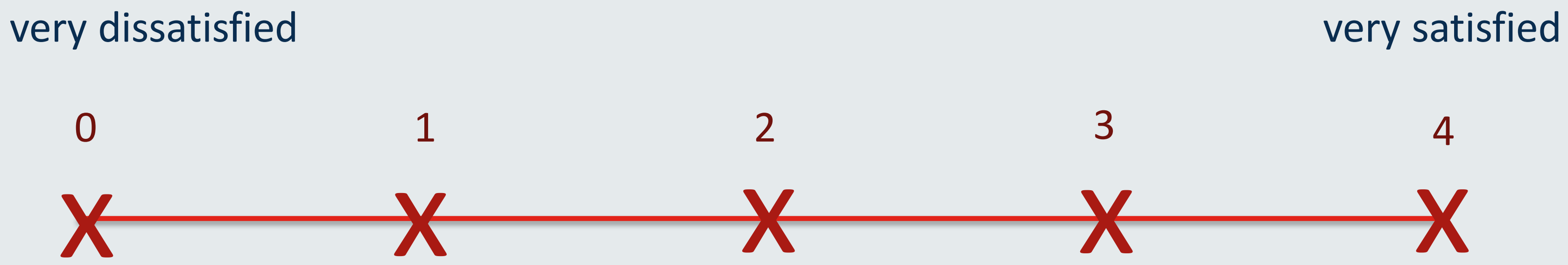


Ordinal (categorical) data or Interval (numeric data?)

Ordinal data take on numerical values and those numbers represent the **order** of the categories. However they lack mathematical meaning as the spacing between categories is not necessarily equal.

But if the variable is structured in a way that it is clear that the spacing is equidistant, and differences between then are meaningful, then the data are **interval** data (numerical data). An example:

How satisfied are you from this lecture from 0 to 4?



That is because it now makes sense to say 4, is double as 2 and the distance between 1 and 3 is the same as, say, 2 and 4. There is a mathematical underpinning in the numbers now.

Summarisng

Ordinal variables and **interval** variables are very often used in Mental Health to measure individuals' perceptions, feelings, agreement, intensity, frequency of symptoms. Actually they are the most often used ones on **psychometric scales**.

It can be tricky sometimes to know how to analyse ordinal data (that should be treated as categorical) from interval data (scale, numerical data). But here is some tips:

- If a variable has **four categories or fewer** then always treat it as **categorical**. Even if the points are equidistant the information we have (4 points) is too small to approximate the underlying variable.
- If a variable **has five or more categories** and these can be assumed to be **equidistant**, then the data can be treated as continuous data – that is, we essentially treat them **as the underlying variable that determines the order (an approximation).**
- If a variable has ordered values where the **difference between two values is meaningful** then these data are interval data and follow the rules of numerical data.

Nominal, ordinal and interval data differences and similarities

Some examples of ordinal data:

Agreement
1. Strongly disagree
2. Disagree
3. Agree
4. Strongly agree

Frequency
1. Almost never
2. Sometimes
3. Often
4. Almost always
5. Always

But it would be interval data if:

On a scale of 1 (strongly disagree) to 10 (strongly agree) , how much do you agree?

How often do you…
1. 1-5 days per month
2. 6-10 days per month
3. 11-15 days per month
4. 16-20 days per month
5. 21-25 days per month
6. 26-30 days per month
7. Every day of the month

Easy to spot though that these are nominal data

Agreement
1. I am not sure
2. I agree to some extent
3. Depends on the occasion
4. I am not informed

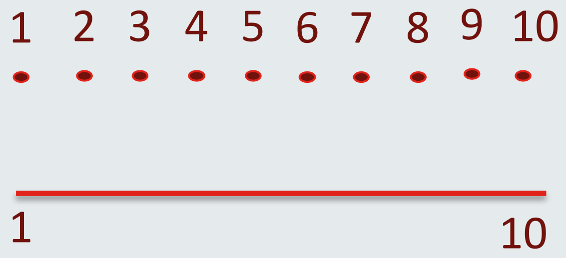
Frequency
1. More than I would want to
2. Less than most people
3. I have not noticed

Numerical Data

Sometimes it can also be tricky to tell apart **discrete** and **continuous** data. Discrete data take only very specific (and pre-specified) **set** of values. Continuous data can take all values in a prespecified **interval**.

Discrete {1, 2, 3, 4, 5, 6, 7, 8, 9,10}

Continuous [1,10]



Typically, discrete data are **counts** and continuous data are **measurements**.

How many children?	Weight
How many cars?	Height
How many times?	Age

Numerical Data

A general rule to tell apart discrete from continuous data, is to remember that continuous data can have any number of decimals while discrete data do not. But this can be proven tricky, as there are exceptions. For instance

- Age in your dataset takes values 19 22 33 56 44 15 22 37 89 61. Is your variable discrete?

  The answer is no, age is continuous. For convenience age was rounded up to zero decimal points but it is clear that any point in between 'years' is a plausible value (could be decimals representing months, weeks, days, minutes etc…)

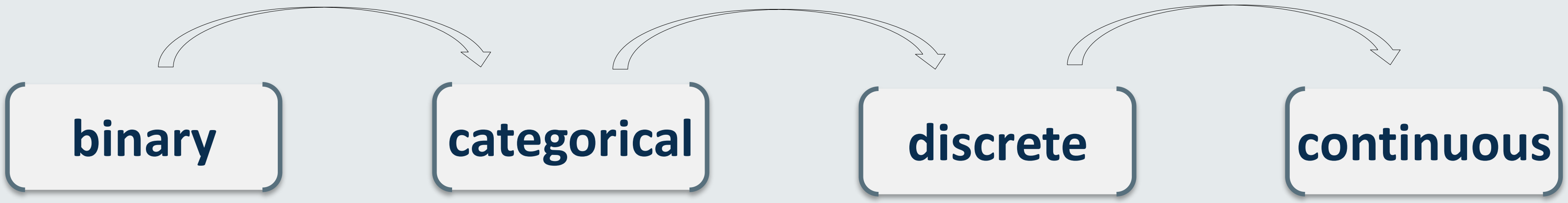
Any value in the interval (0,120) works

- UK shoe numbers include halves, that is 4, 4.5, 5, 5.5 etc. Is your variable continuous since it has decimals?

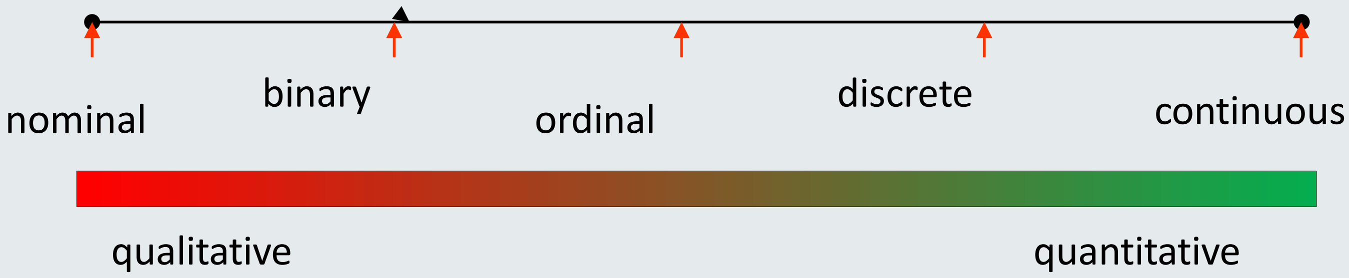
  The answer is no, the shoe size is a discrete variable. Even though there are decimals, these are very specific (you cannot for instance have a shoe of size 4.6). This means that shoe size takes values from a predefined set of numbers, not an interval.

Shoe sizes are a pre-defined finite set of values

Data on a scale



It is useful to think of data on a scale:



Quantitative (numerical) data are generally more useful than qualitative (categorical) data and so if possible choose a 'green' rather than 'red' scale! (WHY?)

Knowledge Check

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

Q1. Which of the variable(s) are classified as **quantitative** variable(s)?

Q2. Which of the variable(s) are classified as **qualitative** variable(s)?

Q3. Which of the variable(s) are classified as **nominal** variable(s)?

Q4. Which of the variable(s) are classified as **ordinal** variable(s)?

Q5. Which of the variable(s) are classified as **discrete** variable(s)?

Q6. Which of the variable(s) are classified as **continuous** variable(s)?

Knowledge Check Solutions

1. **Which of the variable(s) are classified as quantitative variable(s)?**
   Age, Height, LDL, Number of Children
   These variables take numerical values only and the values reflect the actual measurement (with units) of the subjects or objects we are measuring.
2. **Which of the variable(s) are classified as qualitative variable(s)?**
   Blood Group, Gender, Feeling Happy, Smoke, Social class
   These variables are represented by categories and each category represents a particular characteristic of interest within a group of subjects or objects.
3. **Which of the variable(s) are classified as nominal variable(s)?**
   Gender, Blood Group, Smoke
   These variables consist of categories that are mutually exclusive but have no ranked order, e.g. Male / Female.
4. **Which of the variable(s) are classified as ordinal variable(s)?**
   Feeling Happy, Social Class
   These variables consist of categories that are mutually exclusive and have a ranked order. Thus, for example, the category “strongly agree” may precede “agree”. Note that the “interval” between categories may not be numerically equal.
5. **Which of the variable(s) are classified as discrete variable(s)?**
   ID, Number of Children
   These variables take integer values. ID is the subject or case number and Number of Children are counts.
6. **Which of the variable(s) are classified as continuous variable(s)?**
   Age, LDL, Height
   These variables can take any value within an interval, including decimal parts. The precision of the measurement will depend on the measuring device used.

Reflection

- Write down a list of 5 different variables you might come across in your own research.

- Write next to them what types of variables they are.

Reference List

For more details of the concepts covered in Session 1, see Chapters 1- 3 of the book:
Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences (4th Edition), Prentice Hall Inc. 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.

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**

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Summarising categorical data

Topic title: Measurement and graphical representations of data

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Learning Outcomes

- To understand the descriptive indices suitable for categorical data
- To understand the descriptive charts suitable for categorical data
- To be able to use a software package to create descriptive indices and charts

Recap: Types of Variables

What type of variables we may have

- **Variables** *(data)*
  - **Categorical** *(Qualitative)*
    - **Nominal** *(no ranking)*
    - **Ordinal** *(ranking)*
  - **Numerical** *(Quantitative)*
    - **Discrete** *(set of values)*
    - **Continuous** *(all values in an interval)*

Ethnicity:
1= Asian
2= Black
3= White

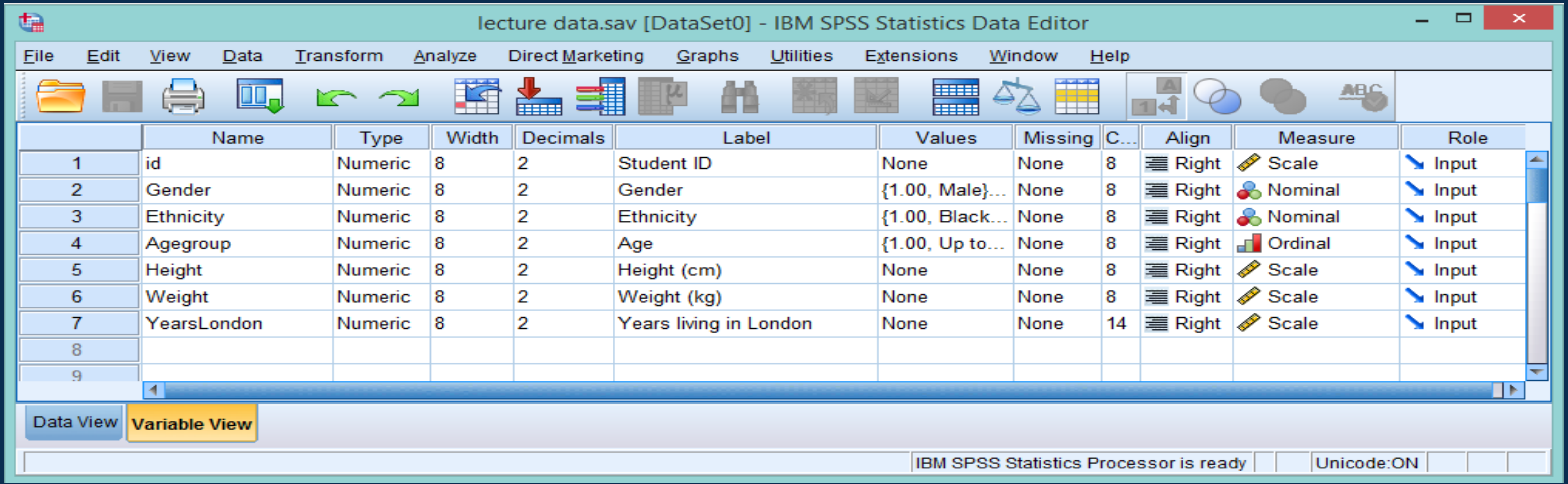
Economic status:
1= Poor
2= Middle
3= Rich

Number of children:
0, 1, 2, 3,…
(very specific values)

Height, weight, age:
(any value within range)

