1495
UNIVERSITY OF
ABERDEEN
CELEBRATING
525 YEARS
1495 – 2020

ABERDEEN 2040

Exploratory Data Analysis

Data Mining & Visualisation

Lecture 3

2025

Recap

- Levels of measurement
  - Nominal, ordinal, interval, ratio

Today...

- Exploratory data analysis
- Descriptive statistics

Exploratory Data Analysis

What is Exploratory Data Analysis (EDA)?

Data mining will typically require some form of familiarity or understanding about the data that you are analysing.

However, importantly, the data won't always be in a clean, usable, and well-documented format for you to work with!

What is Exploratory Data Analysis (EDA)?

EDA is the process of exploring your data, often in an unplanned and investigative manner.

This might involve summarising or visualising aspects of the data, identifying outliers, cleaning, pre-processing, etc.

It can help determine how best to manipulate data, help generate hypotheses, and help guide the analysis process.

What is Exploratory Data Analysis (EDA)?

EDA is often a crucial first step in gaining a more in-depth understanding of the data before and during analysis.

The aim is to become more familiar with the data!

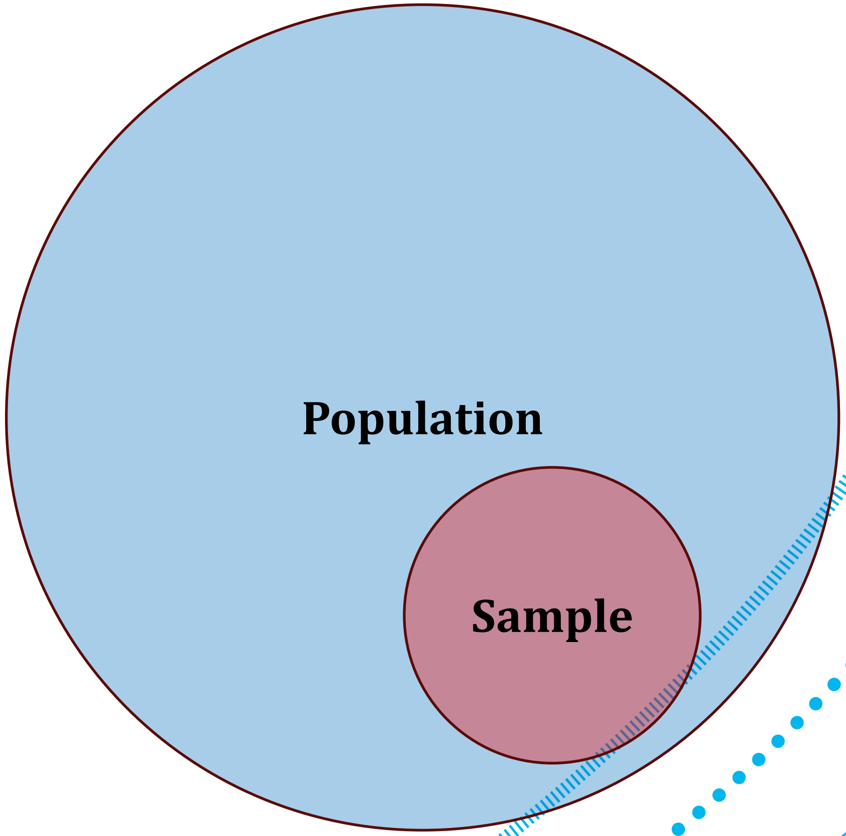
It is a crucial set of skills that will be widely applicable across data science, machine learning, and broader AI topics.

Descriptive Statistics

Note: Population Vs Sample

In statistics, there is an important distinction between **populations** and **samples**.

A given dataset can represent some population, or a sample of that population.

