Statistics for CSAI II

3 – Hypothesis Testing/ Intro to Correlation

Dr. Travis J. Wiltshire |

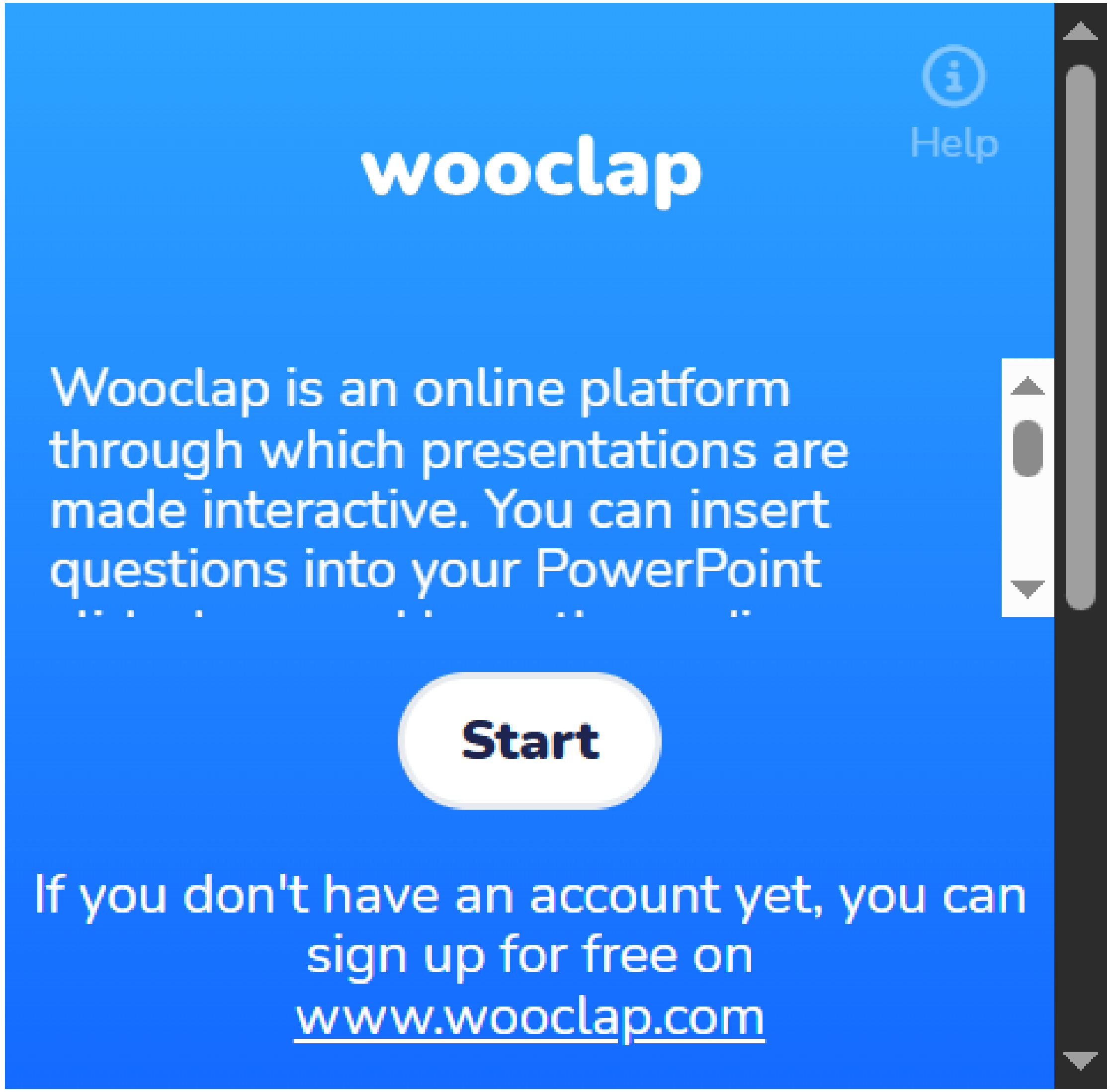
Tilburg University

Modules

1. Introduction and Probability
2. Sampling Theory
3. *Revisiting Hypothesis Testing & Intro to Correlation*
4. Correlation
5. Intro to Regression
6. More Regression Centering and Checking Assumptions
7. Multiple Regression and Assumptions
8. Interactions
9. Multiple Regression with Categories
10. Multiple Regression with Polynomials
11. Mixed Models
12. Growth Curve Analysis

Outline

1. Hypothesis Testing Revisited
2. Correlation


Help
wooclap
Wooclap is an online platform through which presentations are made interactive. You can insert questions into your PowerPoint
Start
If you don't have an account yet, you can sign up for free on www.wooclap.com

What is a hypothesis?

- A formal statement predicting the relationship between two or more variables.
  - The **number of hours** spent studying will improve **student's grades** in a course.
  - Playing video games more **often** will lead to stronger social **relationships** in teenagers.
  - Higher **instances** of communication breakdowns in teams will lead to poorer teamwork **performance**.
  - Increased **word frequency** leads to shorter **reading times**.

Hypothesis Testing

- The general goal of a hypothesis test is to **rule out chance** (sampling error) as a plausible explanation for the results from a research study.
- Hypothesis testing is a technique often used to help determine whether a specific effect or relationships exists in a population
  - Testing a difference between groups based on a treatment or conditions
  - Testing associations, relationships, etc. between variables