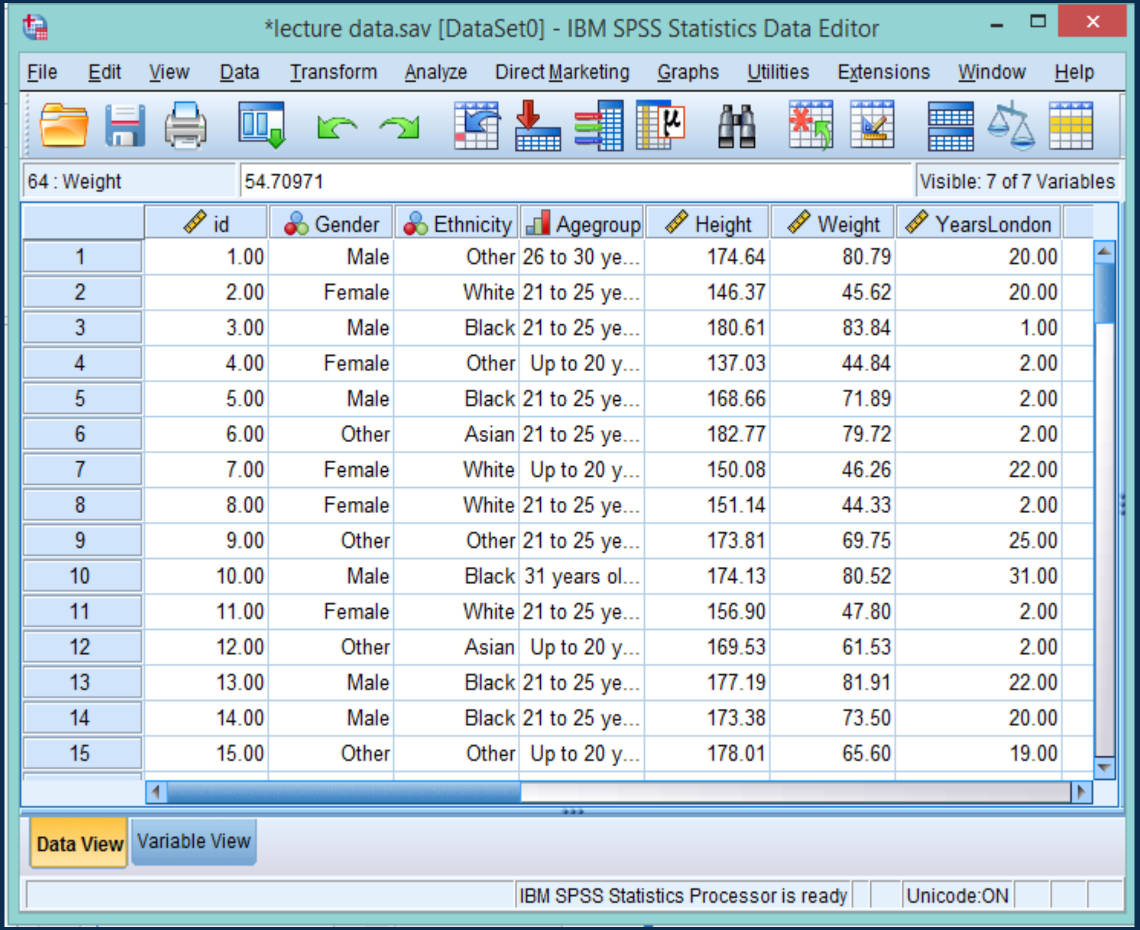
SPSS Slide

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



	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											

Cleaning & Describing Data



	id	Gender	Ethnicity	Agegroup	Height	Weight	YearsLondon
1	1.00	Male	Other	26 to 30 ye...	174.64	80.79	20.00
2	2.00	Female	White	21 to 25 ye...	146.37	45.62	20.00
3	3.00	Male	Black	21 to 25 ye...	180.61	83.84	1.00
4	4.00	Female	Other	Up to 20 y...	137.03	44.84	2.00
5	5.00	Male	Black	21 to 25 ye...	168.66	71.89	2.00
6	6.00	Other	Asian	21 to 25 ye...	182.77	79.72	2.00
7	7.00	Female	White	Up to 20 y...	150.08	46.26	22.00
8	8.00	Female	White	21 to 25 ye...	151.14	44.33	2.00
9	9.00	Other	Other	21 to 25 ye...	173.81	69.75	25.00
10	10.00	Male	Black	31 years ol...	174.13	80.52	31.00
11	11.00	Female	White	21 to 25 ye...	156.90	47.80	2.00
12	12.00	Other	Asian	Up to 20 y...	169.53	61.53	2.00
13	13.00	Male	Black	21 to 25 ye...	177.19	81.91	22.00
14	14.00	Male	Black	21 to 25 ye...	173.38	73.50	20.00
15	15.00	Other	Other	Up to 20 y...	178.01	65.60	19.00

The dataset contains data from 80 students, with respect to their:

reported gender

ethnicity

age group

height in cm

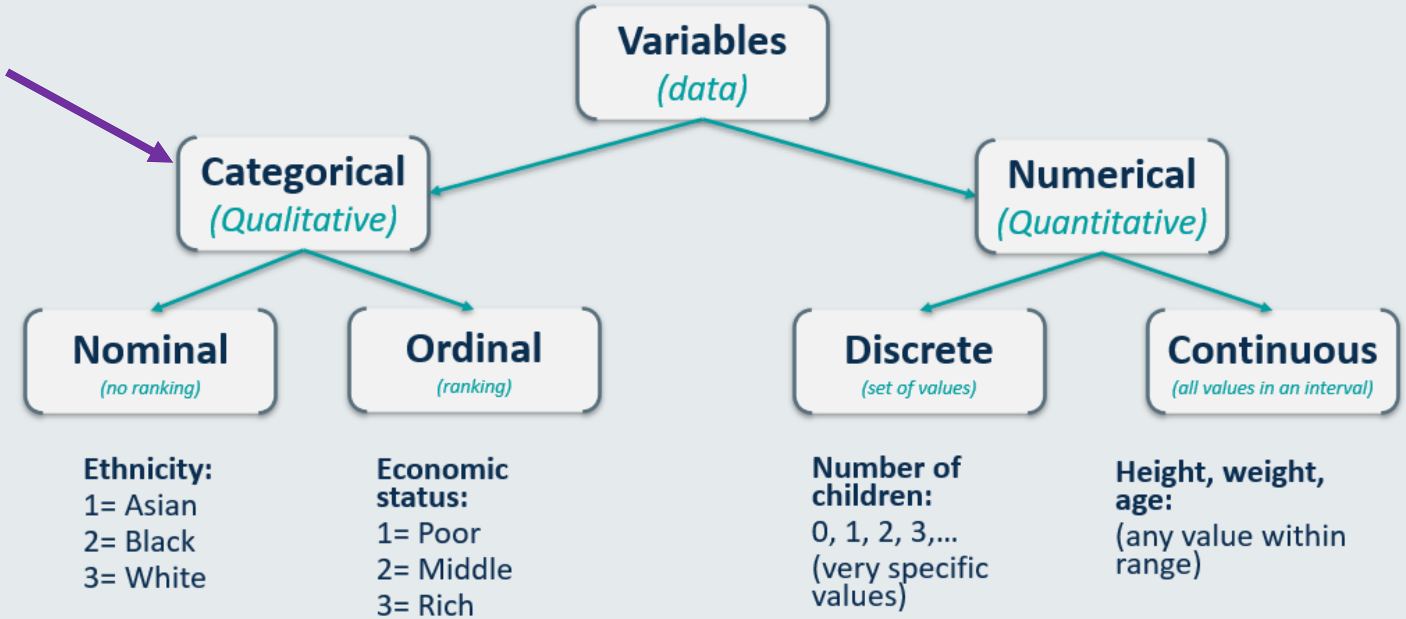
weight in kg

years living in London

Should we start scrolling up and down to spot typos or to see how many females we have? What if we had 800 students?

Types of Variables

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



- Descriptive indices ? ?
- Charts/plots ? ?

Qualitative (Categorical) Data

In categorical data, one would be interested in how many people are in each category and in total. We call this the '**frequency** of each category' and we use 'N' to symbolise the number of people. We also express these frequencies as **percentages** (%). Let's look at Gender (nominal data) as an example

Table 1: SPSS Frequency table for Gender

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	28	35.0	36.4	36.4
	Female	30	37.5	39.0	75.3
	Other	19	23.8	24.7	100.0
	Total	77	96.3	100.0	
Missing	System	3	3.8		
Total		80	100.0		

categories

missing values

Number of people in each category

Totals with and without missing values

% with and without missing values

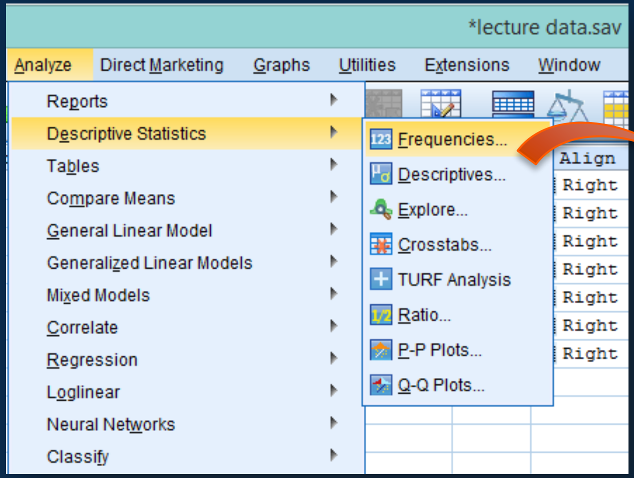
- 35% of the individuals in the sample (N=80) identified themselves as males
- 36.4% of the individuals who responded (N=77) identified themselves as males
- 75.3% of the individuals who responded (N=77) identified themselves as either males or females.

The cumulative % makes more sense in ordinal data

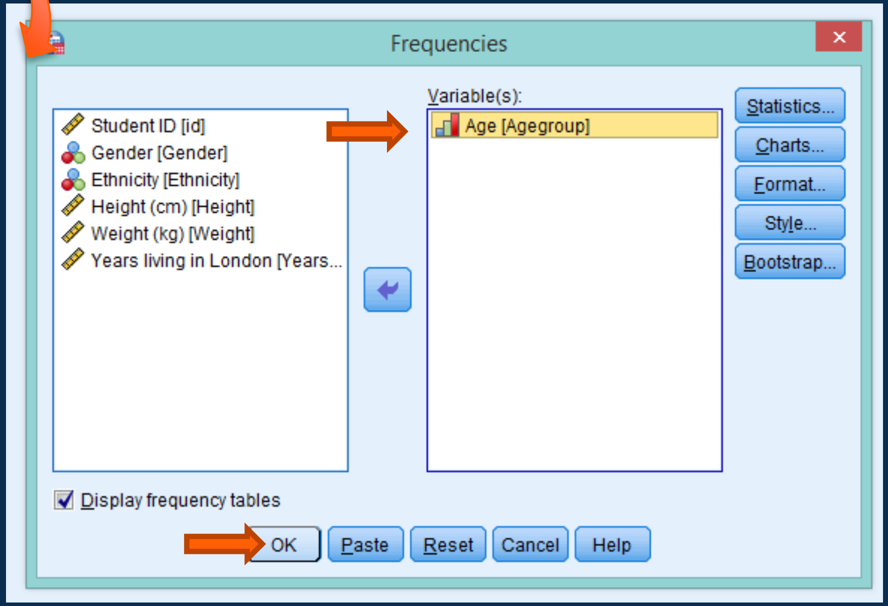
SPSS Slide: ‘How to’ Steps

You can create the **frequency table** for agegroup (ordinal data) using the following steps:

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



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



Click on ‘OK.

Output and Interpretation

Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Up to 20 years old	19	23.8	23.8	23.8
	21 to 25 years old	45	56.3	56.3	80.0
	26 to 30 years old	12	15.0	15.0	95.0
	31 years old and above	4	5.0	5.0	100.0
	Total	80	100.0	100.0	

INTERPRETATION: In our sample, most people belong to the 21-25 years old **age** group (N=45, 56.3%). The vast majority of the individuals in our sample were up to 25 years old (N=64, 80.0%). Only 4 people (5.0%) were 31 years old or above.

Ethnicity

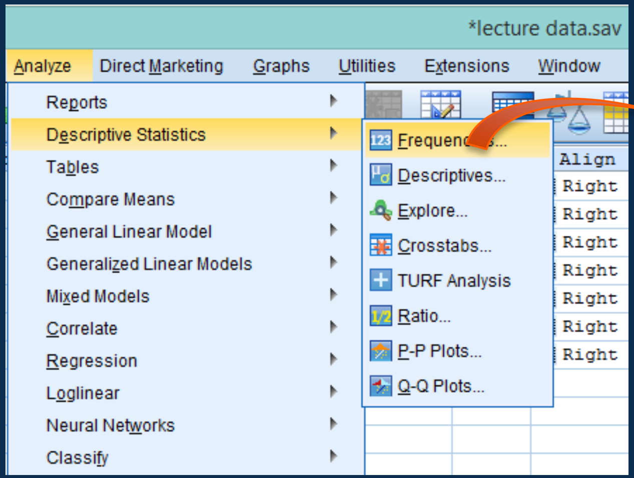
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Black	11	13.8	13.8	13.8
	White	19	23.8	23.8	37.5
	Asian	17	21.3	21.3	58.8
	Mixed	18	22.5	22.5	81.3
	Other	14	17.5	17.5	98.8
	22.00	1	1.3	1.3	100.0
	Total	80	100.0	100.0	

Typo spotted

By creating a frequency table for Ethnicity we were able to spot a typo/error in the data.

