

Tutorial – Statistics Part 1 (Overview)

1. Briefly state the difference between

Categorical Data and Continuous Data

Categorical Data – each data item can take a small set of prescribed values e.g.

- True / False
- bad / neutral / good (-1, 0 1 if represented numerically!)
- Nike / Adidas / Puma

Continuous Data – can have any value in a particular range e.g.

- Heights of people in a room
- Temperature readings
- Amount of rainfall in mm

Favourite Sports Brand

Index	Sports Band
1	Nike
2	Nike
3	Puma
4	Adidas
5	Adidas

Total monthly rainfall (mm) Dublin Airport 2022

Month	Rainfall (mm)
Jan	14.4
Feb	88.5
Mar	45.6
Apr	28.1
May	48.4

Categorical data

Continuous data

Primary Data and Secondary Data

Primary Data – data collected from our original sources for the purpose of a study or experiment – have more control over the data collected and the conditions under which it is collected. E.g. surveys (created by the person/team carrying out the study)

Secondary Data – data compiled for another purpose (e.g. census data) but utilised for our own study

Random Sampling and Stratified Random Sampling

Population - all subjects possessing a common characteristic that is being studied.
Depending on the data – can be quite large

A **sample** is a subgroup or subset of the population. Have different methods of selecting a sample from a population

Random Sampling

Sampling in which the data is collected using chance methods or random numbers. Every member of the population has an equal chance of being in the sample (see example in lecture slides).

Stratified Random Sampling

Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques. More sophisticated form of random sampling (see example in lecture slides).

Observational Data and Experimental Data

Observational Data:

Data in which subjects are observed and studied, but no attempt is made to manipulate or modify the subjects.

- a. e.g. trying to determine the effects that eating strictly organic foods has on overall health. If we were to take 200 people, where 100 have eaten organically for the past 3 years and the other 100 have not eaten organically in the past 3 years. Then give each subject (person) an overall health assessment. We can analyse the data and use it to draw conclusions on how eating organically can affect a person's overall health.

Experimental Data:

Data in which a treatment is applied (have control), and then its effects on the subjects are studied.

- e.g. getting 200 random people that do not eat organically and then have 100 eat organically for the next 3 years and have 100 not eat organically for the next 3 years. At the end of the study assess/analyse each person's overall health.

More Examples – Khan Academy

- we can discover **association** between variables from observational data
 - and determine **cause and effect** from experimental data
2. When performing ***data analysis*** why sometimes is a **sample** of the data used and not the entire **population**?

Entire population may be too large to work with! Take a sample (using sampling technique e.g. stratified random sampling) and work with that data. It is important that the sample is *representative* (same trends and patterns exist) of the original dataset/population and that the sampling method doesn't introduce any bias into the data.

3. Complete the table below and use a bar chart to display the data:

A bar chart is used for categorical data ...

Relative Frequency: Is the proportion of the data which falls in each class

- We calculate the relative frequencies by dividing by the total number of items of data.
- The relative frequencies add up to 1.

Country	Frequency	Relative Frequency
Ireland	10	$10/70 = 0.14$
England	12	$12/70 = 0.17$
Scotland	11	$11/70$
Wales	9	$9/70$
France	13	$13/70$
Spain	15	$15/70$

Total Frequency = 70

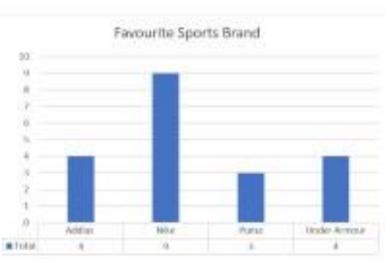
Bar Chart

Students asked about their favourite sports brand (20 students
very small sample!)

At a glance, a bar chart can expose important structures in data e.g. which categories are common and which are not

Index	Brand	Index	Brand
1	Nike	11	Under Armour
2	Puma	12	Nike
3	Nike	13	Under Armour
4	Nike	14	Addias
5	Nike	15	Addias
6	Puma	16	Under Armour
7	Addias	17	Nike
8	Addias	18	Nike
9	Nike	19	Puma
10	Under Armour	20	Nike

- Simple way to present data is in a table
- However, a table can be hard to read & draw conclusions from especially where there is a lot of data!



A bar chart

- set of bars, one per category where the height of each bar is proportional to the number of items in that category.
- used for categorical data

Draw the bar chart for Q3.....

