

KEY TERMINOLOGY

Dataset: Your collection of data (e.g., survey responses, lab test results, etc.)

Population: The entire group of people (or animals or whatever unit) that you’re interested in researching. E.g. the population of a city or country.

Sample: The subset of the population that you can collect data from/about.

Variable: Something that changes in value and can be directly measured.

Hypothesis: An assumption that can be supported or rejected by your data.

Descriptive statistics: Statistics that help you describe or summarise each variable in your dataset - i.e., that describe your **sample**.

Inferential statistics: Statistics that help you test hypotheses, assumptions or predictions about an entire **population** using your dataset.

KEY CONCEPT

How closely your sample reflects your population of interest is referred to “representativeness”.

Generally speaking, you’ll want your sample to be as representative as possible, as this allows you to draw conclusions about the population more confidently.

DESCRIPTIVE STATISTICS



Easy 15-Minute Explainer

FREE Stats Cheat Sheet



TYPES OF DATA

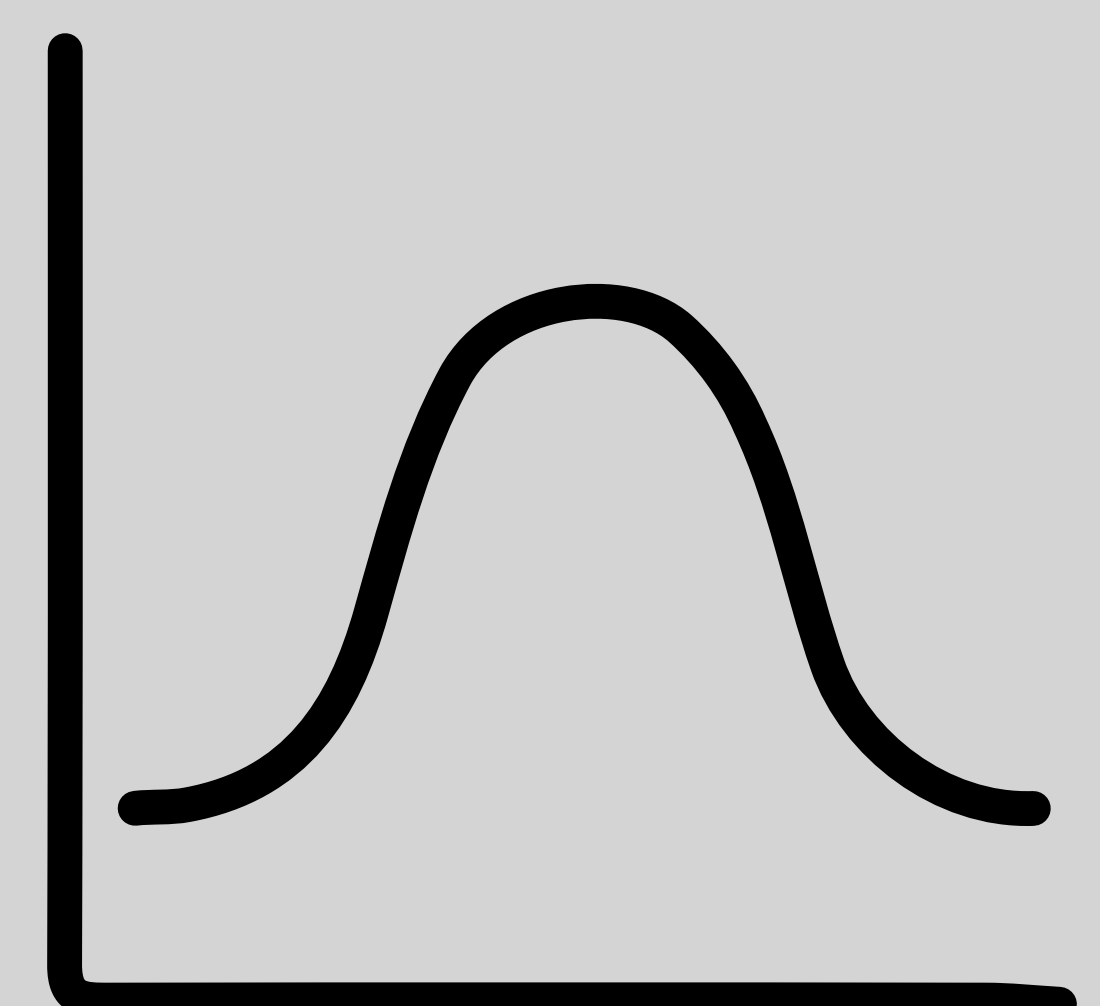
Data can come in all shapes and sizes, and it’s important to understand what type of data you’re working with, as this will influence what types of statistical tests (descriptive and inferential) you can apply to your dataset. At the broadest level, the two types of data are categorical (category-based) and numerical (numbers-based).

