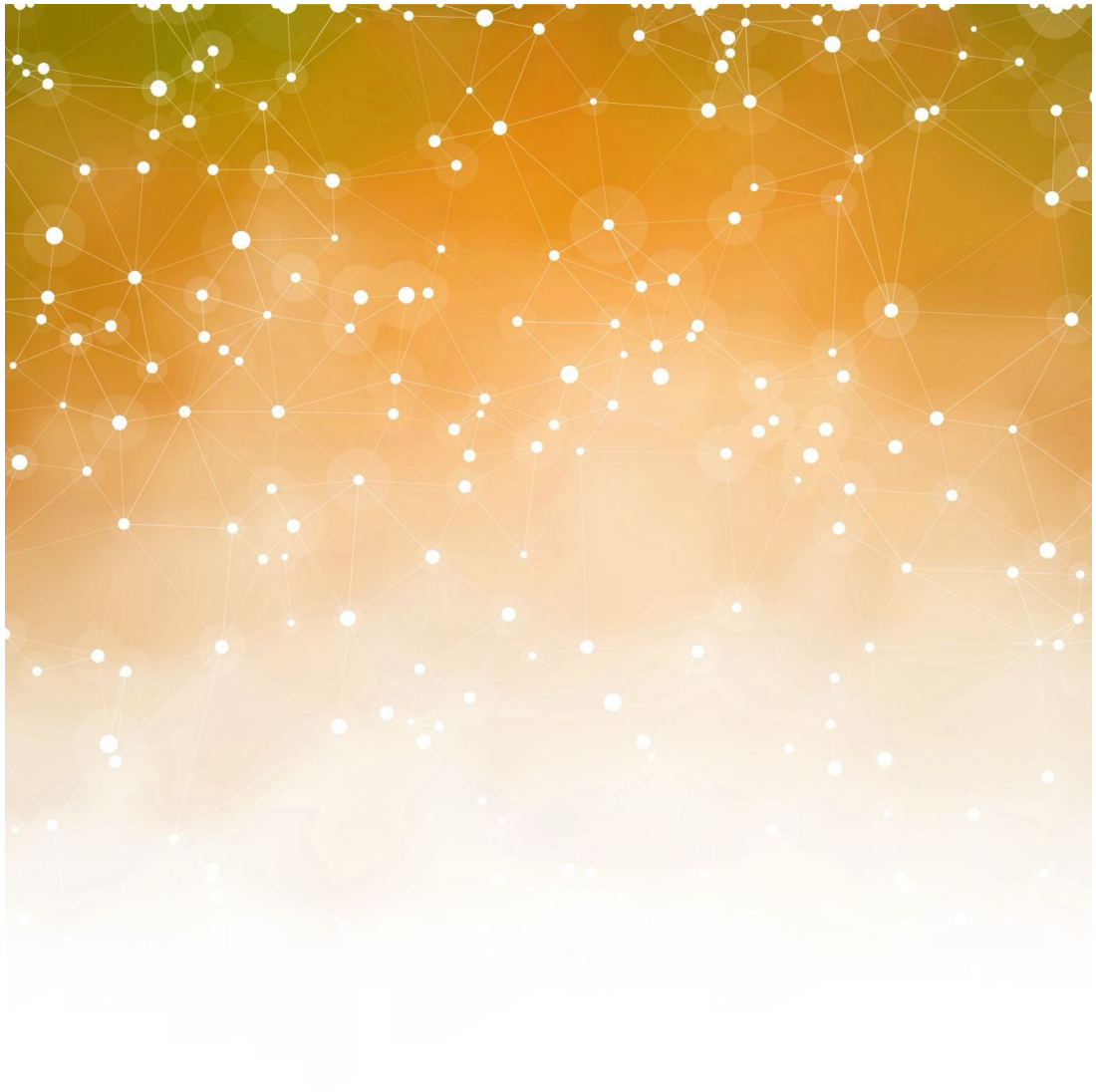
