

# Session 7.2: Introduction to Statistics

Dr Annabel Bradbury, TRL Limited

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

- *Numbers can help make sense of the world*
- Simple does not mean trivial: simple numbers can help answer difficult questions
- Poorly researched numbers can mean bad policies, bad government decisions, missed opportunities, messed up lives
- Scientific understanding depends on:
  - Observation
  - Measurement (in terms of numbers)=>Analysis of numbers and their interpretation

- *Numbers are often all that we have – ignoring them is not an alternative*
- But.....
- Unexpected and extreme findings:
  - May tell us something new and extraordinary
  - Might be curious
  - Might be irrelevant to the bigger picture
- Outlying numbers (outliers):
  - Number is amazing
  - Number is incorrect
  - It has been misinterpreted

## Data – what is it?

- **Data** are values of qualitative or quantitative variables, belonging to a set of items.
- Data are typically the results of measurements and can be visualised using graphs or images.
- **Raw data**, i.e., unprocessed data, refers to a collection of numbers, characters and is a relative term; data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next.
- **Field data** refers to raw data collected in an uncontrolled environment.
- **Experimental data** refers to data generated within the context of a scientific investigation by observation and recording, including sampling.

## Data and data sets

- *“Data” is the plural of “datum”*
- In mathematics, engineering, and statistics, the terms given are used interchangeably.
- *A **data set** (or **dataset**) is a collection of data, usually presented in tabular form.*
- Each **column** represents a particular variable.
- Each **row** corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object.
- Each value is known as a **datum**.
- The data set may comprise data for one or more members, corresponding to the number of rows

## Variables

- A **variable** is a value that may change within the scope of a given problem or set of operations.
- In contrast, a **constant** is a value that remains unchanged, though often unknown or undetermined.
- Variables are either a **dependent variable** or an **independent variable**.
- **Independent variables** are inputs and may take on different values.
- **Dependent variables** are those values that change as a consequence of changes in other values (e.g. independent variables) in the system.

## Functions

- A **function** is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output.
- When one value is completely determined by another, or several others, it is called a **function** of the other value or values.
- In this case the value of the function is a dependent variable and the other values are independent variables.
- The notation  **$f(x)$**  is used for the value of the function  $f$  with  $x$  representing the independent variable.
- Notation such as  **$f(x, y, z)$**  may be used when there are several independent variables that are not the same

- What do we want to count?
- *Define the target population*
- Definitions tempt us to be specific but neatness can obscure vital information
- False clarity can be misleading
- *Numbers are often just “mushy peas” - rather than diamond fact!*
- Simplify numbers and they become clear; clarify numbers and you stand apart with rare authority

## Example: Football

- Find out the average height of Premier League footballers
- What is the population?
  - All who are in a current PL squad?
  - All who have played for a PL team this season?
  - All who have ever played for a PL team?
  - All who have ever played for a PL team for a full 90 minutes?
- What is a good sample?
  - All Manchester United players?
  - All goalkeepers?
  - All strikers

## What is a big number?

- Size matters:
  - Millions, billions or trillions?
- The best unit for comparison - yourself
- Use common sense - is it realistic or meaningful?
- Useful approach - divide by population
- Ask: *does that number seem to be a "big" number?*
- *It all depends on the context*

## Example: Traffic growth in Ethiopia

Table 3.1 Average Daily traffic Over Years

Year	90/91	91/92	92/93	93/94	94/95	95/96	96/97	97/98	98/99	99/00	00/01	01/02	02/03
AADT	23499	25199	30569	36831	40858	43987	46223	51078	55923	55672	72467	78686	83003
Growth rate	-	7.2	21.3	20.5	10.9	7.7	5.1	10.5	9.5	-0.4	30.2	8.6	5.5

Source: PPD, Report on Annual Rural Traffic Movements in Ethiopia. (1991-2003)

