# Advanced Data Science

## Topic 11b – Part 3

# 1. What We'll Cover



**This topic will introduce…**

- **What is data science.**
- **Key concepts – the scientific method.**
- **Useful terminology.**

- **Important tools - Statistics.**
- **Data collection & Experiment Design.**

**Part 3**

- **Probability basics.**
- **Data distributions.**
- **Hypothesis testing.**

**The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.**

# 2. Data

- We seek to answer questions using statistical methods and a collection of observations.
- Observations may be obtained in a variety of ways.
- Data is a collection of observations described using variables.
- We use some simple notation to describe variables: $x_i$

Variables $x_i$

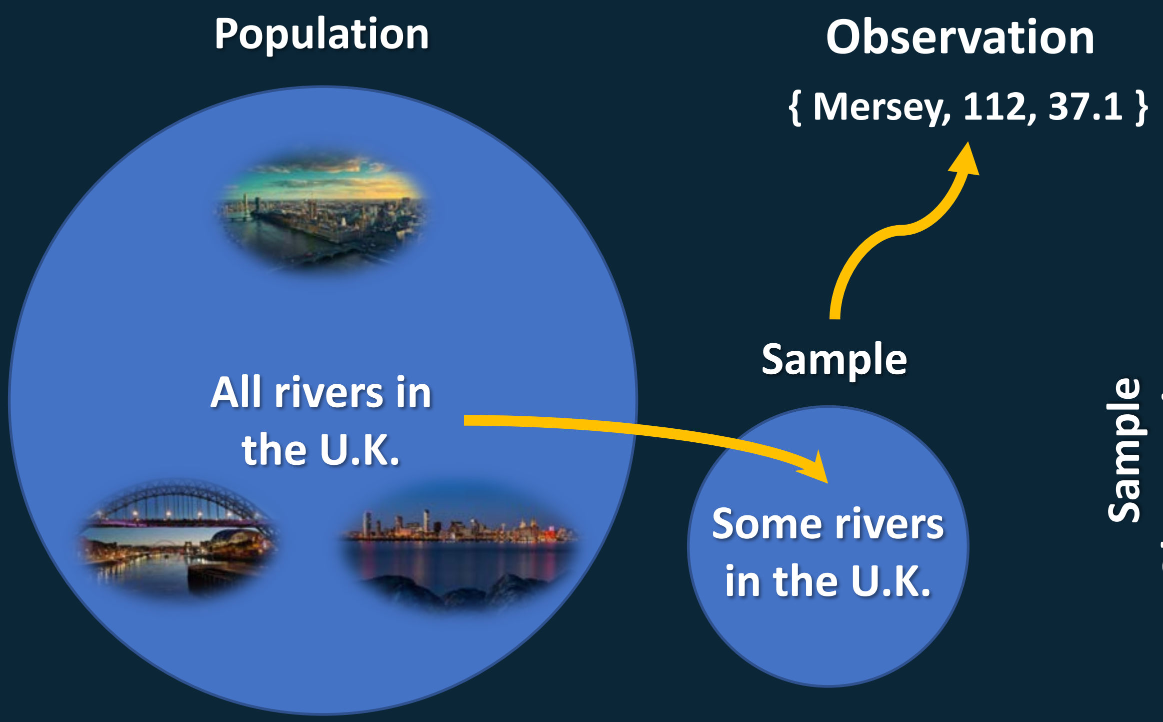
Observations

| ID | Name | Length (km) | Flow $m^3/s$ |
|---|---|---|---|
| 1 | Mersey | 112 | 37.1 |
| 2 | Tyne | 118 | 45.2 |
| 3 | Tay | 188 | 179.0 |
| 4 | Severn | 354 | 107.4 |
| … | … | … | .. |

$x_1$ $x_2$ $x_3$

# 3. Data

- **A population** is a complete dataset that contains all potential observations of an event or phenomena.
- **A sample** represents a subset of a population chosen in some way.
- **An observation** is an individual example from a population or sample.




Variables $x_i$

Sample Observations

| ID | Name | Length (km) | Flow $m^3/s$ |
|---|---|---|---|
| 1 | Mersey | 112 | 37.1 |
| 2 | Tyne | 118 | 45.2 |
| 3 | Tay | 188 | 179.0 |
| 4 | Severn | 354 | 107.4 |
| … | … | … | .. |

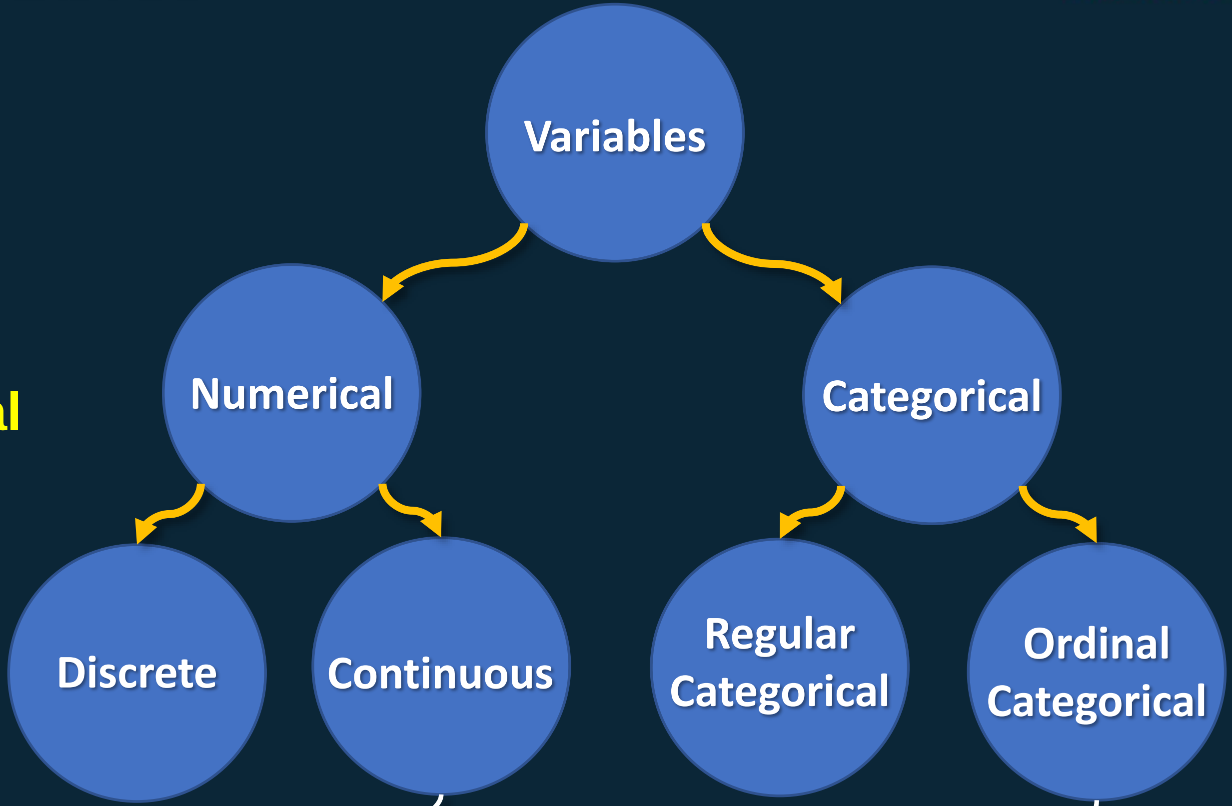
$x_1$ $x_2$ $x_3$

Edge Hill University

</TECHUP_WOMEN>

# 4. Variables

- Variables may be **numerical** – which includes:
  - o **discrete** variables - whole numbers
  - o **continuous** variables – decimal components
- Variables may also be **categorical:**
  - o **Regular categorical** variables describe categories.
  - o **Ordinal categorical** variables have a natural ordering.




$$x_1 = \{Mersey, Tyne, Tay, Severn\}$$

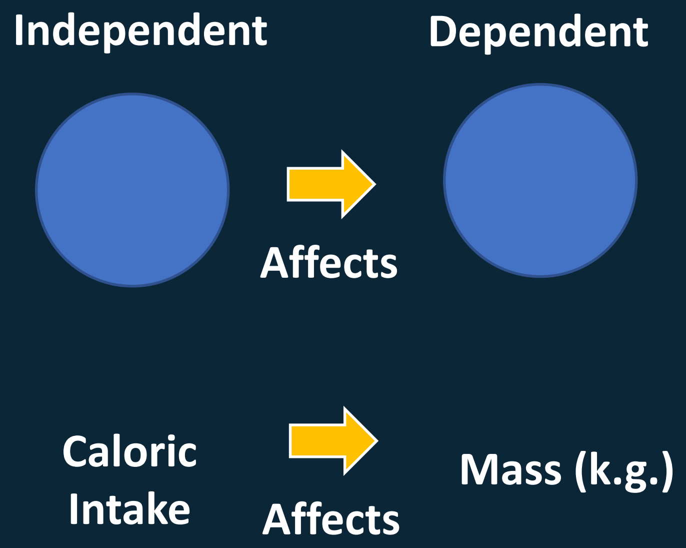
$$x_2 = \{112, 118, 188, 354\}$$

$$x_3 = \{37.1, 45.2, 179.0, 107.4\}$$

$$x_4 = \{Good, Good, Excellent, Very\ good\}$$

# 5. Variables

- Relationships may exist between variables.
- May need to verify the relationship, or prove no relationship exists.
- If variables are related or "associated", then changing the value of one impacts the other.
- The variable that creates change is known as the independent variable.
- The variable being affected is the dependent variable.



# 6. Variables & Correlation

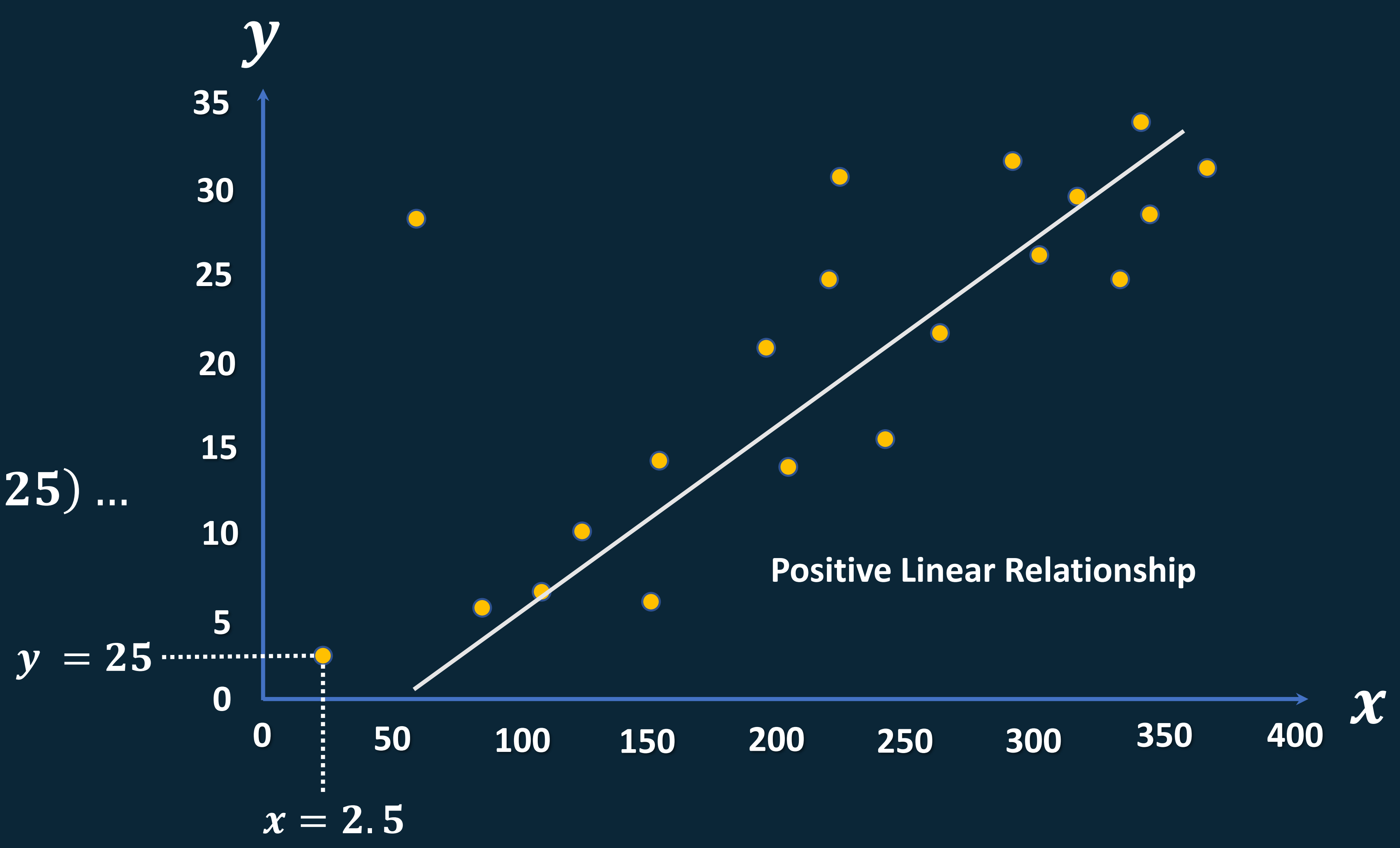
| ID | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | 2.5 | 25 | 32 |
| 2 | 6 | 80 | 12 |
| 3 | 10 | 125 | 12 |
| 4 | 5 | 150 | 23 |
| … | … | … | .. |