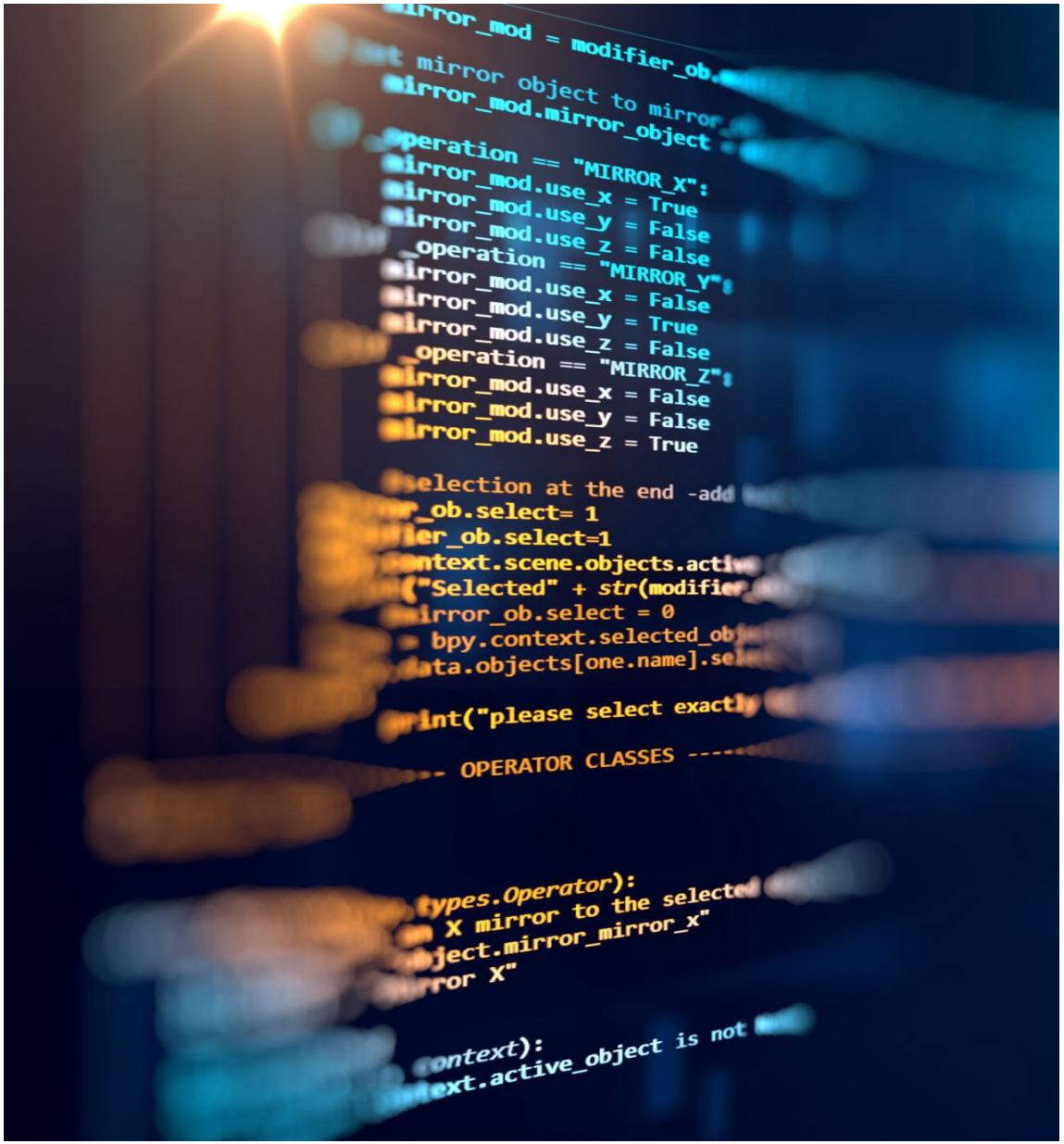
Quantitative Analysis
Week 06