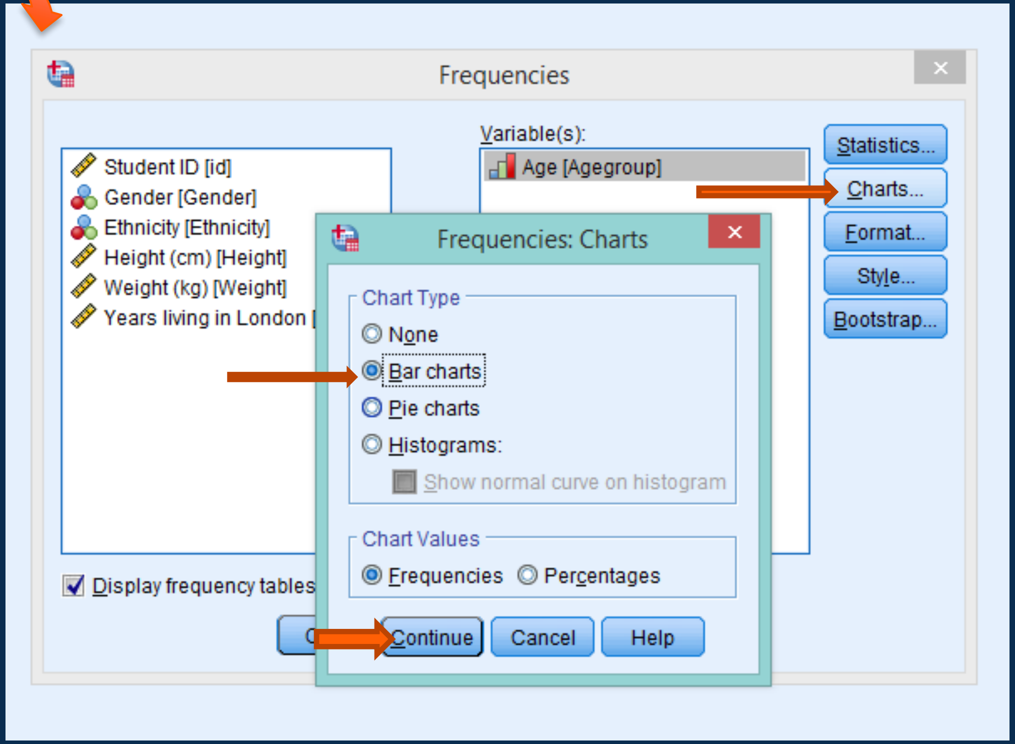
SPSS Slide: 'How to' Steps

You can create the **charts** for agegroup (ordinal data) using the following steps:



Click on the 'Analyse Tab'→ 'Descriptive Statistics' → 'Frequencies'

Add the variable of interest (agegroup) into the 'Variable(s)' box



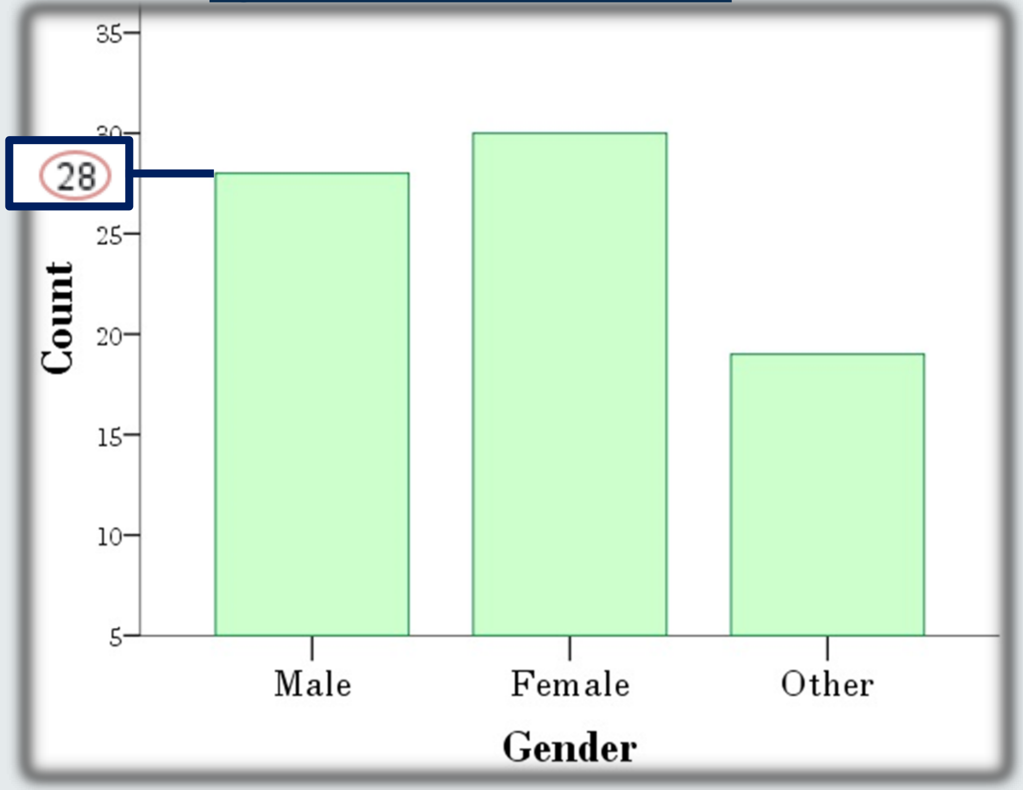
Click on 'Charts'

Choose 'Bar Chart' or 'Pie Chart'
Click 'Continue'
Click on 'OK'.

Describing Categorical Data using Charts

- To depict categorical data, most often we use a **Bar Chart** or a **Pie Chart:**

Figure 2: SPSS Bar Chart of Gender



Gender

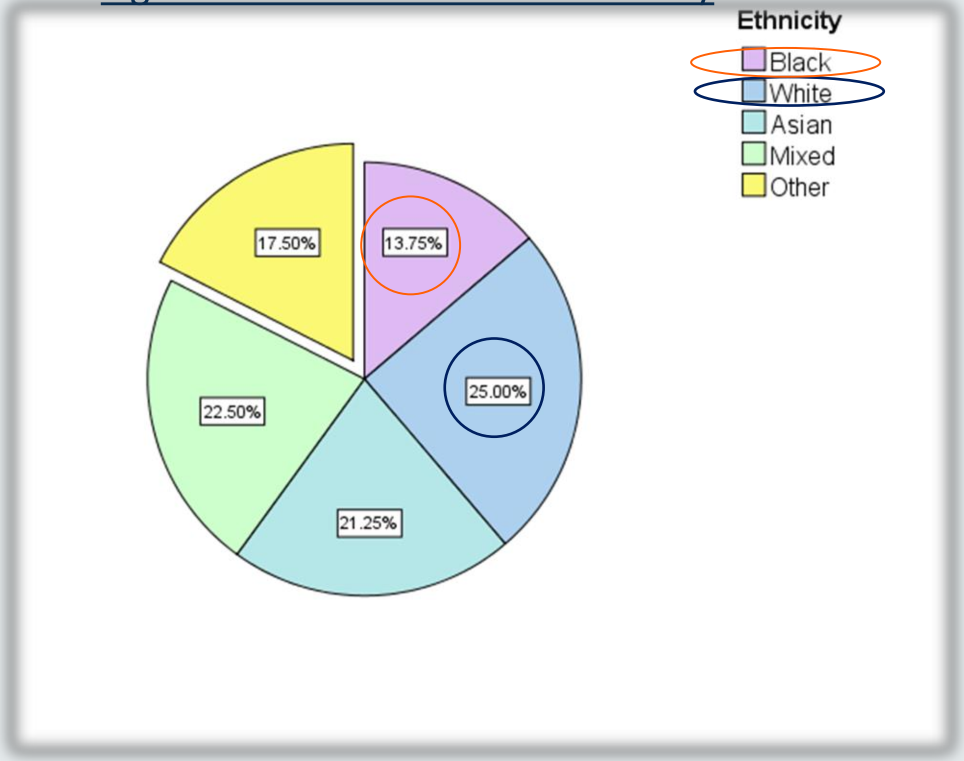
	Frequency	Percent
Male	28	35.0
Female	30	37.5
Other	19	23.8

In a bar chart, the height of the bars represents the frequency of each category.

Describing Categorical Data using Charts

- To depict categorical data, most often we use a **Bar Chart** or a **Pie Chart:**

Figure 1: SPSS Pie Chart of Ethnicity



Ethnicity

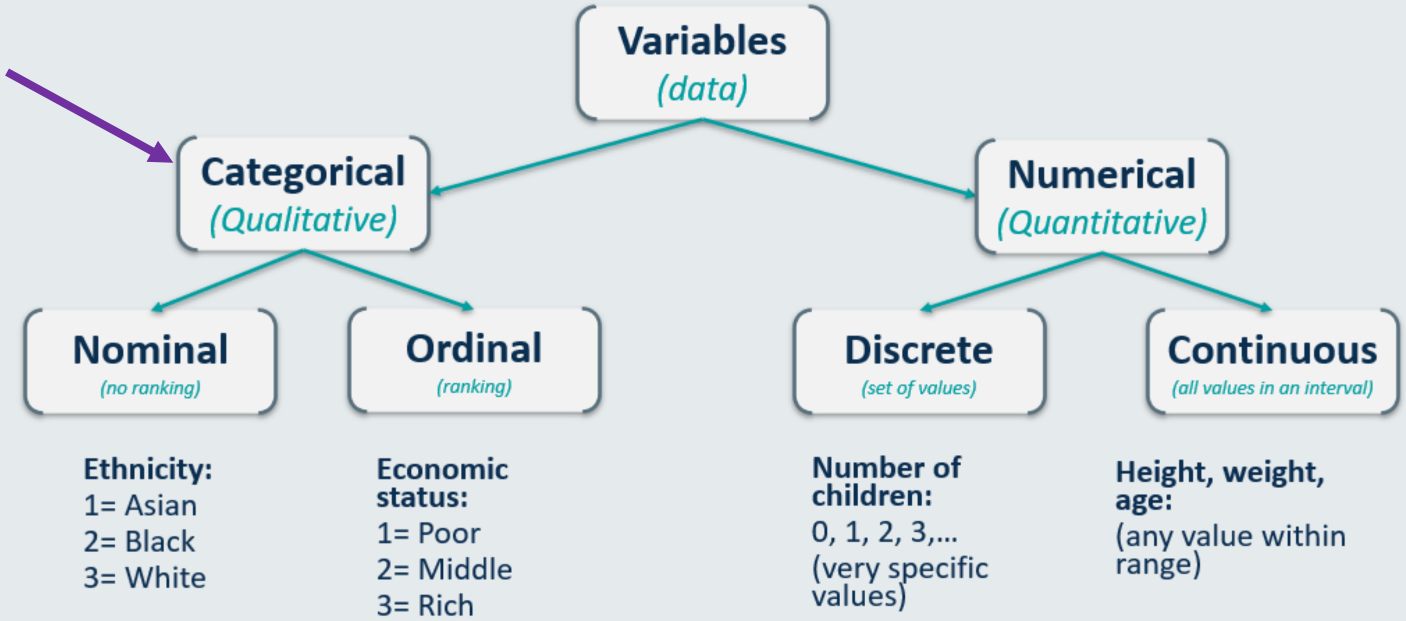
	Frequency	Percent
Black	11	13.8
White	20	25.0
Asian	17	21.3
Mixed	18	22.5
Other	14	17.5

only for nominal data

In a pie chart, the size of the sector represents the frequency of each category. More people, more pie.

Types of Variables

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



- Descriptive indices — Frequencies (Percentages %)
- Charts/plots — Pie Chart (only for nominal)
  Bar Chart

Knowledge Check

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

Q1. Which of the variables would you describe using **frequencies (percentages %)**

Q2. Which of the variable(s) would you use a **pie chart?**

Knowledge Check

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
:	:	:	:	:	:	:	:	:	:

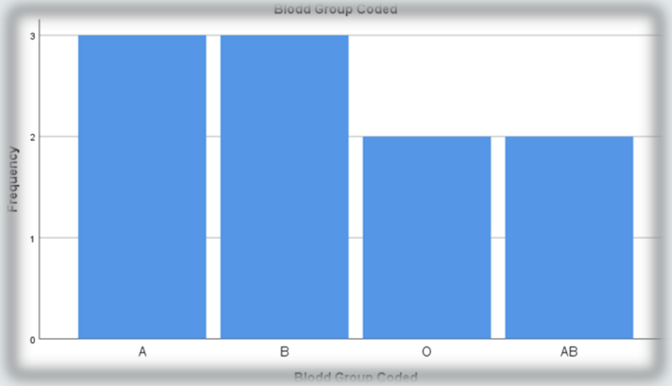
† LDL =Low Density Lipoprotein

Q3. Below is a frequency distribution for the variable social class give an interpretation of this information.

Social Class Coded

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	I	4	40.0	40.0	40.0
	II	1	10.0	10.0	50.0
	III	2	20.0	20.0	70.0
	IV	2	20.0	20.0	90.0
	V	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Q4. Below is a bar chart of the variable 'Blood Group' what does the chart show us?



Knowledge Check Solutions

Q1. Which of the variables would you describe using **frequencies (Percentages %)**

Blood Group, Gender, Feeling Happy, Smoke, Social class.

All of these variables are qualitative (categorical) variables and would be described by frequencies and percentages.

