Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

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

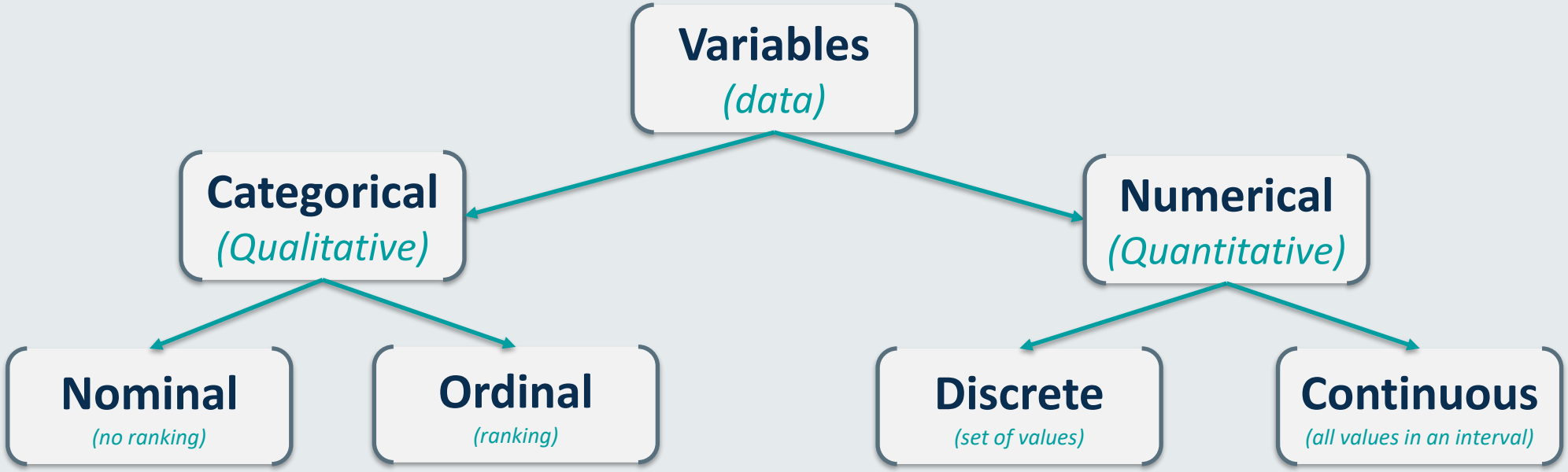
Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

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



Ethnicity:
1= Asian
2= Black
3= White

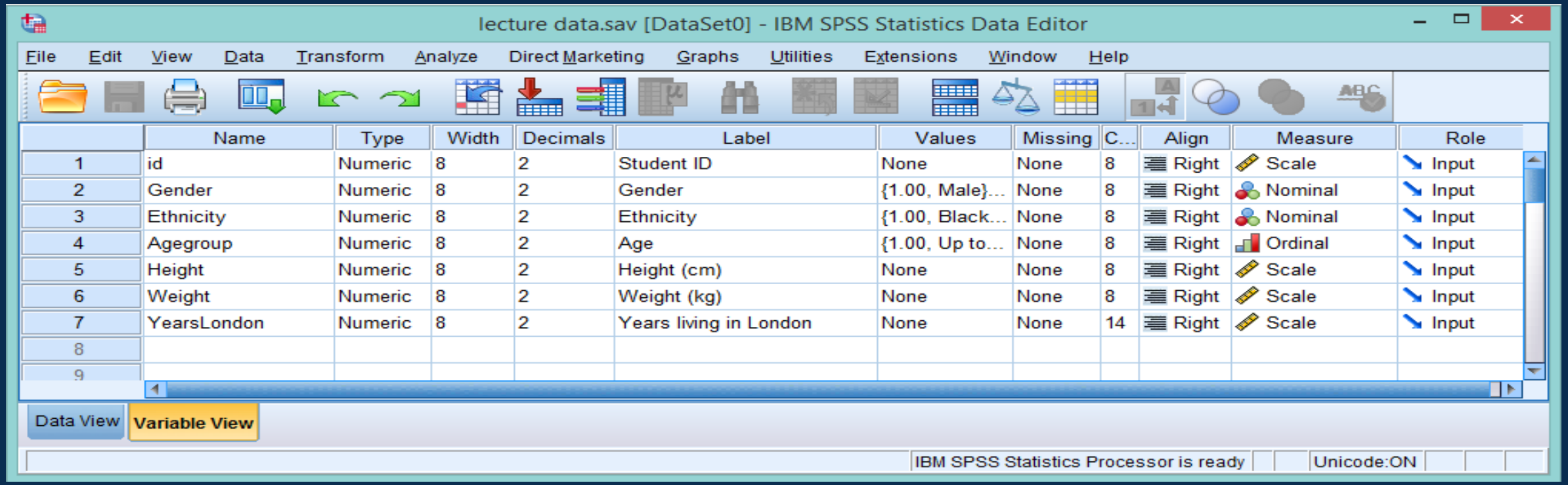
Economic status:
1= Poor
2= Middle
3= Rich