More Statistical Tests





Some notes on Assignments

Your assignment submissions should just be one R file!

Do not change the format of this file (*don't turn it into an Rmd file, or a PDF, or a Word document, or a series of screenshots…*).

Don't worry about the directory you include in the `setwd()` command – the grading software replaces this automatically, so it'll run on my computer regardless.

Include any details about issues you ran into and how you fixed them as comments in the R file – these are lines starting with a # symbol.

Review

Last week, we looked at one of the most important and fundamental statistical techniques used in quantitative analysis – the **t-test**, which is used to determine whether two **samples** are likely to come from a population with the same mean / distribution.

t-tests,
normal distributions,
& confidence levels

The t-test is based on assumptions about the **distribution** of our data – notably, that the data has a **normal distribution**, with most cases close to the mean and few outliers.

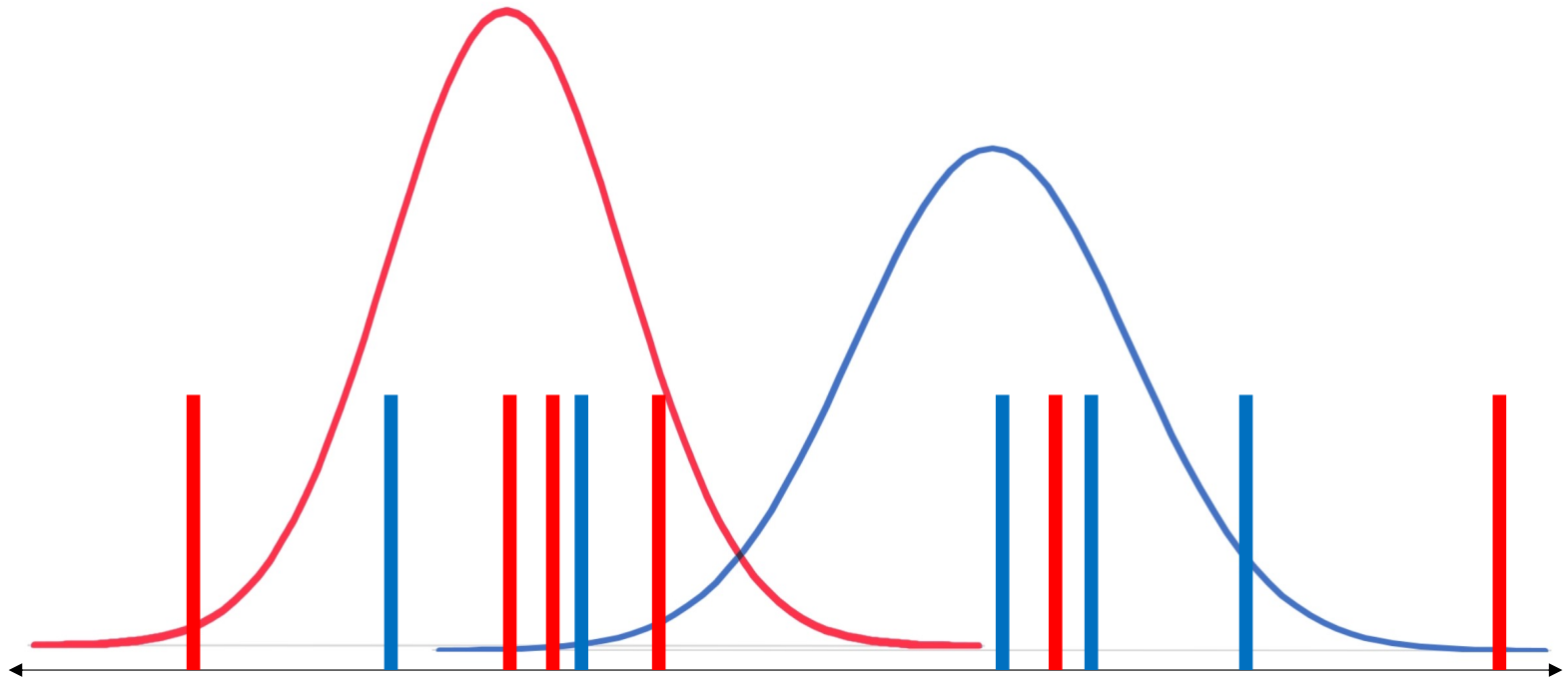
The **assumption of normality** lets us make robust estimates of the **probability** of the population having certain characteristics, based on a small sample.

We state these findings with **confidence levels**. 95% is common; but remember, 95% confidence means a 1 in 20 chance of being wrong due to random sampling!

In other words, given two different samples of data....

Are these samples most likely to come from the **same** population?

Or from **different** populations?



Hypothesis Testing

We discussed t-tests in terms of comparing two samples of observations to see if they are statistically the same. We described this as **hypothesis testing**. T-tests are one of a large family of statistical tests that we can use to test simple hypotheses.

