Business and Local Government Data Research Centre

Introduction to Statistics

EXPLORING DATA
ENHANCING KNOWLEDGE
EMPOWERING SOCIETY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of Essex

Cutting-edge training delivered by leading experts in the field of data analytics brought to you by the Business and Local Government Data Research Centre

- Exploring data
- Enhancing knowledge
- Empowering society

Offering grant funded data analytics research projects, training, events, webinars and data consultation services.

To find out more please contact Laura.brookes@essex.ac.uk

Agenda

1. Introduction
2. Data Types and Levels of Measurement
3. Central Tendency and Dispersion
4. Hypothesis Testing
5. T-tests
6. Covariance and Correlation
7. Linear Regression

What is Statistics?

- The analysis, interpretation and presentation of data
- Statistics help us solve a problem: we usually are not able to study the population of interest directly
    - Population – entire set of items or subjects one wishes to study (e.g. counties, boroughs, UK citizens)
    - Sample – subset of a population chosen for study
- We use statistics to make predictions (or inferences) about a population based on data from a sample of that population

Types of Statistical Analysis

- Descriptive statistics – summarize data
- Inferential statistics – make predictions within a sample to make inferences about a wider population

Introduction to Measurement

- Measurement is essential for quantitative research – the assignment of numbers to objects or events
- Different data types will have different measurement levels
  - Continuous i.e. interval values such as population
  - Count – counts vs. uncountable
  - Categorical – but not ordered
  - Ordinal – ordered categories, i.e. deprivation index
  - Binary – two categories, i.e. yes/no

What is R?

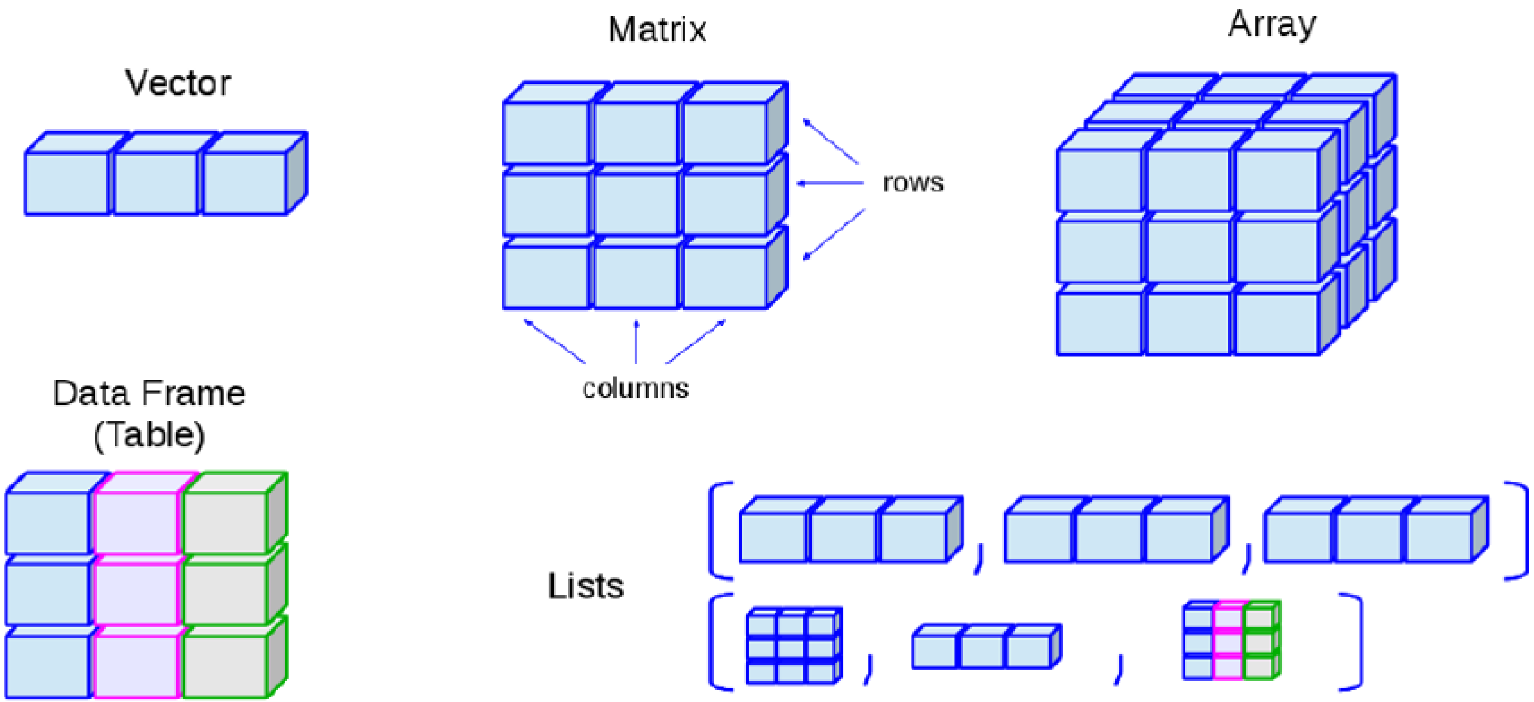
- An environment for statistical computing and graphics
- It is free
- It packs powerful graphical facilities
- It is a simple and effective programming language
- Most statistical models are already implemented
- New models are often implemented in R first