Null and Alternative Hypotheses

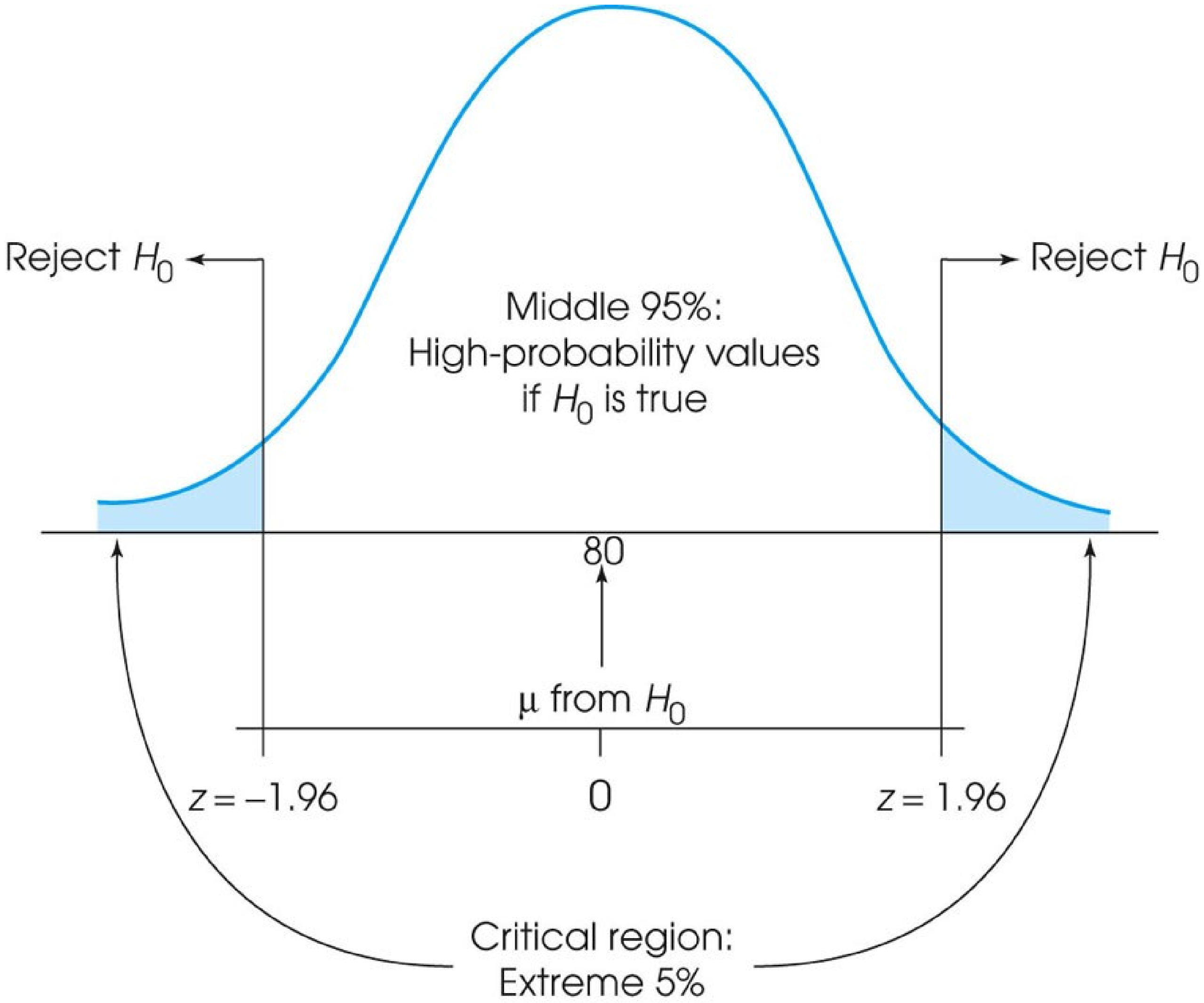
- Convert the research question to null and alternative hypotheses
  - The **null hypothesis (*H_o*)** is a claim of "no difference in the population" (Or an effect is zero)
  - The **alternative hypothesis (*H_a*)** claims "H_o is false"
- Collect data and seek evidence against H_o as a way of bolstering H_a (deduction)
- How much evidence is enough to *reject the null hypothesis?*

Calculating a test statistic

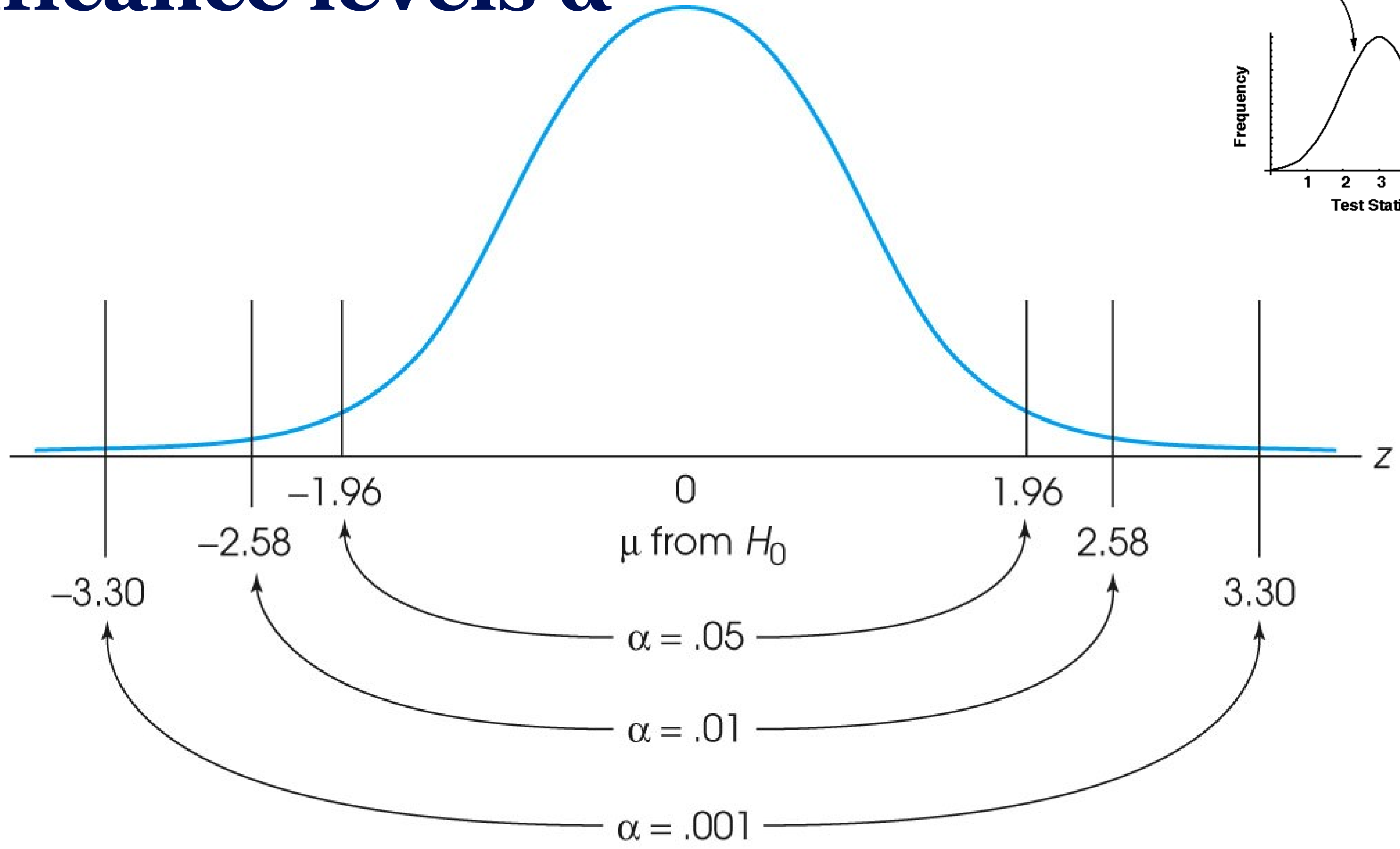
- A test statistic is a statistic with a known distribution, when H is true, that allows calculation of p-$_{o}$values
- Z is a test statistics that has the standard normal distribution

