Technical Panel Interview

Diego Maximiliano Macall, Ph.D.

04-05-2026

diegomacall.github.io diego-macall 0000-0002-9436-8740

Presentation Outline

About me

About me

Barcelona, Spain

- Ph.D. in Environmental Science & Technology, M.Sc. in Agricultural Economics, Agronomic Engineer:
  - Prizes:

About me

Barcelona, Spain

- Ph.D. in Environmental Science & Technology, M.Sc. in Agricultural Economics, Agronomic Engineer:
  - Prizes:
    - *"la Caixa"* Doctoral INPhINIT Fellowship.
    - GIZ Essay Contest Winner, COFOCA VII participation.

About me

Barcelona, Spain

- Ph.D. in Environmental Science & Technology, M.Sc. in Agricultural Economics, Agronomic Engineer:
  - Prizes:
    - *"la Caixa"* Doctoral INPhINIT Fellowship.
    - GIZ Essay Contest Winner, COFOCA VII participation.
  - Chemistry, Organic Chemistry, & Biochemistry.
  - Econometrics, AI, ML, LP.

About me

Barcelona, Spain

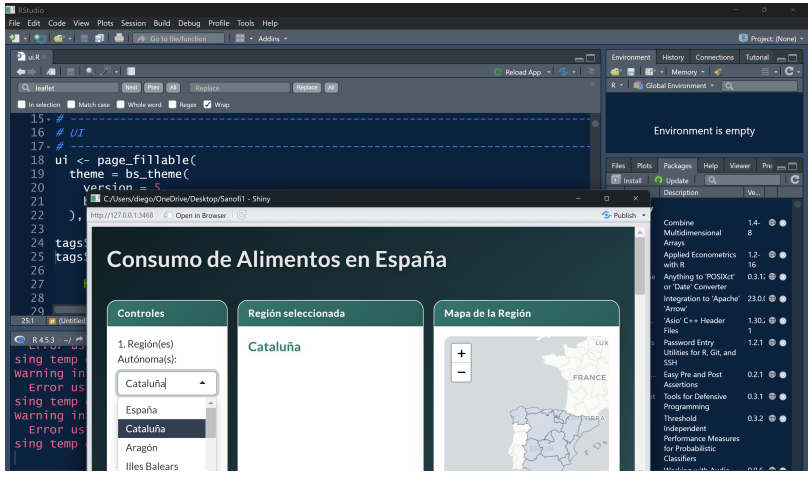
- Strong technical writing skills
  - 17 peer-reviewed articles, 7 as lead author.
  - Policy briefs, blogs, social media.

About me

Barcelona, Spain

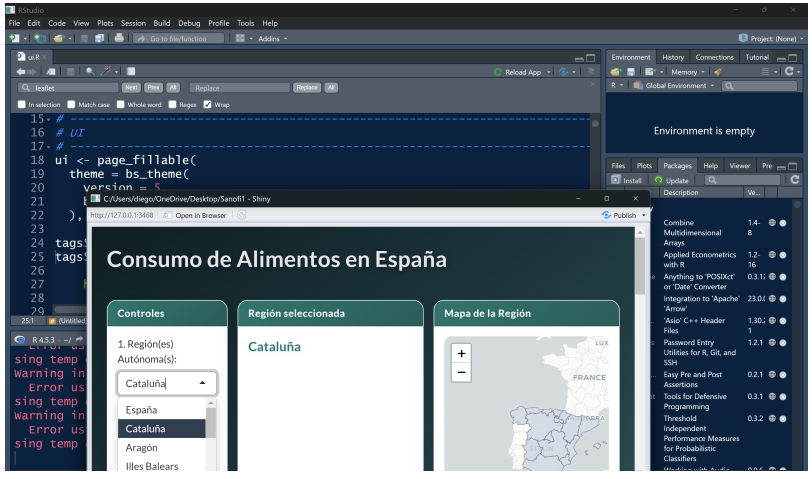
- International professional experience:
  - Worked in Brazil, Canada, El Salvador, Spain, & U.S.
  - Collaborated in multidisciplinary teams.

What am I working on right now?



Spain Food Consumption

What am I working on right now?



Spain Food Consumption

ICABR

Statistics for Chemical Manufacturing Control (CMC Statistics)

Introduction to Experimental Design

- Experimentation is expensive in terms of time, human capital, and resources.

Introduction to Experimental Design

- Experimentation is expensive in terms of time, human capital, and resources.
- It can be defined as "the investigation of a defined area with a firm objective, using appropriate tools and drawing conclusions that are justified by the experimental data so obtained" (Armstrong, 2006).
- Conclusions that can be drawn from the data depend, to a large extent, on the manner in which the data were collected.