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

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

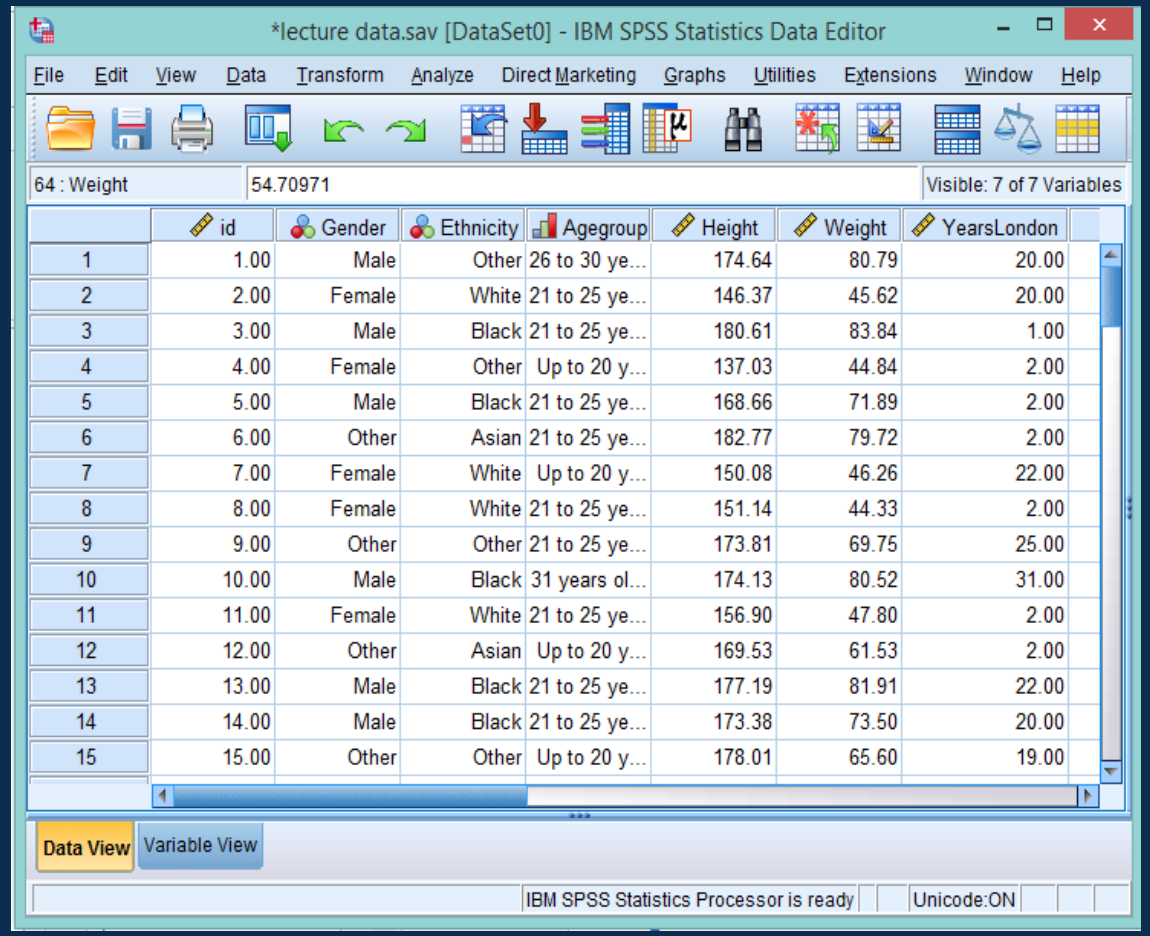
SPSS Slide

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