- Percentages are important in statistics
- Percentage is a fraction of 100
- *Percentage changes (up or down) depend entirely on where you start from*
- Natural frequency – number of people in every hundred who are affected

## Uncertainty

- *Uncertainty is a fact of life*
- *Numbers may be precise but they do not overcome uncertainty*
- Many numbers are uncertain
- Numbers can clarify uncertainty but cannot beat it
- *Do not over-interpret numbers*
- BUT do not throw them all away
- Being fallible does not make numbers useless
- *Numbers narrow the scope of our ignorance – they do not remove it*
- The issue is whether numbers are so wrong as to be misleading
- Statisticians admit that their numbers may be wrong by giving a range of uncertainty

## Comparisons

- Comparisons – be careful – *is the basis/ definition the same?*
- Comparisons – 4 issues:
  1. Who – which group is selected?
  2. When – what is the period for comparison?
  3. What – what is the parameter being compared?
  4. How – was the methodology the same?
- *International comparisons are difficult and generally meaningless*
- *Most comparisons are subjective*
- *Need to ensure comparisons are like-for-like*
- Meaningful comparisons are rarely found in individual figures

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

- **Statistics** is the study of the collection, organization, analysis, interpretation, and presentation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.
- Statistics is a **scientific discipline**.
- **Mathematical statistics** studies statistics mathematically.
- A **statistician** is someone who is well versed in the successful application of statistical analysis.
- **Statistic** refers to a quantity (such as mean or median) calculated from a set of data, whose plural is *statistics*.

## Statistics

*Statistics is not just the dry collection of facts; it is the science of making sense of the facts we have*

Statistics could be described as:

- Designing appropriate ways of collecting data (numbers) and extracting information from them
- Exploring, analysing and summarising data
- Constructing and testing models which can be used as a basis to make inferences and to draw conclusions