4. The system administrator in a college is interested in determining the amount of time that computing students spend logged in for each day. To do this the administrator monitors 98 randomly selected computing students for an entire day. The study provides the following information:

Logged In Time	Frequency (No. of Students)
< 2 hours	6
≥ 2 hours and < 4 hours	14
≥ 4 hours and < 6 hours	34
≥ 6 hours and < 8 hours	26
≥ 8 hours and < 10 hours	10
≥ 10 hours and < 12 hours	8

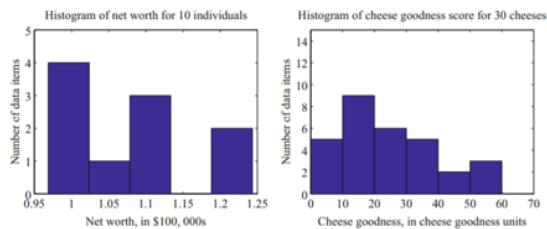
- (i) Draw a histogram of the data
- (ii) What information does the histogram present and what insights to the data under study does it provide?

A histogram is used for continuous data ...

Histograms

Index	Net worth	Index	Taste score	Index	Taste score
1	100, 360	1	12.3	11	34.9
2	109, 770	2	20.9	12	57.2
3	96, 860	3	39	13	0.7
4	97, 860	4	47.9	14	25.9
5	108, 930	5	5.6	15	54.9
6	124, 330	6	25.9	16	40.9
7	101, 300	7	37.3	17	15.9
8	112, 710	8	21.9	18	6.4
9	106, 740	9	18.1	19	18
10	120, 170	10	21	20	38.9

To ‘draw a picture’ of a set of continuous data we must first split an interval enclosing the smallest and largest values into several non-overlapping classes of equal width.



- Usually shows the distribution of values of a single variable
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

The class intervals / bins are drawn on the x -axis and the frequency on the y -axis.

Draw the histogram for Q4.....

What insights into the data do we get?

Note: Although a bar chart and a histogram look very similar, it is important to realise that a bar chart has a finite number of categories (categorical data) along the axis, whereas a histogram has a continuous numerical scale (continuous data).

5. Explain what is meant by the terms ‘mean’, ‘median’, ‘mode’ and ‘standard deviation’. Using the following data: 38, 40, 55, 60, 65

Compute the

- Mean (Average)
- Standard deviation
- Median

Explain what is meant by the terms ‘mean’, ‘median’, ‘mode’ and ‘standard deviation’. – all describe the distribution of data in a dataset. *Mean or Average, median and mode* all describe the central tendency of the data.

Mean is the average of the data.

Median is the mid-point in the data and is concerned with the order of the dataset. Median is a better measure of central tendency when the dataset contains outliers (i.e., extreme values)

Mode is the value that occurs the most frequently in a data set.

Standard deviation used with the mean is a measure of how spread out (or distributed) the values are about the mean/average.

Calculations:

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

$$\bar{x} = \frac{38+40+55+60+65}{5} = \frac{258}{5} = 51.6$$

$$\text{Standard Deviation} \quad s^2 = \frac{\sum_i x_i^2 - n(\bar{x})^2}{n-1}$$

$$s^2 = \frac{38^2 + 40^2 + 55^2 + 60^2 + 65^2 - 5(51.6)^2}{5-1}$$

$$s^2 = \frac{13894 - 13312.8}{4} = \frac{581.2}{4}$$

$$s^2 = 145.3$$

$$\therefore s \text{ (standard deviation)} = \sqrt{145.3} = 12.05$$

Median

Sort the values in ascending order 38, 40, 55, 60, 65

↑
Median

Complete the remaining questions in class...

6. Using the following data: 10, 12, 7, 6, 3, 15, 21, 4, 9, 5, 13, 19
Compute the
 - i. Mean (Average)
 - ii. Standard deviation
 - iii. Median
 - iv. Quartiles
7. a) Compute the *mean*, *median* and *standard deviation* for the following list of ages:
17, 21, 17, 18, 17, 19, 18, 17, 63

b) Arrange the values in ascending order and remove the largest value in the list.
Recalculate the *mean* and *median* for the modified list of values. Discuss the effect of adding or removing an extreme value on the *mean* and *median* values.

Formula Sheet

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

$$\text{Standard Deviation (1): } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\text{Standard Deviation (2): } s^2 = \frac{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}{n-1}$$