Business and Local Government Data Research Centre

What is R Studio?

- Working environment R
  - While you do not need R Studio to run R, it makes working with R much easier
  - Just like R, it works on PC, Mac & Linux
- Home of all things R: https://cran.r project.org/
- Keep R updated check for new versions
- To get help: Google the question or error message is always a good start
- https://stackoverflow.com/questions/tagged/r

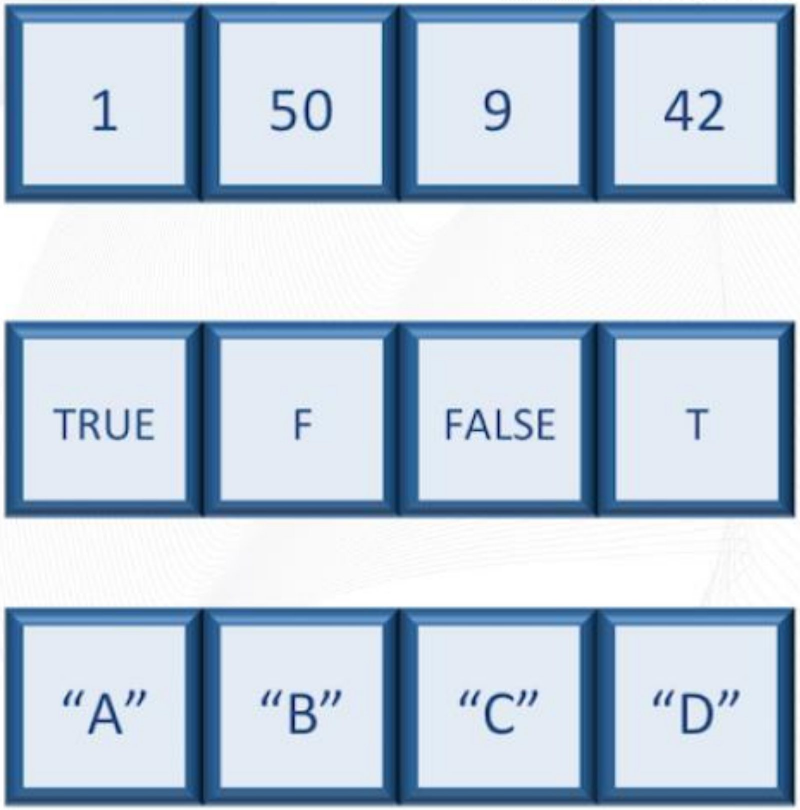
R Syntax, data structures and types



http://venus.ifca.unican.es/RIntro/dataStruct.html

Vector

- Numeric vectors
  - a <- 5
  - a <- c(1, 50, 9, 42)
- Logical vectors
  - b <- a < 10
- Character vectors
  - a <- "this is text"



Matrix

- Matrix: two dimensional, store data of same mode; two coordinates to identify a unique matrix element

90	5	137	9
87	40	2	52
4	102	32	41

Data Frame

- Each row is a vector and ***observation***
- Each column is a vector and ***variable***
- Rows and column must be of equal length
- Missing values listed as NA (numerical) and "" (string)