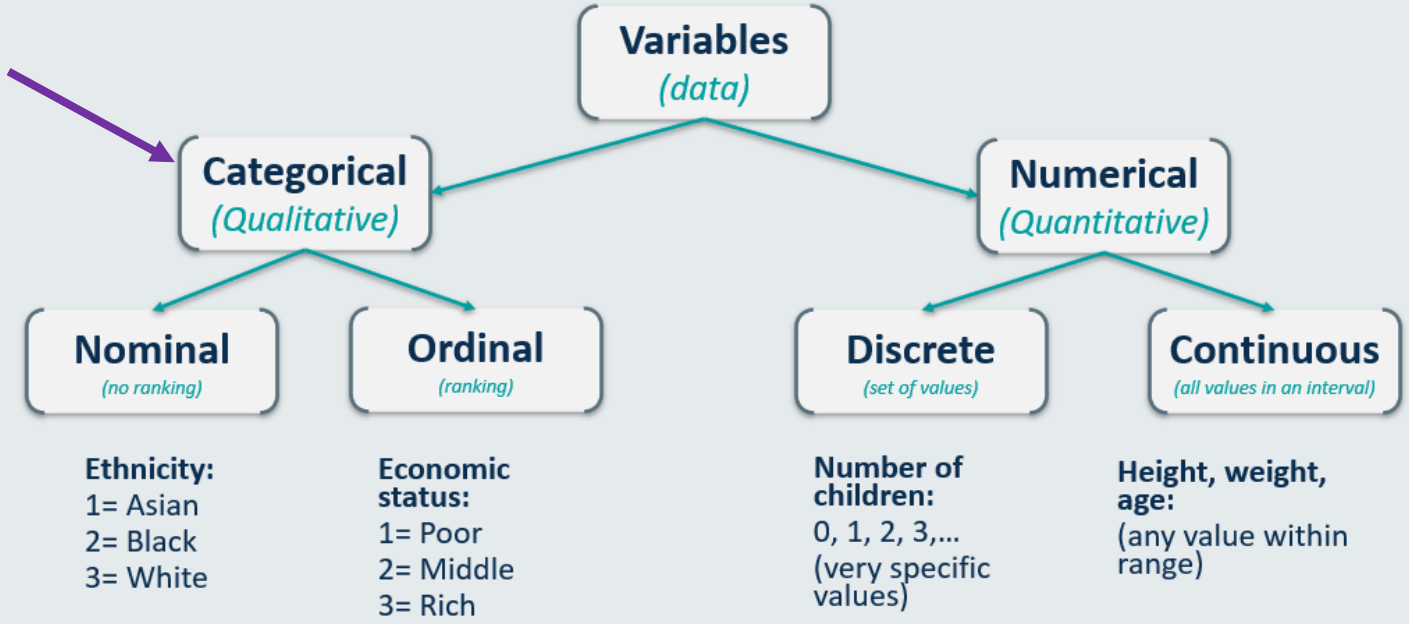
The dataset contains data from 80 students, with respect to their:

- 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 <u>how many</u> 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