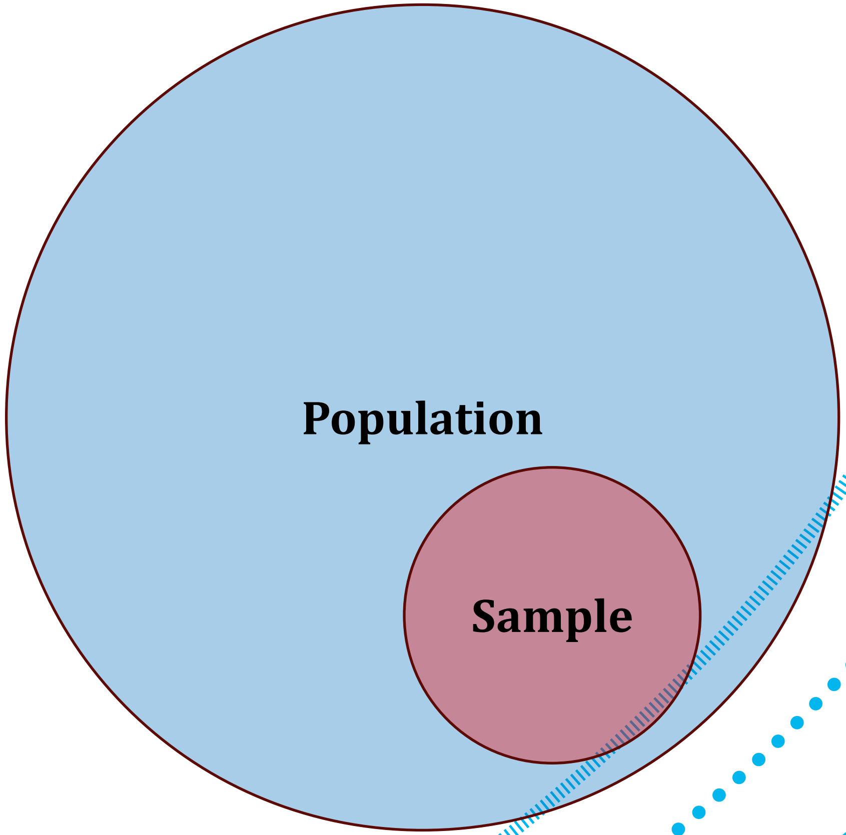


Note: Population Vs Sample

A population is an entire group of people/objects/items etc. of interest.

However, it will often be impractical or impossible to get population-level data.

Think of trying to collect demographic data for every person in China!

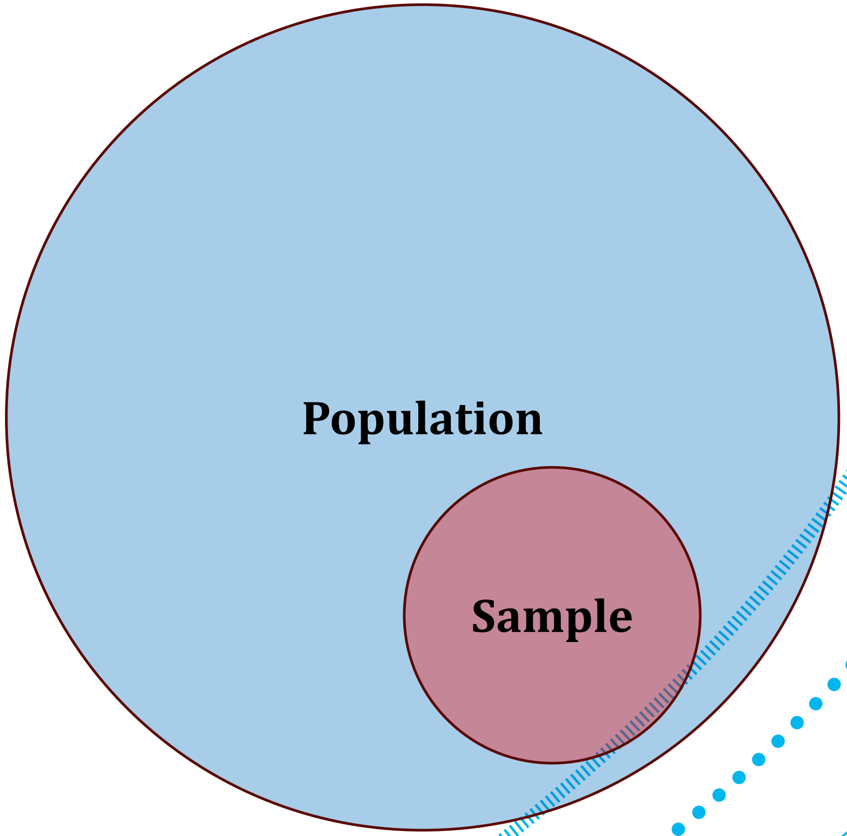


Note: Population Vs Sample

A sample is a subset of the population we're interested in studying.