Introduction to Experimental Design

- Experimentation is expensive in terms of time, human capital, and resources.
- It can be defined as "the investigation of a defined area with a firm objective, using appropriate tools and drawing conclusions that are justified by the experimental data so obtained" (Armstrong, 2006).
- Conclusions that can be drawn from the data depend, to a large extent, on the manner in which the data were collected.
- Thus, experimental design must include not only the proposed experimental methodology, but also the methods whereby the data from the experiments is to be analyzed.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.
3. Selection of a suitable response.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.
3. Selection of a suitable response.
4. Choice of the experimental design. This is often a balance between cost and statistical validity.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.
3. Selection of a suitable response.
4. Choice of the experimental design. This is often a balance between cost and statistical validity.
5. Performance of the experiment: the data collection process.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.
3. Selection of a suitable response.
4. Choice of the experimental design. This is often a balance between cost and statistical validity.
5. Performance of the experiment: the data collection process.
6. Data analysis.

Stages of Experimental Process

1. Statement of the problem. What is the experiment supposed to achieve? What is its objective?
2. Choice of factors to be investigated, and the levels of those factors that are to be used.
3. Selection of a suitable response.
4. Choice of the experimental design. This is often a balance between cost and statistical validity.
5. Performance of the experiment: the data collection process.
6. Data analysis.
7. Drawing conclusions.

Parametric vs Nonparametric Statistics

Parametric vs Nonparametric Statistics

- **Parametric:** assume a distribution and use full numerical information
- **Nonparametric:** fewer assumptions, rely on ranks or relative comparisons
- **Trade-off:**
  - Parametric $\rightarrow$ efficient under ideal conditions
  - Non-parametric $\rightarrow$ robust under real-world conditions
- **Focus here:**
  - Nonparametric methods, as they are often more applicable to complex, real-world data settings

How do we decide whether to use a parametric or nonparametric test?

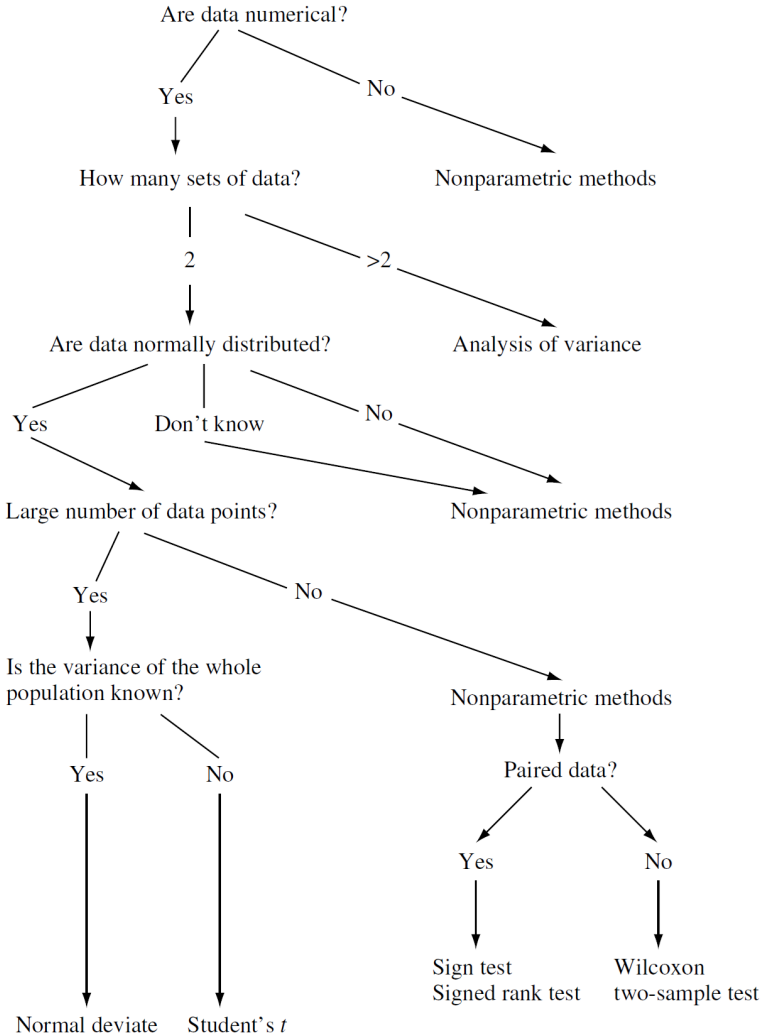


Figure 1: Decision diagram for identifying the correct statistical test for mean comparisons across groups.

