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: Types of data

Topic title: Measurement and graphical representations of data

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:

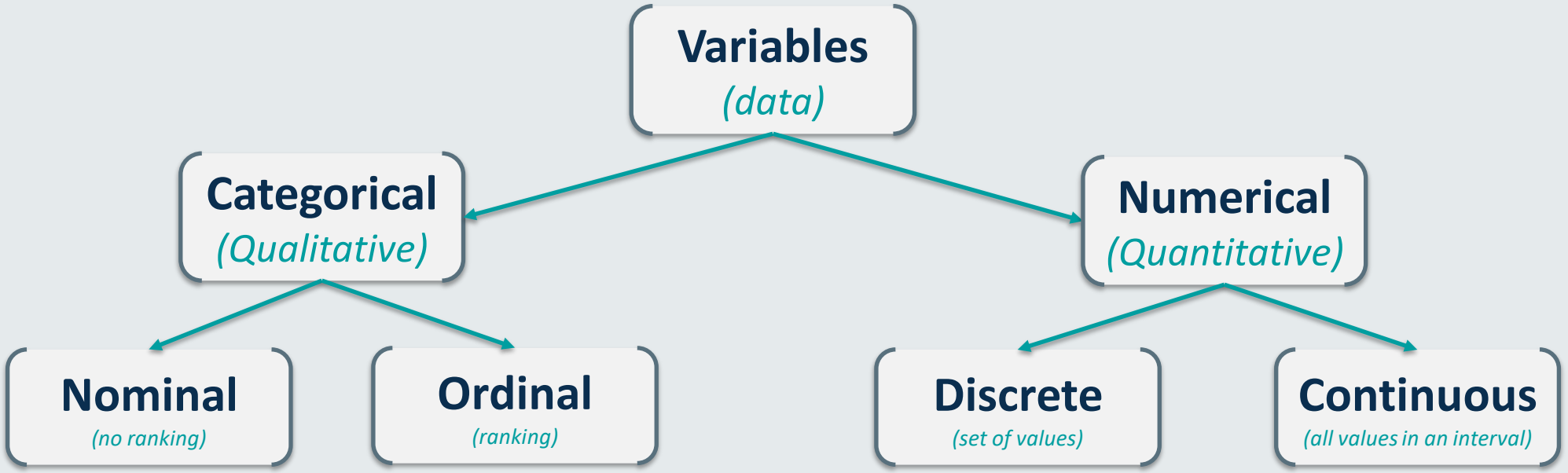
- make sure you do not have errors or typos in your data (**data cleaning**)
  You need to make sure that your data are cleaned, otherwise

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

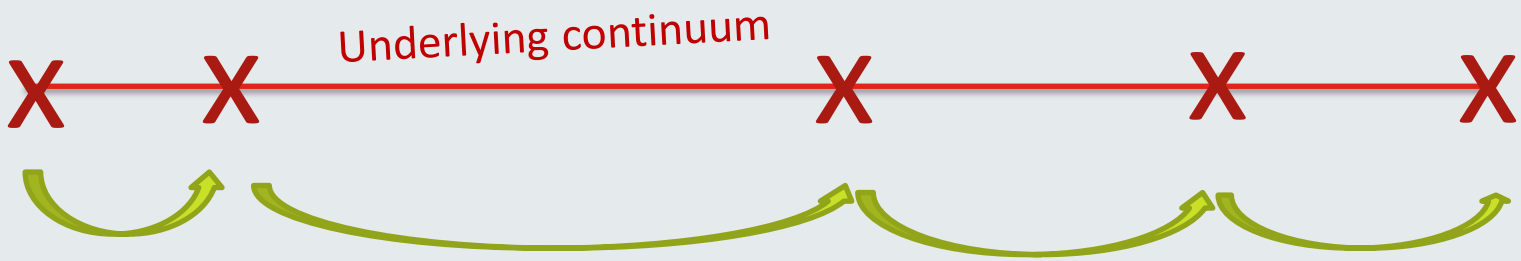
I. Buddhism
II. Christianity
III. Hinduism
IV. Islam
V. No religion

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?