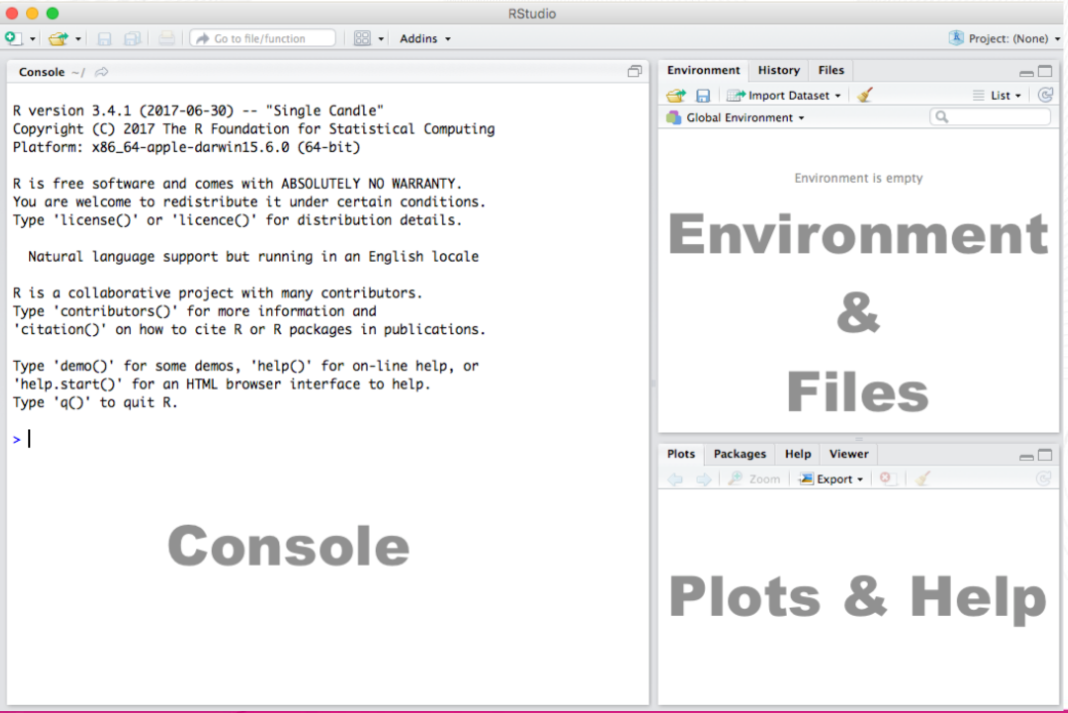
"A"	102	"Hela"	TRUE
"B"	40	"BHK"	F
"C"	12	"hESC"	T

Business and Local Government
Data Research Centre

Now we can open R Studio!

- For the session materials visit the following web page:

https://esrc-blg.github.io/stats101



Descriptive Statistics

- Descriptive statistics are a good way to get a “feel” for the data
- The first step to data analysis
- Most interested in two qualities of the variables we are working with:
    - Central tendency
    - Dispersion

Central Tendency

- Working value of a "typical" observation, or the value of the observation at the center of a variable's distribution
- Measure of central tendency depends on type of variable:
    - Categorical – mode
    - Ordinal – median
    - Continuous/count/**binary** – mean