*Statistics aim to cope with uncertainty not in producing certainty*

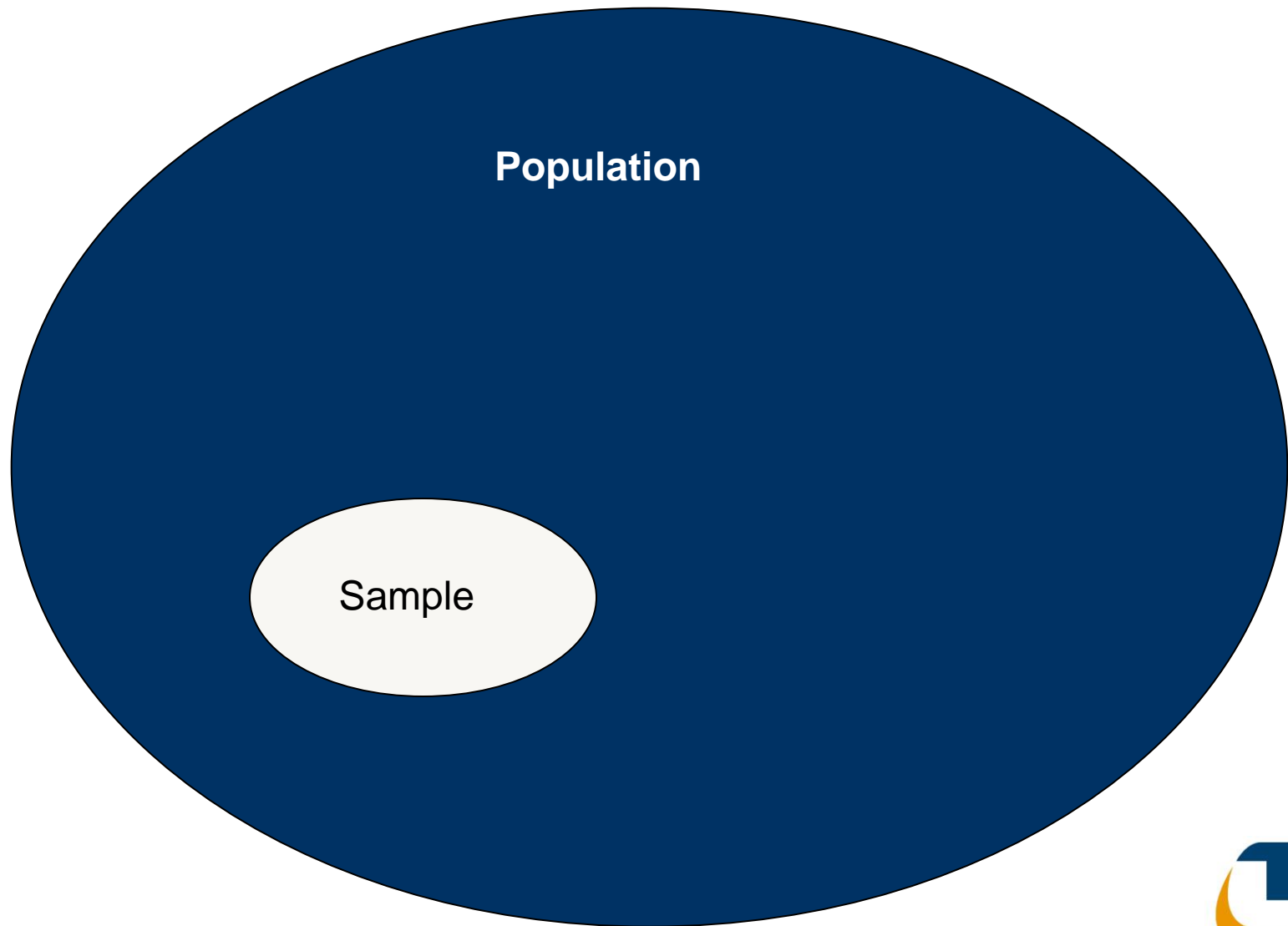
*Precision is often unlikely*



*Making assertions about population(s) from sample(s)*

- A **population** is the collection of items under investigation
- A **sample** is a subset of the population
- Population characteristics are unknown but sample characteristics are measurable.
- Sample values are estimates of population values.

*If a sample is to represent the population then the sample must be unbiased*



If we study the time it takes a man to drive from Dar es Salaam to Morogoro different types of analysis are possible e.g.

<b>Population</b>	<b>Sample</b>
All possible journeys from Dar es Salaam to Morogoro	Sample the time from 100 of those journeys and calculate the average
All drivers driving from Dar es Salaam to Morogoro	Sample the time taken for 100 drivers and calculate the average

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

- *Quantitative – numerical values*
  - Discrete data
  - Continuous data
- *Qualitative – data made up of categories*
  - Nominal data
  - Ordinal data

- *Quantitative – numerical values*
  - Discrete data:
    - "A", "B", "AB" or "O", for blood type
    - less than 5, between 5 and 10, or greater than 10
  - Continuous data:
    - 1.1, 1.2, 1.3 etc (and all intervening numbers i.e. a range)
- *Qualitative – data made up of categories*
  - Nominal data:
    - Rocks can be categorized as igneous, sedimentary and metamorphic
    - Transport data could be driver, passenger, cyclist, pedestrian
  - Ordinal data:
    - Ranked data: excellent, good, average, below average, poor

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

- A Pivot Table (an un-weighted table) is an interactive way to quickly summarize large amounts of data.
- You can use pivot tables to create frequency tables and crosstabs (a cross-tabulation or contingency tables).