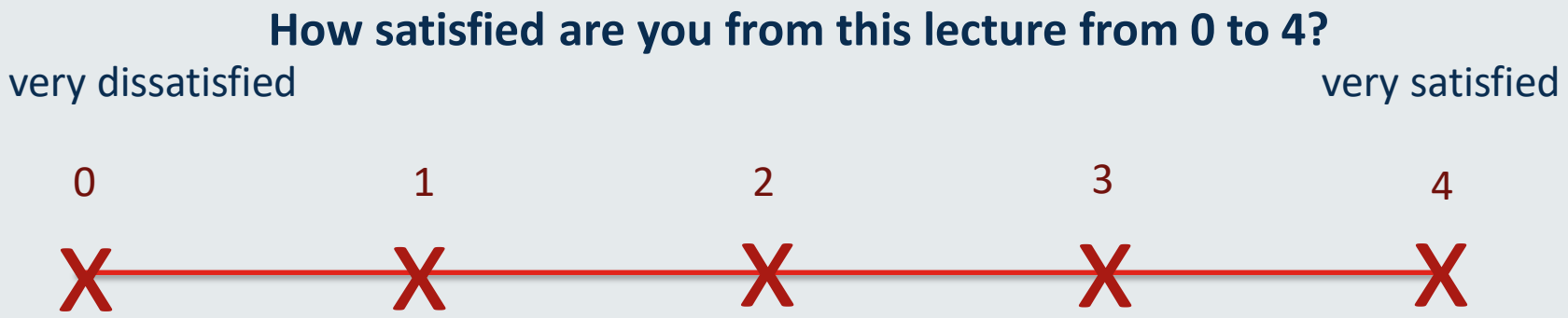
very dissatisfied dissatisfied neutral satisfied very satisfied



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:



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

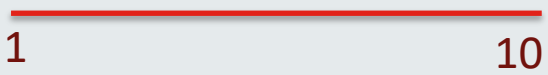
Frequency
1. More than I would want to
2. Less than most people
3. I have not noticed

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?

How many cars?

How many times?

Weight

Height

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…)

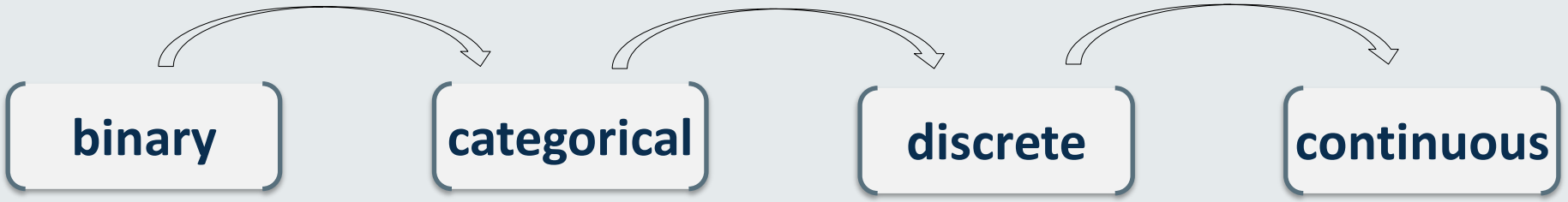
Any value in the interval (0,120) works

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

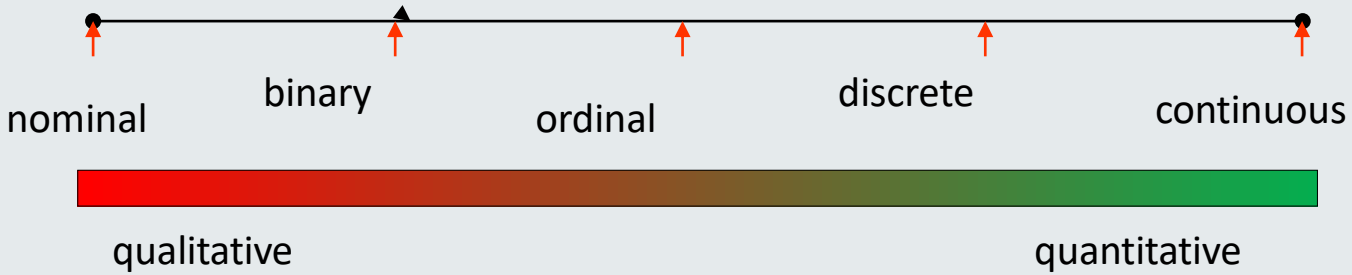
  The answer is no, the shoe size is a discrete variable. Even though there are decimals, these are very specific (you cannot for instance have a shoe of size 4.6). This means that shoe size takes values from a predefined set of numbers, not an interval.

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**