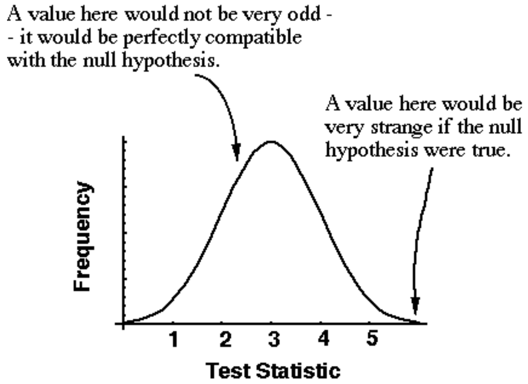
$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

$$z_{\bar{X}} \sim \text{Normal}(0, 1)$$



Significance levels α





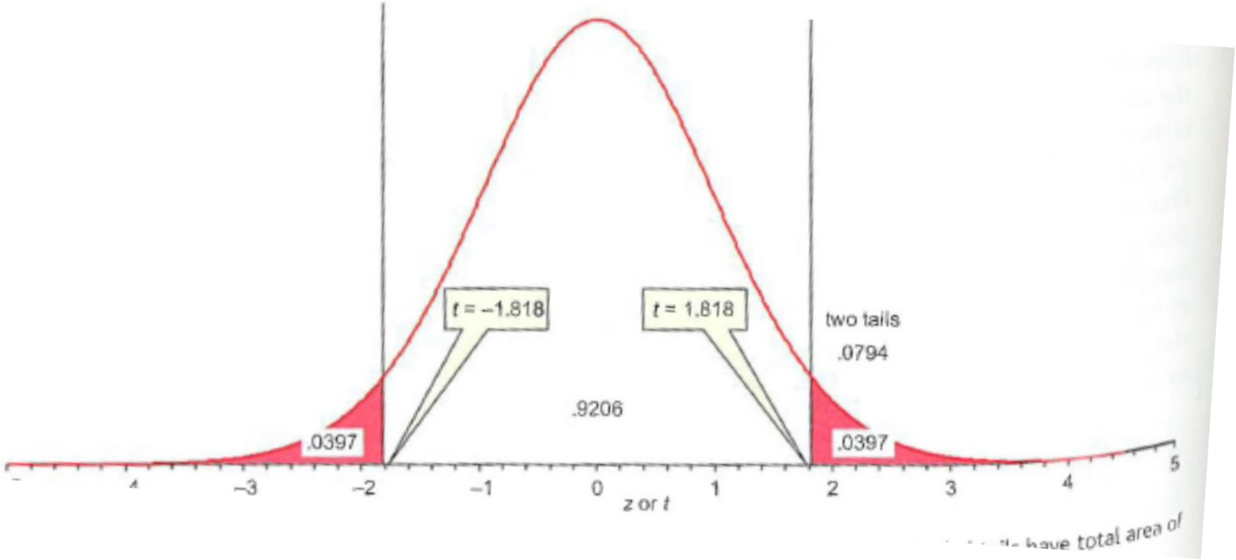
Z test in R

1. Randomly generate a vector in R from a normal distribution with mean of 55.4, n=30,sd=5.
2. Install and load the BSDA package
3. Check out the arguments for the z.test() function
4. Come up with a null hypothesis and alternative hypothesis regarding your randomly generated vector using the following population parameters: $H_{0}: \mu_0 = 50, \sigma = 20$
5. Use the z.test() function to test your hypothesis
6. Read the output and make a decision about the null hypothesis

Using the t distribution when population SD is unknown

- S is our sample SD

$$\text{t-statistic} = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$



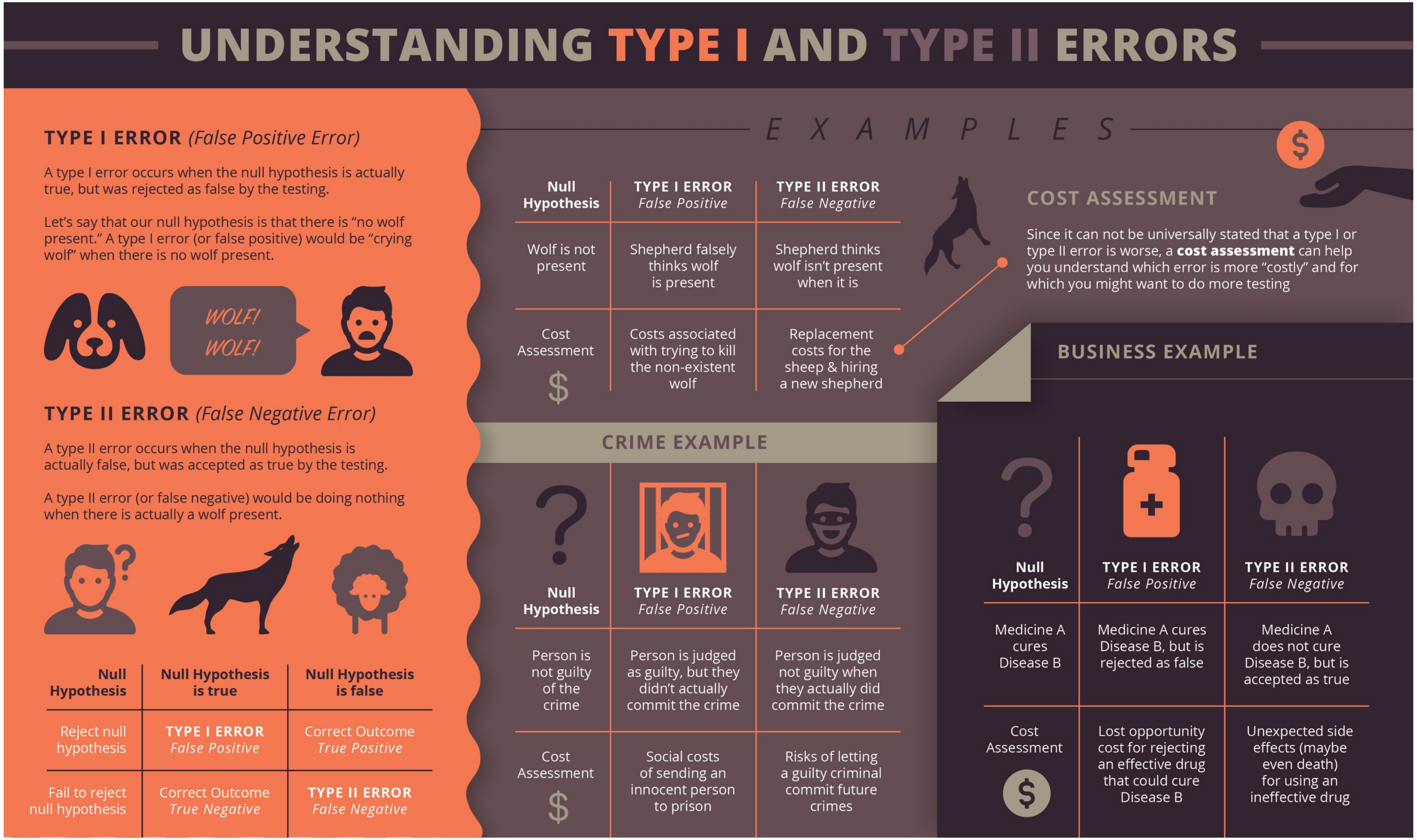
T-test in R

1. Use your vector in R from a normal distribution with mean of 55.4, n=30,sd=5.
2. Check out the arguments for the t.test() function
3. Come up with a null hypothesis and alternative hypothesis regarding your randomly generated vector using the following population parameter: $H_{0}: \mu_0 = 50$
4. Use the t.test() function to test your hypothesis
5. Read the output and make a decision about the null hypothesis
6. How does this compare to your z.test?
7. What about the CIs?

Errors in Hypothesis Tests

- Just because the sample mean (following treatment) is different from the original population mean does not necessarily indicate that the treatment has caused a change or that there is an effect.
- You should recall that there usually is some discrepancy between a sample mean and the population mean simply as a result of sampling error.

Errors in Hypothesis Tests (cont.)