Specifically, t-tests are designed to test hypotheses where the **predictor variable** (independent variable) is **categorical** (specifically, a **binary category**), and the **outcome variable** (dependent variable) is **quantitative**.

The hypothesis tested by a t-test is: "***Whether a case is a member of Category A or Category B predicts a change in the outcome variable X***."

Other Statistical Tests

As the t-test is for a **binary categorical predictor variable** and a **quantitative outcome variable**, it logically follows that there must be other tests we can use for different kinds of data.

Multi-Category Predictor →
Quantitative Outcome
ANOVA

Quantitative Predictor →
Quantitative Outcome
Pearson's R,
Spearman's rho

Category Predictor →
Category Outcome
Chi-Square (χ^2) Test

Where does Regression fit into this?

You may notice that on the previous slide I left off one major category:

Quantitative Predictor → Categorical Outcome

This is a type of statistical test that is primarily done using a **regression model**, specifically **logistic regression**; for more than two categories, you might use a **multinomial logistic regression** or even a **machine learning classification model**.

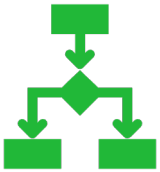
We'll move on to talking about regression models next week. This week, we'll cover the other statistical tests listed on the previous page.

ANOVA
(ANalysis Of VAriance)

Multi-Category Predictor,

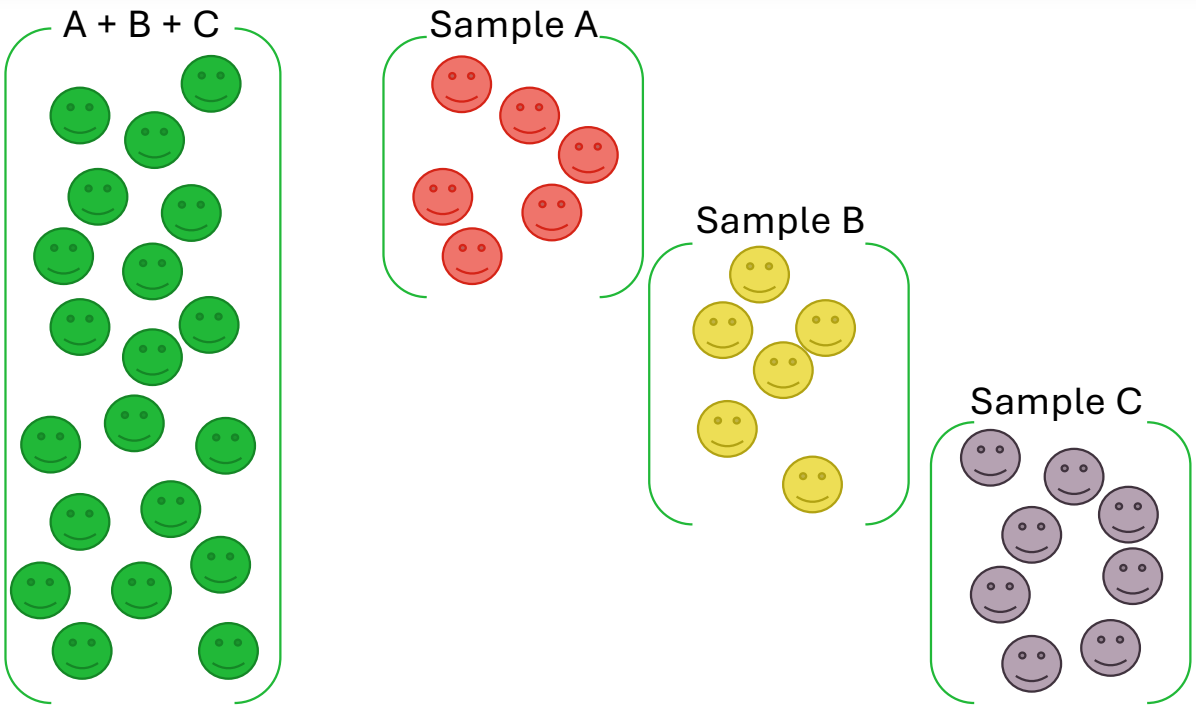
Quantitative Outcome

Used when you have multiple categories (e.g. samples from several groups) and want to test for significant differences among them.