We can use it to estimate the population, which is more feasible…

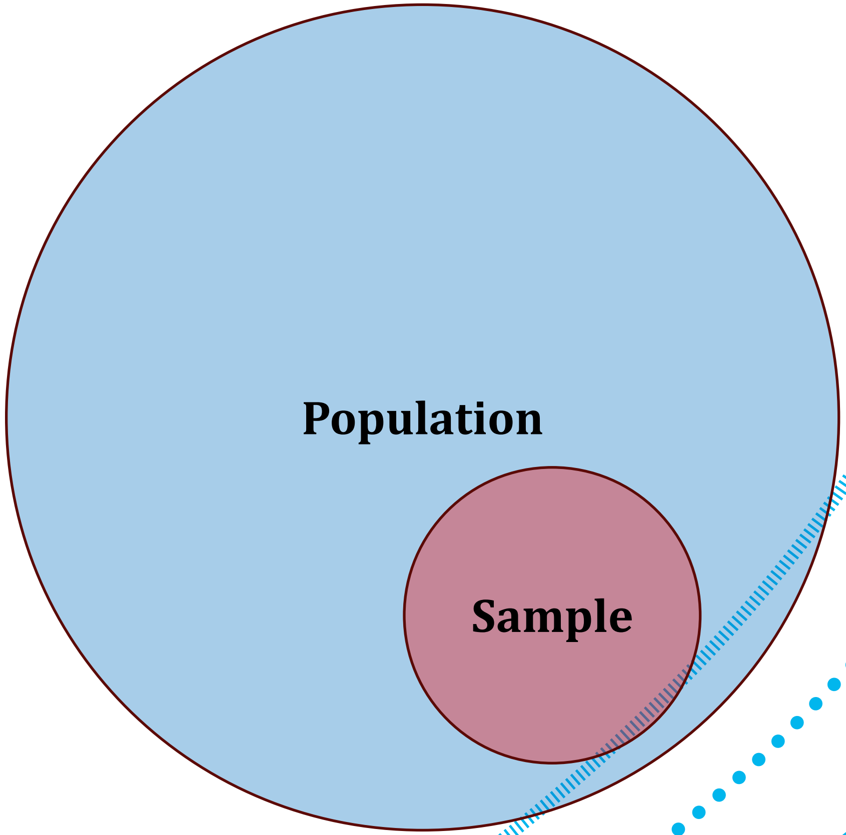
…But it will typically have some unknown sampling error.



Note: Population Vs Sample

Whether you are studying a **population** or a **sample**, some statistical formulas will change.

This is not something you have to be overly concerned with, but it is something that you need to be aware of!



Descriptive Statistics

Descriptive statistics provide a way of summarising our data.

They give us a way of understanding the properties and characteristics of our dataset, and what the data 'looks' like.

Often, we're interested in the **central tendency** of the data, and its **variability**.

Measures of Central Tendency

The central tendency of a set of data (e.g. a variable) attempts to describe the data via its central position.

The aim is to provide an accurate description of the entire set of data with one value.

One of the most common measures of central tendency is the average of the data, i.e. the mean.

Arithmetic Mean

The average of the data. Note the different symbols for population and sample.

It gives us an effective measure of central tendency, however, it is highly susceptible to skewed data and outliers.

It also may not accurately reflect the data (e.g. discrete variables, ordinal data).

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

(Population mean)

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

(Sample mean)

Median

The median value is the value separating the upper half from the lower half of a set of values

$$\text{If } n \text{ is odd, } \operatorname{med}(x) = x_{(n+1)/2}$$

$$\text{If } n \text{ is even, } \operatorname{med}(x) = \frac{x_{(n/2)}+x_{((n/2)+1)}}{2}$$

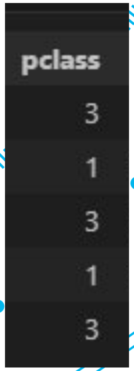
Note that the median value may not actually be present in the dataset

Median

The median is particularly useful when dealing with ordinal data (i.e. that which has ordered classes, such as rankings).

In these cases, the mean would not be suitable, since ordinal data will not have equal intervals.

E.g. think about 'passenger class' in the Titanic dataset.



pclass
3
1
3
1
3

Mode

The modal value within a set of data is the most frequently occurring value.