$x$ $y$

$$(2.5, 25), (6, 80), (10, 125) \ldots$$

Independent → Dependent

$x$ Increases $y$



# 7. Variables & Correlation



$x$ $y$

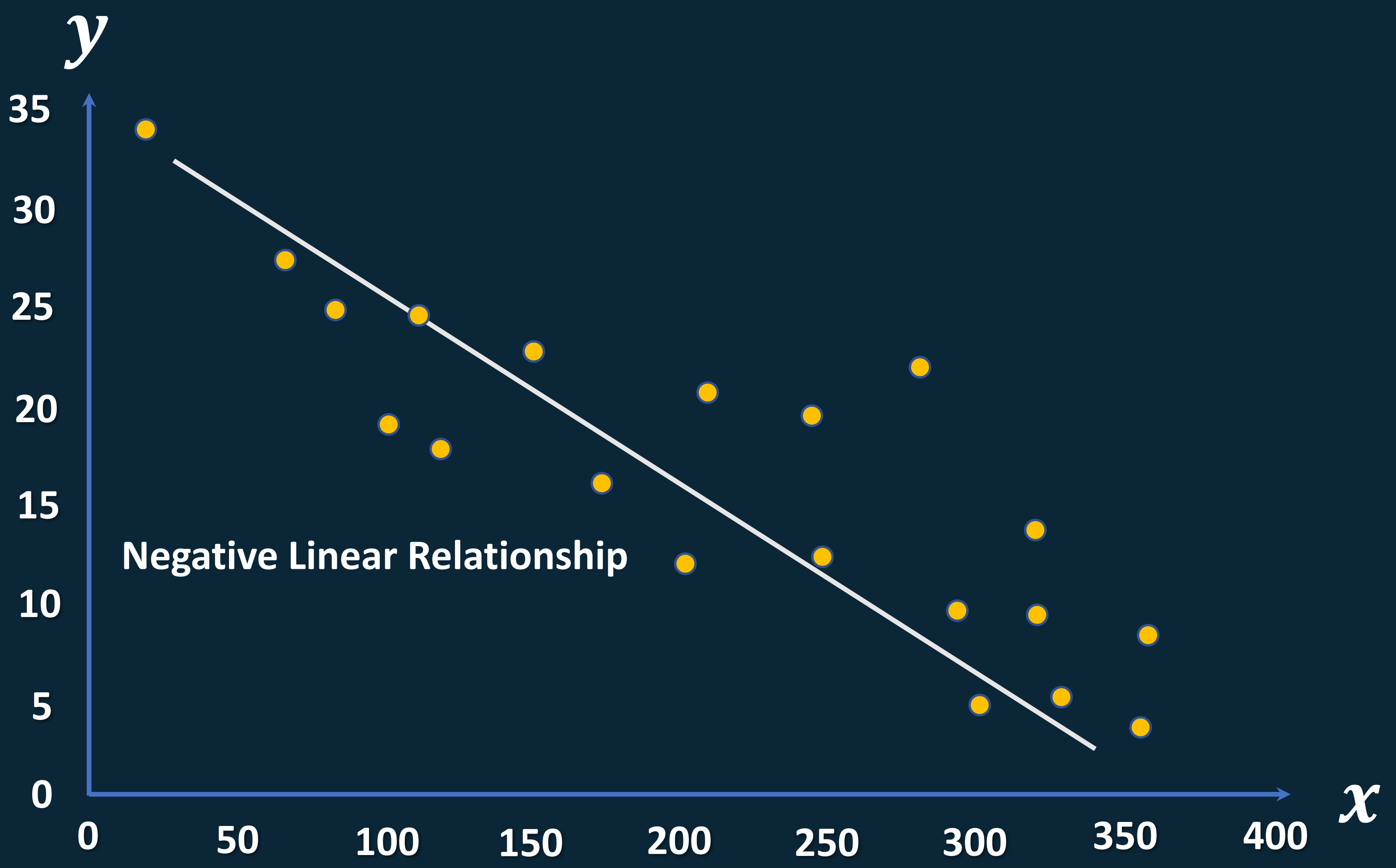
| ID | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | 2.5 | 25 | 32 |
| 2 | 6 | 80 | 12 |
| 3 | 10 | 125 | 12 |
| 4 | 5 | 150 | 23 |
| … | … | … | .. |

$(2.5, 25)$ , $(6, 80)$ , $(10, 125)$ …

Independent — Dependent

$x \rightarrow y$

Decreases



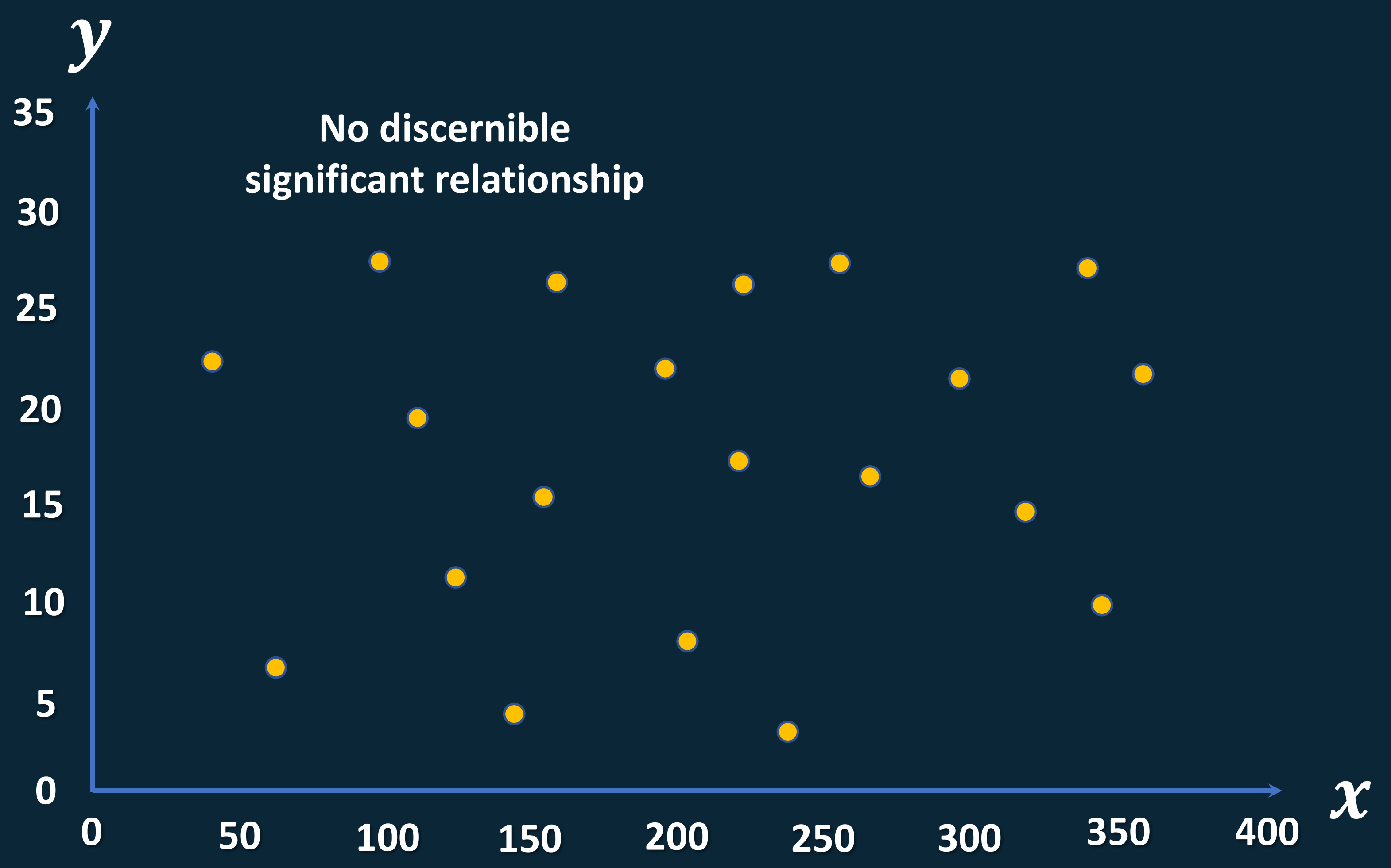
# 8. Variables & Correlation



$x$ $y$

| ID | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1 | 2.5 | 25 | 32 |
| 2 | 6 | 80 | 12 |
| 3 | 10 | 125 | 12 |
| 4 | 5 | 150 | 23 |
| … | … | … | .. |

$$(2.5, 25), (6, 80), (10, 125) \ldots$$



# 9. Variables



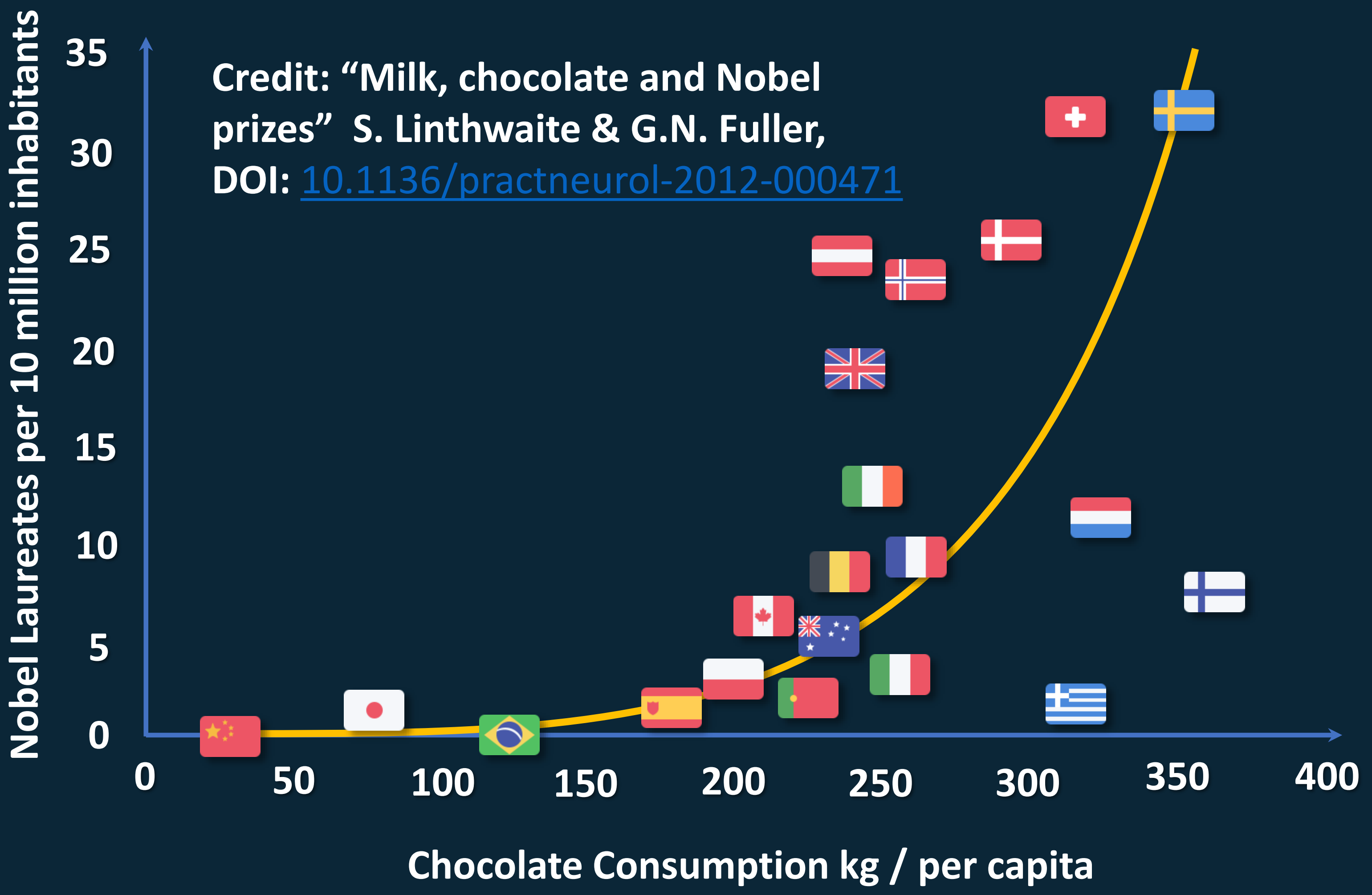
- If there appears to be a relationship between two variables, this doesn't mean there is.
- Correlation does not imply causation.
- Just because the independent variable seems to affect the dependent variable, does not mean it is responsible for the change.