Think of it as a generalisation of the t-test to multiple-category problems.

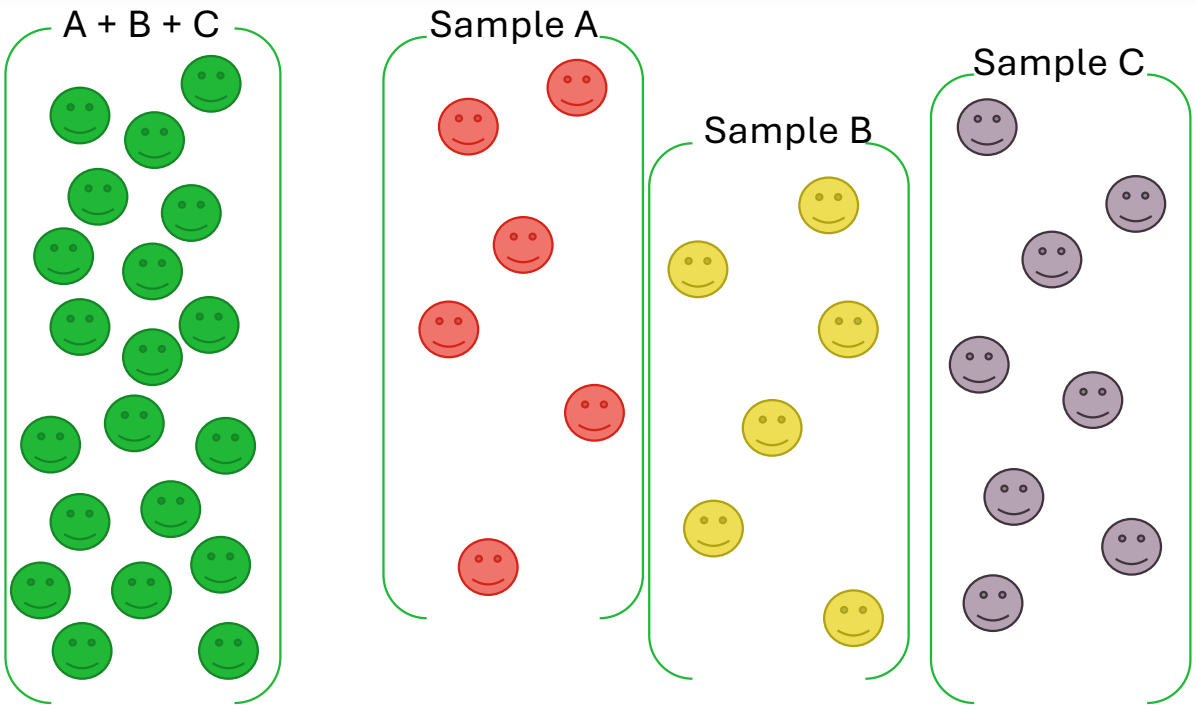
How ANOVA Works



In this case, Samples A, B, and C have high **between-group variance** (i.e. different means) and low **within-group variance** – i.e. they aren't too spread out.

This means that dividing the A+B+C data into three has helped to explain a lot of the variance in the overall data, so the division into categories is statistically meaningful.

How ANOVA Works ②



In this case, however, the samples low **between-group variance** and high **within-group variance**.

Intuitively, we can tell that dividing up the sample this way hasn't explained much of the sample variance, if any; there's no clear connection between the sample groups and the variance.

What is Variance?

We have an intuitive notion of what “variance” is – how spread out the values in a sample are. For research, though, we need to define variance clearly.

Standard Deviation is a measurement of deviation from the mean in a sample:

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

… while **Variance** is standard deviation squared and measures average difference from the mean.

Standard Deviation is a useful tool; it measures how spread out observations are from the mean. In normally distributed data, most observations shouldn’t be more than two or three standard deviations from the mean.

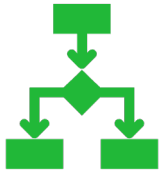
Chi-Square Test

Categorical Predictor,

Categorical Outcome

Two types of Chi-Square test exist.