Dispersion

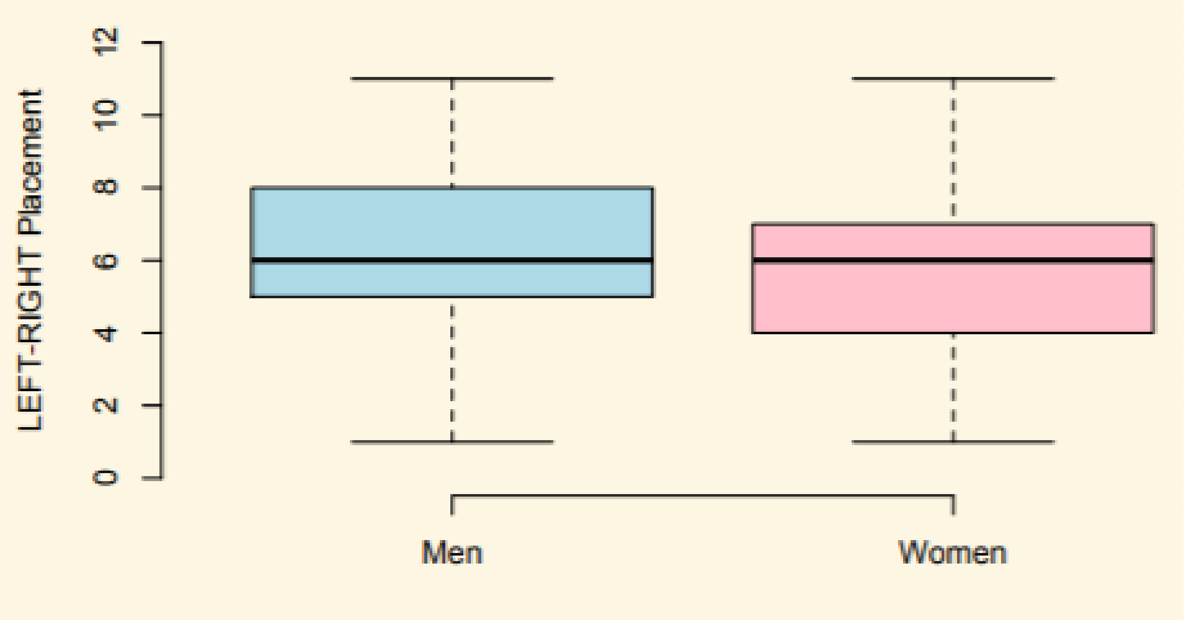
- Dispersion – measures the spread of values for a variable
- Again, the precise measure of dispersion depends on the level of measurement:
  - Categorical/**binary** – proportion of each category
  - Ordinal – range or interquartile range
  - Continuous/count – variance/standard deviation

Range and interquartile range

- Range – difference between the lowest and highest value of ordinal variable
- Interquartile range is better – the difference between the 25th percentile and the 75th percentile

Range and interquartile range

- Interquartile range is better – the difference between the 25th percentile and the 75th percentile



Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

- Approximately the average of the squared deviations (from the mean)
- Variance is zero if all observations are identical to mean and increases as observations become further from mean
- Not used frequently – hard to interpret as it is squared!
- Squared because of negative numbers and outliers

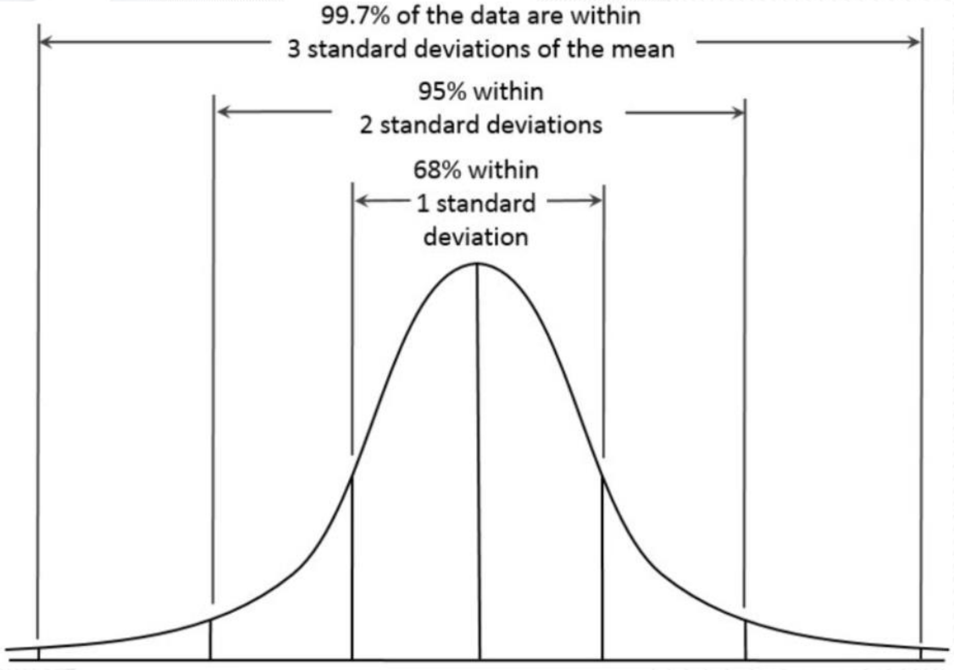
Standard Deviation

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

- The typical distance an observation is from the mean of all observations (square root of variance)
- The greater the variability around the mean, the greater is standard deviation
- Interpretation is easier – reverts back to original unit