Q2. Which of the variable(s) would you use a **pie chart?**

You could use a pie chart or bar chart to visualise any of the above variables, but it may be more meaningful, visually, to do a pie chart for where we have more than 2 categories like blood group. For the ordinal variables is best to use the bar charts (feeling happy, social class)

Q3. Below is a frequency distribution for the variable social class give an interpretation of this information.

In our sample, half of the individuals were in social classes III to IV (N=6, 50%).

Q4. Below is a bar chart of the variable 'Blood Group' what does the chart show us?

The majority of subjects belong to blood groups A and B (N_A = 3, N_B = 3, 60%) with the rest of the subjects split evenly between blood groups O and AB (N_O = 2, N_{AB} = 2, 40%)

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 Inc. 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**

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

Institute of Psychiatry, Psychology and Neuroscience

Module Title: Introduction to Statistics

Session Title: Summarising numerical data

Topic title: Measurement and graphical representations of data

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdula

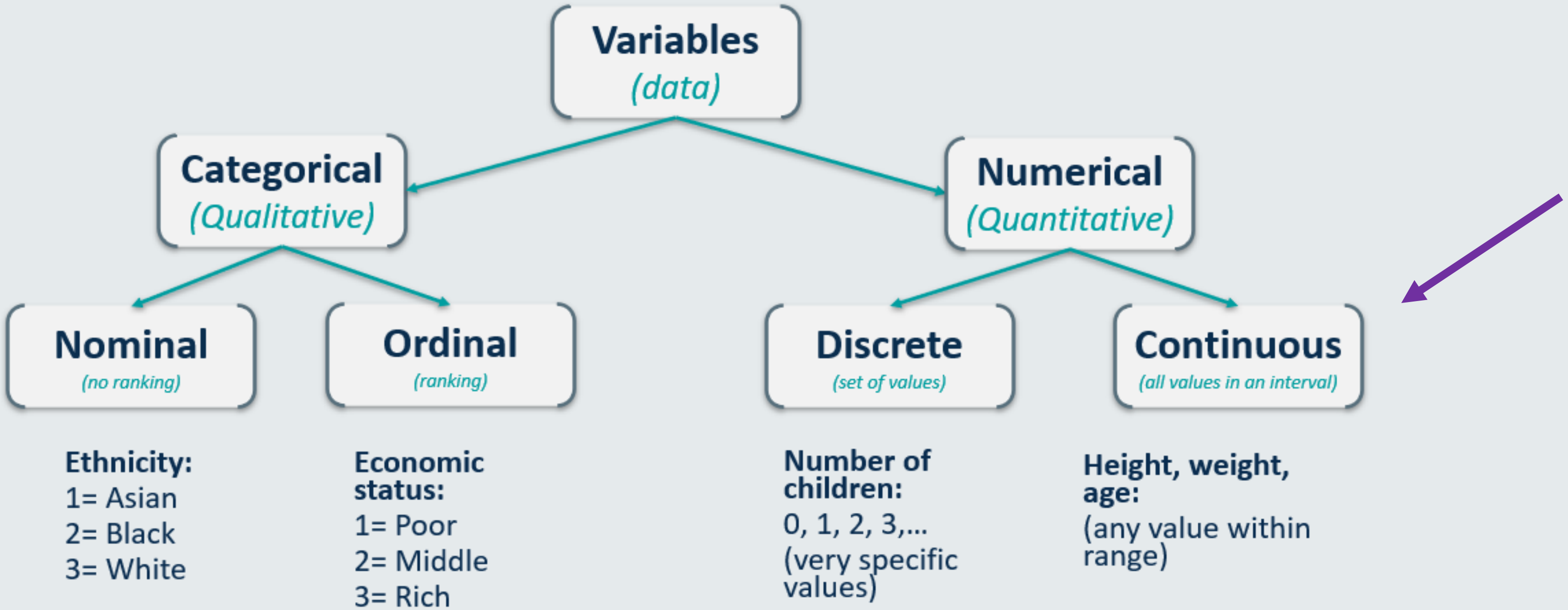
Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

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

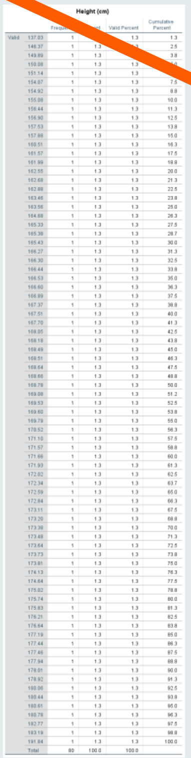
Based on the type of each variable, we use different ways to describe the data.

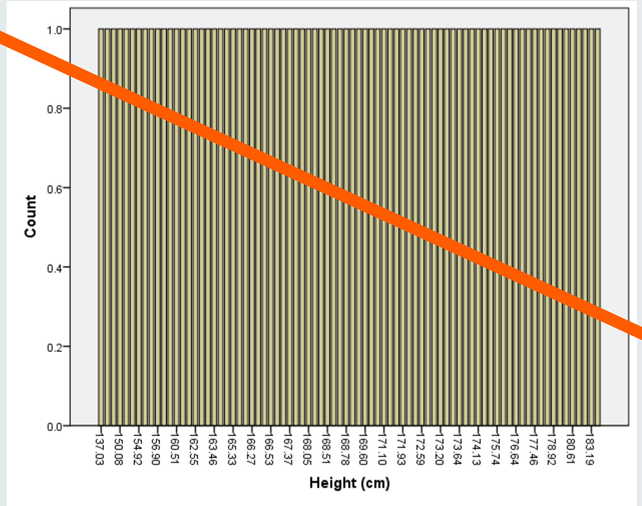


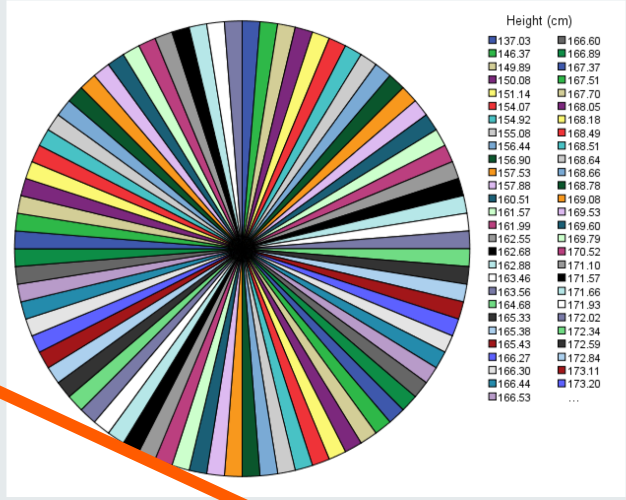
- Descriptive indices — Frequencies (Percentages %) — ?
- Charts/plots — Bar Chart — ?

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.

$$\frac{37+41+18+21+17+86+31+33+21+55}{10} = 36 \text{ years old on average}$$

$\updownarrow$

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

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**	**Observations (x_i)**
+1	**+5**	**-18**	**-15**	**-19**	**+50**	**-5**	**-3**	**-15**	**+19**	**Distance ($x_i - \bar{x}$)**
1	**25**	**324**	**225**	**361**	**2500**	**25**	**9**	**225**	**361**	**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…

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

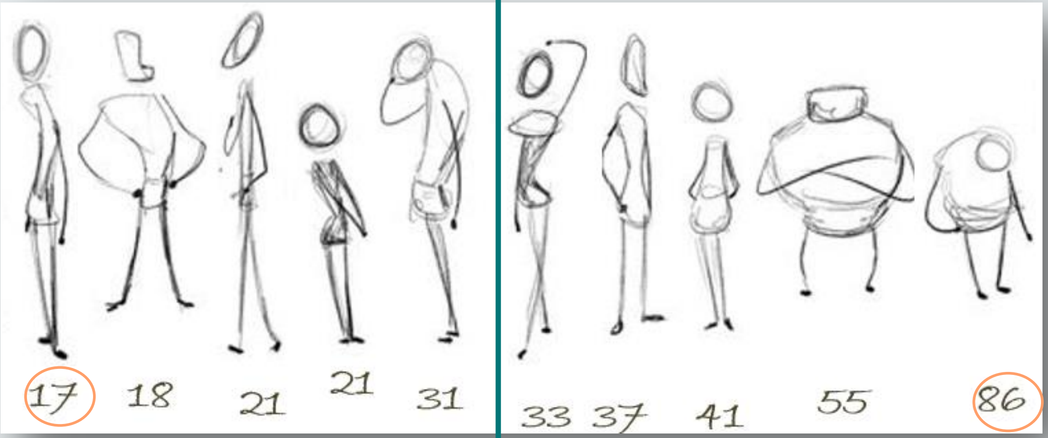
The standard deviation (SD) is
how spread out a group of numbers is from the mean,
by looking at the square root of the variance…

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

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.



32

50% of the people younger

(31+33)/2

50% of the people older

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)

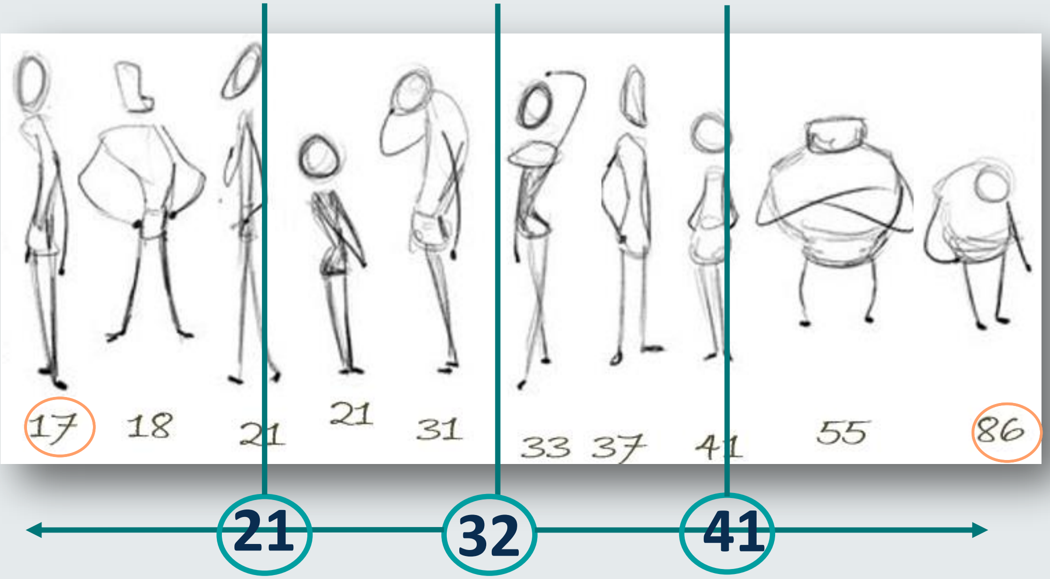
Median = 32

Minimum = 17, Maximum = 86

Range = 86 – 17= 69

Quantitative (Numerical) Data

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



Min

Q1: Lower quartile

Q2: Median

Q3 Upper quartile

Max

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
	150.08	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
	155.08	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
	162.88	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
	168.19	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
	168.66	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
	173.38	1	1.3	1.3	70.0
	173.48	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
	173.81	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
	175.02	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
	178.92	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)

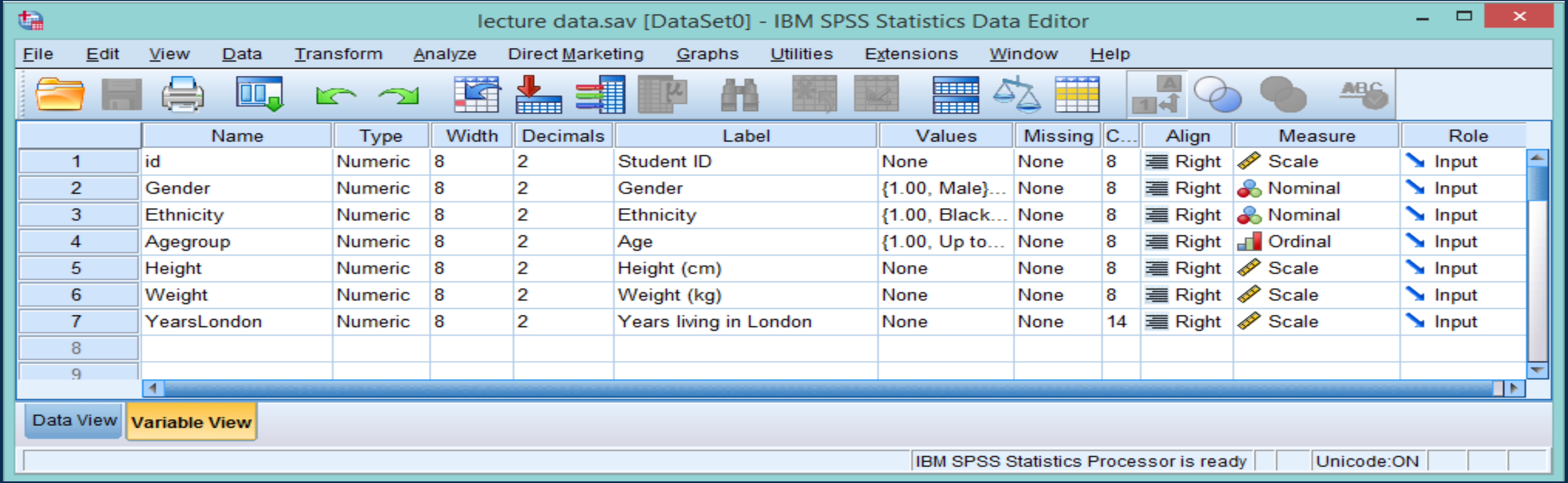
- Mean value: the **average** height of the students was 168.5cm (5.5ft)
- Median value: half of the students **were taller** than 169cm (5.5ft)
- Mode value: the height **most often** reported was 173cm (5.7ft)

