Introduction to R

Antonia Santos

Program

PART I

Introduction to R and base R programming

PART II

Data manipulation

PART III

Data visualisation

PART IV

Introduction to modelling in R

Bibliography

- R Manual (`https://cran.r-project.org/doc/manuals/R-intro.html`)
- R for Data Science (2e) (`https://r4ds.hadley.nz`)
- Fundamentals of Data Visualisation (`https://clauswilke.com/dataviz/`)

Part IV

Data visualisation

1. HYPOTHESIS TESTING
2. T-TESTS
3. ANOVA
4. LINEAR MODELS

HYPOTHESIS TESTING

Tonight, you're going to a party. The weather forecast says there's an 80% chance of rain. Do you take an umbrella?

HYPOTHESIS TESTING

A statistical method used to make decisions or inferences about a population based on sample data.

HYPOTHESIS TESTING

A statistical method used to make decisions or inferences about a population based on sample data.

H_0 : Null hypothesis
H_1 : Alternative hypothesis

HYPOTHESIS TESTING

A person comes into court charged with a crime. A jury must decide whether the person is innocent or guilty. What is the null hypothesis?

HYPOTHESIS TESTING

A person comes into court charged with a crime. A jury must decide whether the person is innocent or guilty. What is the null hypothesis?

H_0 : The person is innocent.
H_1 : The person is guilty.

HYPOTHESIS TESTING

- A company claims that their new energy drink increases productivity. A researcher wants to test whether the drink has a significant effect on productivity.
- A scientist is testing a new medication to reduce blood pressure. They want to determine if the medication is effective in lowering blood pressure compared to a placebo.
- A school wants to know if there is a difference in average test scores between students who attend tutoring sessions and those who do not.
- A factory claims that their light bulbs last an average of 1000 hours. A quality control team tests whether the average lifespan of the bulbs is different from 1000 hours.

HYPOTHESIS TESTING

Type of errors

Since the decision to accept or reject H_0 is based solely on information from a sample of the population, it is possible to commit one of the following errors:

- Rejecting H_0 when H_0 is true (Type I error);
- Failing to reject H_0 when H_0 is false (Type II error).

Decision	Reality	
	H_0 true	H_0 false
Do not reject H_0	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	Correct decision

HYPOTHESIS TESTING

Type of errors

H_0 : It will rain tonight.

H_A : It will not rain tonight.

- **Type I Error:** You reject H_0, believe it won't rain, go without an umbrella, and get wet.
- **Type II Error:** You do not reject H_0, believe it will rain, take an umbrella, and spend the whole night carrying it without needing to use it.

Decision	Reality	
	H_0: It rains	H_A: It doesn't rain
Takes umbrella	Correct decision	Type II error (β)
Doesn't take umbrella	Type I error (α)	Correct decision

Since these errors are inevitable, a good test should aim to minimise the probability of committing them. As Type I Error is the more serious one, the main concern is generally with the probability of committing this type of error, and the test is then referred to as a significance test.

HYPOTHESIS TESTING

Basic concepts

- **Hypotheses**: These establish the beliefs (statements) to be tested. They are defined based on the knowledge of the problem and can be either simple or composite.
- **Level of significance (α)**: Associated with the decision rule. It is the probability of committing a Type I Error:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}).$$

- **Test statistic**: A statistic that depends on the parameter of interest, but has a known distribution independent of this parameter.
- **Decision rule**: A rule that, based on the data obtained and the significance level α, establishes when H_0 will be rejected.
- **Descriptive level (p-value)**: The probability of obtaining more extreme statistics for the rejection of H_0 than the one provided by the sample.

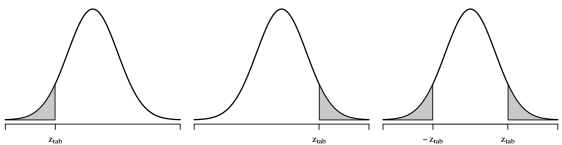
HYPOTHESIS TESTING

Types of tests

- Simple Hypotheses

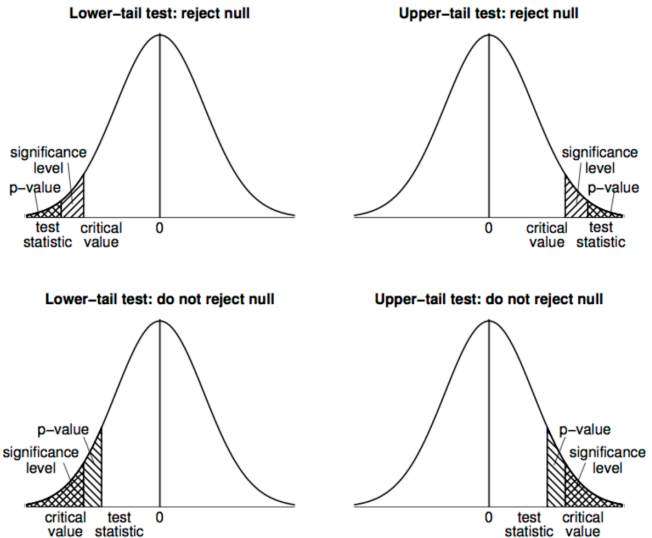
$$H_0 : \theta = a \text{ versus } H_A : \theta = b.$$

- Composite Hypotheses
  - Unilateral (left-tailed): $H_0 : \theta = a$ versus $H_A : \theta < a$;
  - Unilateral (right-tailed): $H_0 : \theta = a$ versus $H_A : \theta > a$;
  - Bilateral: $H_0 : \theta = a$ versus $H_A : \theta \neq a$.