### Exercise 7.1

**Open** the Excel file 'Session 1 & 2 Excel data.xls'

**Create** a pivot table for 'cohort data 2' showing a frequency table of 'Instrg – number of instructors' (it should resemble the frequency table shown in the notes)

**1** General overview - numbers

**2** Statistics

**3** Data types

**4** Pivot tables in Excel

**5** Tables and distribution

**6** Histograms in Excel

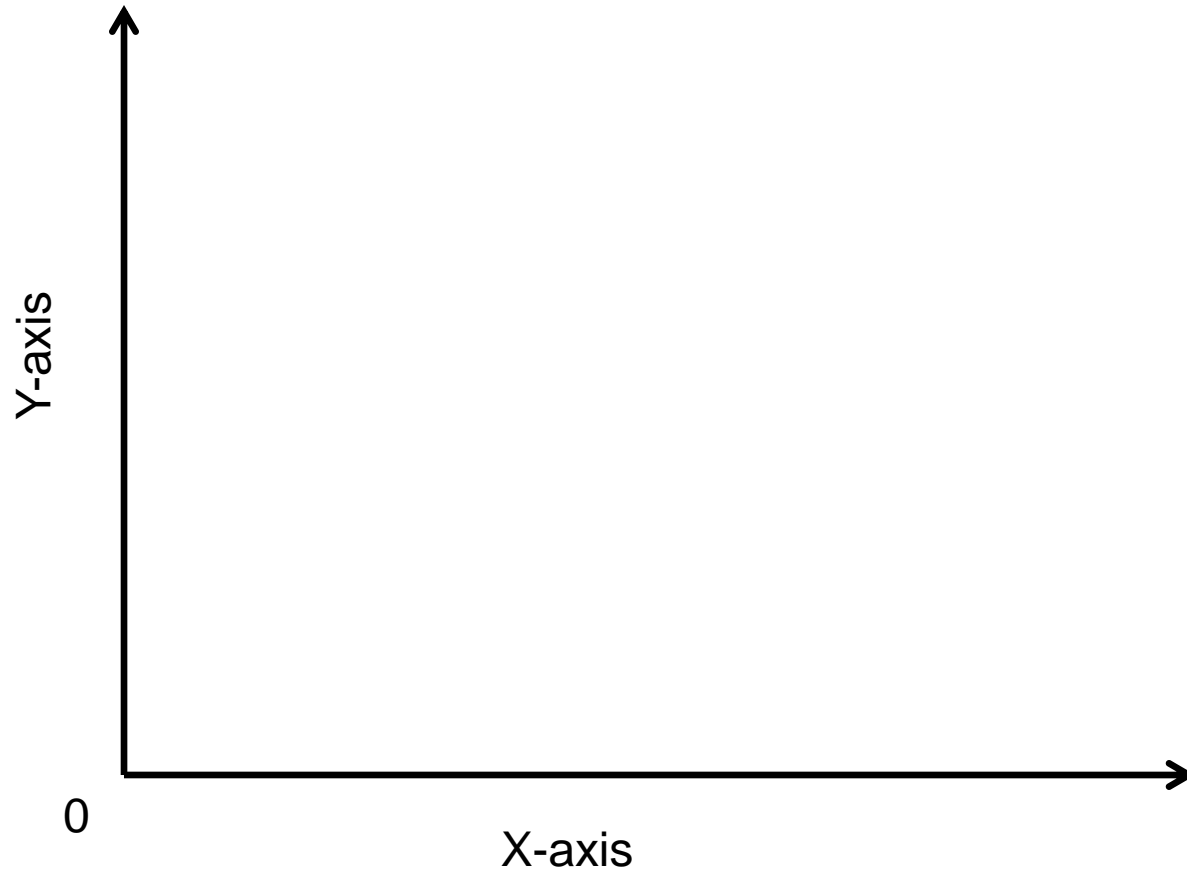
### Tables

- Look at the data!
- Discrete quantitative or restricted qualitative data can be presented in tables
  - Fewer than 10 groups
- Include totals and percentages
- Present variables together