	Categorical	Numerical
What is it?	Data that reflect categories of things. For example, male/female, left/right, etc. You can assign a number to reflect each category.	Data that are naturally numbers-based. For example, age, height, salary, etc.
Common descriptive statistics used	<ul style="list-style-type: none"> • Count/Frequency • Proportion/Percentage 	<ul style="list-style-type: none"> • Mean (average) • Median • Standard deviation
Common graphs	Bar chart, pie chart  	Box plot, stem and leaf, scatter plot

CONNECTED CONCEPT: NORMAL DISTRIBUTION

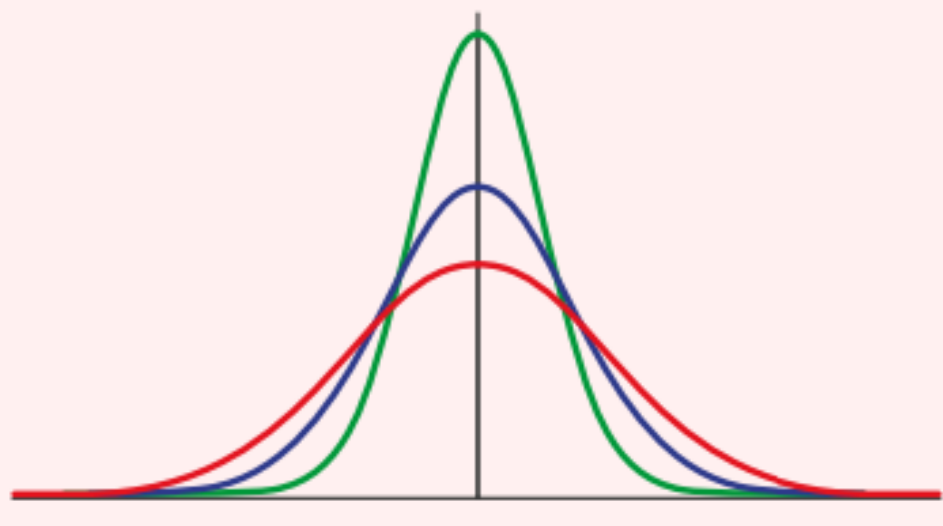
A normal distribution, or bell curve, is a graph showing how variables like heights or scores are spread out. When a distribution is “normal”, most values are near the average, creating a high peak in the middle, and fewer values are far from the average, making the graph taper off like bell ends on both sides.

You’ll likely encounter many bell curves within your research, but it’s useful to note that data can take many other shapes as well.