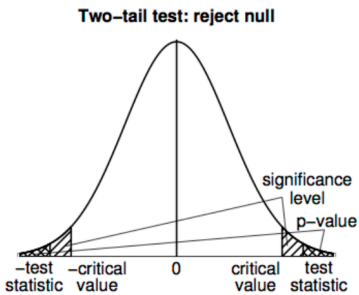
HYPOTHESIS TESTING

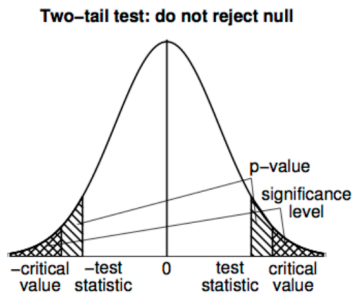
Types of tests



HYPOTHESIS TESTING

Types of tests





HYPOTHESIS TESTING

Steps in hypothesis testing

1. State the hypotheses: Define H_0 and H_A.
2. Choose a significance level (α): Typically 0.05.
3. Select the appropriate test: t-test, ANOVA, etc.
4. Calculate the test statistic: Using sample data.
5. Determine the p-value: Compare to  .
6. Make a Decision: Reject or fail to reject H_0.
7. Draw conclusions: Interpret the results in context.

HYPOTHESIS TESTING

Test for the mean of a population

- **Null Hypothesis (H_0)**: The population mean is equal to a specific value, μ_0.

$$H_0 : \mu = \mu_0$$

- **Alternative Hypothesis (H_A)**: The population mean is different from ($\neq$), greater than ($>$), or less than ($<$) μ_0, depending on the nature of the test (two-tailed or one-tailed).
  - For a two-tailed test:

$$H_A : \mu \neq \mu_0$$

  - For a right-tailed test:

$$H_A : \mu > \mu_0$$

  - For a left-tailed test:

$$H_A : \mu < \mu_0$$

HYPOTHESIS TESTING

Test for the mean of a population

- If the population variance is unknown and the sample size is small ($n < 30$), we use the **t-statistic**:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

  Where:
  - $\bar{x}$ is the sample mean,
  - s is the sample standard deviation,
  - n is the sample size.
- For large samples ($n \geq 30$) or known population variance, the **z-statistic** can be used:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

  Where σ is the population standard deviation.
- For a **two-tailed test**, reject H_0 if the test statistic falls outside the critical values at both tails.
- For a **one-tailed test**, reject H_0 if the test statistic is greater (right-tailed) or smaller (left-tailed) than the critical value.

HYPOTHESIS TESTING

Your turn.

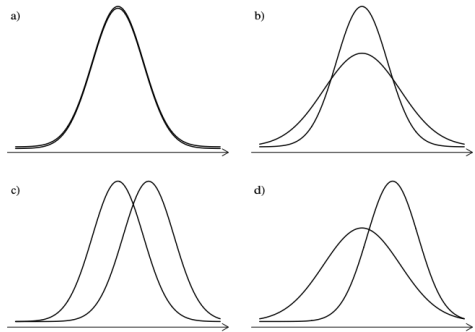
Question 1:

- *airquality data: Is the mean ozone level (Ozone) in the airquality dataset significantly different from 50 ppb*
- *mtcars data: Test if the mean miles per gallon (mpg) of cars with 6 cylinders is less than 20.*
- *Iris data: Test if the mean sepal length of the species setosa is different from 5.0 cm.*

HYPOTHESIS TESTING

Test for comparing means

Given two populations, characterised within the same family of distributions, the objective in this case is to test comparative statements about the parameters of the two populations.



HYPOTHESIS TESTING

Test for comparing means

Paired data (dependent): These are cases where it is reasonable to assume that there is a correlation between the observations from different populations. Examples:

- Before and after experiments;
- Different measurements on the same sampling unit.

Unpaired data (independent): These are cases where it is reasonable to assume independence between the observations from different populations.
A t-test for two populations (two-sample t-test) is used to determine whether the means of two independent groups are significantly different from each other. It assumes that the data in both groups follow a normal distribution.

HYPOTHESIS TESTING

Your turn.

Question 2:

- *iris data: Test if the mean sepal length of the species setosa is different from that of versicolor.*
- *mtcars: Test if the mean miles per gallon (mpg) of cars with 4 cylinders is different from that of cars with 6 cylinders.*
- *airquality data: Is there a significant difference in mean temperature (Temp) between May and September?*

Analysis of Variance (ANOVA)

- **ANOVA** stands for **Analysis of Variance**.
- A method used to test the equality of three or more population means, based on the analysis of sample variances.
- The sample data are divided into groups according to a characteristic (factor).
- **Factor (or treatment)**: is a characteristic that allows distinguishing different populations from one another. Each factor contains two or more groups (classifications).