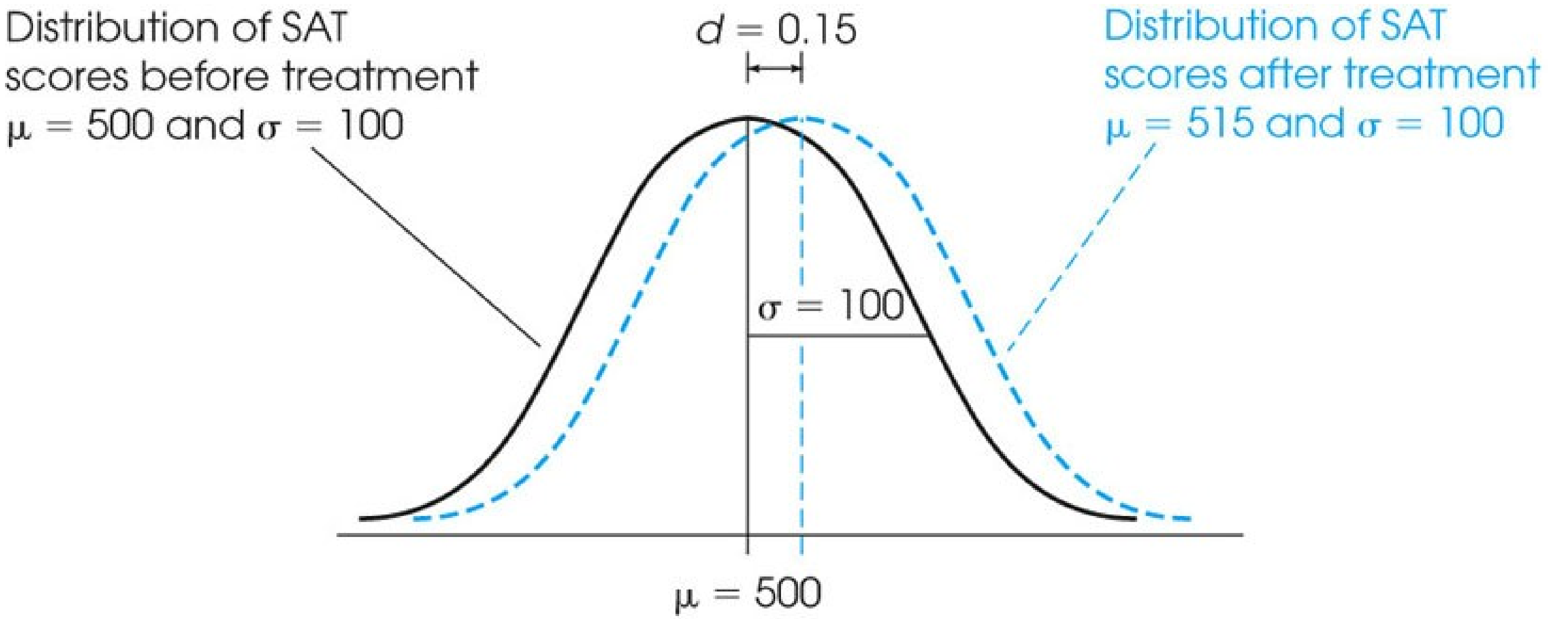
Measuring Effect Size

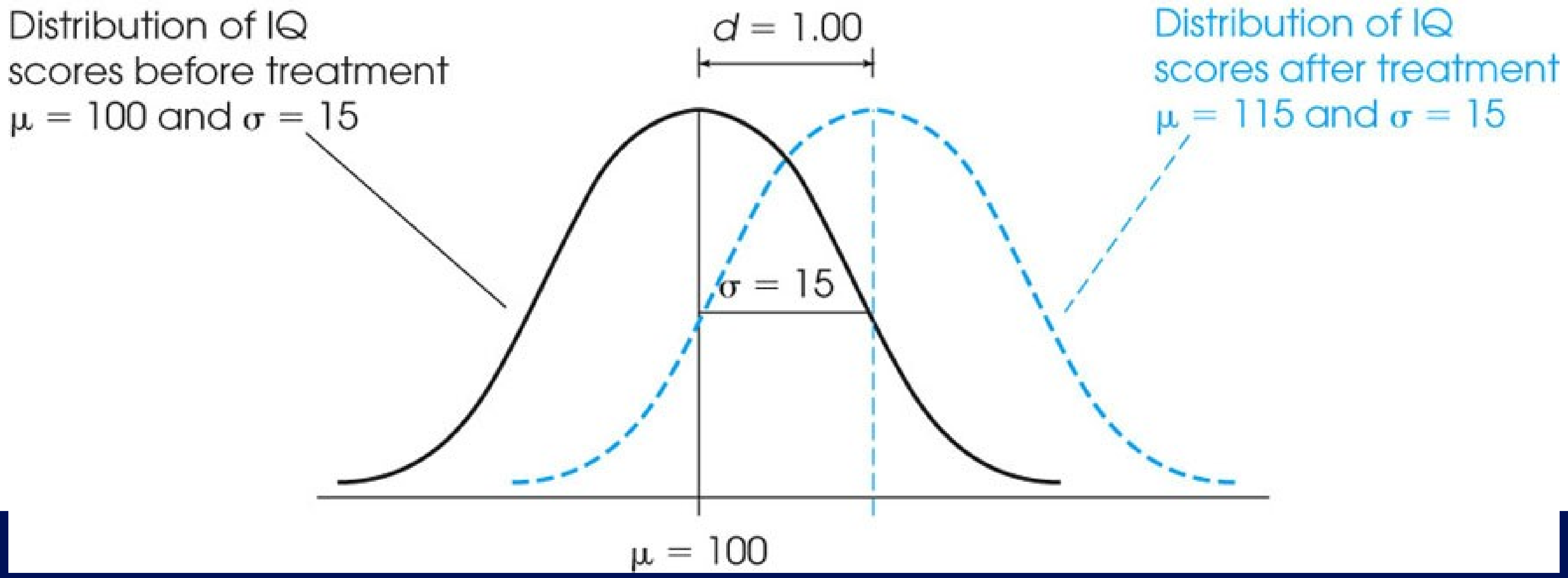
- Because a significant effect does not necessarily mean a large effect, it is recommended that the hypothesis test be accompanied by a measure of the **effect size**.
- For mean comparisons we typically use Cohen's d as a standardized measure of effect size.
- Much like a z-score, **Cohen's d** measures the size of the mean difference in terms of the standard deviation.

TABLE 1: Thresholds for interpreting effect size

Test	Relevant effect size	Effect size threshold			
		Small	Medium	Large	Very large
Standardized mean difference	d, Δ, Hedges' g	.20	.50	.80	1.30
Correlation	r	.10	.30	.50	.70

Notes: The rationale for these benchmarks can be found in Cohen (1988) at the following pages: d (p.40) and r (pp.79-80). Supplementing Cohen's (1988) original small, medium and large effect sizes, Rosenthal (1996) added a classification of very large, defined as being equivalent to, or greater than d = 1.30 or r = .70.