Independent

Chocolate Consumption

Dependent

Nobel Laureates



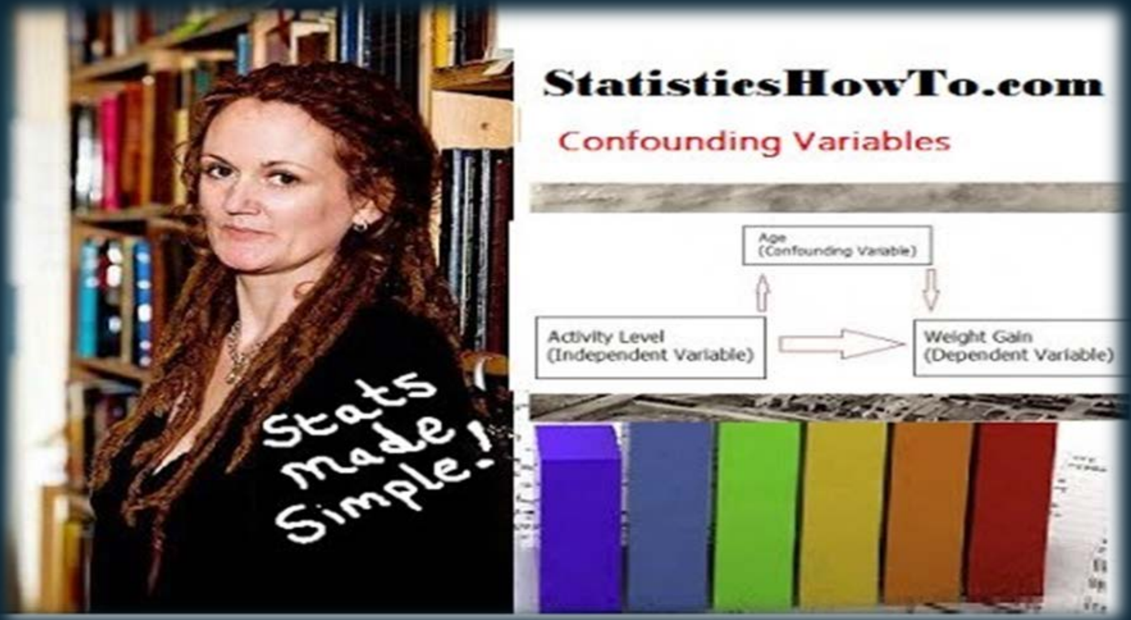
# 10. Correlation & Causation




**Credit: Crash Course**

# 11. Confounding Variables

- We sometimes encounter what we call confounding variables.
- These can be correlated with both the independent and dependent variables leading to spurious associations.
- Such variables are often unaccounted for (or not understood) when designing our experiments.
- We must ensure we aren't being influenced by such variables.
- We can mitigate the impact of confounding variables by controlling for them.
- This involves ensuring that these variables don't change during experimentation.




Credit: Stephanie Glen

# 12. Random Variables

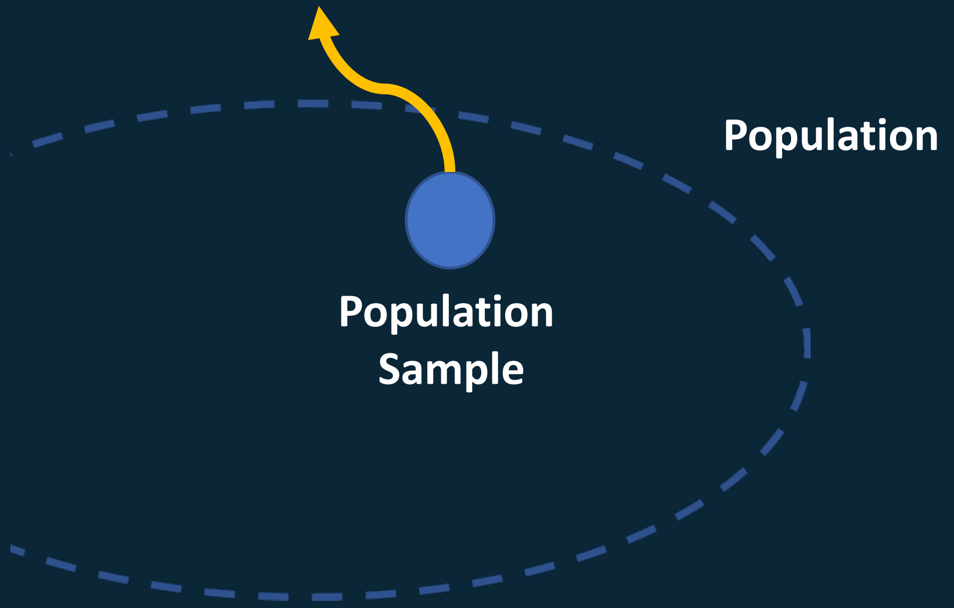
- The variables considered so far, can belong to a particular class of variables : random variables.
- Random variables are numerical variables whose values are determined via the outcome of a random event or phenomena.
- Most variables we'll encounter will be random.
- Random variables can be discrete or continuous.
  - Discrete random variables take on exact integer values (that's whole numbers) e.g. 1, 2, 3, etc.
  - Continuous random variables on the other hand take on real values, e.g. 1.2, 3.14, -45.2.


Credit: Khan Academy

# 13. Studies

## Sample Studies

Estimate population average, spread, minimum or maximum



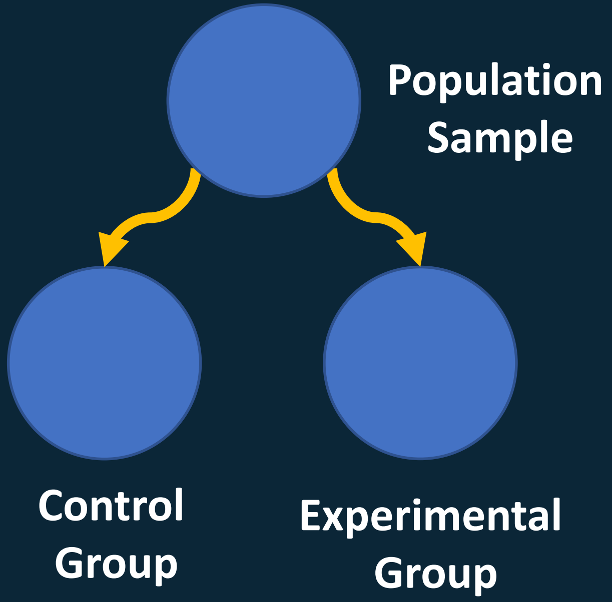

Estimate a population parameter

## Observational Studies


- Prospective studies
- Retrospective studies

Look for trends / Correlations

## Experimental Studies

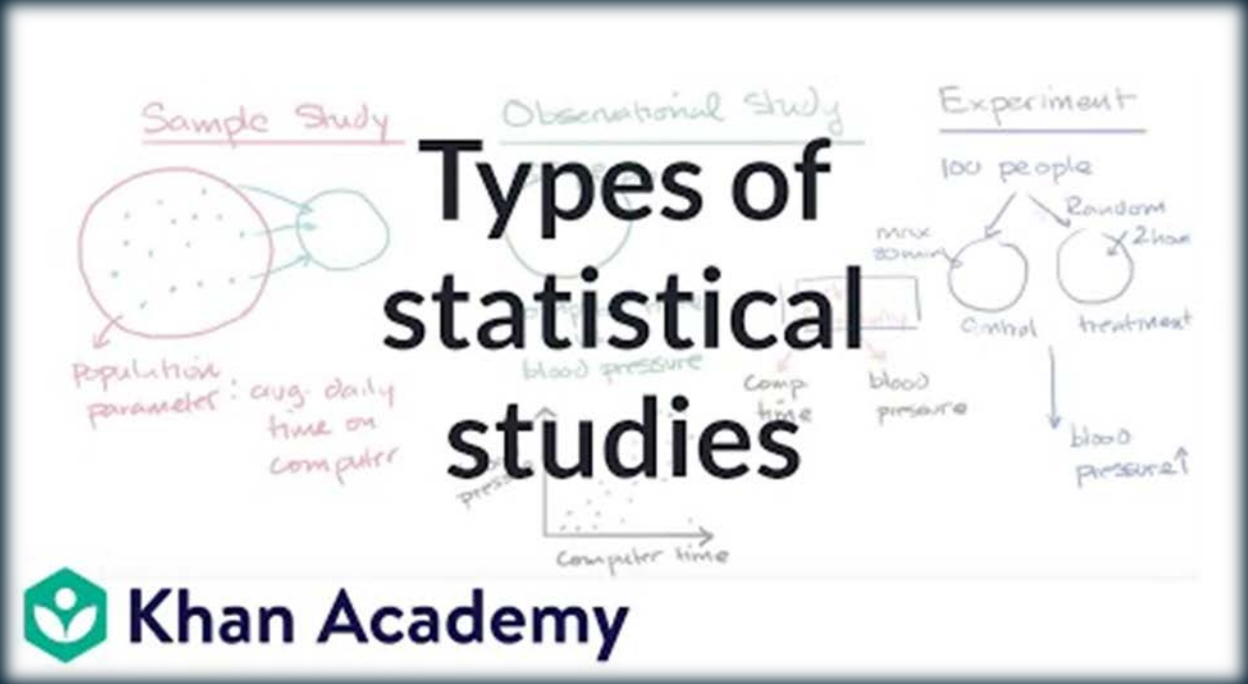



Answer specific question

Edge Hill University

</TECHUP_WOMEN>

# 14. Studies




**Credit: Khan Academy**

# 15. Experiment Design 1



- Let’s consider the main principles of experimental design, crucial for designing sound experimental studies.
- When designing an experiment, we do our best to control for any differences between the experimental and control groups.
- It can be difficult to control for everything.
- Randomization helps protect us from issues arising from variables we forgot to control for.
- We must randomize the cases we choose from a population to account for any variables not controlled.

# 16. Experiment Design 2

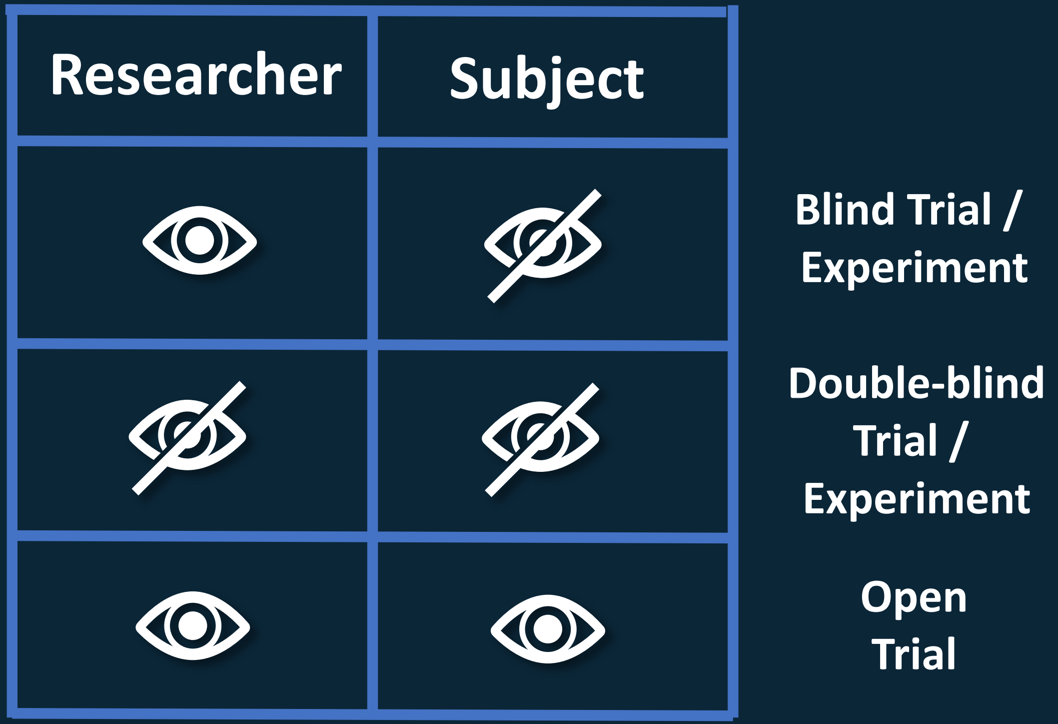
- We must design experiments that are **replicable**.
- If an experiment cannot be reproduced, we cannot validate the original results.
- Experiments become non-replicable when the data or tools used during an experiment are discarded.
- Good experiments come with **logs**, that describe what was done in sequence allowing for reproducibility.
- Sometimes we may know, or suspect, that variables other than the independent variable, affect the dependent variable.
- Under such circumstances we may group cases from the population based on this variable into **blocks** and then randomly select cases from each block to form the experimental and control groups.
- This strategy is known as **blocking**.