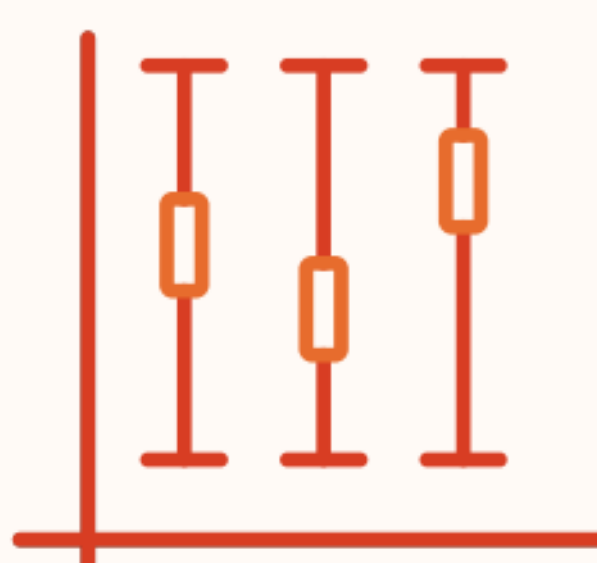

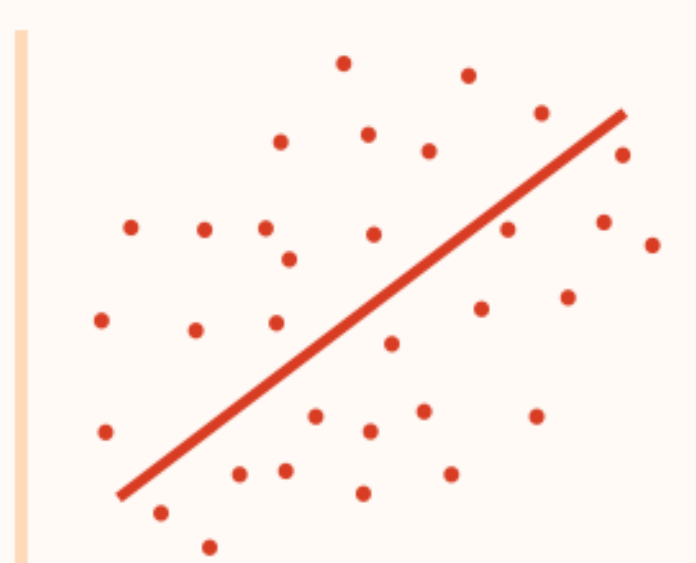
WHY DESCRIPTIVES MATTER

Descriptive statistics allow you to provide a simple summary of large amounts of data, making it easier for your reader (and yourself!) to understand patterns, trends, and key points at a glance. They also help you identify potential issues or errors within your dataset, so that you can address those before doing further analysis or drawing conclusions.

Test	Definition	Example
Mean	The average value of the variable in the sample.	The average height of a British woman is 161.6cm.
Median	The “middle value” if you were to order your data from smallest to largest.	The median income in the US is \$31,100 per year.
Standard deviation	A measure of the spread of your data. (i.e., how “widely” your data spreads from the mean). 	The standard deviation of the height of British woman is 5.9 cm.
Frequency	A number that expresses the size or occurrence of a specific category or event in a variable, in relation to the total sample.	The sample consists of 65 women and 20 men.
Proportion (percentage)	A fraction or ratio that expresses the size or occurrence of a specific category or event in a variable, in relation to the total sample. This can also be expressed as a percentage.	The sample consists of 65/85 women (76.5%).

WHY INFERENCE MATTER

Inferential statistics help you to draw conclusions about a larger population of interest, based on a sample of data (your dataset). This approach is not only more practical and cost-effective than studying the entire population, but it also allows for testing hypotheses. Three common inferential tests include the t-test, chi-squared and correlation (see below).