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											

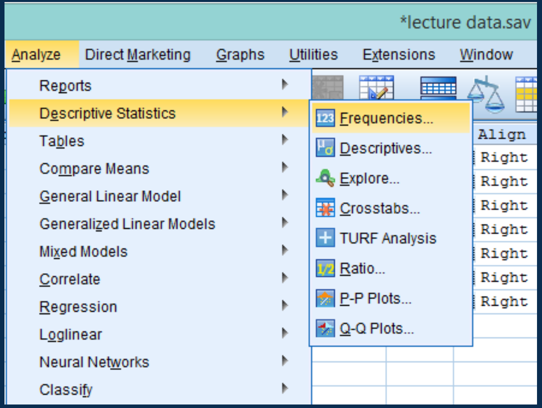
SPSS Slide: ‘How to’ Steps

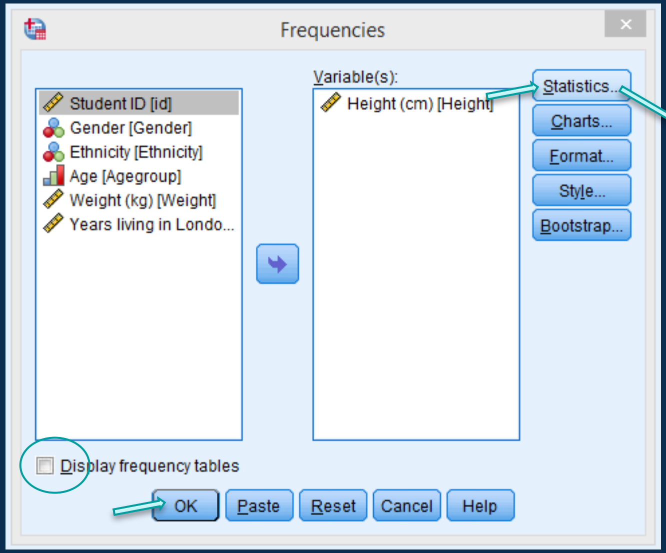
You can create the descriptive indices for height using the following steps:

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

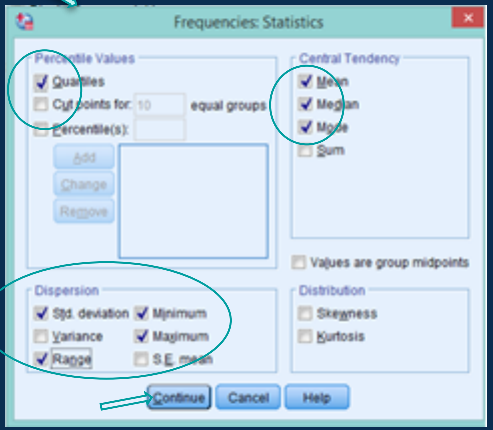
Add the variable of interest (height) into the ‘Variable(s)’ box

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’

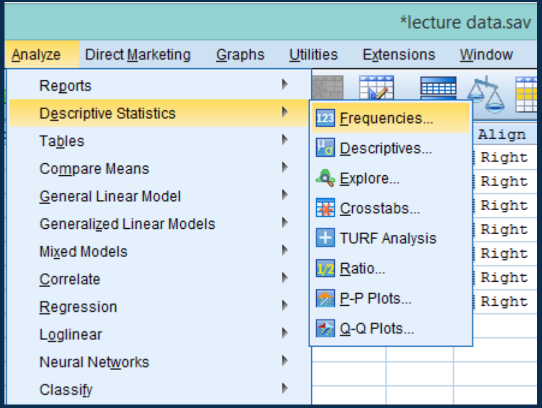
Click on ‘OK

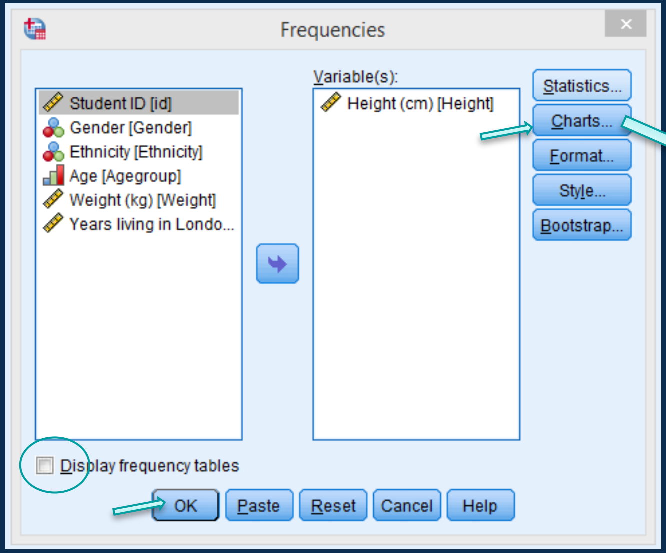
SPSS Slide: ‘How to’ Steps

You can create a chart using the following steps:

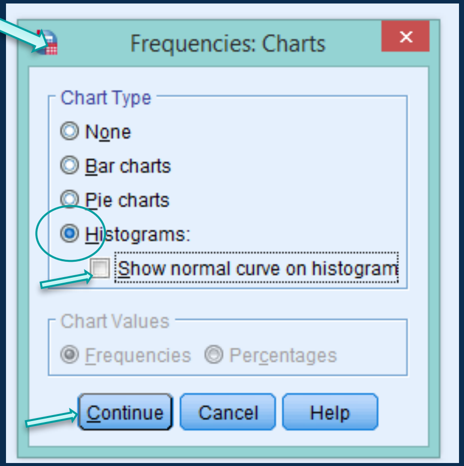
Click on the ‘Analyse Tab’→ ‘Descriptive Statistics’ → ‘Frequencies’

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

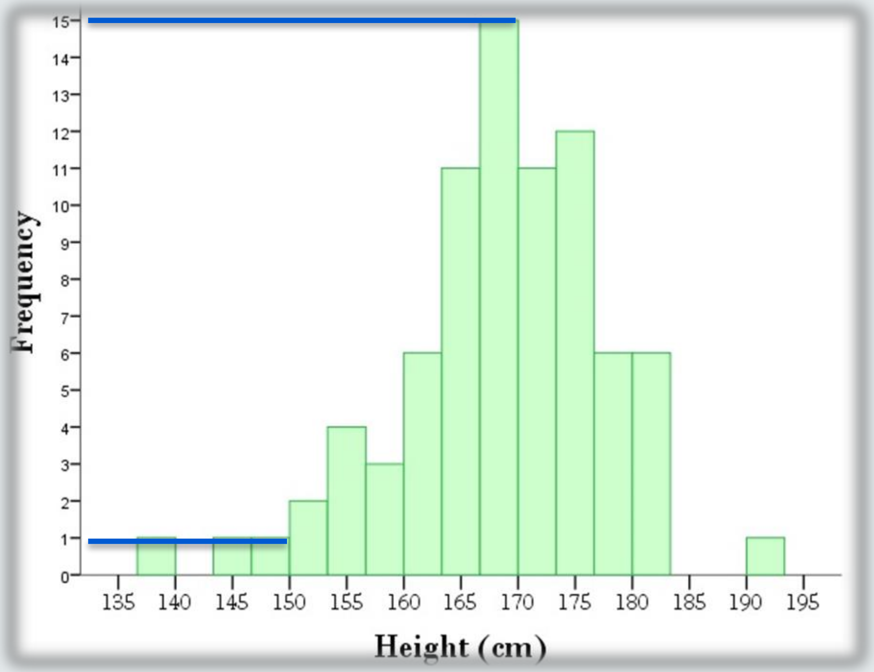
Describing Quantitative (Numerical) Data using Charts

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



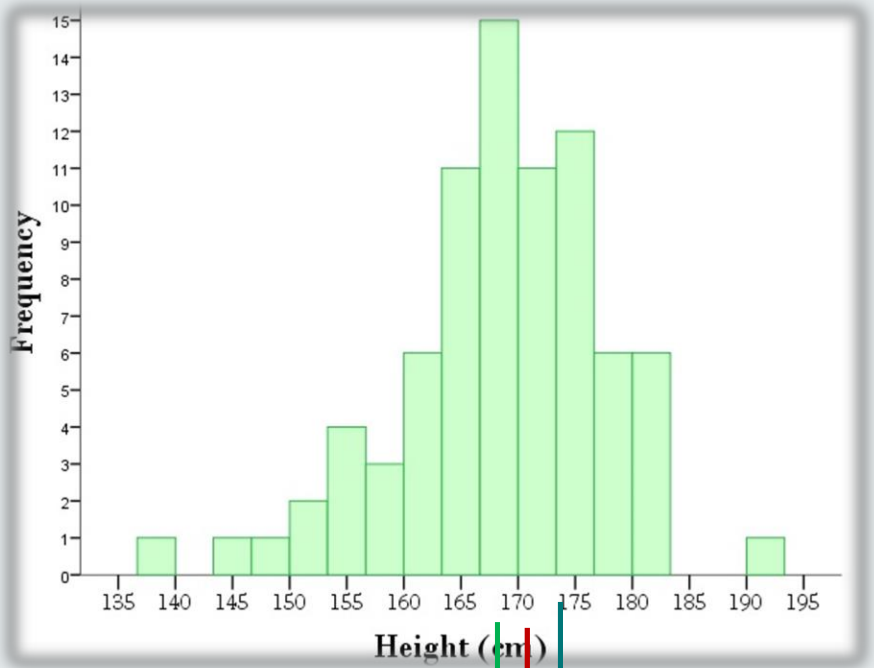
Describing Quantitative (Numerical) Data using Charts