Distribution of SAT
scores before treatment
μ = 500 and σ = 100
d = 0.15
Distribution of SAT
scores after treatment
μ = 515 and σ = 100
σ = 100
μ = 500


Distribution of IQ
scores before treatment
μ = 100 and σ = 15
d = 1.00
Distribution of IQ
scores after treatment
μ = 115 and σ = 15
σ = 15
μ = 100

Cohen’s D effect size

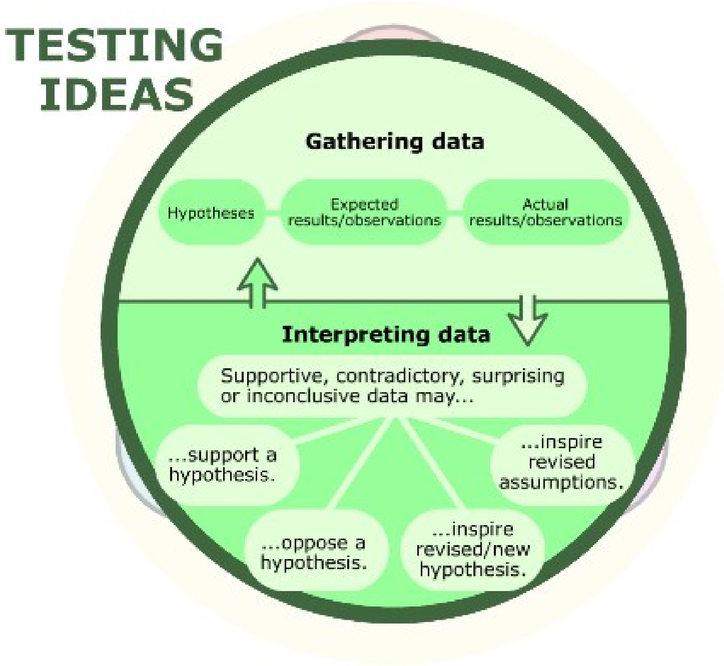
1. Install/load the ‘psych’ package
2. Read about the cohen.d() function
3. Load the sat.act data from the psych package into your global env and inspect it
4. Calculate the effect size that gender has on each of the variables in the sat.act dataset (gender is coded as males = 1, females = 2)
5. What do the effect sizes tell you?

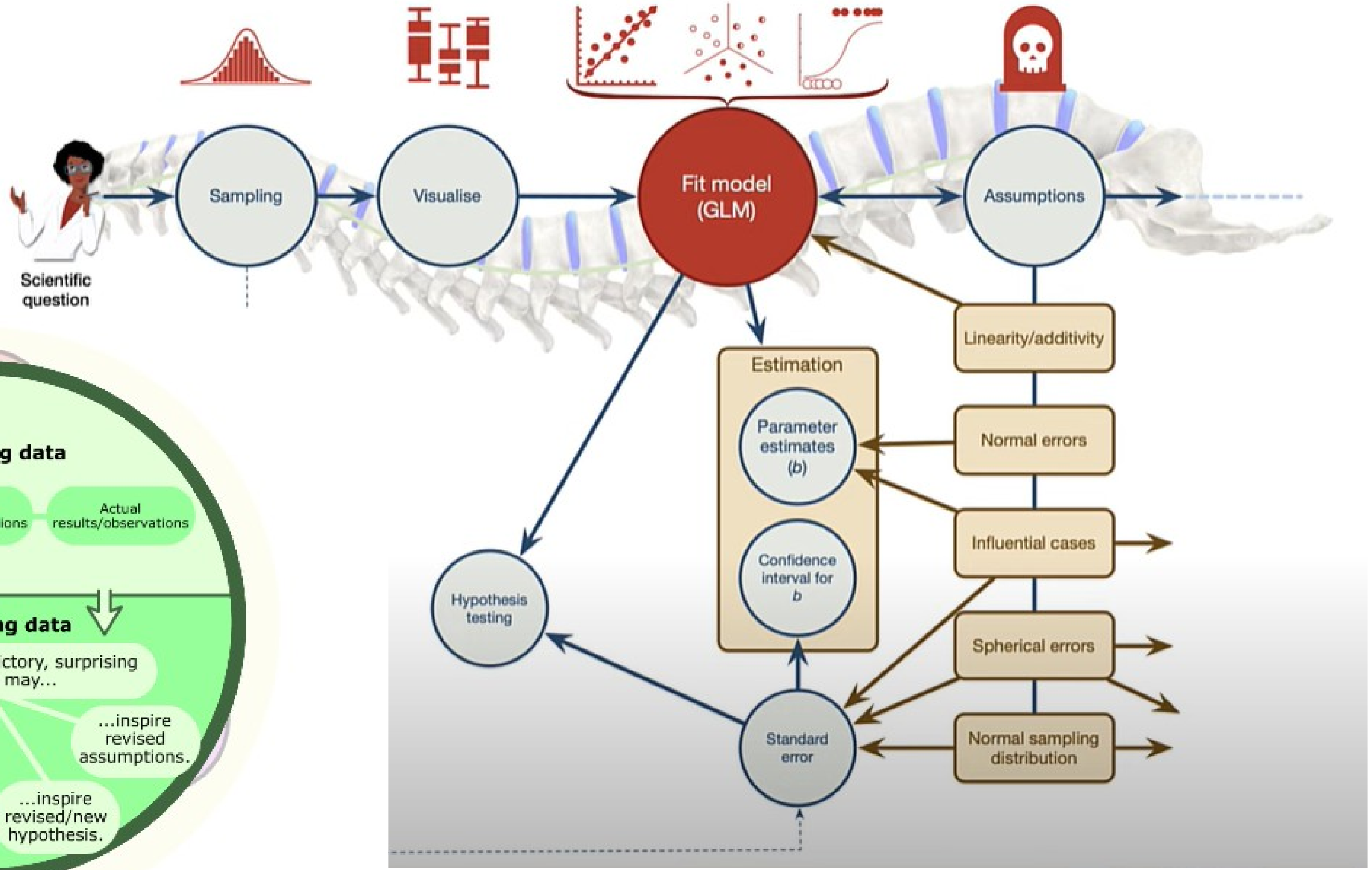
Hypothesis Testing and a Model Based Approach

Data = Model predictions + Error

- Comparison (relationship between variables, difference between conditions)
- Prediction (predicted outcomes vs predicting relationship vs using the model for prediction)
- Model fit