	T-test	Chi-squared Test	Correlation
Dependent variable	Numeric (e.g., mathematics test scores)	Categorical (e.g., food preference: pizza and salad choice)	Numeric (e.g., height)
Independent variable	Categorical (two groups) (e.g., control and experiment)	Categorical (e.g., men and women)	Numeric (e.g., weight)
Example hypothesis	"Mathematics scores will differ between the control and experimental group."	"Men are more likely to choose pizza over salad."	"Height is significantly related to weight."
Types and alternatives	<ul style="list-style-type: none"> Independent (two different groups) Paired (same group measured twice; e.g., before and after an experiment) Mann-Whitney U-Test (non-parametric data) 	<ul style="list-style-type: none"> Fischer's Exact test (often for small sample sizes) 	<ul style="list-style-type: none"> Pearson (classic correlation) Spearman (non-parametric)
Useful charts	<ul style="list-style-type: none"> Box plot 	<ul style="list-style-type: none"> Stacked bar chart 	<ul style="list-style-type: none"> Scatterplot 

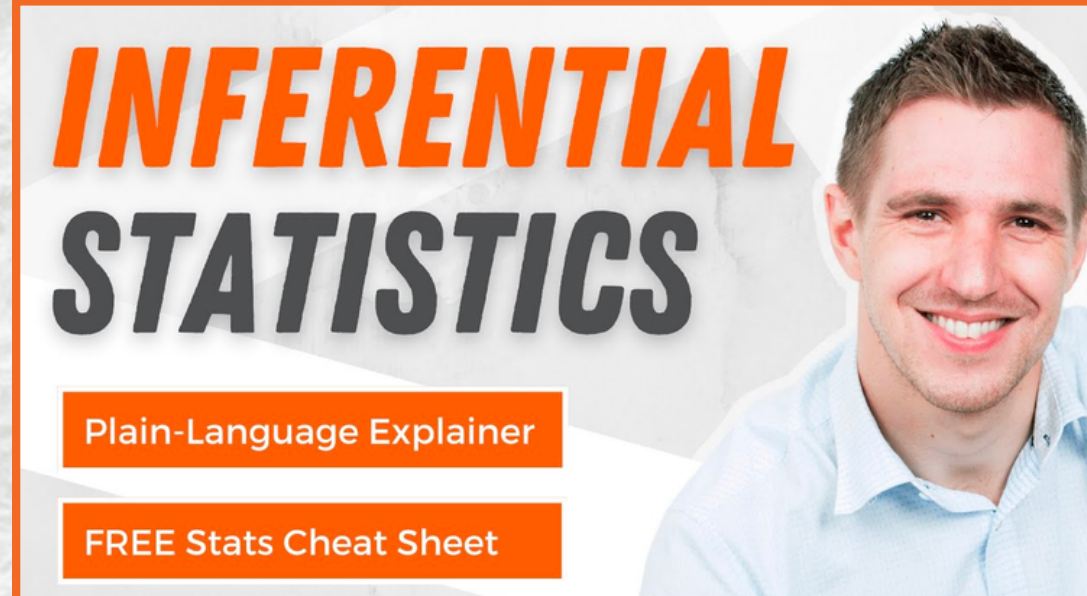
OTHER TESTS

There are many different types of inferential tests, beyond the three common ones we've covered here.

Some other tests you may encounter or use include:

ANOVA: Compares a numeric variable to two (or more) groups (e.g., testing mathematics scores across three schools). Similar to a t-test, but it allows for more than two groups.

Multiple regression: Allows you to predict the value of a dependent variable (e.g., market price for a house) using a combination of other independent variables (e.g., the number of bedrooms, bathrooms, floor space, etc.).



Learn more on our YouTube channel @gradcoach

You'll need to report on **PROBABILITY** and **EFFECT SIZE** when reporting the outcomes of inferential tests.

- Probability: Often reported as the "p-value", is the likelihood of your test being supported by other studies, or that the outcomes are simply a product of chance.
- Effect size: How "big" the effect is (e.g., how strongly variables are correlated or how different the groups actually are from each other).

You'll also need to choose a **SIGNIFICANCE LEVEL** for your tests. This is typically a "cut off" where p-values below this level are seen as "significant" (e.g., significantly related or significantly different) $p < 0.05$ (5%) is the more typical significance level for most fields, but can be lower to provide more cautious interpretations or if you are running multiple tests.