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



mode

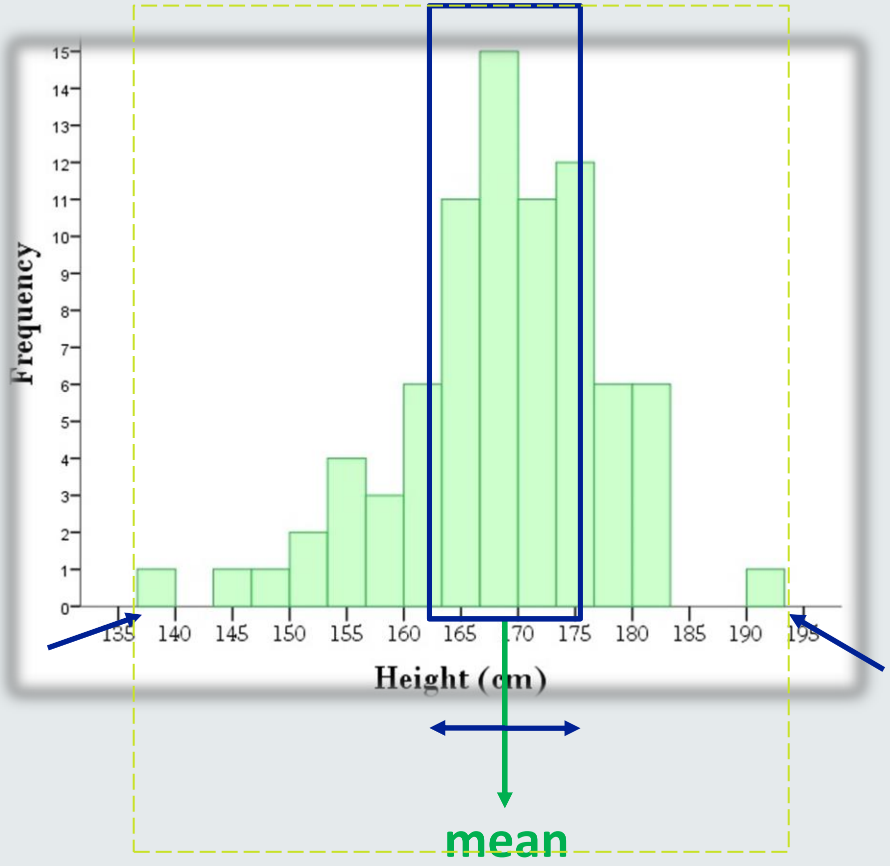
median

mean

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

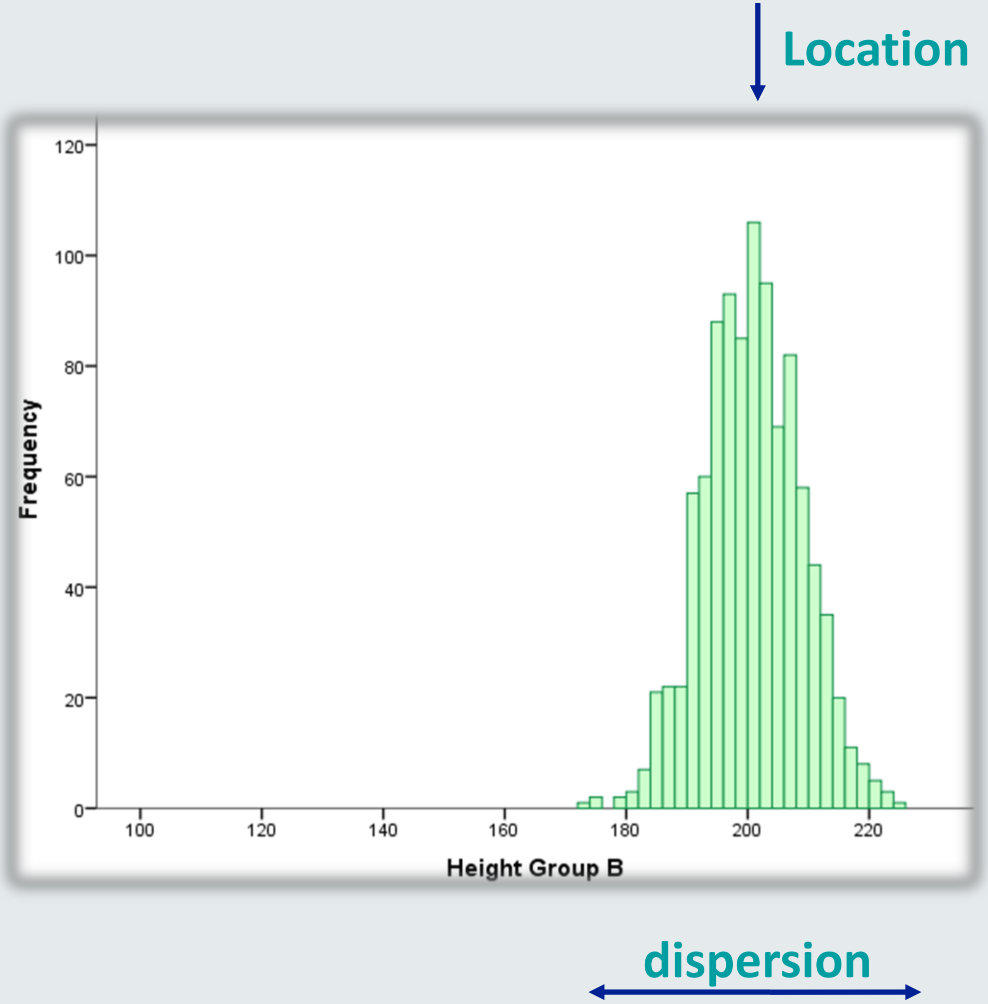
Height (cm)

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?

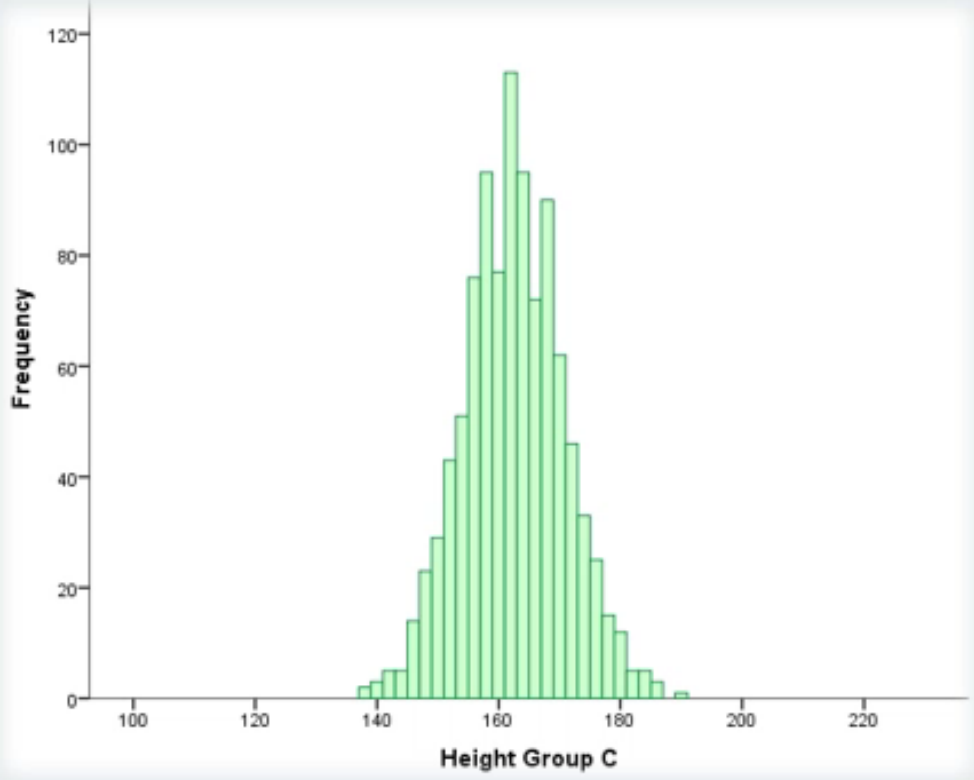


Mean = 200cm
SD = 8cm

Describing Quantitative (Numerical) Data using Charts

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

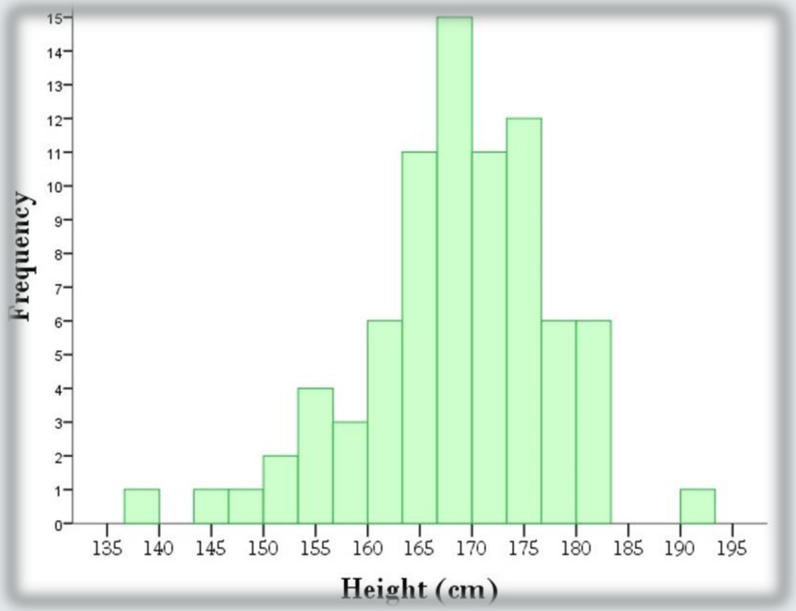
Location



dispersion

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.

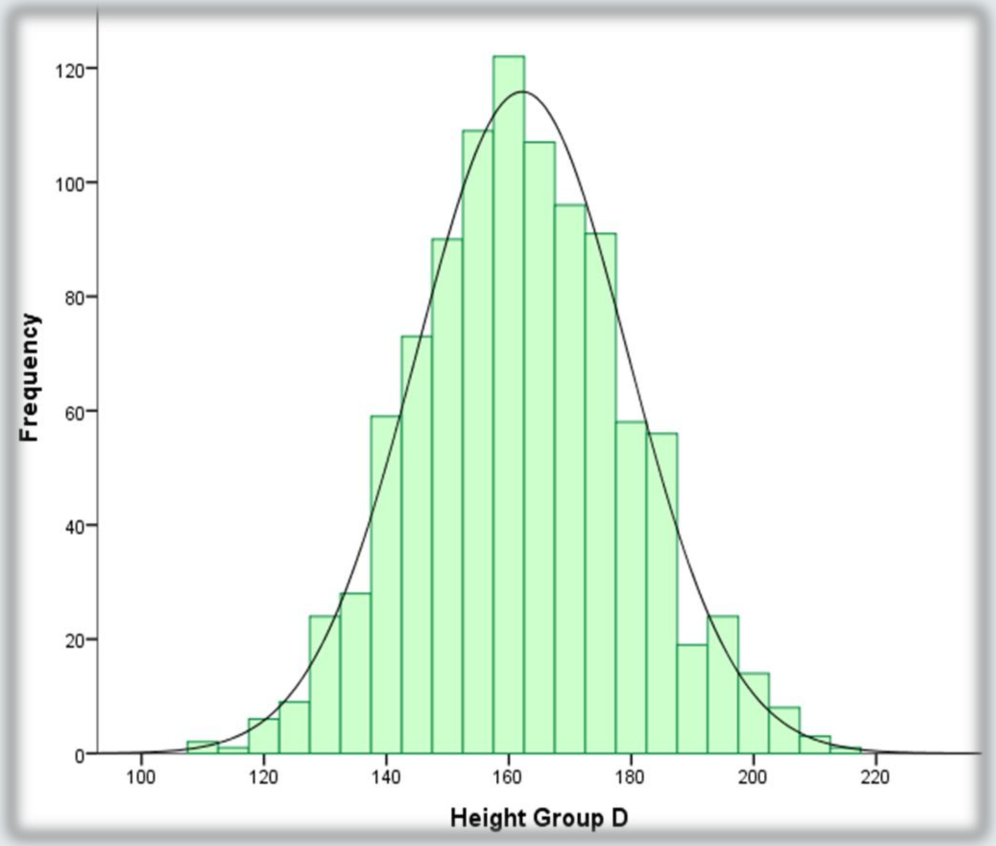
The average height was 168.53cm (SD=9.15cm).

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

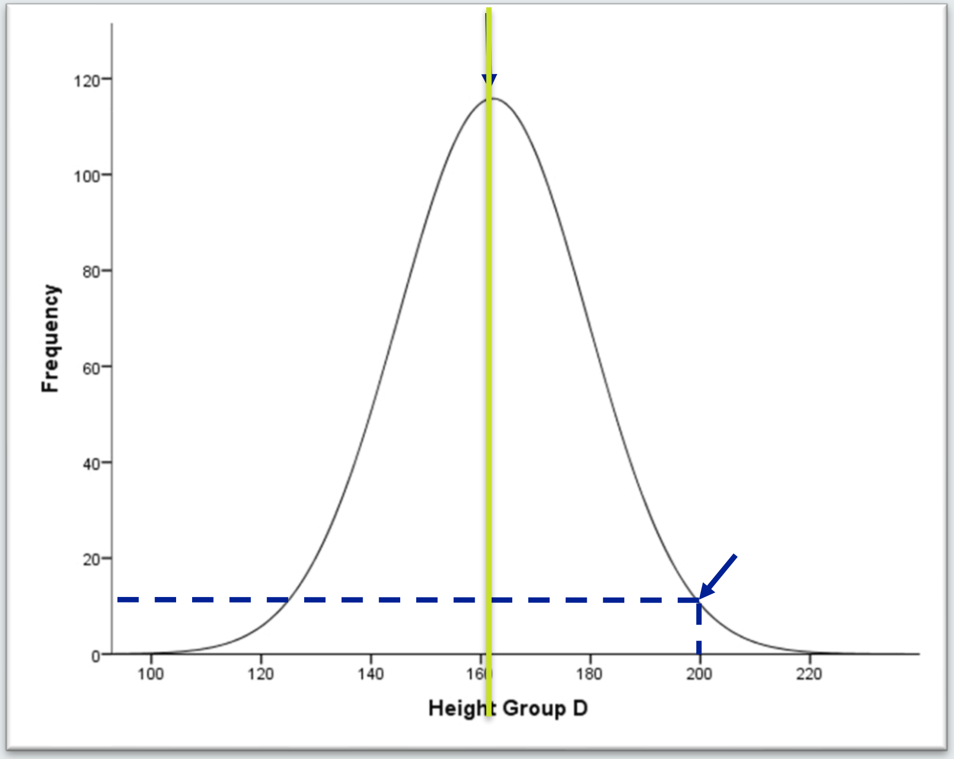
That is, the curve of a normal distribution with the same mean and standard deviation as our data…

The Normal Curve

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

Mean (and median and mode)

Mean = 162
Sd = 28



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

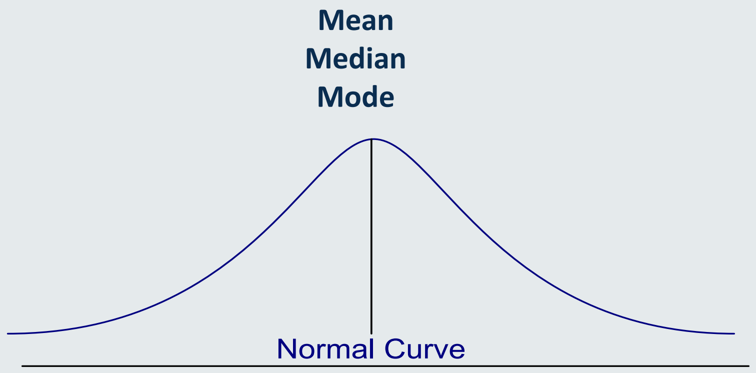
18 294 21 285 26 23 17 18 16 20 25 23 17

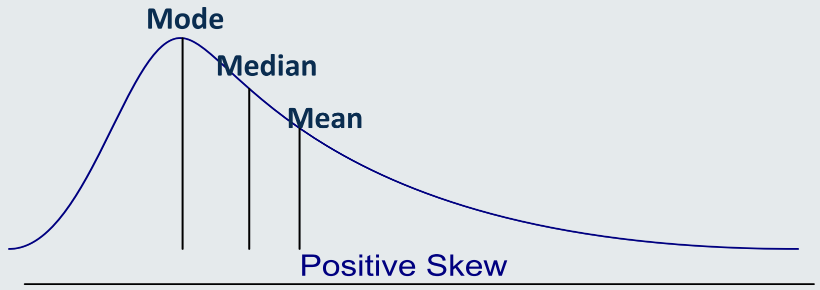
Mean £62K, SD = £101K
Median = £21K
min = £16K / <u>max =294K</u>