2	3	7	7	5	4	3	7	1	9	2	7

Here, 7 occurs 4 times, and is the modal value.

Measures of Variability

Knowing the central tendency of your data is very useful.

However, often, you will also want to know its **variance**; i.e. how dispersed (or spread out) the data are.

Standard Deviation

The average amount of dispersion that our data has from the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}$$

(Population standard deviation)

Note the difference in calculating std. dev. for the population vs a sample!

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

(Sample standard deviation)

In this course, we'll mostly focus on working with **samples**.

Z-Score

The Z-Score (or 'standard score') is the number of standard deviations that **a given data point** differs from the mean.

$$z = \frac{(x - \bar{x})}{s}$$

Z-Score

- A Z-Score of **0** means that the value is **identical to the mean**

- A Z-Score of **1** means that the value is **1 standard deviation above the mean**

- A Z-Score of **-2** means that the value is **2 standard deviations below the mean**

Z-Score

Let's say we have a sample of six datapoints:

1	3	6	7	9	10

Z-Score – Step 1: Calculate the Mean

Let's say we have a sample of six datapoints:

1	3	6	7	9	10

$$\bar{x} = 6$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(1+3+6+7+9+10)}{6} = \frac{36}{6} = 6$$

Z-Score – Step 2: Calculate the Std. Dev

Let's say we have a sample of six datapoints:

1	3	6	7	9	10

$$\bar{x} = 6$$
$$s = 3.464$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$$

$$s = \sqrt{\frac{(1-6)^2+(3-6)^2+(6-6)^2+(7-6)^2+(9-6)^2+(10-6)^2}{6-1}}$$

$$s = \sqrt{\frac{25+9+1+9+16}{5}} = \sqrt{\frac{60}{5}} = \sqrt{12} = 3.464$$

Z-Score – Step 3: Calculate the Z-Scores

Let's say we have a sample of six datapoints:

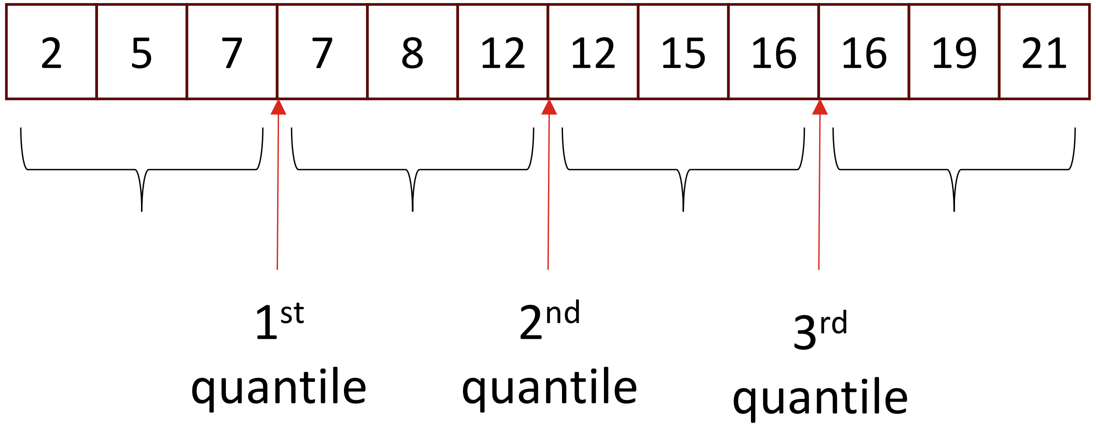
1	3	6	7	9	10

$$
\begin{aligned}
z = \frac{(x - \bar{x})}{s} &= \frac{(1-6)}{3.464} = -1.443 \\
&= \frac{(3-6)}{3.464} = -0.866 \\
&= \frac{(6-6)}{3.464} = 0 \\
&= \frac{(7-6)}{3.464} = 0.289 \\
&= \frac{(9-6)}{3.464} = 0.866 \\
&= \frac{(10-6)}{3.464} = 1.15
\end{aligned}
$$

$$
\begin{aligned}
\bar{x} &= 6 \\
s &= 3.464
\end{aligned}
$$

Quantiles

Quantiles are a way of dividing your sample into *q* equal groups, where the *kth* quantile represents the value below which *k* / *q* of the data fall.