Standard Deviation

- Tells you how large the dispersion is of continuous, count and binary variables
- Important are we know how of population falls within standard deviations



Business and Local Government Data Research Centre

Maths functions in R

`log(x)`	Natural log.	`sum(x)`	Sum.
`exp(x)`	Exponential.	`mean(x)`	Mean.
`max(x)`	Largest element.	`median(x)`	Median.
`min(x)`	Smallest element.	`quantile(x)`	Percentage quantiles.
`round(x, n)`	Round to n decimal places.	`rank(x)`	Rank of elements.
`signif(x, n)`	Round to n significant figures.	`var(x)`	The variance.
`cor(x, y)`	Correlation.	`sd(x)`	The standard deviation.

…but we can also use the summary () function

…let’s return to R Studio with some real data

Exploring Relationships

- After understanding the data, we often want to understand the relationships between two (or more) variables

    i.e. between a predictor (independent variable) and an outcome of interest (dependent variable)

- Summary statistics are the first steps in trying to understand relationships in the data

Has a Relationship occurred by Chance?

- The Lady Tea Tasting Test (Fisher, 1925) – assess chance through **hypothesis testing**
- 1920s Tea party, Dr. Bristol, claims to be able to distinguish whether the milk or the tea had been poured into the cup first (**hypothesis**)
- A test was arranged for 8 cups, 4 of each type in random order
- This is then tested against a **null hypothesis**, which states our hypothesis is false, and that chance is to blame!

The Lady Tea Tasting Test (Fisher, 1925)

- To figure out the frequency of different possibilities, we assess many ways there are to pick 4 cups out of 8? (70)

Successful guesses	Selected Possibilities	Unselected Possibilities	Total Possible Combinations
0	MMMM	TTTT	1×1
1	MMMT, MMTM, MTMM, TMMM	TTTM, TTMT, TMTT, MTTT	4×4
2	MMTT, MTMT, MTTM, TMTM, TTMM, TMMT	TTMM, TMTM, TMMT, MTMT, MMTT, MTTM	6×6
3	MTTT, TMTT, TTMT, TTTM	TMMM, MTMM, MMTM, MMMT	4×4
4	TTTT	MMMM	1×1
		Total	70

- Perhaps Dr. Bristol cannot tell the difference – **null hypothesis**
- What is the probability of observe what we observe?
- Dr. Bristol correctly identifies 4 out of 4 – is the **test statistic**

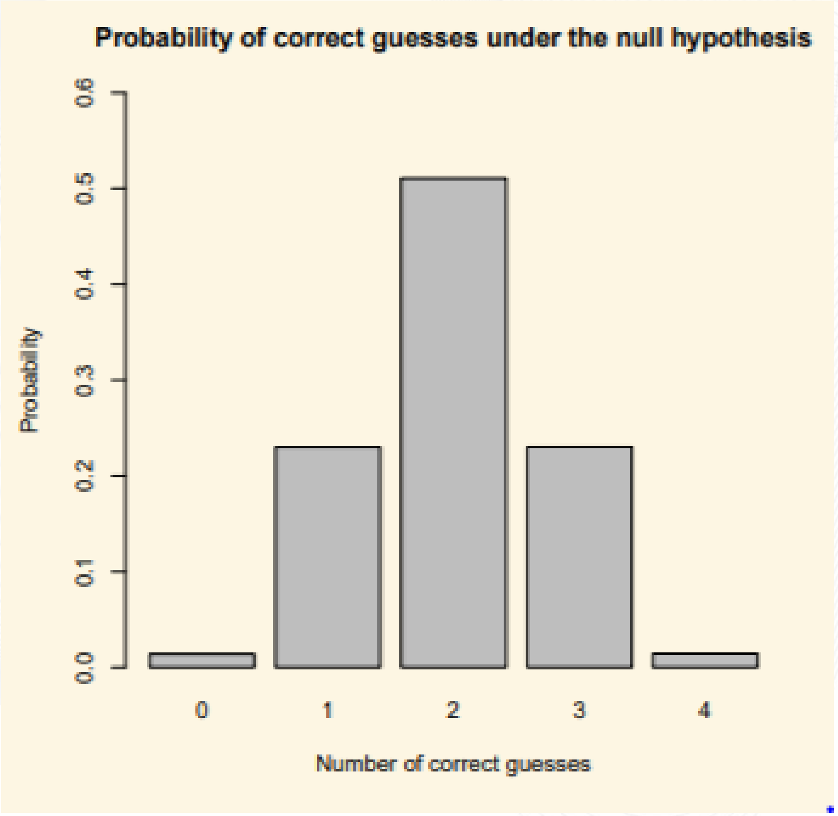
Business and Local Government Data Research Centre

What were the chances?