Edge Hill University

</TECHUP_WOMEN>

# 17. Experiment Design 3

- For instance in the medical domain, researchers may keep patients in the dark about their treatment.
- In this case the patients are said to be “blind”, thus this is a blind experiment.
- This helps researchers to avoid influencing patients simply by telling them about the medication they’re given – this helps avoid the placebo effect.
- In some cases experiments are double-blind. Here the researchers don’t know about the treatments patients are receiving. Any subconscious hints they may give off about the treatment will be avoided.
- These approaches are important for data science.
- Experiments with no “blinding” are known as open trials.



# 18. Samples

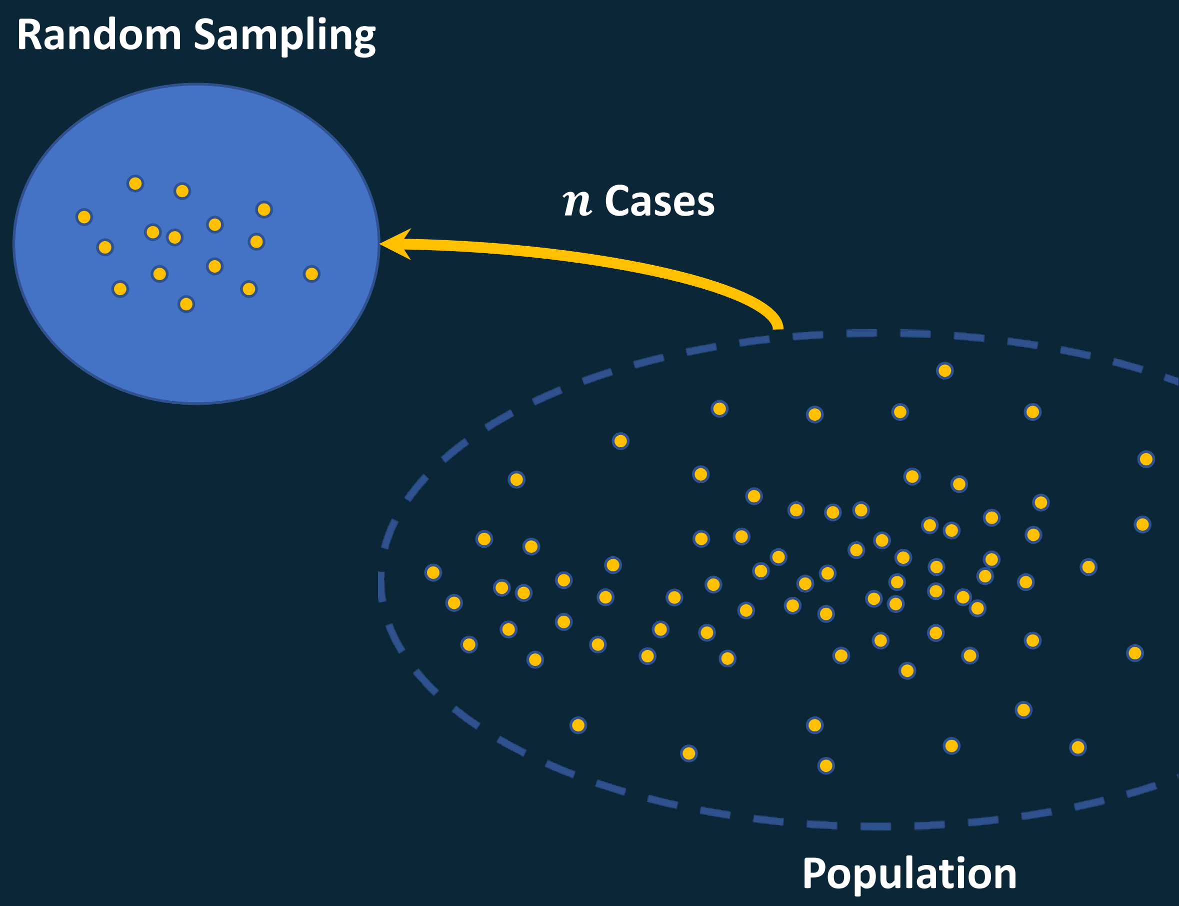


- We've now learned about the different types of study we can undertake (sample, observational & experimental studies).
- All three require the acquisition of data from some population.
- Data science questions, and research questions in general, are usually targeted toward a specific population.
- There could be very <u>many</u> examples in a population - too expensive to collect all.
- Instead we collect data from an unbiased sample of the population.
- Ideally we aim to undertake data science investigations on large representative samples of data.
- We can create samples by applying a sampling methodology to population data.

# 19. Random Sampling

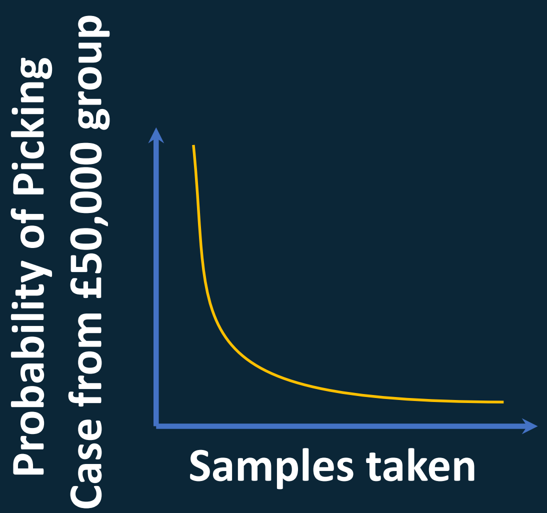


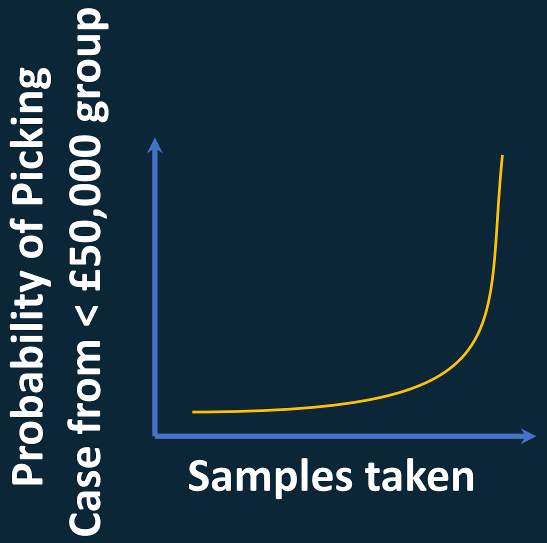
- The most commonly used methodology is random sampling.
- This method simple chooses $n$ cases from the population at random. Also known as random sampling without replacement.
- This type of sampling can be very effective.
- However, consider the following situation: choose the default credit limit for customers based on a random sample of 100 cases.
- Detail – population skewed toward those with incomes > £50,000. That is, 90% of cases belong to that category.



# 20. Random Sampling

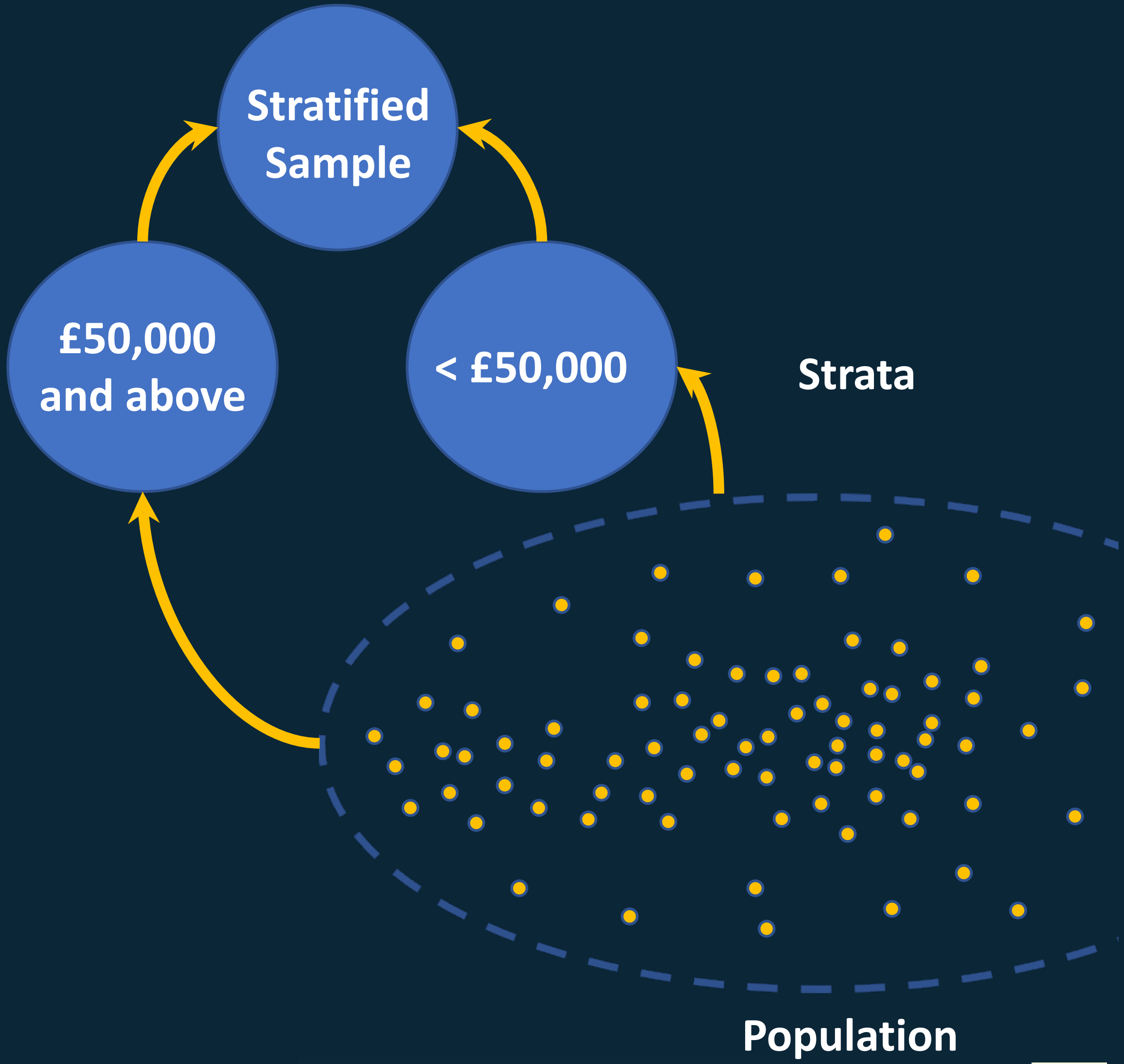
- If 90% of the samples belong to customers earning over £50,000, there is a 90% (4 in 5) chance, of picking a case from this group.
- Since samples are not replaced after being chosen this probability drops over time.
- With each sample draw from the population, the probability of picking a case from the £50,000 group diminishes.
- So if we take enough samples, eventually we'll start to get cases from the less than £50,000 group. But to get to this point we may need to set $n$ to a very high value!