2	5	7	7	8	12	12	15	16	16	19	21

1st quantile

2nd quantile

3rd quantile

Here, we've split the data into four groups, i.e. q = 4.

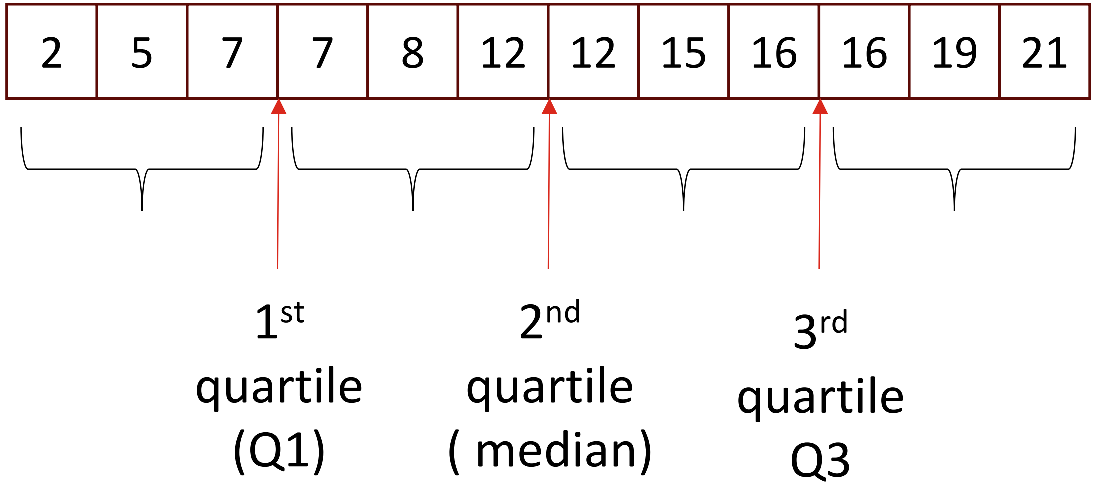
Quantiles

Some commonly used quantiles have their own names, including **percentiles** (100 groups), **deciles** (10 groups), **quintiles** (5 groups), and **quartiles** (4 groups).

Referring to the **median** of a dataset is the same as referring to the **50th percentile**, the **5th decile**, and the **2nd quartile**.

Interquartile Range (IQR)

The **interquartile range (IQR)** is a measure of variability using **quartiles**, measuring the difference between the 1st quartile (Q1) and the third quartile (Q3) – i.e. IQR = Q3 - Q1.



2	5	7	7	8	12	12	15	16	16	19	21

This dataset has a median of 12 and an IQR of 9.

Skew & Kurtosis

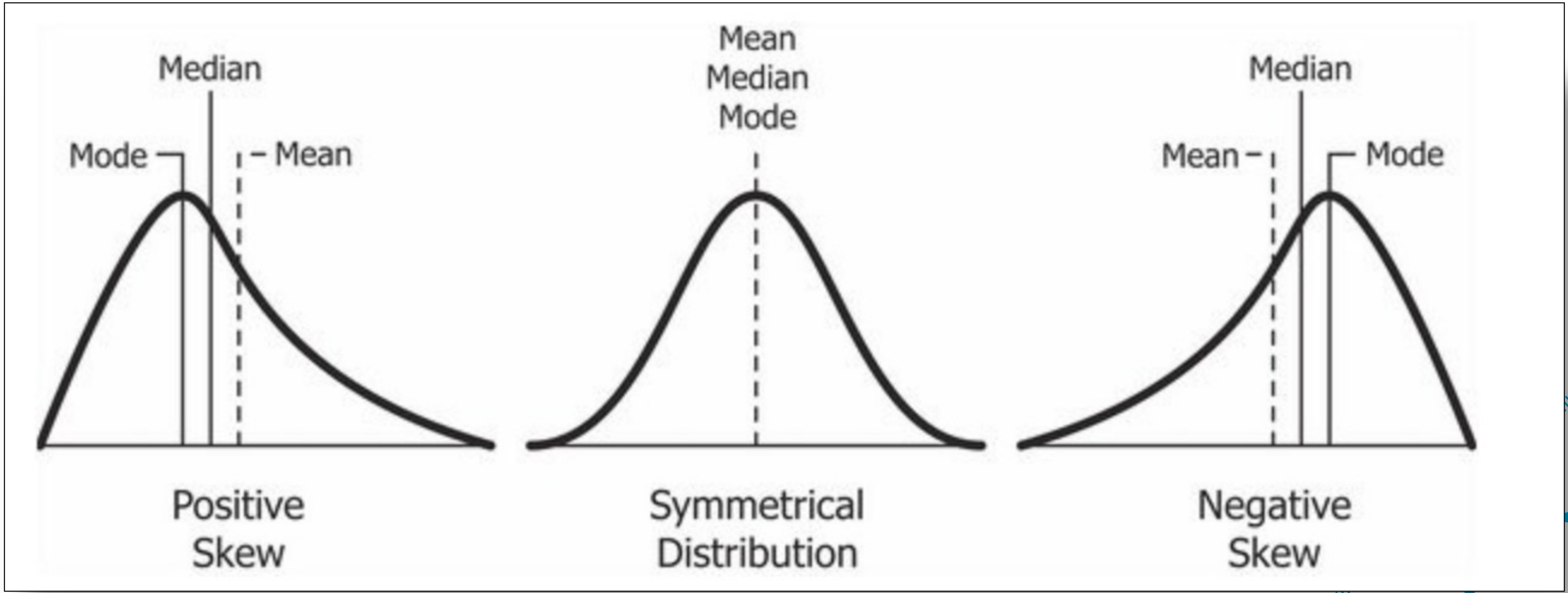
Another way of understanding the characteristics of your data is through the shape of its distribution.