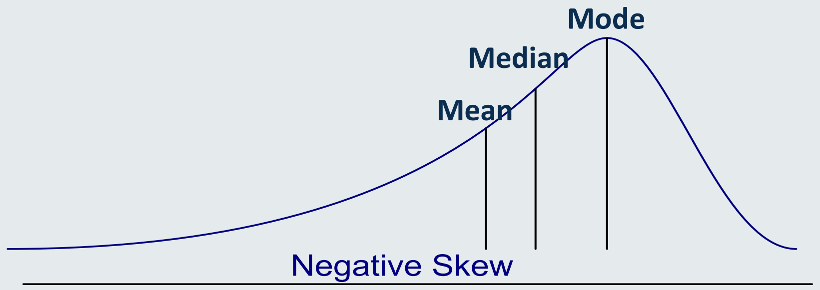
Describing Quantitative (Numerical) Data using Charts

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

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





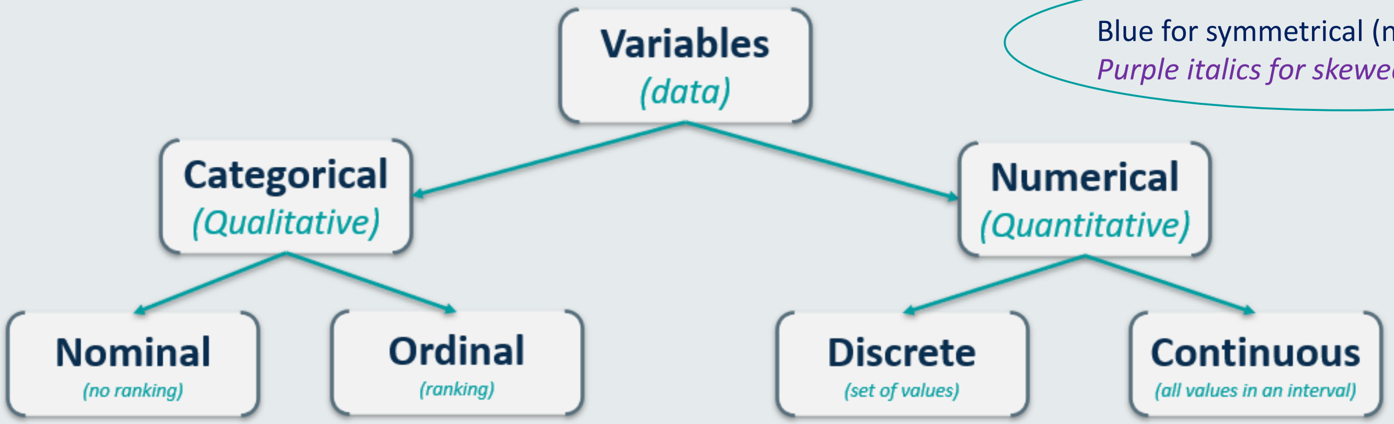


If the data are skewed, typically report on: **median and min-max and IQR**

Types of Variables

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

Blue for symmetrical (normal)
Purple italics for skewed or discrete data



Ethnicity:
1= Asian
2= Black
3= White

Economic status:
1= Poor
2= Middle
3= Rich

Number of children:
0, 1, 2, 3,...
(very specific values)

Height, weight, age:
(any value within range)

- 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

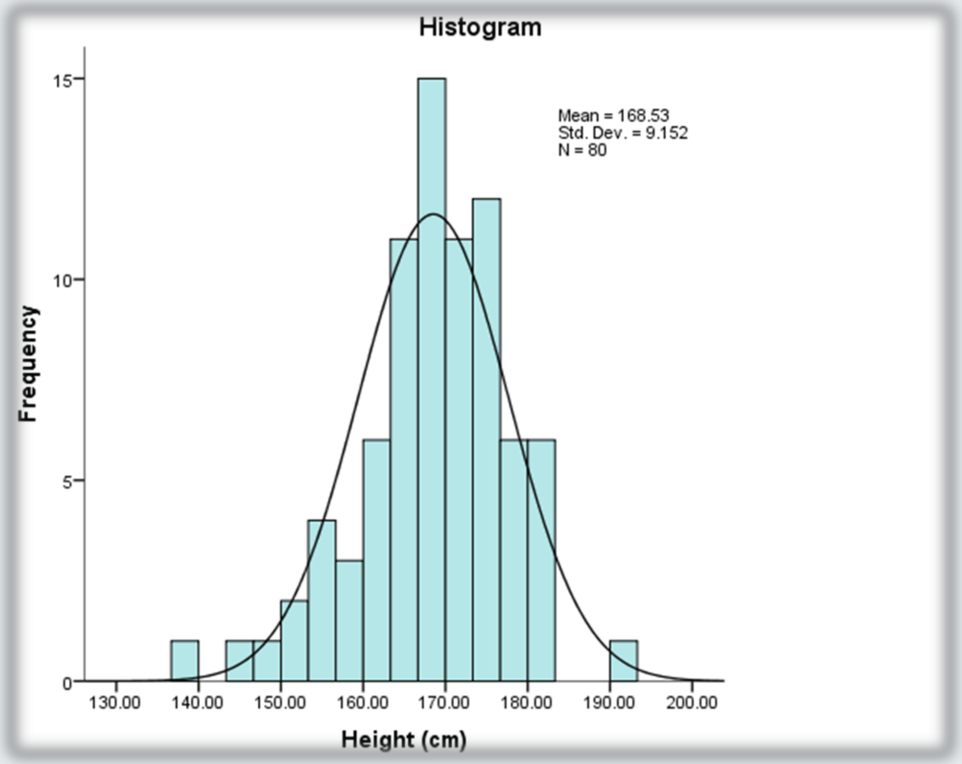
A chart has all the information we need and is easier to understand

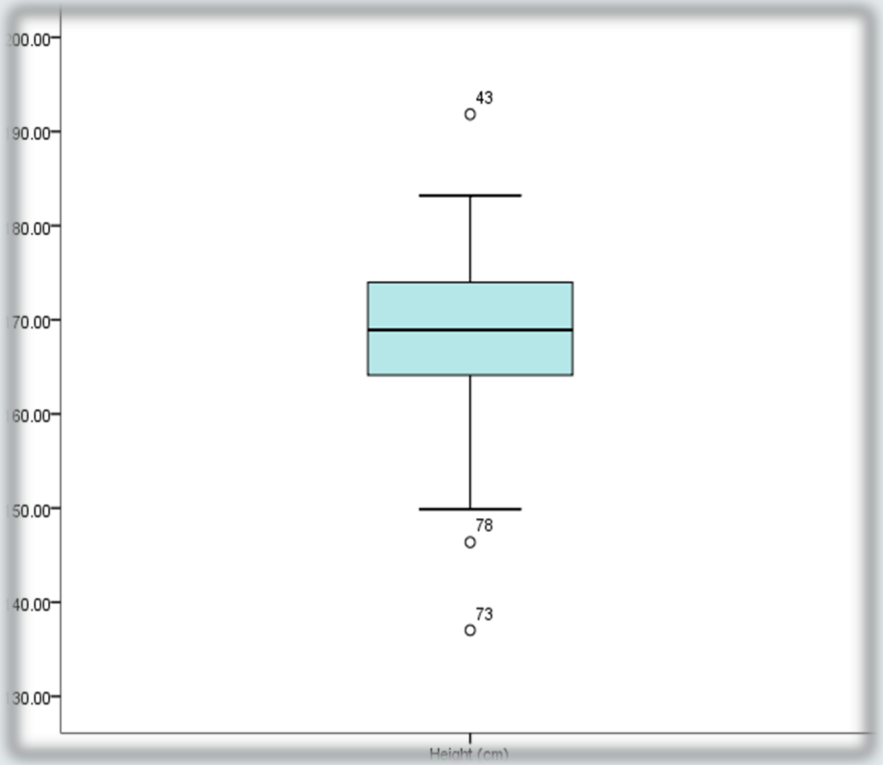
Statistics

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

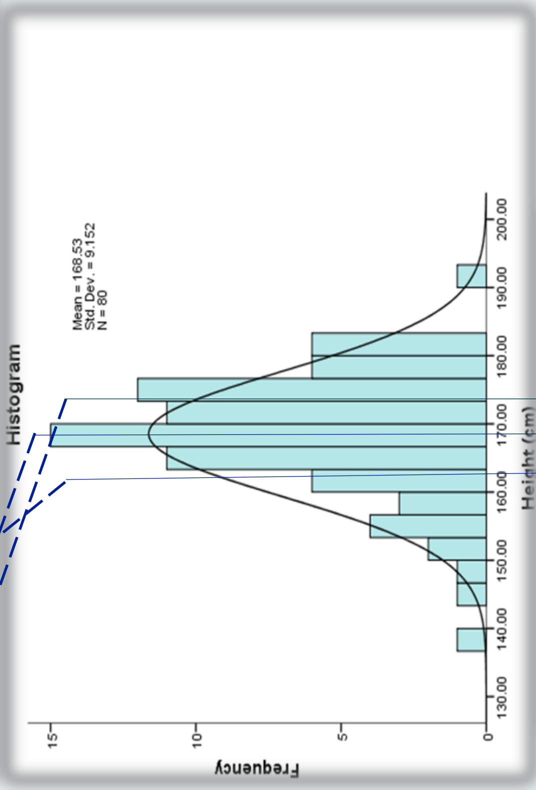
To describe a numerical variable, we need to properly **summarise** it.

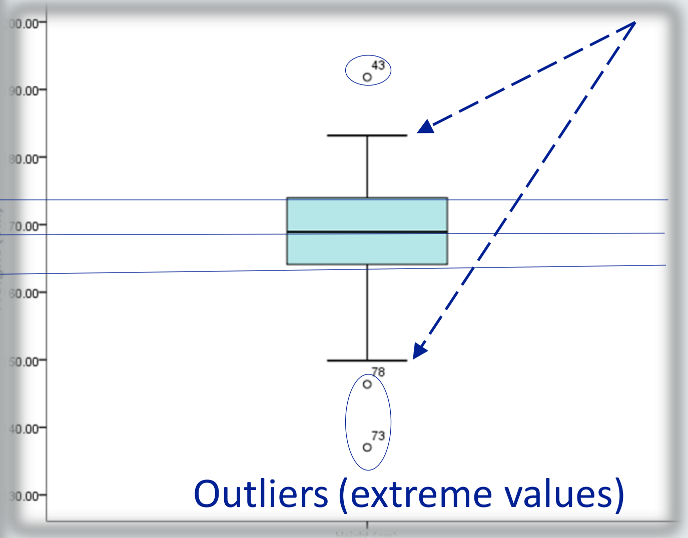
Statistics

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





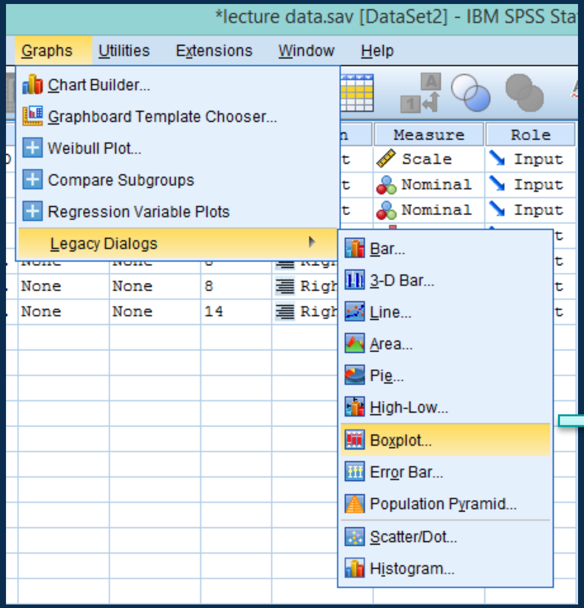
Min and max if outliers were omitted

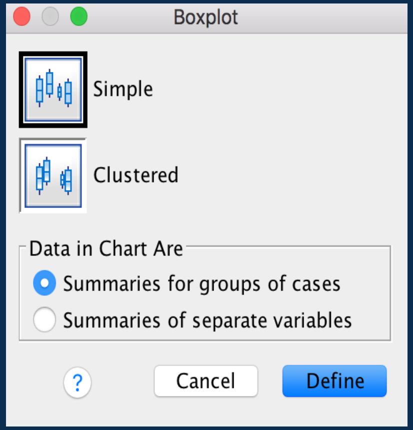
Outliers (extreme values)