- If Dr. Bristol was only really guessing, the probability that she would have correctly identified all four cups of tea:

  1/70 = 0.014

- This is the **p-value**, the probability of observing the data we observe
- Convention is that if less than 0.05 we can reject the null

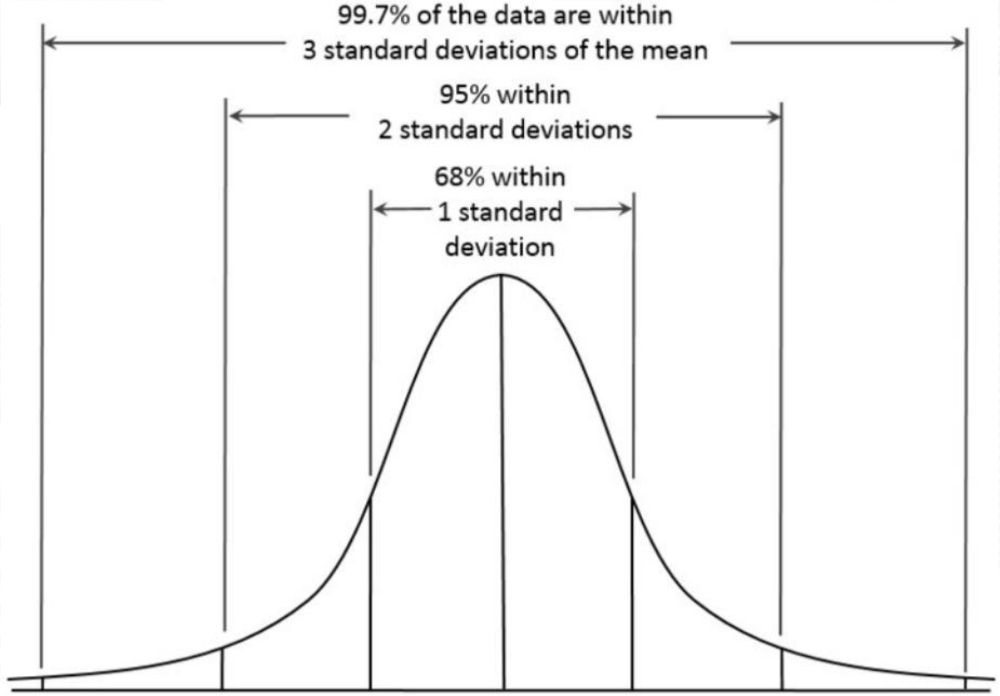


Hypothesis Testing

- To recap, there are several main elements to any hypothesis test of relationships
  - Must develop a hypothesis and null hypothesis
  - Calculate test-statistic
  - Which is used to calculate a p-value
  - From this we can decide whether to reject null hypothesis, i.e. reject whether the result occurred simply due to chance

t-statistic (is a test statistic)