Hypothesis Testing in a Model-Based Framework





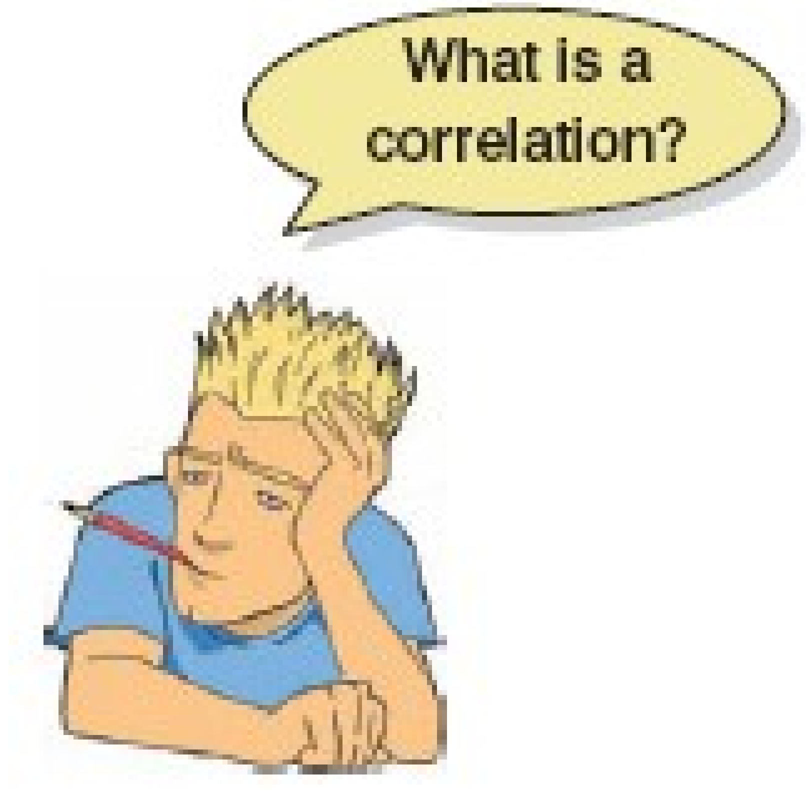
Correlation

Outline

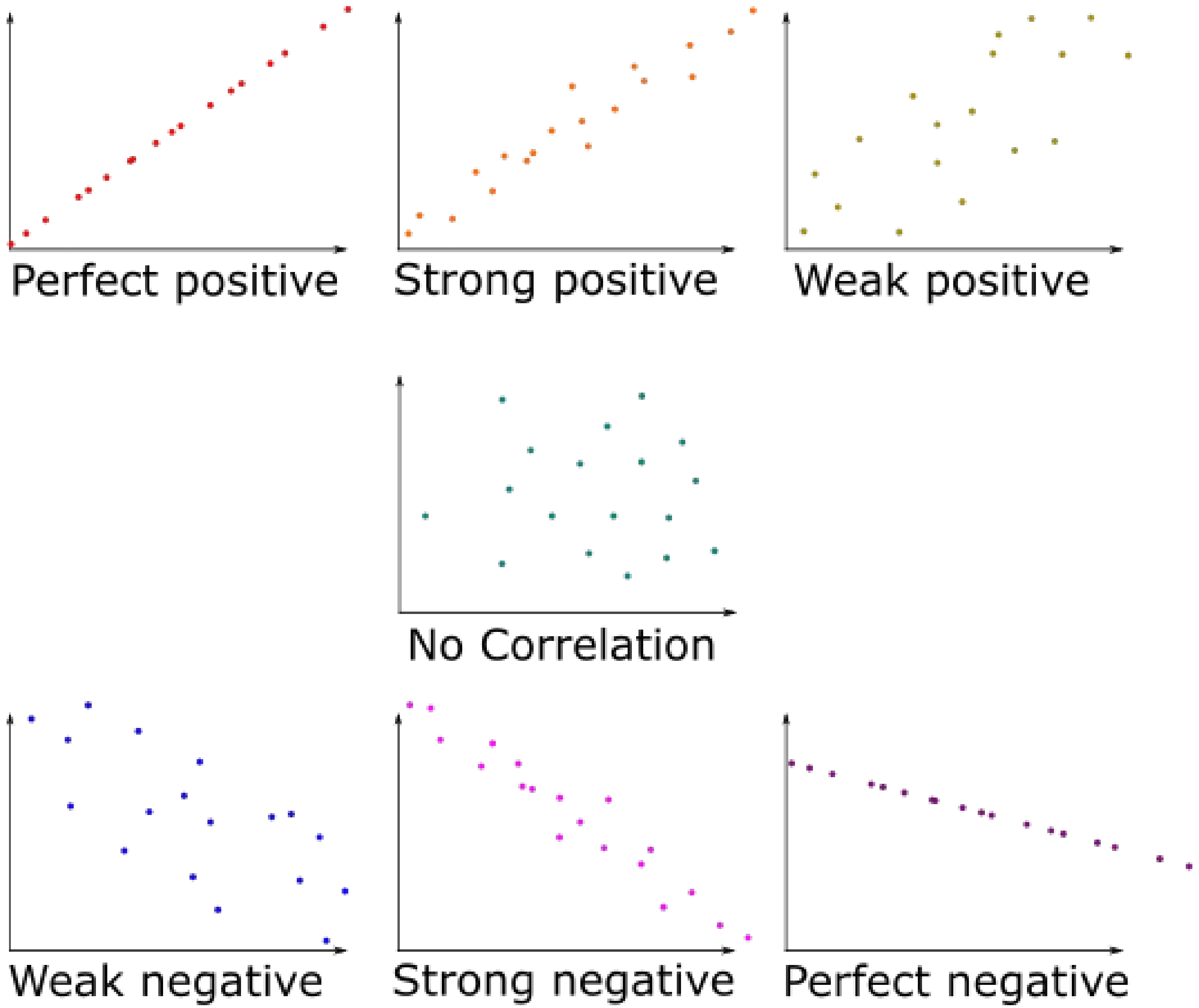
1. Measuring relationships
   - Scatterplots
   - Covariance
   - Pearson's correlation coefficient
2. Nonparametric measures
   - Spearman's rho
   - Kendall's tau
3. Interpreting correlations
   - Causality
4. Partial correlations

What is a Correlation?

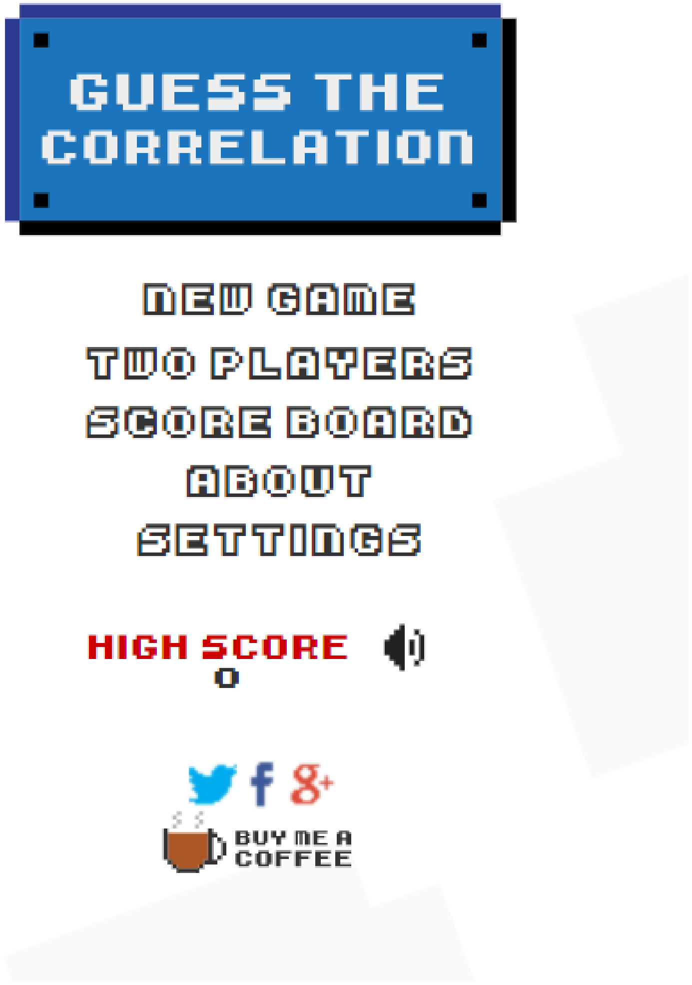
- It is a way of measuring the extent to which two variables are related.
- It measures the pattern of responses across variables.
- What correlation is NOT:
  - a measure of causation