### Exercise 7.2

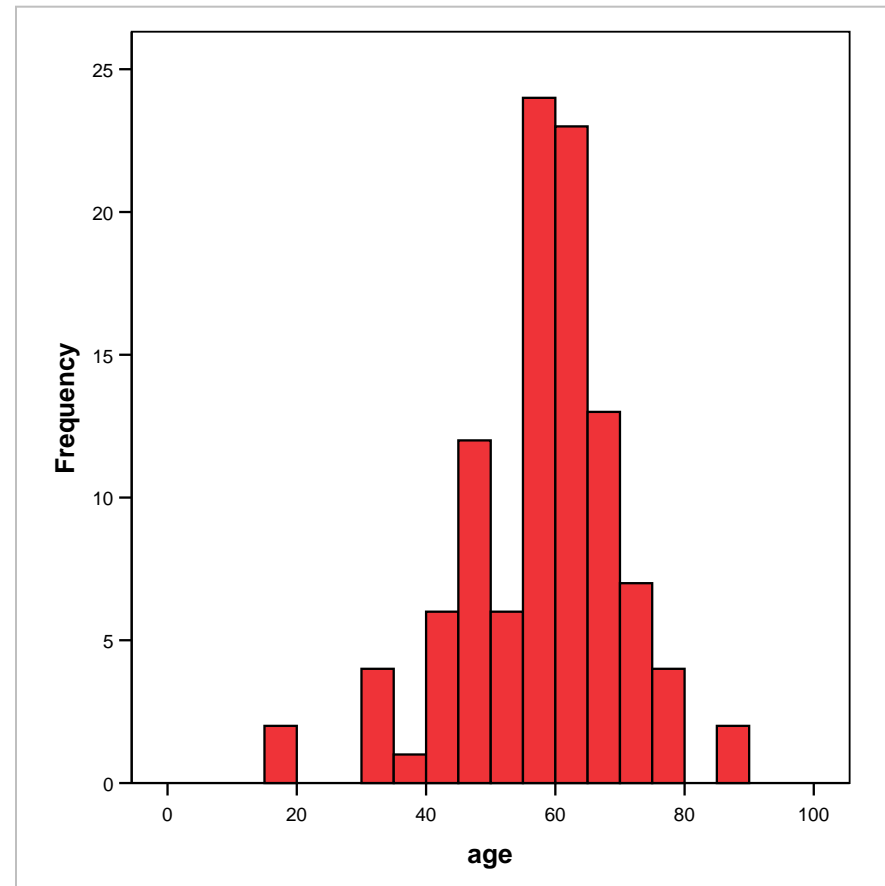
- Open SAR
- Prepare a frequency table for *sex* using **Pivot Tables**
- Prepare a frequency table for *age* using **Pivot Tables**
- Prepare a frequency table for *sex* against *agegrp* by putting *sex* in columns and *agegrp* in row values
- Prepare a frequency table with totals and row percentages (add Counts and Row %):
- Prepare a frequency table with totals and column percentages:
- What is the difference between the percentages?

## Axes



## Distributions

- Continuous data
  - Variables with more than 10 groups becomes unmanageable
  - Group data and present in tables
  - Plot in a Histogram



1 General overview - numbers

2 Statistics

3 Data types

4 Pivot tables in Excel

5 Tables and distribution

6 Histograms in Excel

- Excel has the capability of doing many statistical analyses e.g. histograms, statistical inference etc.
- If loaded the **data analysis** command is available under the **data** tab.
- If not then it can be easily installed:
  - Click the **Office Button** , and then click **Excel Options**.
  - Click **Add-Ins**.
  - In the **Manage box**, click **Excel Add-ins**, and then click **Go**.
  - In the Add-Ins available box, select the **Analysis ToolPak** and the **Analysis ToolPak - VBA** check box, and then click **OK**.

### Exercise 7.3

- Install the analysis toolpak if the data analysis command is not available under the data tab
- Open the Excel file '*session 1 & 2 Excel data.xls*'
- Open cohort data 2
- Create a histogram for 'age'
- **data → data analysis → histogram**
- Leave bin-range blank and select chart output box



**Do You  
Have Any  
Questions?**