# 21. Stratified Sampling

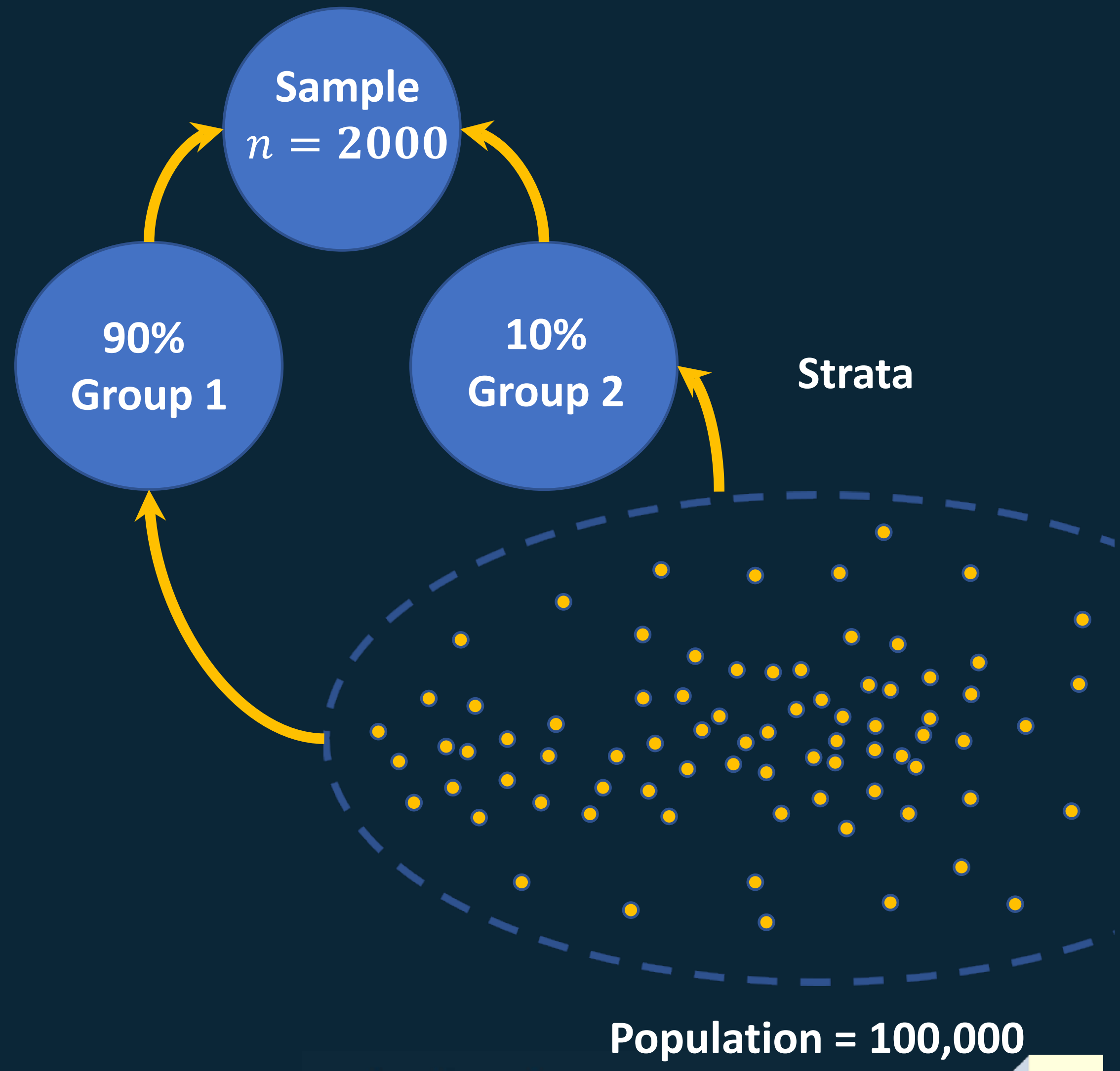
- When randomly chosen samples become intrinsically biased, we can apply a method called stratified sampling to try and get an unbiased selection of cases.
- Here were randomly sample a population as before.
- Yet this time, we maintain the proportion of high and lower earners in the sample by randomly sampling for each group or "strata".
- The resulting the sample is split so that the proportion of cases in each group, reflects the split in the population.
- Stratified sampling is useful, as it allows us to preserve population splits in our samples.
- In other words, it let's us preserve the true population true distribution.



# 22. Stratified Sampling - example



- Suppose we have a population of 100,000, and the population is split so that 90% of cases belong to Group 1, and 10% to Group 2.
- We need to create a sample of the population of size $n$ = 2000.
- How many random samples should we make for group 1 and 2?
    - For group 1 we need to randomly sample $n \times 0.9 = 1800$ times from this strata.
    - For group 2 we need to randomly sample $n \times 0.1 = 200$ times from this strata.
- This will give us a sample representative of the population.



Edge Hill University

# 23. Weighted Sampling

</TECHUP_WOMEN>

- In some cases random and stratified sampling may not help.
- This is true when preserving the population distribution is unhelpful.
- This applies when we're trying to target rarer groups in our populations.
- In these cases we can use weighted random sampling. Works by weighting the sampling so it favors one strata over another.
- Suppose we have a population of 100,000, and the population is split so that 90% of cases belong to Group 1, and 10% to Group 2.
- We need to create a sample of the population of size $n$=2000.
- How many random samples should we make for group 1 and 2? We can use the simple formula $n \times w$ , which is simply the number of samples multiplied by the weighting.
  - For group 1 we set the weighting to $w$ =0.5. We need to randomly sample 2000 $\times w$ = 1000 times from this strata.
  - For group 2 we set the weighting to $w$ =0.5. We need to randomly sample 2000 $\times w$ =1000 times from this strata.
- The weights must add up to 1.

# 24. Sampling & Anecdotal Data

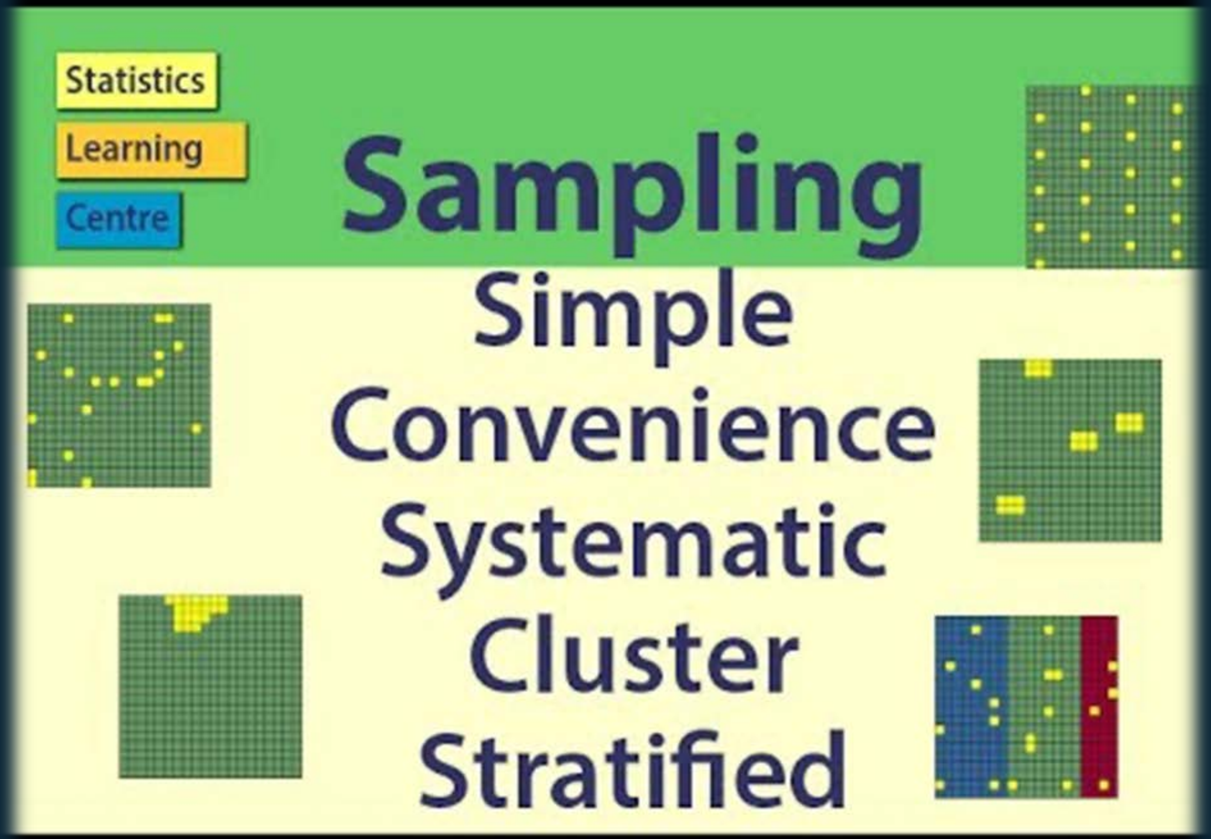


- There are many times of sampling methods available.
- The best one to use depends on the question you're trying to investigate.
- It is therefore up to you as the data scientist to choose an appropriate method.
- When a sample contains very few examples, any investigation we undertake can only yield what we call anecdotal evidence.
- This type of evidence may be true, yet can be <u>very</u> dangerous to use.
- I would strongly caution against using anecdotal data for anything other than providing general impressions.

# 25. More Sampling



- There are some real-world issues to consider when thinking about sampling.
- Suppose you conduct a customer survey. If only 30% of customers respond, is that sample representative? If the sample is not representative you may have encountered non-response bias.
- Another common issue arises in this scenario, when certain subsets of customers are able to complete the survey because it's simply easier for them to do so.
- This is known as a convenience sample.




Credit: Dr Nic's Maths and Stats

# 26. Checkpoint



**We've reached a checkpoint. Stop here and take a rest if needed. Let's recap what we've introduced so far.**

- **Data sets.**
- **Populations vs. samples.**
- **Different types of variable.**
- **Scatter plots.**
- **Different forms of correlation.**
- **Experimental studies.**
- **Experimental design**
- **Sampling methodologies.**

**That's quite a lot! Take some time to digest that material, then return when you can. When ready, proceed to learn how we apply statistics to data.**

# 27. Summary Statistics



- Once we've collected sample data, we can start studying it.
- The first step almost always involves computing summary statistics that describe the data.
- Such statistics are incredibly useful - they can reveal broad trends, are easy to interpret, and easy to compute.
- Suppose a company wants to know if targeted advertisements lead to an increase in the volume of sales (i.e. number of individual items sold) across a broad range of consumer types.
- They form a research question – does targeted advertising increase sales volume?
- To answer this, the company collects sample data from 450 website customers chosen at random across a range of demographic groups.
- The groups are equally split into two – an experimental group who will be exposed to targeted advertising and a control group.

| Customer | Group | Individual product purchases |
|---|---|---|
| 1 | Experimental | 1 |
| 0 | Experimental | 0 |
| 3 | Experimental | 2 |
| … | … | … |
| 449 | Control | 3 |
| 450 | Control | 0 |
| 451 | Control | 0 |

# 28. Summary Statistics



**Original Data**

| Customer | Group | Individual product purchases |
|---|---|---|
| 1 | Experimental | 1 |
| 0 | Experimental | 0 |
| 3 | Experimental | 2 |
| … | … | … |
| 449 | Control | 3 |
| 450 | Control | 0 |
| 451 | Control | 0 |

**Summary Data**

| Group | Customers | Customers making purchases 1 or more purchases | Total Item Sales |
|---|---|---|---|
| Experimental | 225 | 95 | 140 |
| Control | 225 | 100 | 124 |
| Total | 450 | 195 | 264 |

# 29. Summary Statistics



- In total 195 customers made 1 or more item purchases.
- The proportion of customers making a purchase overall = 195/450 = 0.433333 = ~43%.
- For the groups we have the following:
  - Control group proportion = 100/225= 0.444444 = ~44%.
  - Experimental group proportion = 95/225 = 0.4222222 = ~42%.
- Here we see that the control group had a higher ratio of customers making purchases.
- We note that the experimental group did yield more sales overall.
- We can compute the average sales per customer that made 1 or more purchases (a spending customer), to determine if there is a difference.
- Try to compute that now.

**Summary Data**

| Group | Customers | Customers making purchases 1 or more purchases | Total Item Sales |
|---|---|---|---|
| Experimental | 225 | 95 | 140 |
| Control | 225 | 100 | 124 |
| Total | 450 | 195 | 264 |

# 30. Summary Statistics



- The avg. sales per customer who bought something in the control group: = 124/ 100 = 1.24 = ~ 1.2 sales per spending customer.
- For the experimental group we have 140 / 95 = 1.47368421052631 6 = ~1.5 sales per spending customer.
- Summary statistics provide some initial evidence suggesting that targeted advertising did increase the sales per customer.
- Summary statistics are useful, but be careful with them. They may not generalize well past the data you currently have.
- Consider if they would still apply with respect to a much larger customer sample.