The test for **goodness of fit** asks how likely it is that we would randomly get a certain sample from evenly distributed categories.

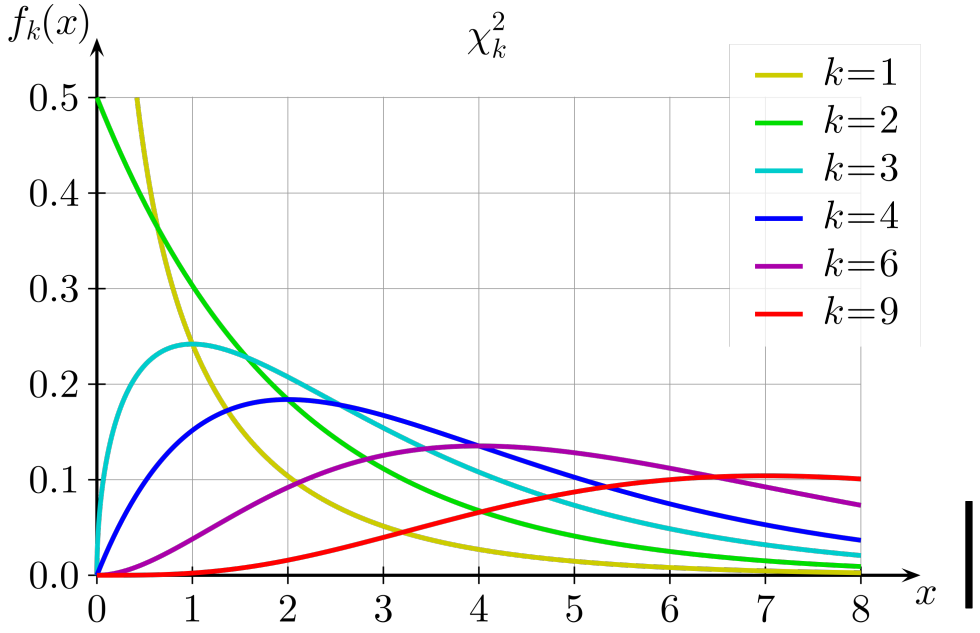


The **test of independence** checks if one set of categories is independent from another – i.e., does membership of one category influence the likelihood of being a member of another category?

How Chi-Square Works

Chi-Square is named for the X^2 distribution. Like the normal distribution, its shape changes and evolves according to a variable *k*, which is the degrees of freedom *df* of the data set.

This table lets us calculate the probability of a categorical distribution being sampled from an *expected, even* distribution.



Fisher’s Exact Test

Chi-Square tests are **nonparametric**: they make relatively few assumptions about your data. They simply assume that given independence between samples, they should be evenly distributed across categories.

However, the algorithm for Chi-Square doesn’t work well when you have small **expected values**. A modified version of the test, **Fisher’s Exact Test**, allows for these cases, and is recommended when you have data in which division across categories could create some cells with values <5.

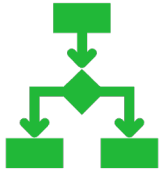
Pearson’s R, Spearman’s rho

Quantitative Predictor,

Quantitative Outcome

Correlation

Used to test for a relationship between two quantitative variables: do the values, on average, move together?



This is the first step towards regression analysis – but also useful in its own right!

Correlation: The Theory

Correlation – or the **correlation coefficient** – is one of the most commonly used statistics to summarise relationships between pairs of variables.

Values are standardised (using **z-scores**, which represent each value as its distance from the sample mean in **standard deviation units**), so correlation is independent of the scale of your variables.

Basic correlation algorithms don't report **statistical significance**, but R's correlation test functions will run t-tests automatically to check for significance levels.

Which Correlation test to use?

Pearson's R is the default correlation test to use in most circumstances.