It answers questions like:

- Do different diets affect animal growth?
- Do various flour types impact bread hardness?

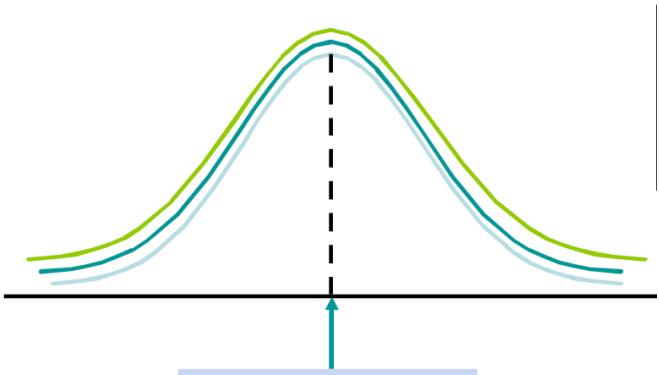
Analysis of Variance (ANOVA)

- We can compare more than two groups at once.
- The null hypothesis (H_0): All group means are equal.
- The alternative hypothesis (H_A): At least one group mean is different.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_t$$

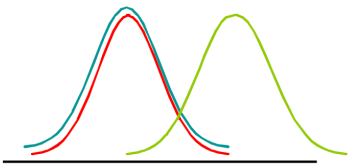
$$H_A : \mu_i \neq \mu_j, \text{ for at least one pair } (i, j)$$

Analysis of Variance (ANOVA)

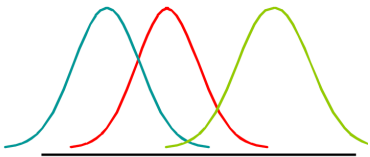


Analysis of Variance (ANOVA)

Assumptions



or



$\mu_1 = \mu_2 \neq \mu_3$

$\mu_1 \neq \mu_2 \neq \mu_3$

Analysis of Variance (ANOVA)

- Populations are normally distributed.
- Populations have the same variance (or the same standard deviation).
  - To test the homogeneity of variances in an ANOVA model, you can use Bartlett's test or Levene's test. Homogeneity of variances is an important assumption in ANOVA, and these tests check whether the variances of the groups are homogeneous.
  - If the hypothesis of equal variances is rejected, another version of the ANOVA can be used: the Welch ANOVA.
  - Welch's ANOVA is typically used for one-way ANOVA when the assumption of equal variances across groups is violated. However, for two-way ANOVA, there is no direct equivalent of Welch's method.
- Samples are random and mutually independent.
- The different samples are obtained from populations classified into only one category.

Analysis of Variance (ANOVA)

Difference between ANOVA and t-test

The main difference between ANOVA (Analysis of Variance) and the t-test lies in the number of groups each compares and the context in which they are applied.
The t-test is used to compare the means of two groups.
Example: Comparing the mean of a control group with the mean of an experimental group.
There are different types of t-tests:

- **Independent samples t-test**: compares the means of two different groups (e.g., men vs. women).
- **Paired samples t-test**: compares the means of the same group at two different times or under two different conditions (e.g., before and after a treatment).

When to use the t-test? When you want to know if there is a significant difference between the means of two groups.

Analysis of Variance (ANOVA)

Difference between ANOVA and t-test

ANOVA is used to compare the means of three or more groups. It assesses whether at least one of the groups is significantly different from the others.
There are different types of ANOVA:

- **One-way ANOVA**: compares the means of groups that differ in only one factor (e.g., three different diets).
- **Two-way ANOVA**: assesses the effect of two factors at the same time (e.g., diet and age).
- **Repeated measures ANOVA**: used when measuring the same group at different times or under repeated conditions.

Analysis of Variance (ANOVA)

Effect of treatments on dried weight of plants

- H_0: The weight is the same for all treatments.
- H_A: The weight is not the same for at least one treatment.

Analysis of Variance (ANOVA)

Your turn.

Question 3: *We are interested in analysing whether the sepal length (Sepal.Length) differs significantly among the three species.*

- *Create a box plot using ggplot2 to visualize the distribution of Sepal.Length for each species. Label the axes appropriately and add a meaningful title.*
- *Perform a one-way ANOVA to test if there is a significant difference in Sepal.Length among the three species.*
- *Write down the null and alternative hypotheses before running the test*
- *Check whether the residuals of the ANOVA model follow a normal distribution using a histogram and a Q-Q plot. Perform a Shapiro-Wilk test as well.*
- *Use Levene's test (from the car package) or Bartlett's test to check if the variance is equal across groups.*
- *Interpret the ANOVA results and discuss whether there is enough evidence to suggest a significant difference in Sepal.Length between species.*
- *If the assumptions are violated, suggest an alternative analysis.*

Thank you!