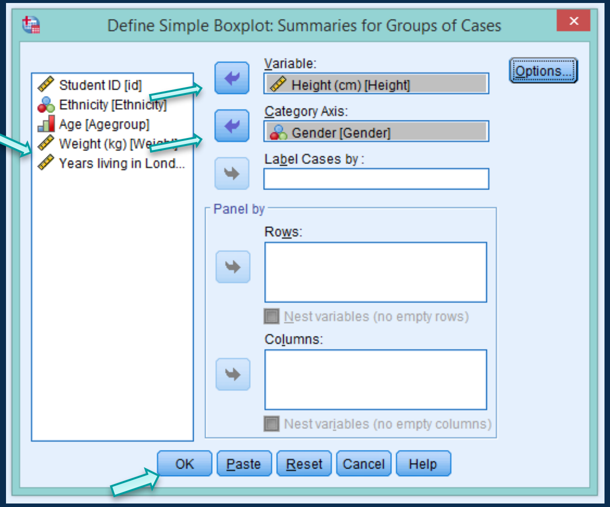
SPSS Slide: ‘How to’ Steps

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

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







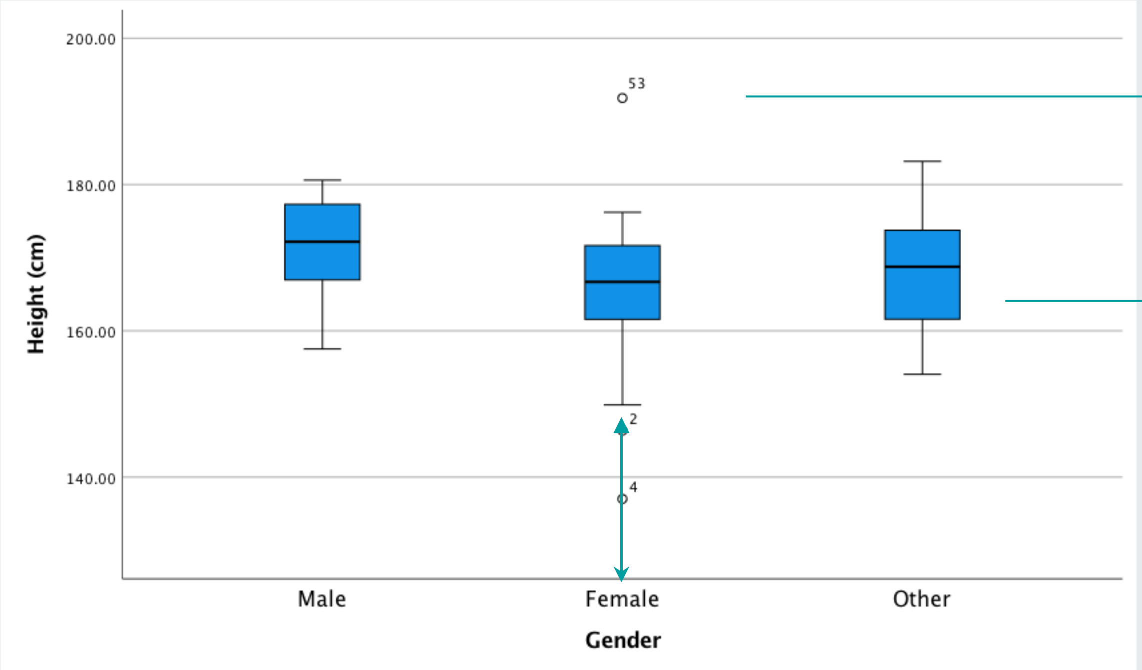
Choose a ‘simple’ layout and click ‘Define’

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.



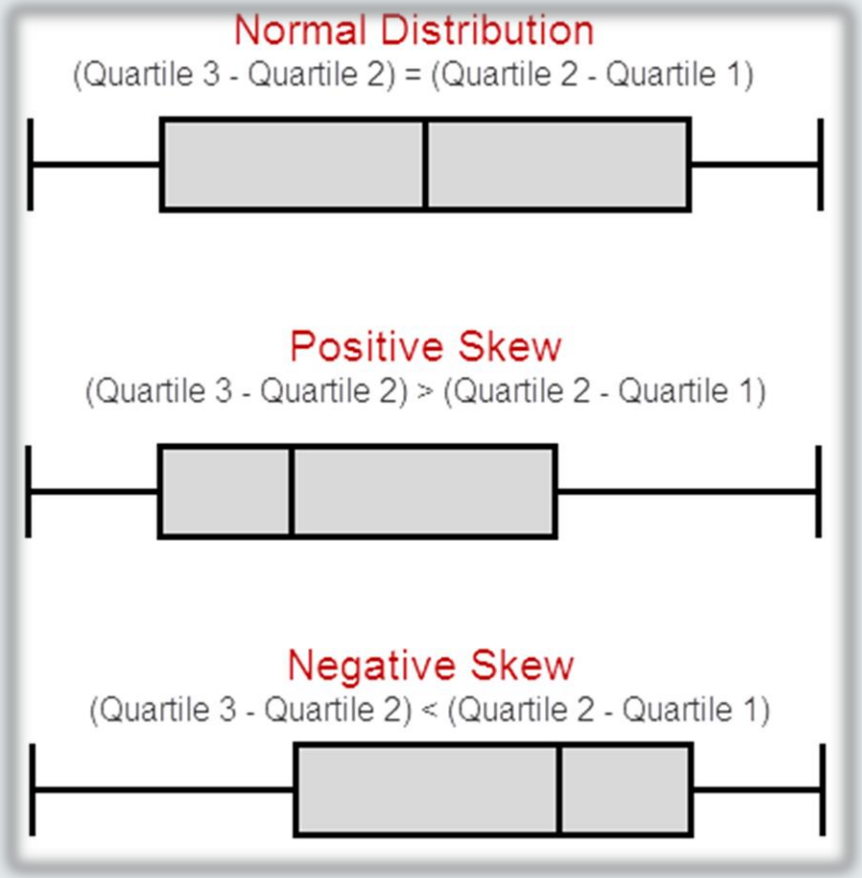
Outliers were ‘females’

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’.

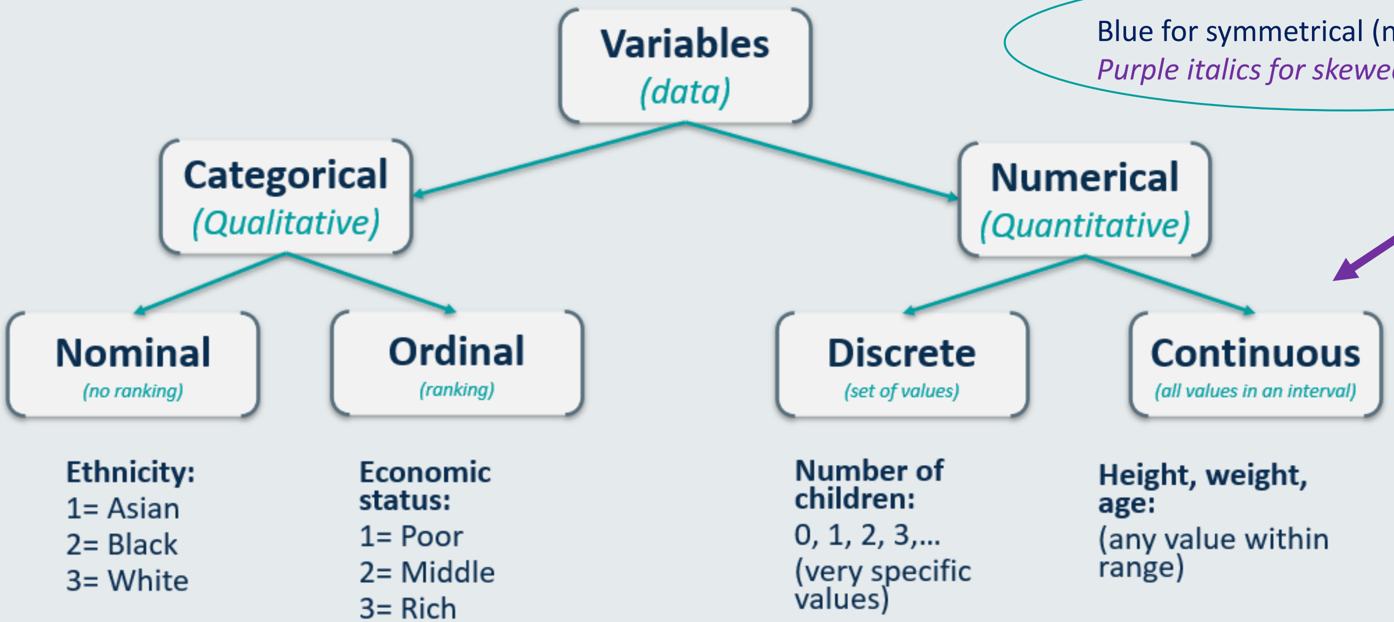
Describing Quantitative (Numerical) Data using Charts

Box Plots and skewness



Types of Variables

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



Blue for symmetrical (normal)
Purple italics for skewed or discrete data

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

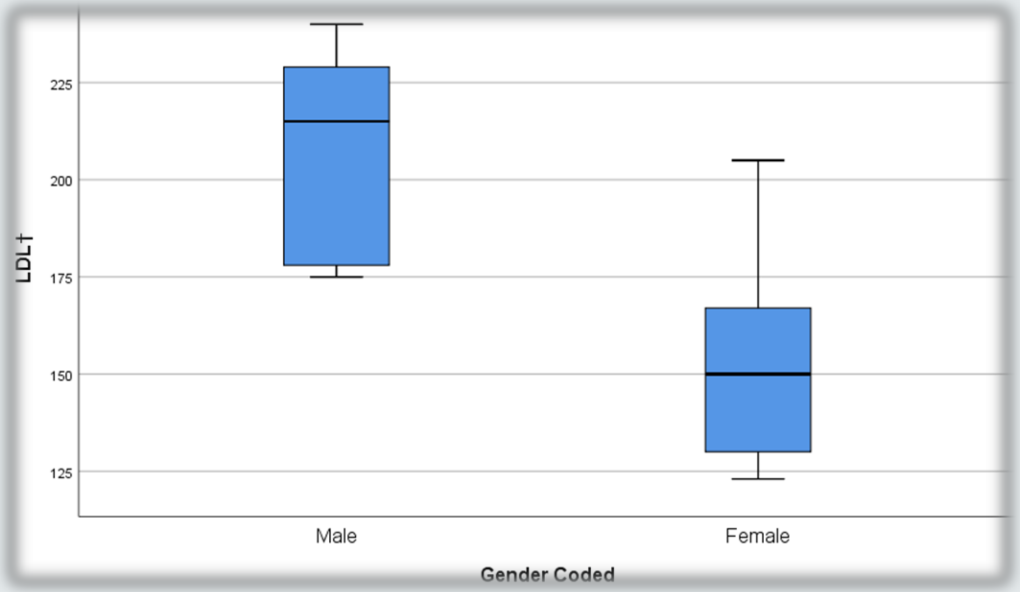
Note: *There appears to be quite a bit of difference between the mean, median and mode and may be indicative that the distribution for LDL may not be normally distributed. You would need to conduct a histogram or further tests for normality to check if this is the case. We can also see that the standard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

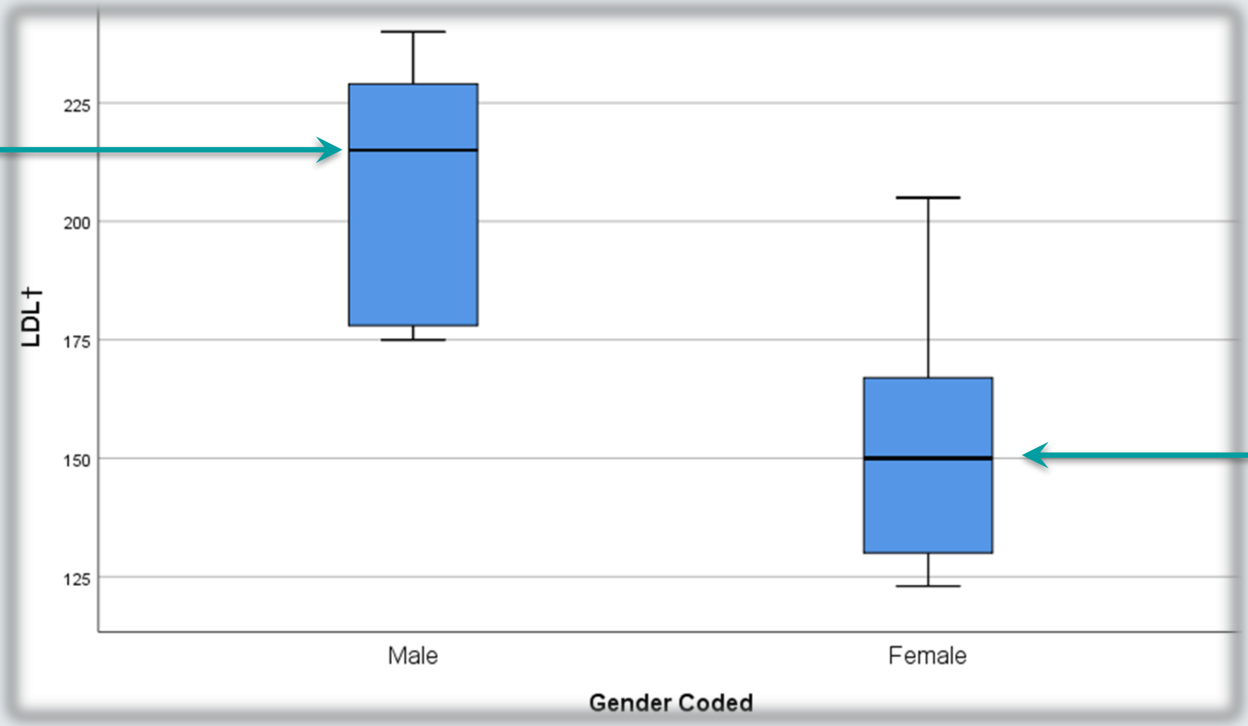
† LDL =Low Density Lipoprotein



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**