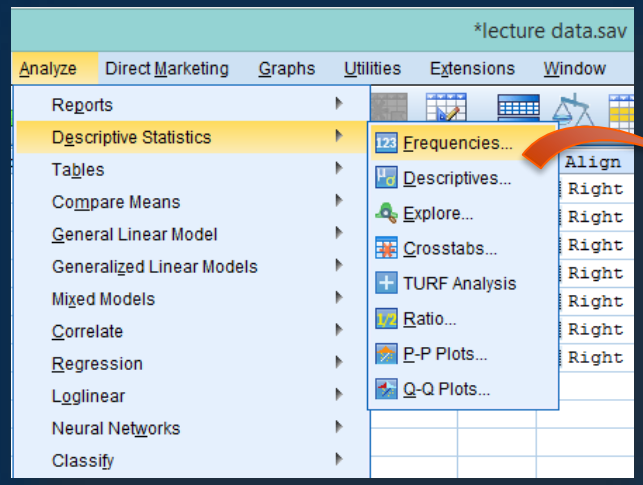
- 35% of the individuals <u>in the sample</u> (N=80) identified themselves as males
- 36.4% of the individuals <u>who responded</u> (N=77) identified themselves as males
- 75.3% of the individuals who responded (N=77) identified themselves as either males or females.

The cumulative % makes more sense in ordinal data

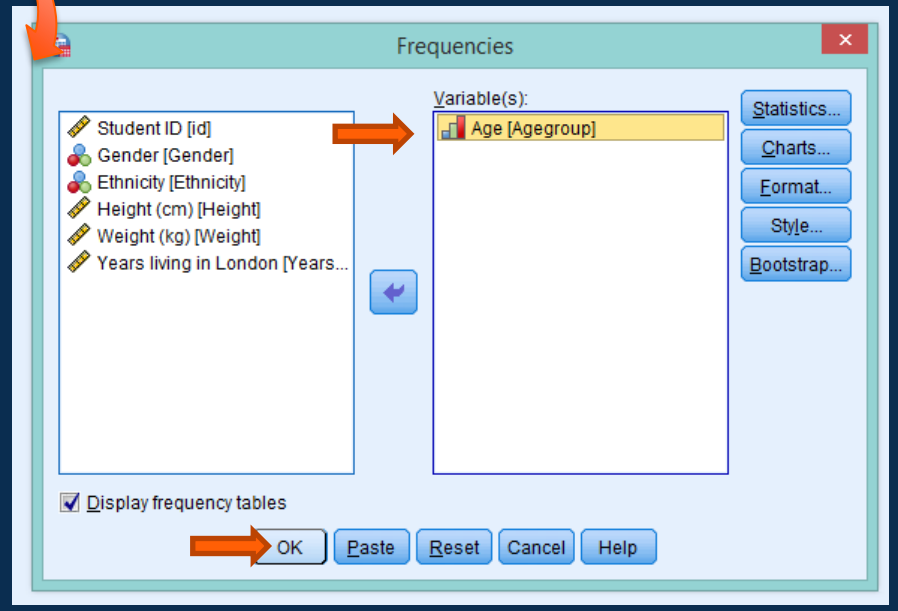
SPSS Slide: ‘How to’ Steps

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

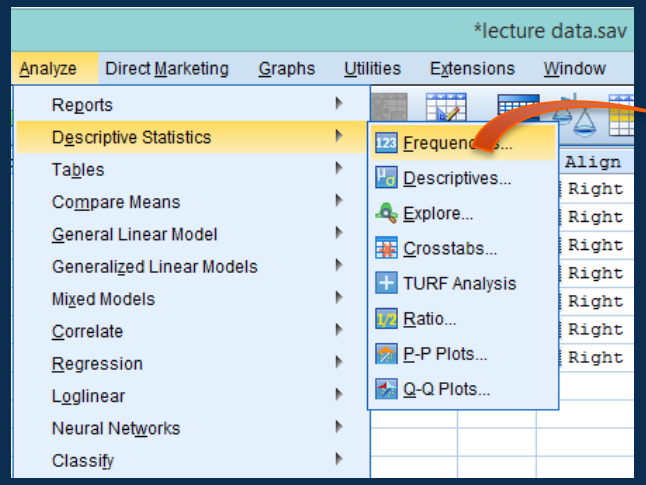
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Black	11	13.8	13.8	13.8
	White	19	23.8	23.8	37.5
	Asian	17	21.3	21.3	58.8
	Mixed	18	22.5	22.5	81.3
	Other	14	17.5	17.5	98.8
	22.00	1	1.3	1.3	100.0
	Total	80	100.0	100.0	