We often think of data distributions in terms of a ’bell curve’, or hear data being described as ’normally distributed’.

However, this isn’t always the case.

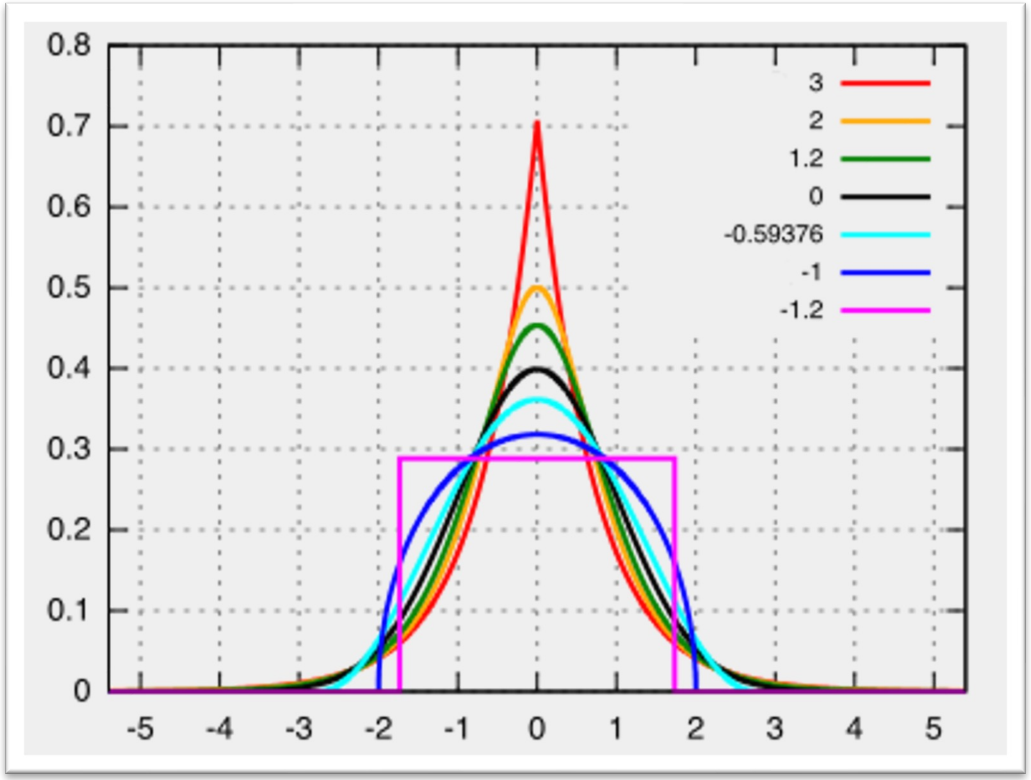
Skew & Kurtosis

Skewness is a measure of the distribution's asymmetry.