**Summary Data**

| Group | Customers | Customers making purchases 1 or more purchases | Total Item Sales |
|---|---|---|---|
| Experimental | 225 | 95 | 140 |
| Control | 225 | 100 | 124 |
| Total | 450 | 195 | 264 |

# 31. Summary Statistics - $\mu$



- The mean - a summary statistic that computes the average (center) of a data.
- Two forms of the mean that we consider as data scientists.
- First there is the **population mean**, mu ($\mu$), the average of the population.
- The population mean is sometimes described with slightly different notation.

$N$ is the number of cases in the population = 5

| Example | Variable $x$ |
|---|---|
| 1 | 3 |
| 2 | 2 |
| 3 | 1 |
| 4 | 2 |
| 5 | 1 |

$$\mu = \frac{3+2+1+2+1}{5} = \frac{9}{5} = 1.8$$

Population

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

Population Mean, $\mu$ = $E(x)$ is the expected value (mean) of $x$

# 32. Sigma (Σ) Diversion



$\Sigma$ means the sum. e.g. $\Sigma\,[1,2,3] = 6$

**Upper limit**

$\sum_i^N i$ means sum the natural integers from $i$ to $N$. e.g. if $i = 1$ and $N = 3$, $\Sigma\,[1,2,3] = 6$

**Lower limit**

$x$ non-indexed variable $x_i$ indexed variable

$\sum_i^N x_i$ means sum the variables in $x$ from position $i$ to $N$.

**Given** $x = [4, 5, 6]$ **Then,**

$x_1 = 4 \quad x_2 = 5 \quad x_3 = 6$

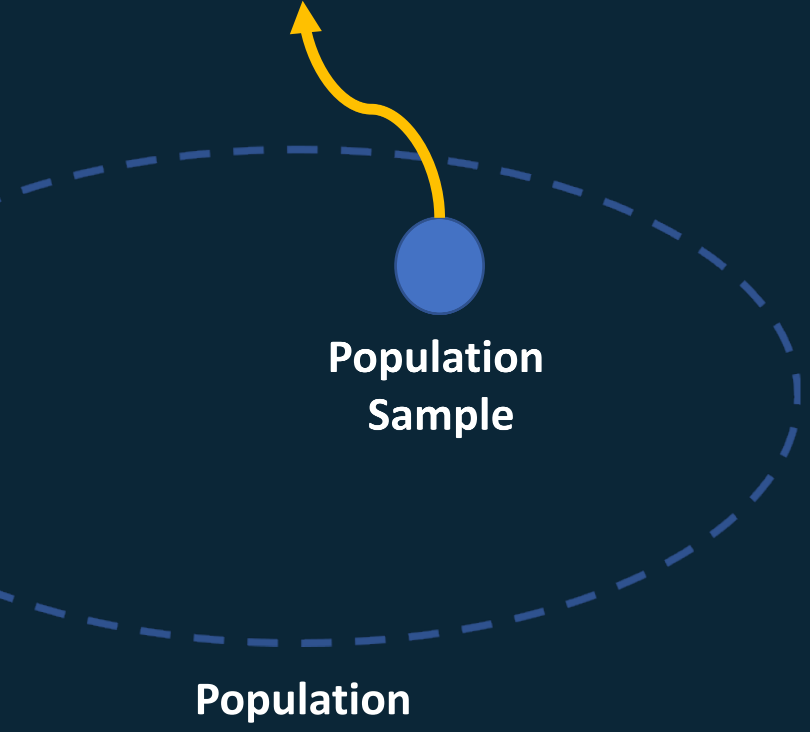
$\sum_{i=1}^{N=3} x_i$

$= x_1 + x_2 + x_3 = 15$

# 33. Summary Statistics - $\bar{x}$

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

Sample Mean, $\bar{x}$




- Second, there is the sample mean.
- This is the mean of a data sample, taken from a population sample.
- The formula for the **sample mean** is almost identical to the population mean. Except it uses a different value for the parameter $n$ (as $N \neq n$). Here little $n$ describes the number of examples in the sample – not the whole population.
- We normally denote the sample mean as x-bar ($\bar{x}$).
- The mean is an important metric, as it allows us to estimate the central value of a dataset. This is an important concept to understand.

# 34. Variance & Standard Deviation



- The mean helps us describe the center of a dataset.
- However it is also important to understand how variable or spread the data is – especially because this may help us determine if outliers are impacting our summary statistics.
- The distance of an observation from the population or sample mean, is called it's deviation.
- We use two variability measures in statistics to measure this deviation – the variance and the standard deviation.
- There

# 35. Variance & Standard Deviation



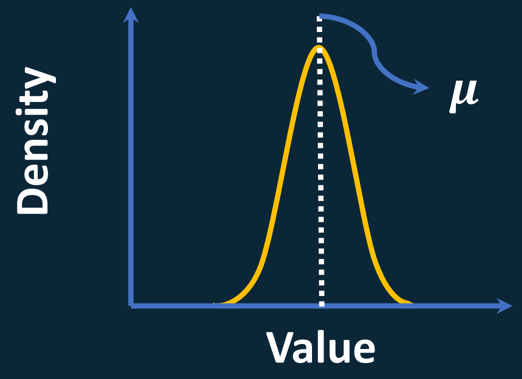
**Sample Variance**

$$s^2 = \frac{\sum_{I=1}^{n}( x_i - \mu)^2}{n-1}$$

Population Sample

Population

**Less variance**




**More variance**


**Population Variance**

$$\sigma^2 = \frac{\sum_{I=1}^{N}( x_i - \mu)^2}{N}$$

- The population variance denoted by Sigma squared ($\sigma^2$).
- The sample variance is given by ($s^2$).
- The correct formula to use depends on whether or not your dealing with a population or a sample – and that's for you to determine.

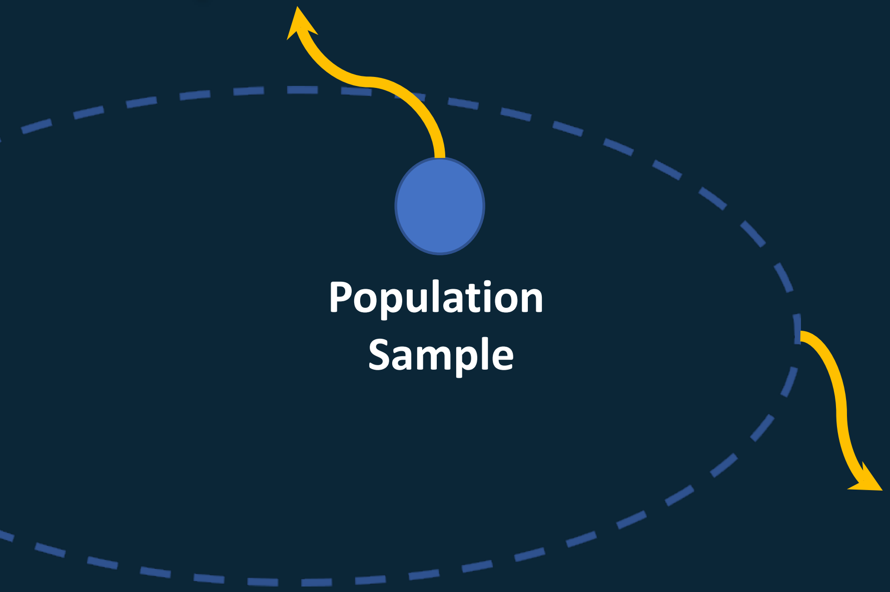
# 36. Variance & Standard Deviation



There

**Sample Standard Deviation**

$$s = \sqrt{\frac{\sum_{I=1}^{n}(x_i - \mu)^2}{n-1}}$$




Population

**Sample Standard Deviation**
**Is sample variance squared.**

$$s = \sqrt{s^2}$$

**Population Standard Deviation**
**Is population variance squared.**

$$\sigma = \sqrt{\sigma^2}$$

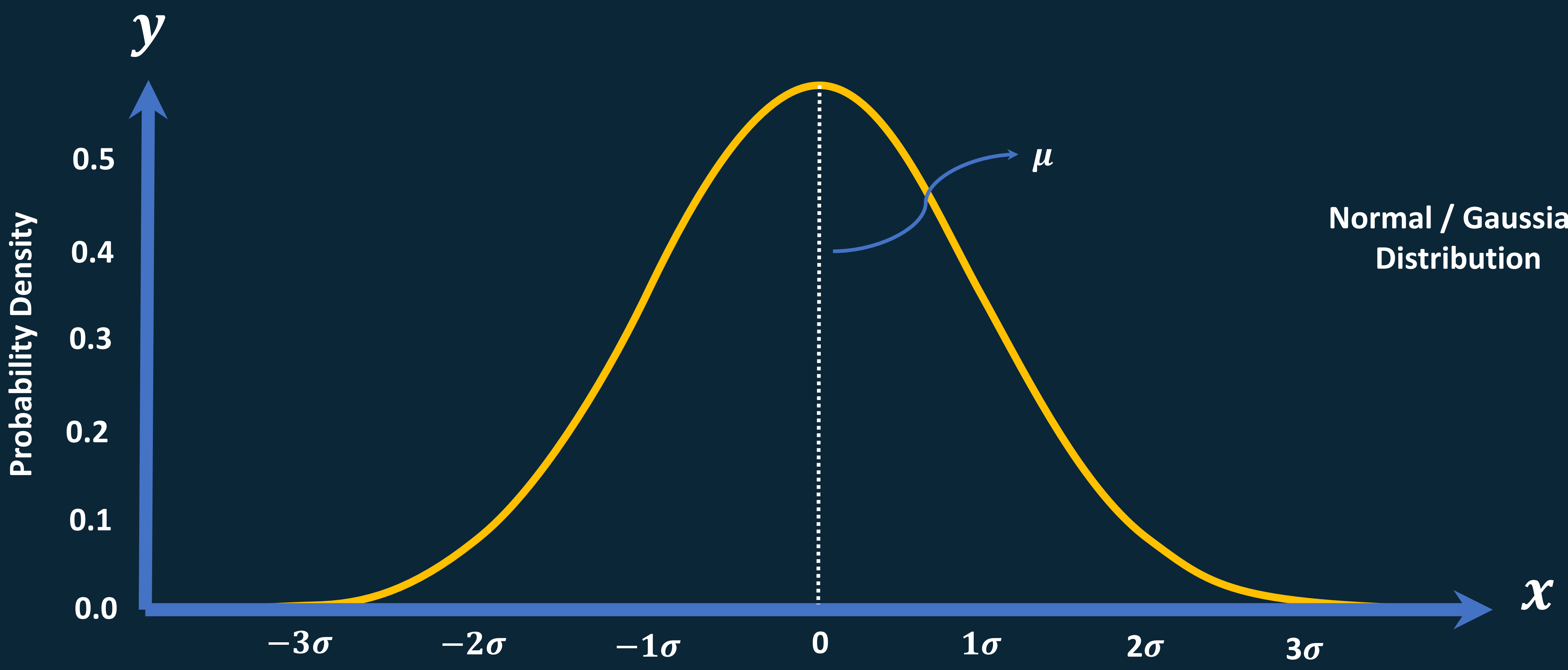
**Population Standard Deviation**

$$\sigma = \sqrt{\frac{\sum_{I=1}^{N}(x_i - \mu)^2}{N}}$$

- **The population standard deviation denoted by Sigma ($\sigma$).**
- **The sample variance is given by ($s$).**
- **Note the relationship between the variance and the standard deviation. The standard deviation, both for a population and a sample, is equal to the square root of the variance.**