Typo spotted

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

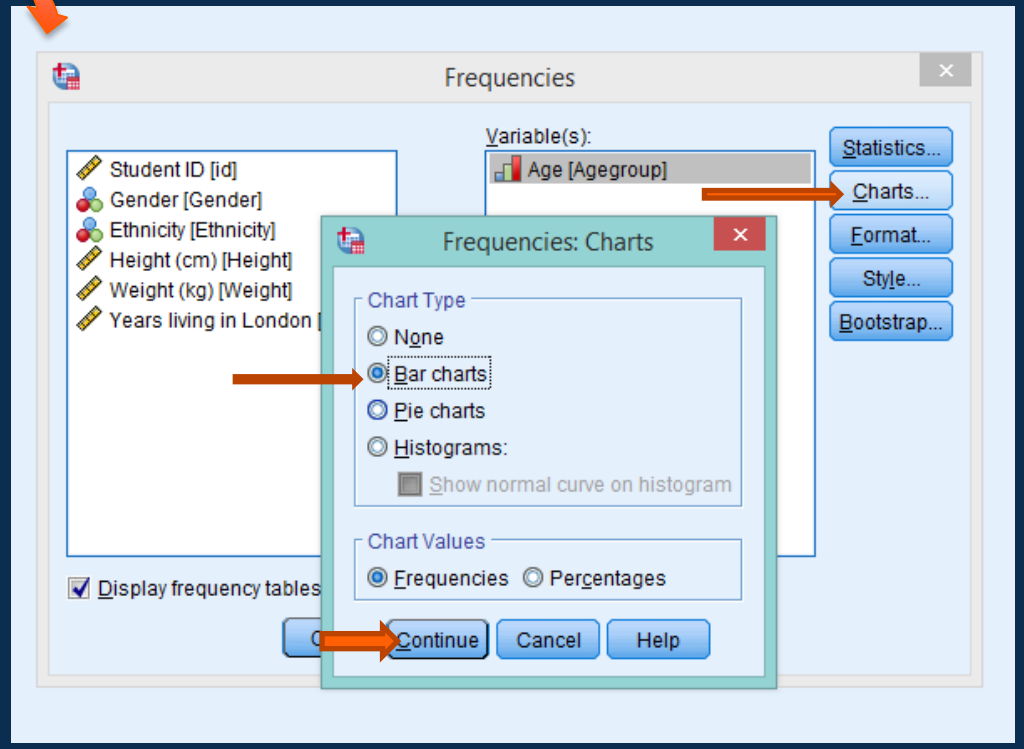
SPSS Slide: ‘How to’ Steps

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