Skew & Kurtosis

Kurtosis is a measure of the ‘tailedness’ of a distribution.



Skew & Kurtosis

From knowing the skew and kurtosis of a variable, we can tell a lot about the characteristics of the underlying data.

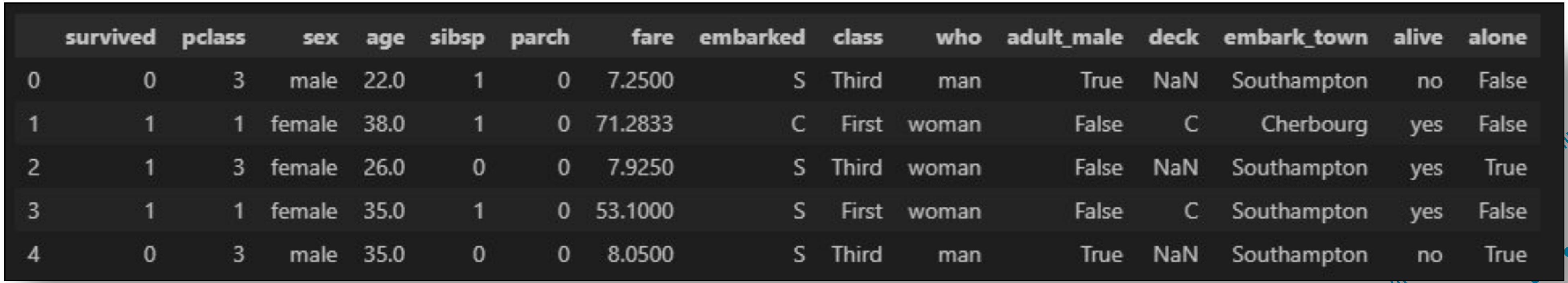
Similarly, from knowing the mean, median, and mode of a variable, we can infer about its skew and kurtosis.

Mean vs Median

Mean vs Median

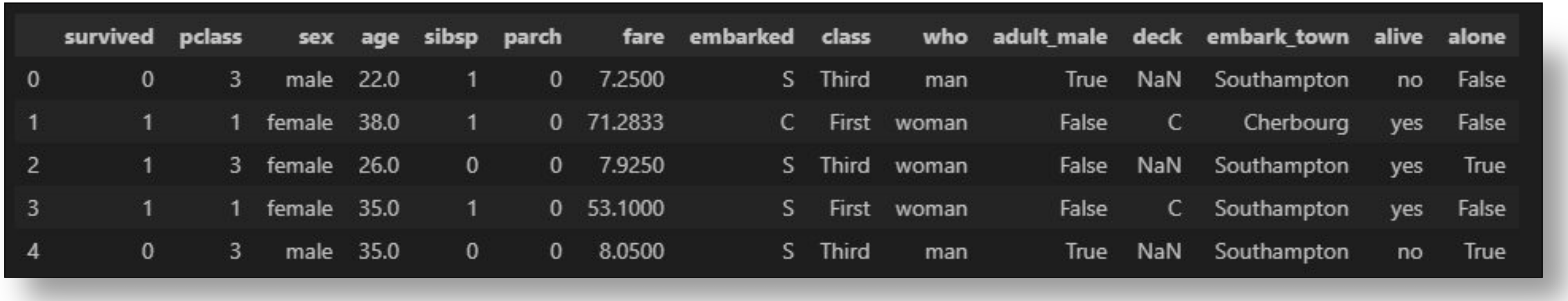
Previously, we discussed how the ‘level of measurement’ of a variable can determine the analysis you can perform.

Let’s take a look at the Titanic dataset again…



	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

Mean vs Median



	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

The ‘pclass’ variable represents the Passenger Class, where 1={First}; 2={Second}; 3={Third}.

Recap: What is its level of measurement?

Mean vs Median

Despite being represented as an integer, pclass is an **ordinal** variable.

It represents ordered groups, however, there is no equal interval between the values.

In other words, two first class passengers does not equate to a second class passenger!

Mean vs Median

When dealing with **nominal** or **ordinal** variables, we tend to use the **median** to measure central tendency.

We can also use the **IQR** to outline its variability.

Mean vs Median

In contrast, when dealing with **interval** or **ratio** variables, we tend to use the **mean** to measure central tendency.

In this case, we can use the **standard deviation** to outline its variability.