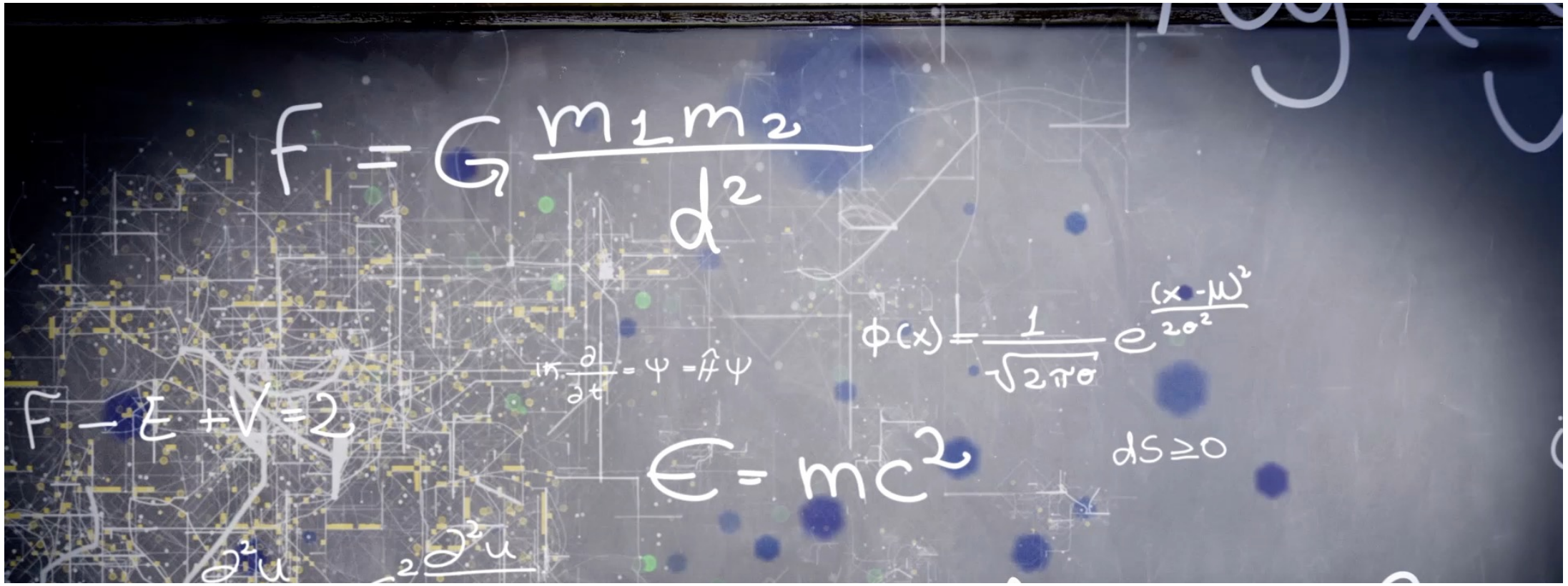
It uses the same assumptions as the t-test – most importantly, that the data are normally distributed.

Spearman's rho is a modified test which can be used in situations where the data are not normally distributed.

Spearman's test also works when the relationship between the data sets is non-linear.

Pearson's R is a **parametric test**, while Spearman's rho is **nonparametric** – a classification that refers to the assumptions (like normality) the test relies upon. Nonparametric tests make fewer assumptions.

Move over to R



Research Project

Research Project: Two Tasks

1) <u>"Poster" Presentation</u>

In the last class of the semester (**Jan 24**), each group will make a five-minute presentation of their project, using a small number of slides (five or six).

(Time will be very tight given the class size, so each group will be held strictly to five minutes!)

2) <u>Research Diary / Lab Notes</u>

As you are doing your research – starting next week with choosing a question – write short (1-2 paragraphs) **personal** diaries / lab notes about progress, decisions made, etc.

Add a brief conclusion and submit this by one week after the final class (**Jan 25**)

Timeline

We are now almost halfway through the semester, so it's time to organise groups for the research project and start coming up with your research themes and questions.

We will organise groups this week. In next week's class (Nov 20), I'll leave some time for the groups to get together and discuss their project, but you'll also need to arrange a meeting time outside class hours.

By the following class on **Nov 27**, each group should have decided upon a research question and be ready to start the literature review.

This Week

<u>Your task this week</u>: I will put an assignment on Moodle this week, which will list a number of possible research topics (based on the interests you told me about at the start of the semester). Respond to the assignment with **first, second, and third preferences** for which projects you'd like to do.

If there is some reason why you cannot work in a group with another student in the class, you may say so in your assignment response. Your choice will be entirely anonymous, and you don't have to explain a reason. I'm trusting you not to abuse this for spurious reasons.