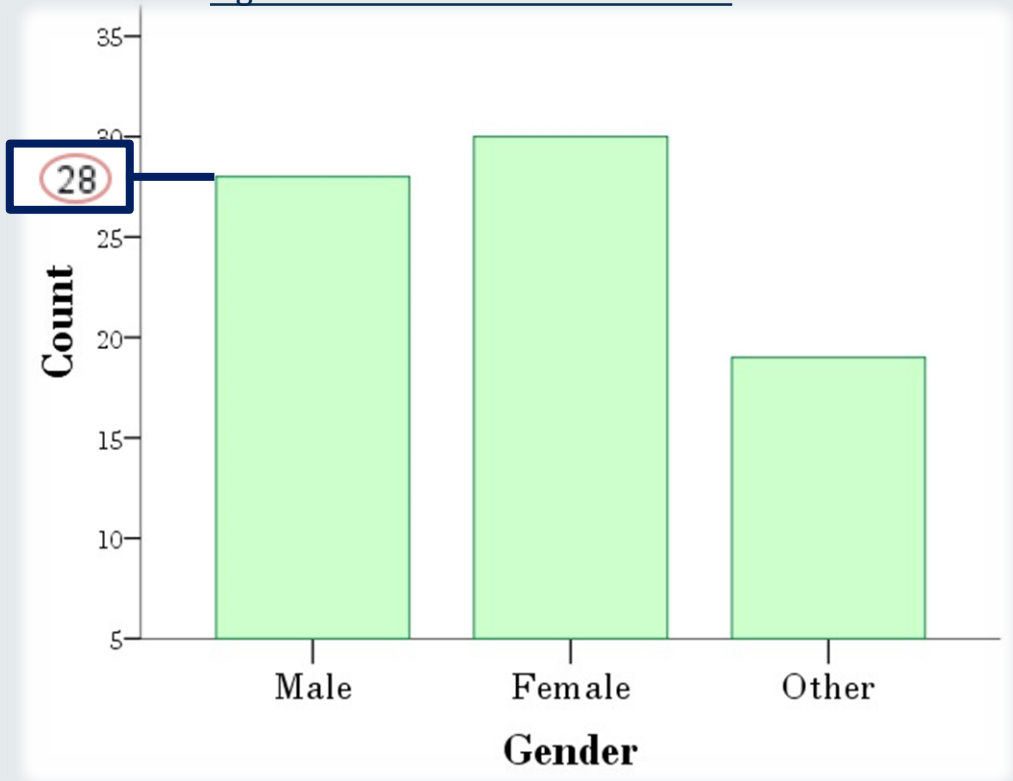
Click on ‘Charts’

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

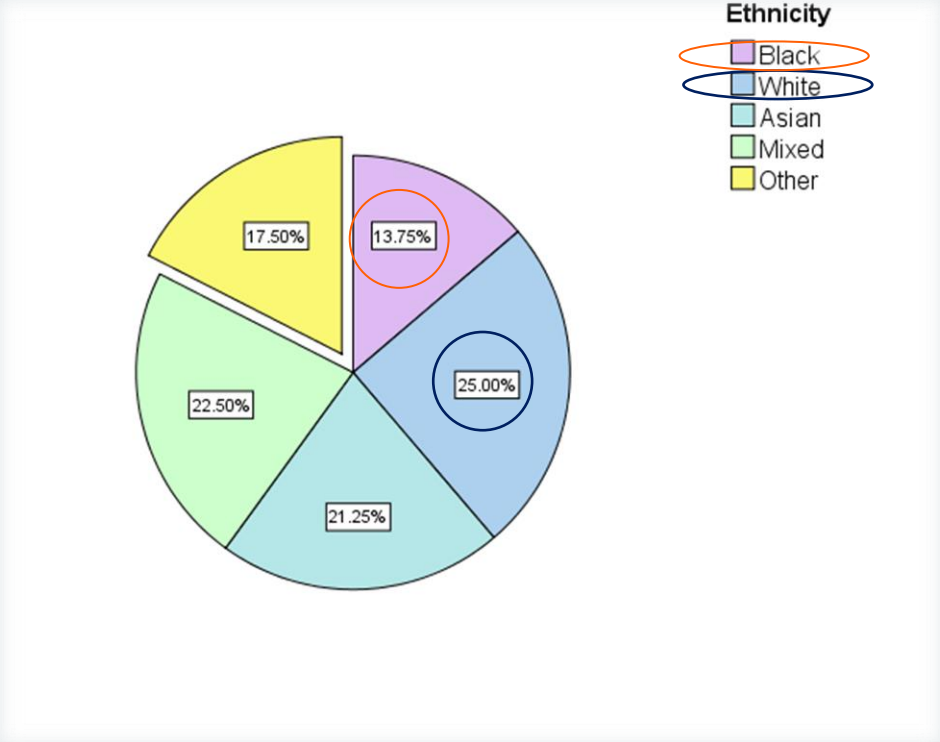
	Frequency	Percent
Male	28	35.0
Female	30	37.5
Other	19	23.8