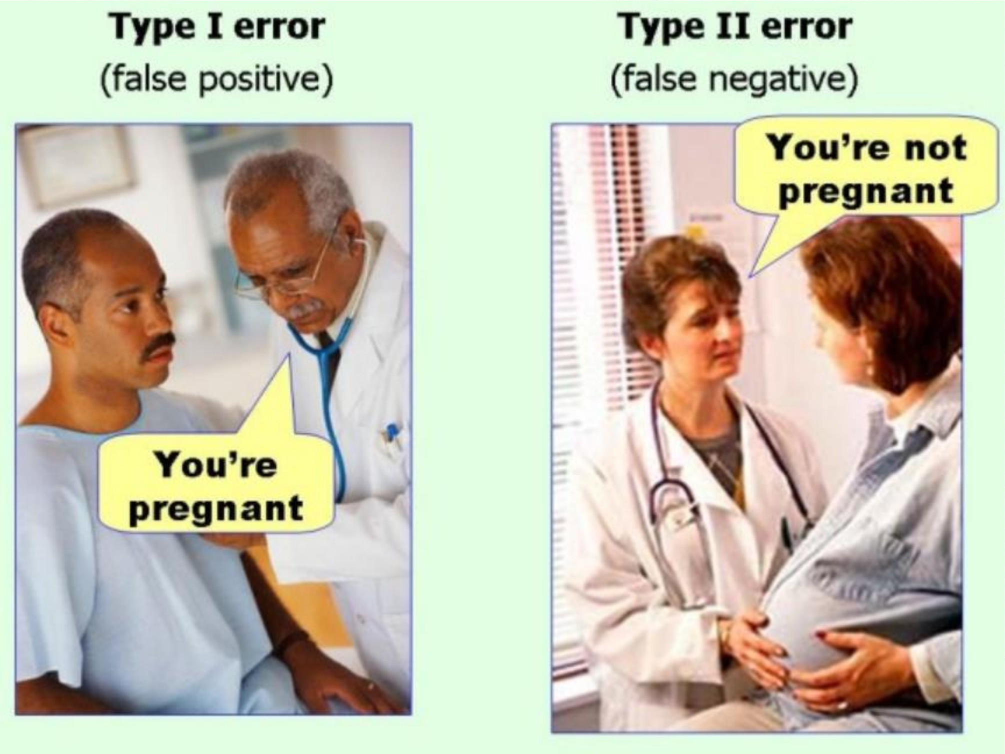
- Remember the standard deviations?
- Well the **t-statistic** is easy to interpret as you are looking for the magic 1.96 deviations from the mean
- Anything above this figure is outside of 95% of the population ($p < 0.05$).
- Too rare to have occurred by chance



https://www.khanacademy.org/math/ap-statistics/two-sample-inference/two-sample-t-test-means/v/two-sample-t-test-for-difference-of-means

To sum up

- A p-value of 0.05 implies we will incorrectly reject the null hypothesis 5% of the time
- A p-value of 0.01 implies we will incorrectly reject the null hypothesis 1% of the time
- With the Tea Tasting test we are 1.4% likely to incorrectly reject the null



Which approach is appropriate for *continuous outcomes*?

- Differences in Means
    - Appropriate when our **independent variable is binary** in nature, i.e. comparing genders
- Correlation and covariance
    - Useful when our i**ndependent variable is a count or continuous** – as you cannot use difference in means
- (and then) Linear Regression
    - Because there are limitations with the above
    - Bivariate – essentially a t-test
    - Multivariate – multiple independent variables

Difference in Means

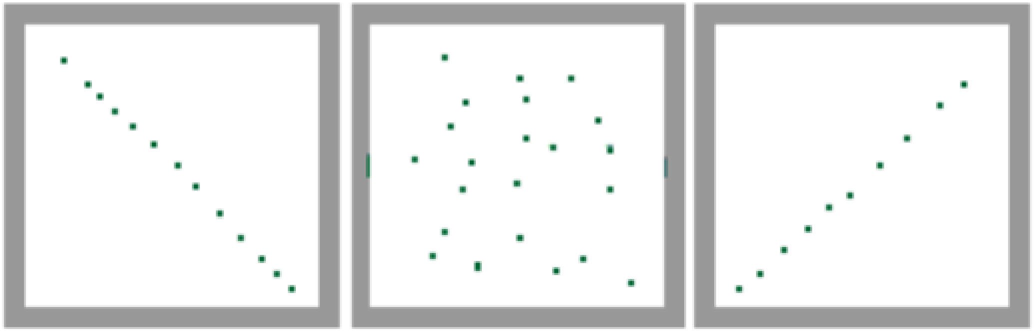
- Binary IV and continuous DV
- We are interested in the difference between two ◦ conditional means and an outcome
  - i.e. is there a difference in crime committed by males and females
- Think of this as exploring "Risk Stratification"

Covariance

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

- Continuous IV and continuous DV
- Assesses whether low and high values match
- Can classify three types of relationship: positive, negative and no relationship.