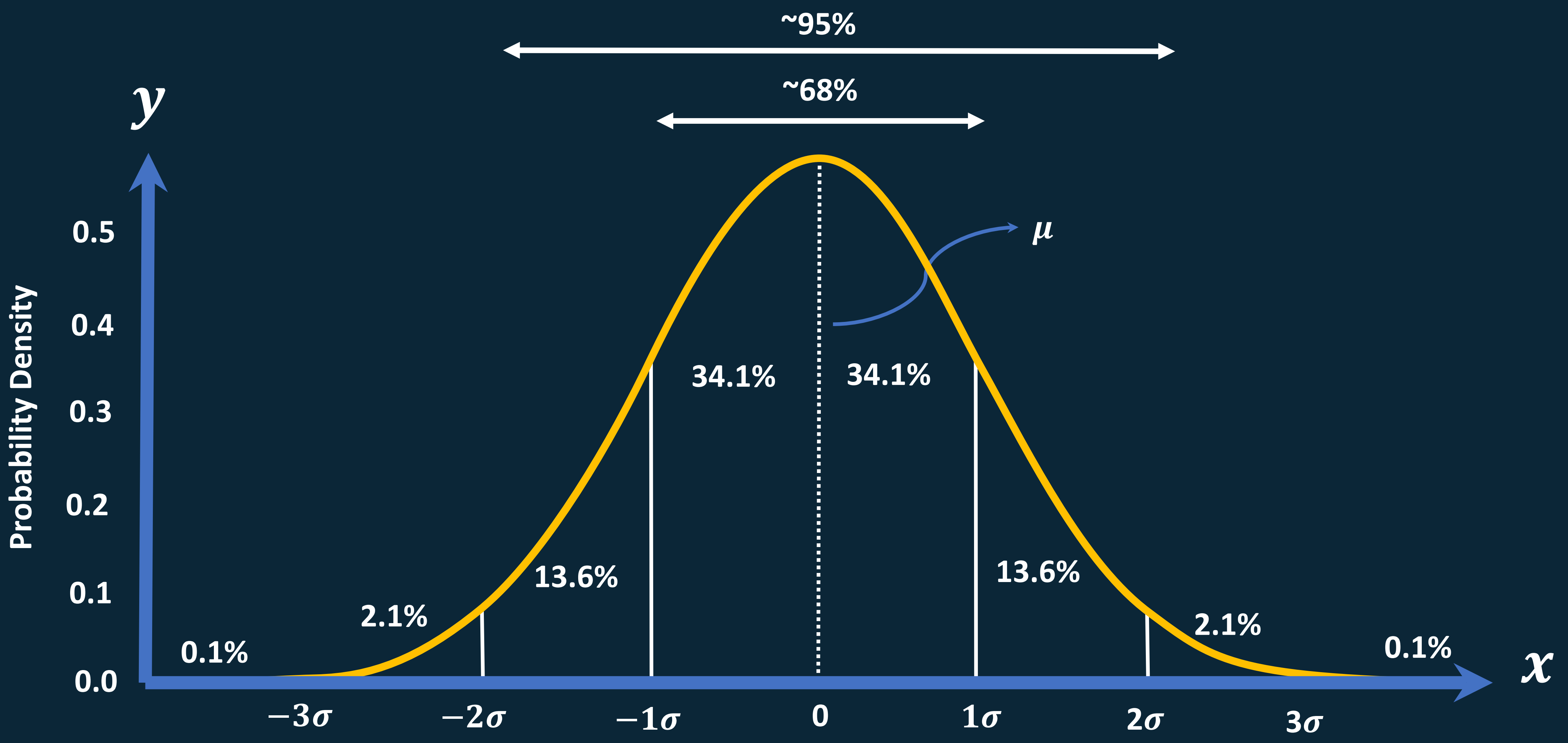
# 37. Understanding Deviation






**Normal / Gaussian Distribution**

# 38. Understanding Deviation

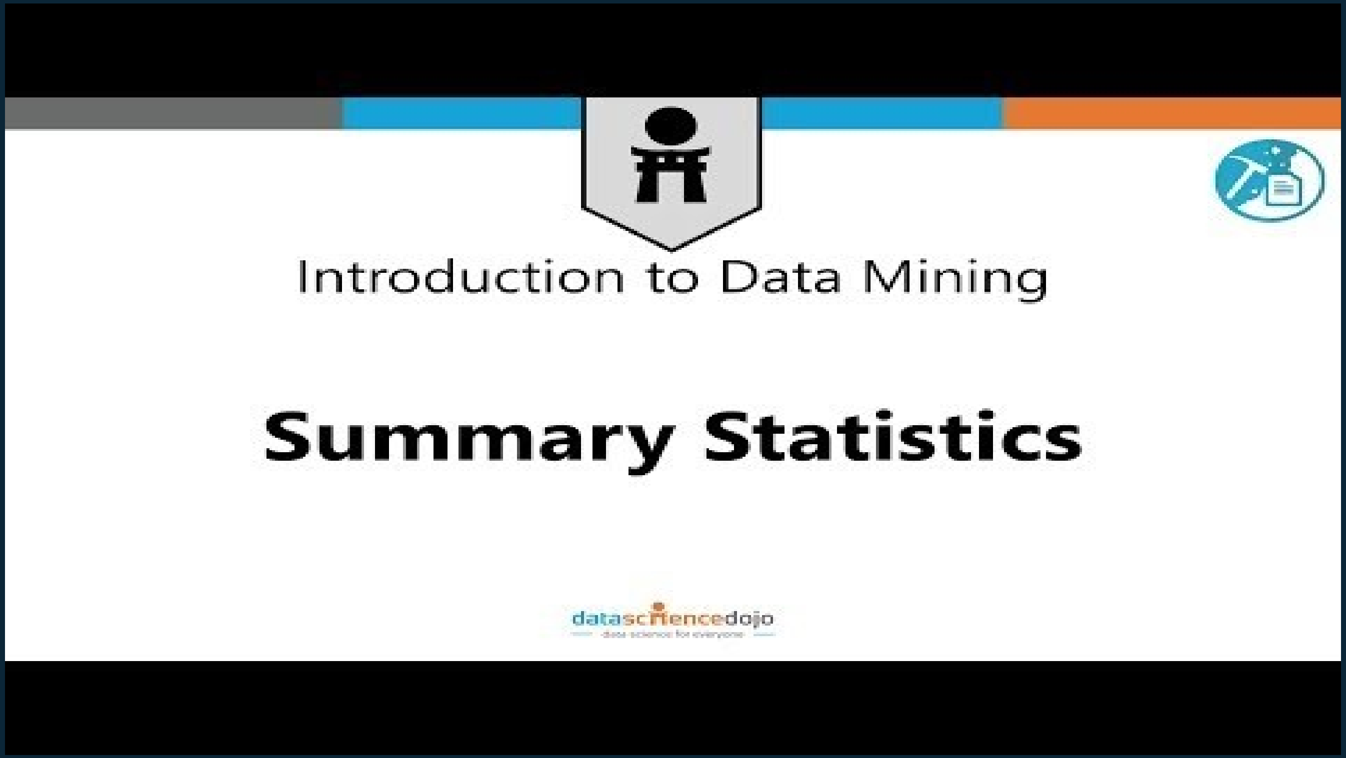


# 39. Spread




Credit: Jeremy Jones

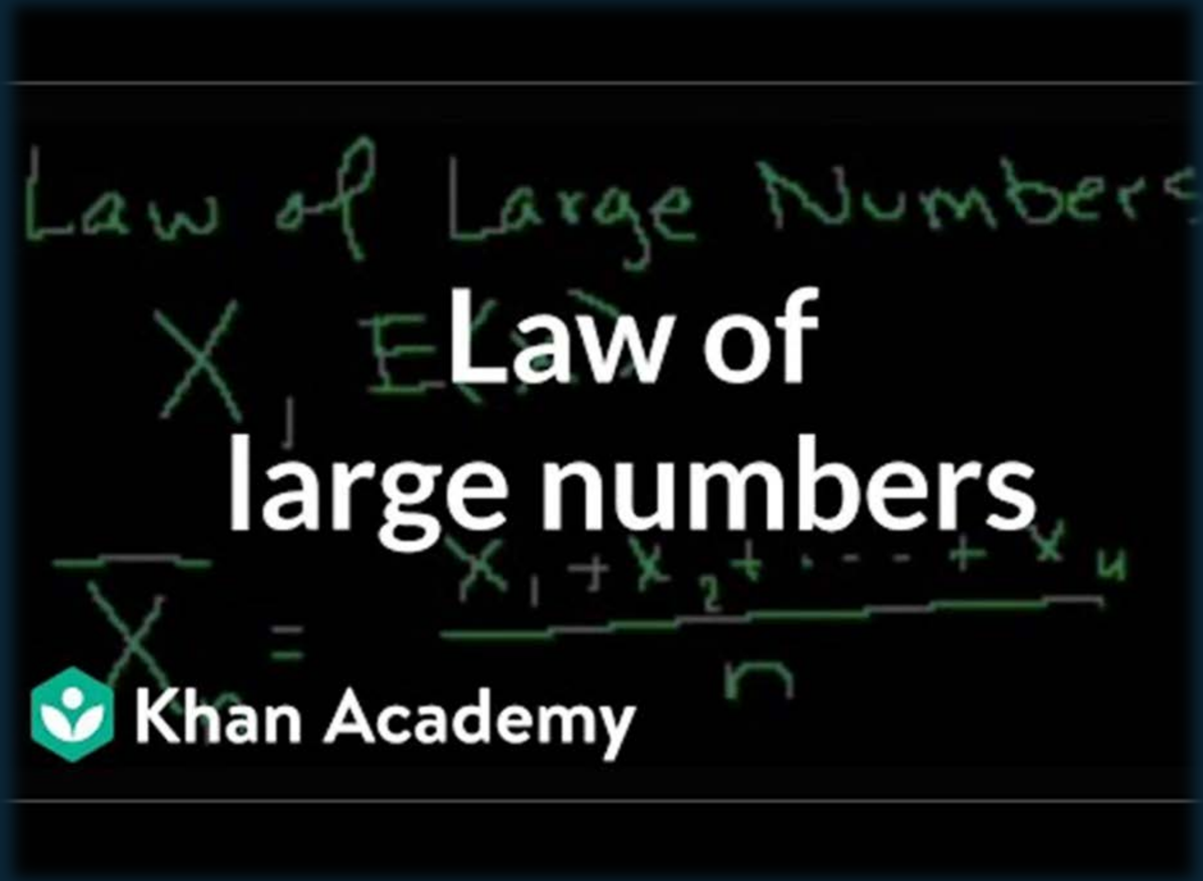
# 40. Summary Statistics - Review




**Credit: Data Science Dojo**

# 41. Law of Large Numbers



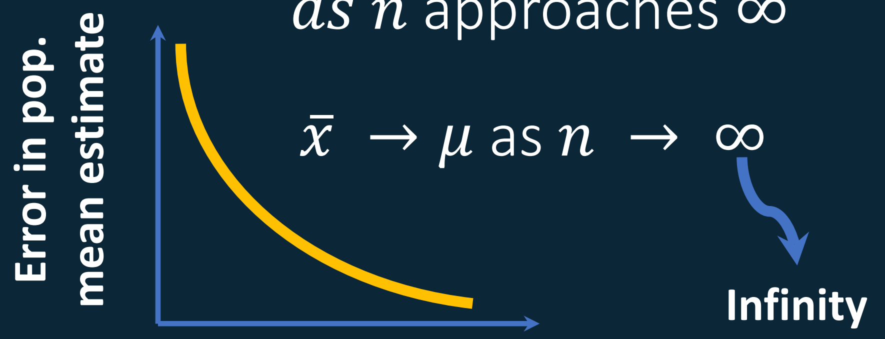



Credit: Khan Academy

- **It can be unusual for us to have access to the population mean.**
- **Sometimes the population mean cannot be computed – without exhaustive data collection.**
- **The sample mean is still very useful – useful for estimating the population mean.**

$\bar{x}$ approaches μ
$as\ n$ approaches ∞

$\bar{x} \rightarrow \mu$ as $n \rightarrow \infty$

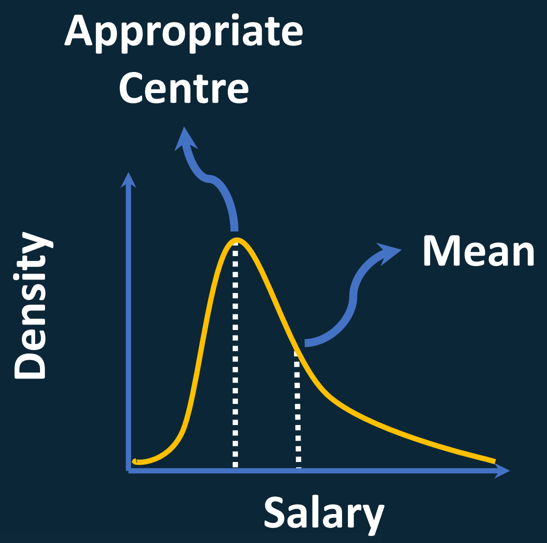


# 42. Outliers

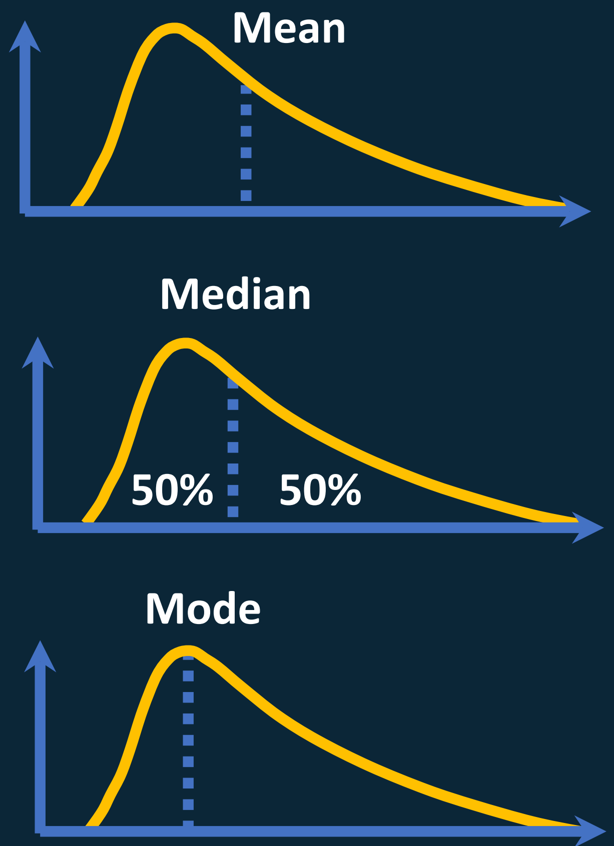


| Example | Variable $x$ |
|---|---|
| 1 | £17,000 |
| 2 | £25,000 |
| 3 | £24,000 |
| 4 | £24,000 |
| 5 | £105,000 |