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

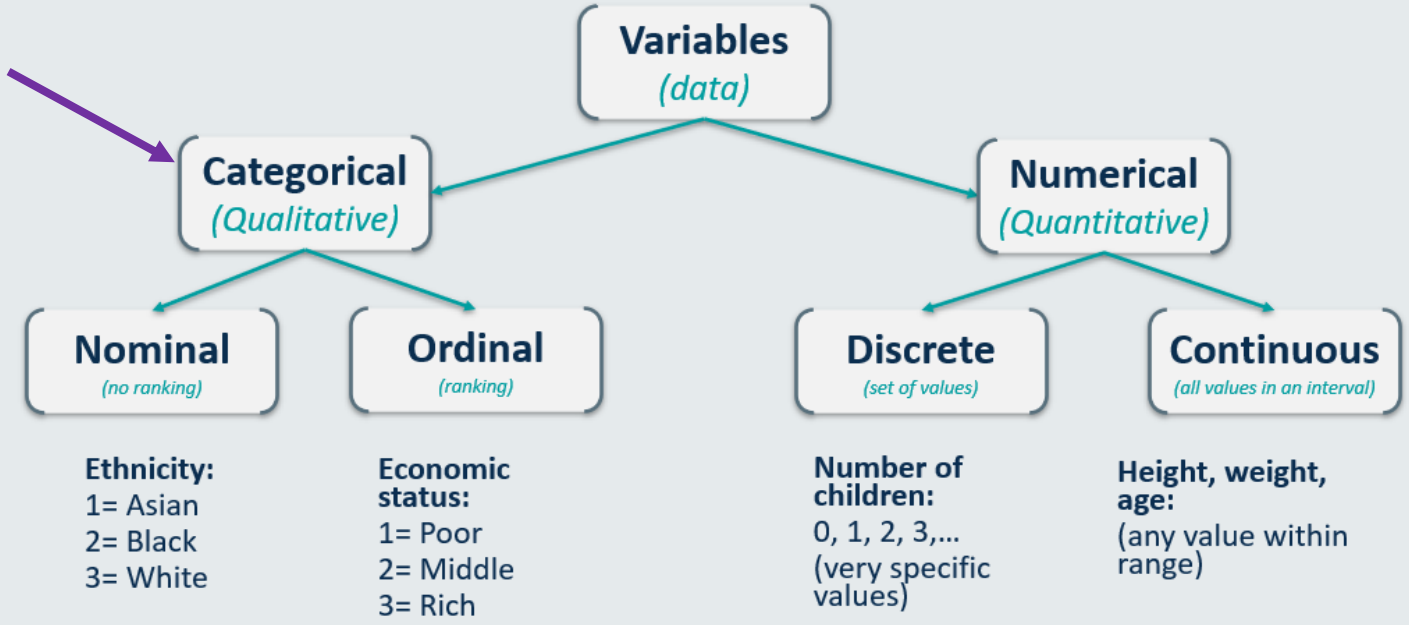
	Frequency	Percent
Black	11	13.8
White	20	25.0
Asian	17	21.3
Mixed	18	22.5
Other	14	17.5

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
:	:	:	:	:	:	:	:	:	:

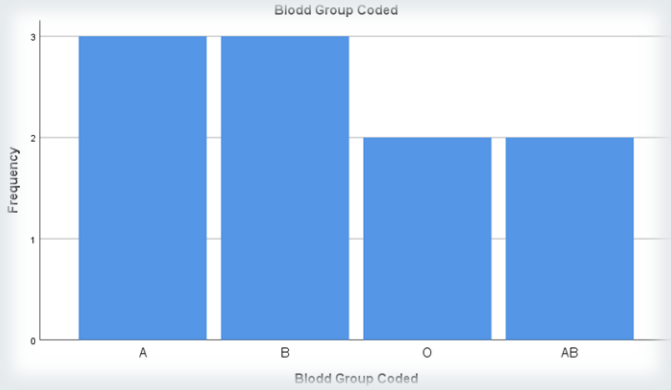
† LDL =Low Density Lipoprotein

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**