Different types of relationships



www.guessthecorrelation.com

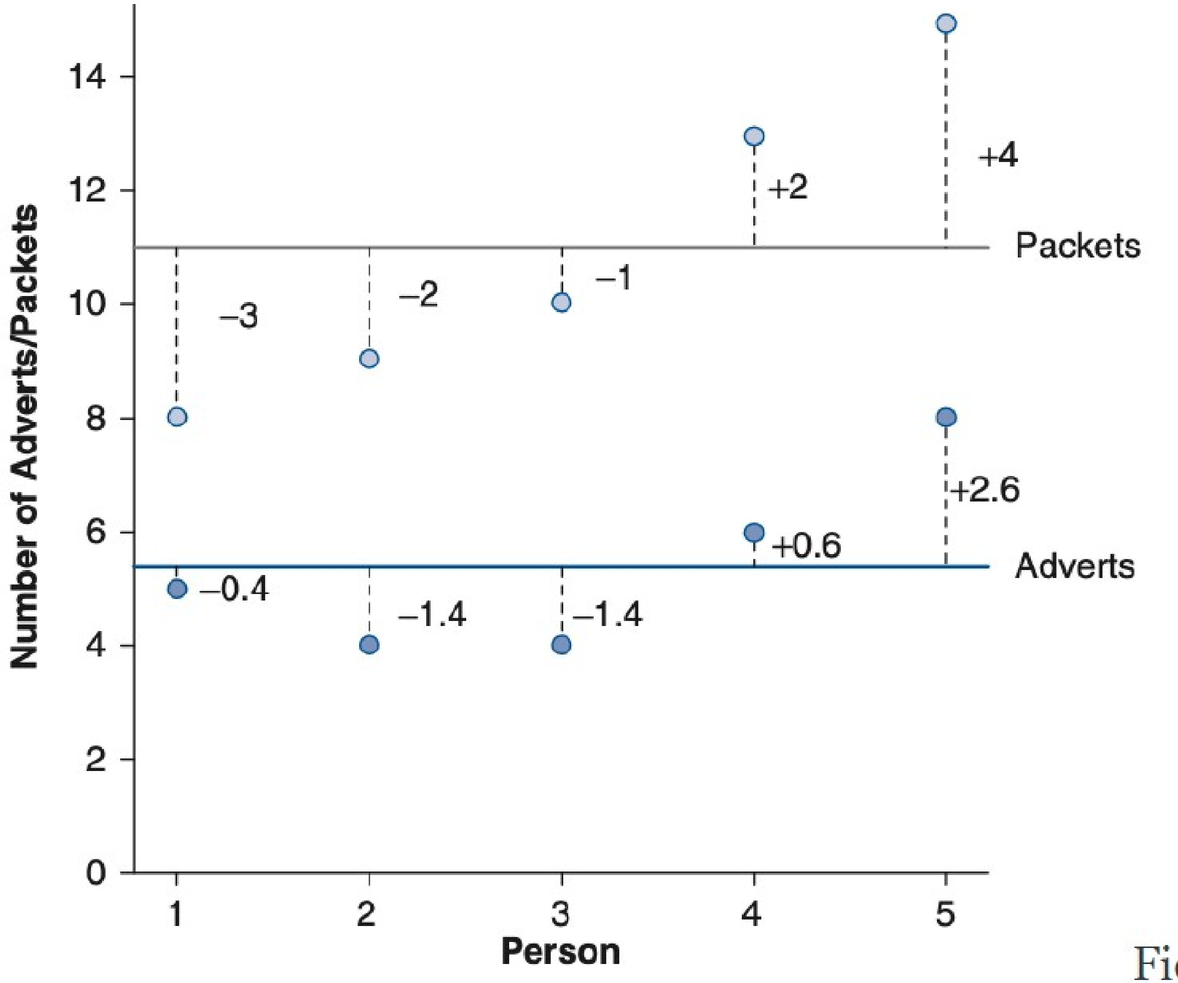


Measuring Relationships

- We need to see whether as one variable increases, the other increases, decreases or stays the same.
- This can be done by calculating the **covariance**.
  - We look at how much each score deviates from the mean.
  - If both variables deviate from the mean by the same amount, they are likely to be related.

Table 6.1 Adverts watched and toffee purchases

Participant:	1	2	3	4	5	Mean	s
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92



Field, Miles & Field (2012). *Chapter 6*

Revisiting the Variance

- The variance tells us by how much scores deviate from the mean for a single variable.
- It is closely linked to the sum of squares.
- Covariance is similar – it tells is by how much scores on two variables differ from their respective means.

Sum of squared errors (deviations from the mean)
Square to avoid values canceling each other out when summing.

$$\text{variance} =$$

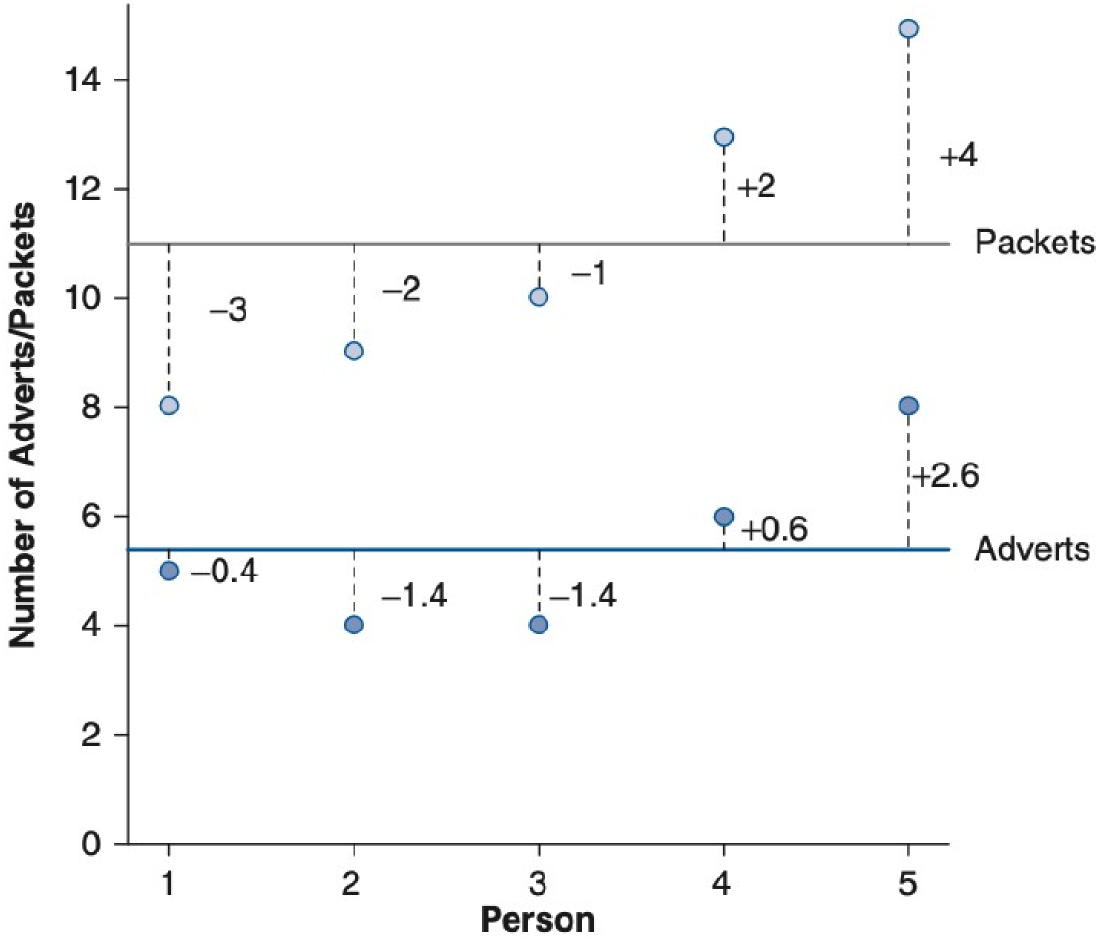
Covariance

- Calculate the error between the mean and each subject's score for the first variable (x).
- Calculate the error between the mean and their score for the second variable (y).
- Multiply these error values.
- Add these values and you get the cross product deviations.
- The covariance is the average cross-product deviations:

$$\mathrm{cov}(x \quad y = \frac{\sum}{N-}$$

Table 6.1 Adverts watched and toffee purchases

Participant:	1	2	3	4	5	Mean	s
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92



$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4}$$

$$= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4}$$

$$= \frac{17}{4}$$

$$= 4.25$$

Covariance in R

- Create two variables in R:
  - Ads with values 5,4,4,6,8
  - Packets with values 8, 9, 10, 13, 15
- Use the cov() function to calculate the covariance of ads and packets

Problems with Covariance

- It depends upon the **units of measurement**.
  - E.g. the covariance of two variables measured in miles might be 4.25, but if the same scores are converted to kilometres, the covariance is 11.
- One solution: **standardize it!**
  - Divide by the standard deviations of both variables.
- The standardized version of covariance is known as the **correlation coefficient.**
  - It is relatively unaffected by units of measurement.
- BONUS: We can manually standardize variables to put them on the same scale by subtracting the mean from each value and dividing by the standard deviation (turning them into z-scores).

Pearson's Correlation Coefficient

$$r = \frac{\mathrm{cov}_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

- Standardized covariance based on standard deviation of the two variables
  - Pearsons product moment correlation coeffecient
  - Ranged between -1 and +1 (where 0 means no correlation)

$$r =$$

$$= \frac{4.25}{1.67 \times \quad 2}$$

$$=$$

Standardizing and Correlation in R

- Create new variables by standardizing both variables below using the scale() function:

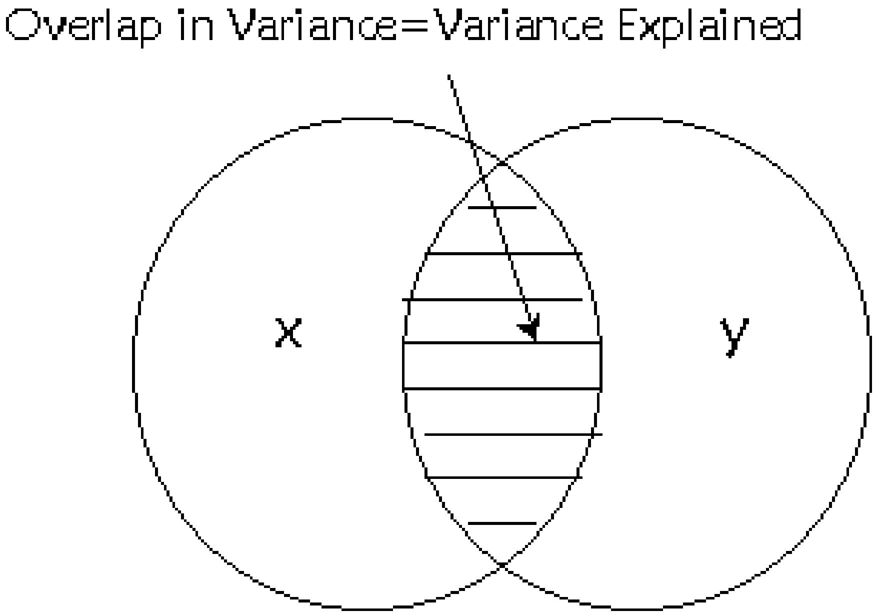
  - Ads

  - Packets

- Use the cov() function to calculate the covariance of the **standardized versions** of ads and packets
- Use the cor() function to calculate the correlation of ads and packets

Things to Know about the Correlation

- It varies between -1 and +1
  - 0 = no relationship
- It is an effect size
  - ±.1 = small effect
  - ±.3 = medium effect
  - ±.5 = large effect
- Coefficient of determination, r^2
  - By squaring the value of r you get the proportion of variance in one variable shared by the other.



General Procedure for Correlations Using R

- To compute basic correlation coefficients there are three main functions that can be used:

  cor(), *cor.test()* and *rcorr()*.

Function	Pearson	Spearman	Kendall	*p*-values	CI	Multiple Correlations?
cor()	✓	✓	✓			✓
cor.test()	✓	✓	✓	✓	✓	
rcorr()	✓	✓		✓		✓

Hypothesis testing with correlations

- We can use cor.test() to test the null hypothesis that the correlation in the population is 0.
- And we can also specify our alternative for whether there should be a negative relationship (alternative ='less') or positive association (alternative='greater')
- Also provides p-values, and CIs

Correlation Testing in R

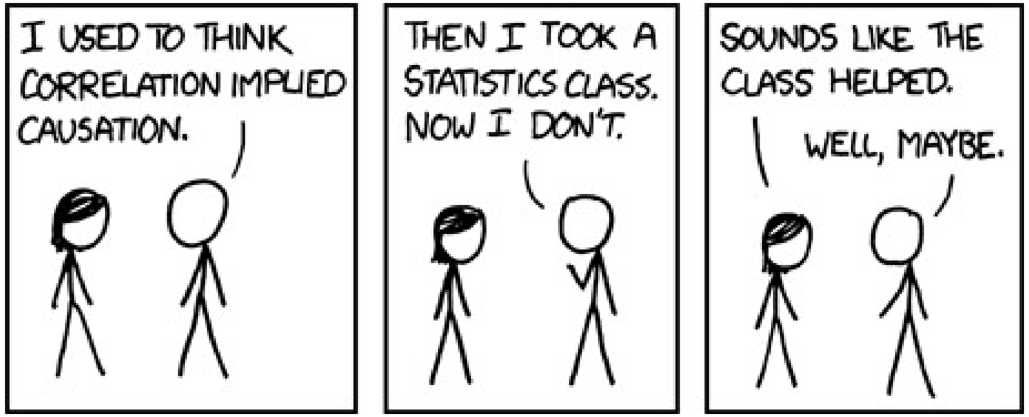
- Use the cor.test() function to calculate the correlation of ads and packets
- Use the cor.test() function to calculate the correlation of ads and packets predict a negative association
- Use the cor.test() function to calculate the correlation of ads and packets predict a positive association

- What do these results suggest to you? Is the correlation between these variables significant?

Correlation and Causality

- The third-variable problem:
  - In any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change.

Check this out for some interesting spurious correlations



Summing Up

- Correlations
  - Positive, negative and range from -1 to 1
  - Small, medium, large effects
  - Not causal

Preparing for Module 4: Correlation

- Field, Miles & Field – CH 6
- Attend Practical Session and Complete Practical Exercsie

Thanks! See you next week! Questions?

Tilburg University