- We can use the sample mean to reasonably estimate the population mean, if we have enough samples.
- However the population mean and the sample mean can be skewed.
- This happens because the mean is not robust to outliers.
- Outliers are "extreme" data points that skew the average providing a misleading impression of the central point of the data.
- Consider the data shown in the table. It describes self-reported salaries for credit card customers in London.
- Aim: to estimate the optimal amount of credit to offer a customer.
- In this example the mean salary is £39,000 - much higher than all but one salary in the data.
- In such cases we can use different measures of centrality, to estimate the midpoint of the data.



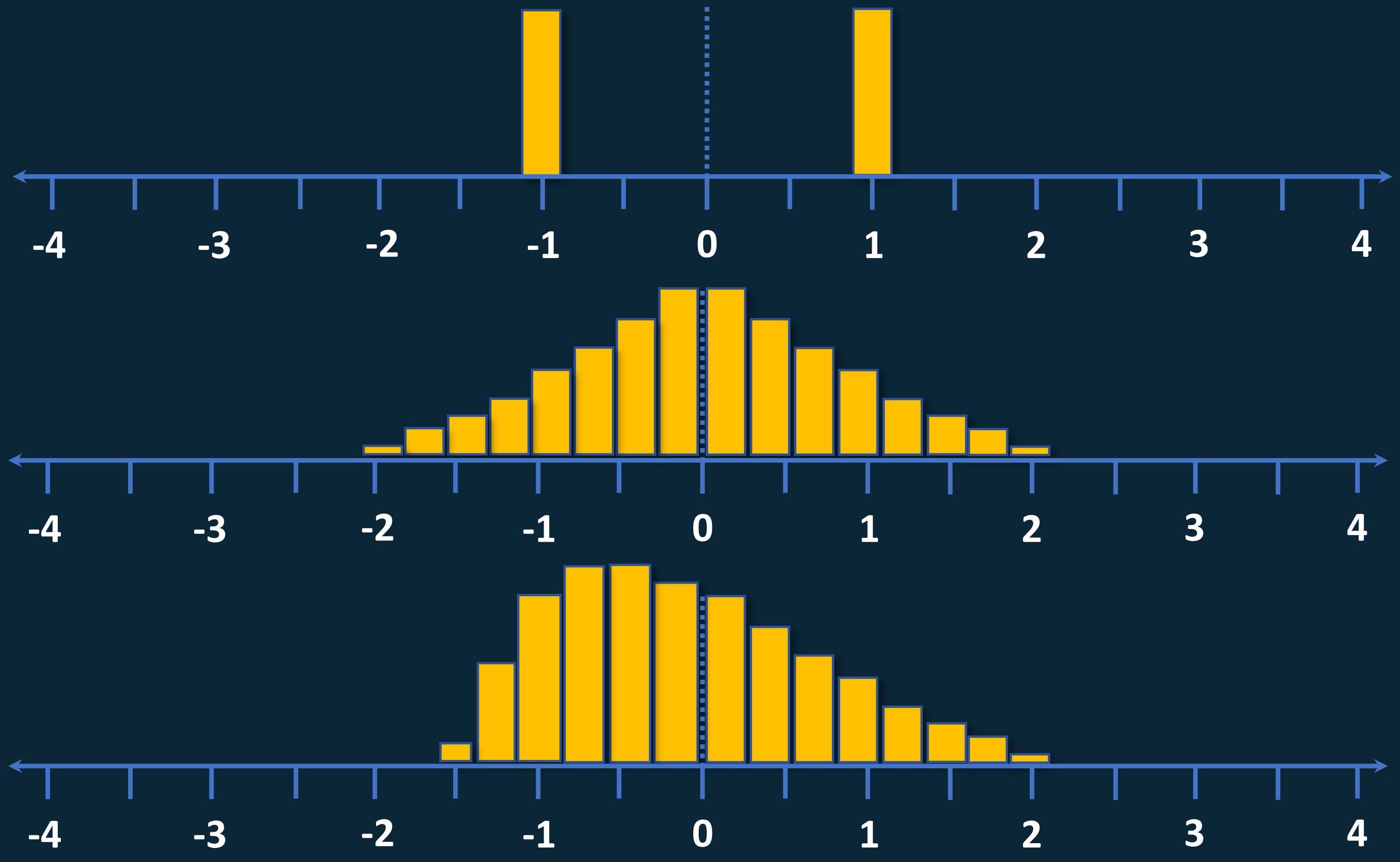
# 43. Outliers




- One way to overcome the influence of outliers on our analyses, is to use more robust statistics.
- For instance, we could use the **median** to estimate the midpoint.
- The median is the central value in the data when ordered – in this case, the value in row 3 (£24,000). If there are an even number of observations, then the median is the central two data points divided by two.
- We could also use the **mode** – this is the most common value found in the data.
- In this case the mode is £24,000.

| Example | Variable $x$ |
|---|---|
| 1 | £17,000 |
| 2 | £25,000 |
| 3 | £24,000 |
| 4 | £24,000 |
| 5 | £105,000 |

# 44. Misleading Statistics



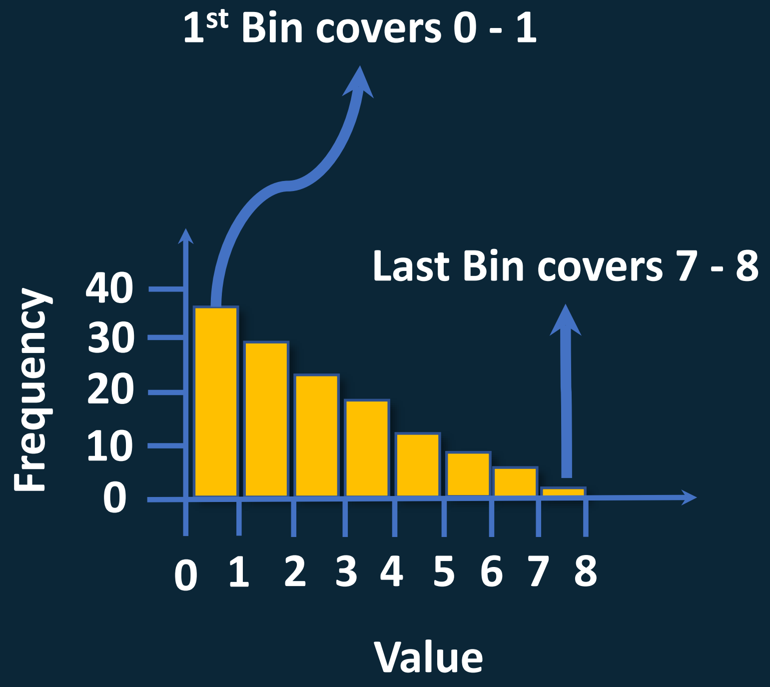

- Data can be confusing however.
- Consider these three fictitious datasets which are summarized by way of a histogram.
- Believe it or not, these data sets have the same mean ($\mu = 0$) and the same standard deviation ($\sigma = 1$).
- This is why it becomes important to visualize our data, to better interpret the data we have.

# 45. Histograms



| Values | Freq. |
|---|---|
| 0 to 1 | 36 |
| 1 to 2 | 29 |
| 2 to 3 | 23 |
| 3 to 4 | 18 |
| 4 to 5 | 10 |
| 5 to 6 | 8 |
| 6 to 7 | 5 |
| 7 to 8 | 1 |
| 8 + | 0 |

- In the previous slide we introduced the histogram.
- This is a type of data visualization tool, that describes the frequencies of observations / outcomes in data.
- Outcomes are first grouped in to bins covering ranges.
- We count the frequencies of values falling in to these ranges.
- Finally, we plot bars where bar height is equal to the frequency of examples falling in the bin range.
- The histogram is a simple but elegant data visualization tool.

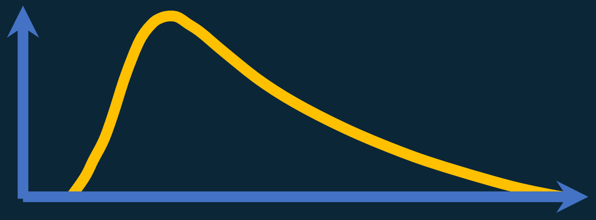


# 46. Extra Terminology

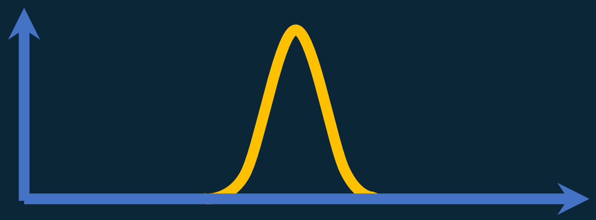


- There is some terminology associated with the characteristics of data we analysis.

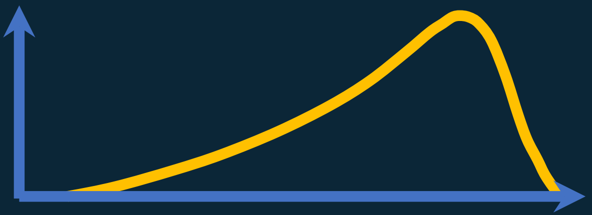
Left Skewed



Unimodal

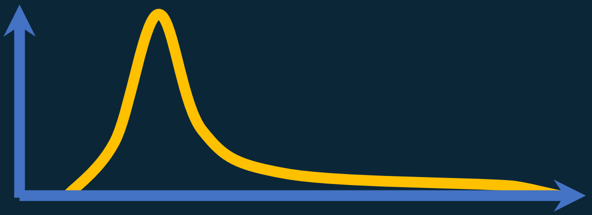


Right Skewed

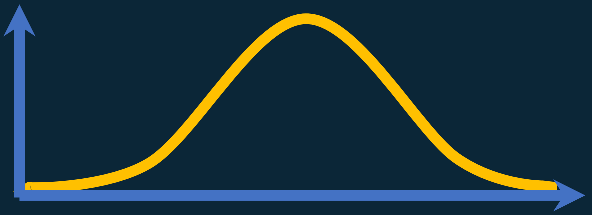


Bi-modal

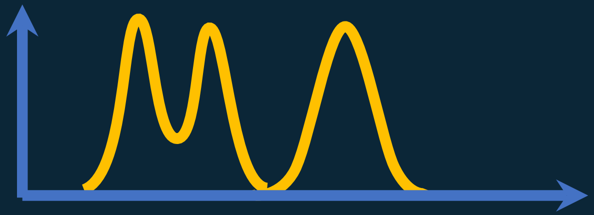
Long-tailed



Symmetrical



Multi-modal



# 47. Checkpoint

**We’ve reached another checkpoint. Let’s recap what we’ve introduced.**

- **Data sets.**
- **Populations vs. samples.**
- **Different types of variable.**
- **Scatter plots.**
- **Different forms of correlation.**
- **Experimental studies.**
- **Experimental design**
- **Sampling methodologies.**
- **Summary statistics (mean, variance, standard deviation, mode and median).**
- **Why such statistics are important.**
- **Sample versus population statistics.**
- **The law of large numbers.**
- **Outliers and their impact on summary statistics.**
- **Histograms.**
- **Terms used to describe dataset characteristics.**

**This puts you in a great place to tackle our next topics - probability and data distributions.**