Covariance



- Positive if values in Y also correspond with X and increase together and negative when Y and X decrease together
- No relationship where corresponding values do not increase/decrease together

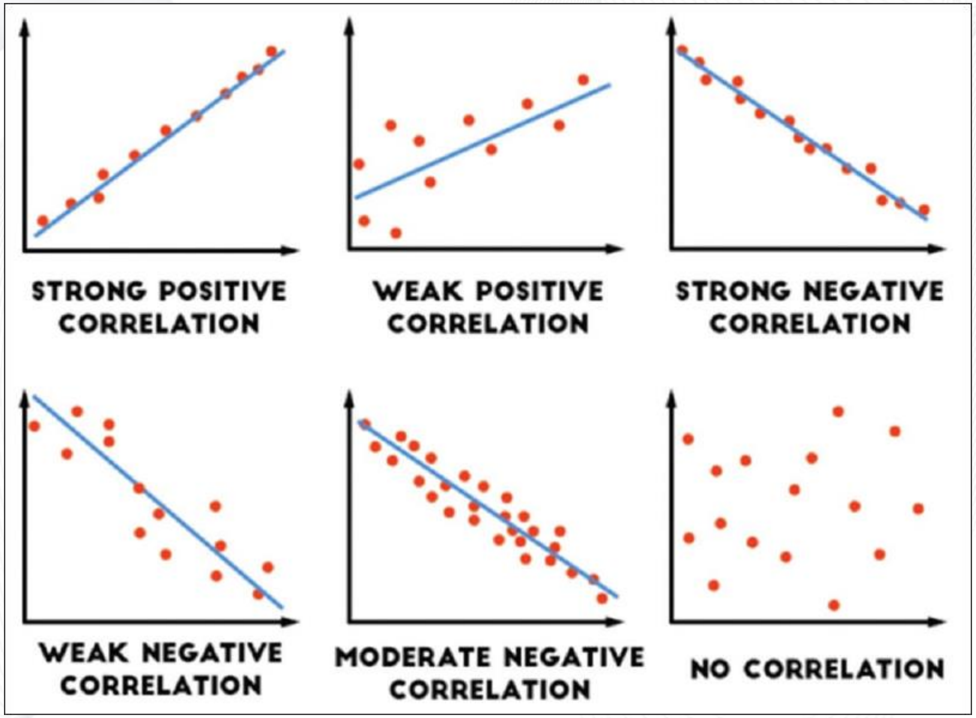
Covariance

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

- Higher values indicate a more dependent relationship
- But hard to interpret:
  - Covariance does not tell us the distance between data points and the covariance line
  - Covariance is sensitive to scale – larger values are further from the mean, leading to higher covariance

Correlation

- Correlation measures direction and how far the points are to the line
- Scaled between -1 (negative relationship close to line) and 1 (positive relationship close to the line)
- 0 – means we cannot reject null hypothesis.
- On same scale so can compare relationships



Correlation

- But there are limitations:
    - Correlation is impacted by sample size
    - Correlation plots straight lines so misses non-linear relationships
    - Also cannot explain the magnitude of the relationship, or slope
- Can be more intuitive to plot two continuous variables
- **Let's do this in R Studio!**

Linear Regression

- A linear regression model approximates the relationship between our independent X and our dependent variable Y
- Essential draws a straight line of best fit through the data, which can be expressed as:

$$Y = \alpha + \beta X$$

- α is the intercept: the value of Y where X = 0
- β is the slope: the amount that Y increases when X increases by one unit – changed by different values

Linear Regression

- The simplest way to summarize the relationship between two variables is to assume that they are linearly related
- We can express this with the bivariate linear regression model:
- $Yi = \beta 0 + \beta 1Xi + ui$

  - Observations $i = 1, \dots, n$
  - Y is the dependent variable
  - X is the independent variable
  - $\beta 0$ is the intercept or constant
  - $\beta 1$ is the slope
  - ui is the error term or residuals (model fit)
  - $\beta 0$ and $\beta 1$ are the coefficients of the regression line

- **Let's run this in R!**

Business and Local Government Data Research Centre

Thank you

Join in the conversation online:

@BLGDataResearch #Data2Life

LinkedIn: ESRC Business and Local Government Data Research Centre

YouTube: ESRC Business and Local Government Data Research Centre

Email: BLGDataResearch@essex.ac.uk