Parametric

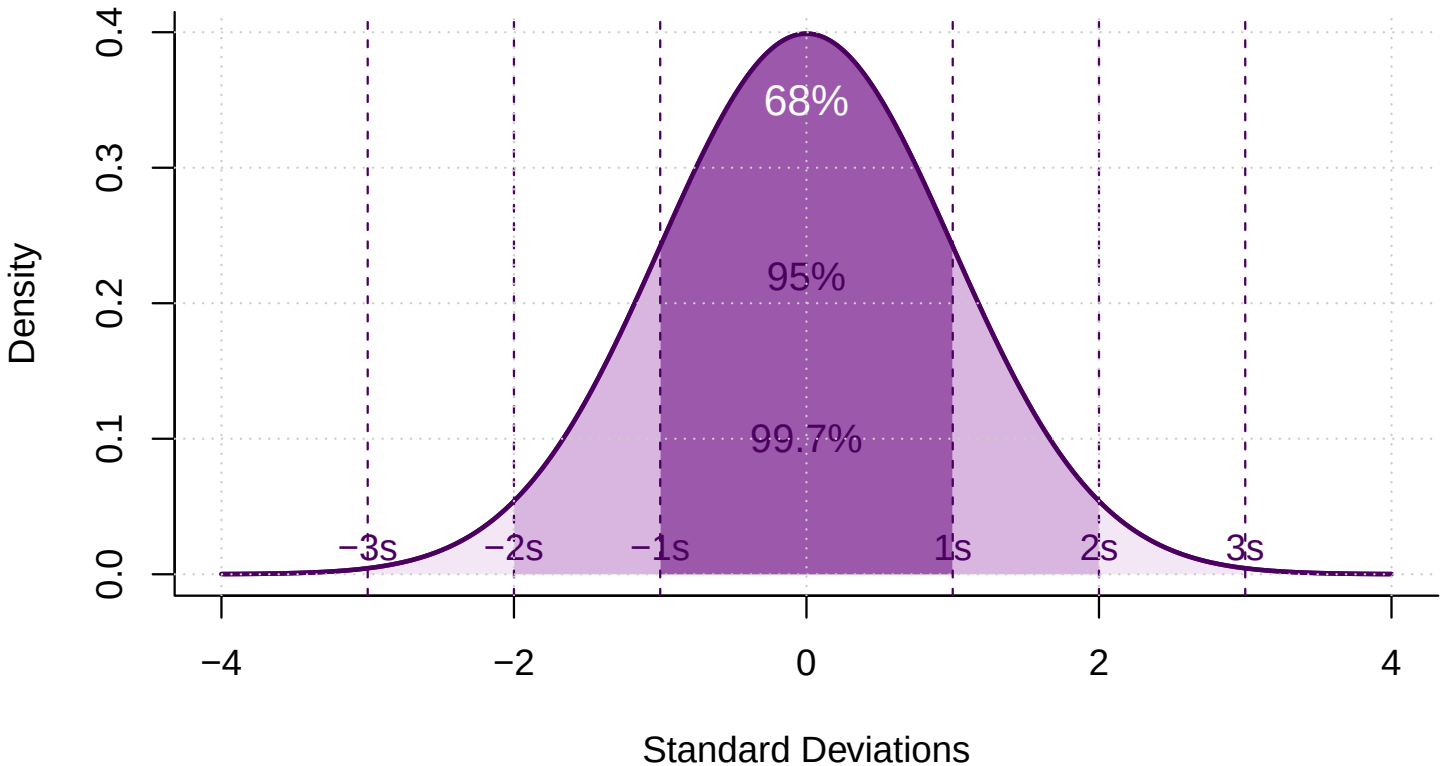
Introduction

- Experiments often compare data across different conditions
- Depend on the assumption that the populations involved are normally distributed.

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

Normal Distribution



Treatment of Outlying Data Points

- There are several possible sources of error in obtaining experimental data; for example, human, instrumentation, and calculation errors may occur.

Treatment of Outlying Data Points

- There are several possible sources of error in obtaining experimental data; for example, human, instrumentation, and calculation errors may occur.
- The use of tests to identify outlying data points is discussed in the 28th edition of United States Pharmacopoeia (Armstrong 2006).
  - *United States Pharmacopeia*, 28th ed., Chapter 1010, United States Pharmacopeial Convention, Rockville, MD, 2005. (Internet Archive)

Comparing Two Means When Population Variance is Uknown (*t-test*)

- Hard-shell capsules are filled with a mixture of active ingredients and diluents (Formulation A).

Comparing Two Means When Population Variance is Uknown (*t-test*)

- Hard-shell capsules are filled with a mixture of active ingredients and diluents (Formulation A).
- A new formulation is devised (Formulation B) which, it is believed, will alter the disintegration times of the capsules.

Comparing Two Means When Population Variance is Uknown (*t-test*)

- Hard-shell capsules are filled with a mixture of active ingredients and diluents (Formulation A).
- A new formulation is devised (Formulation B) which, it is believed, will alter the disintegration times of the capsules.
- The objective of the experiment is therefore to establish whether a significant difference exists between the mean disintegration times of the two formulations.

t-test

1. Choose significance level: Typically $\alpha = 0.05$
2. Number of replicate determinations to be made.

t-test

1. Choose significance level: Typically $\alpha = 0.05$
2. Number of replicate determinations to be made. The disintegration test of the European Pharmacopoeia, which requires six measurements be carried out (Armstrong 2006, p. 14).

t-test

1. Choose significance level: Typically $\alpha = 0.05$
2. Number of replicate determinations to be made. The disintegration test of the European Pharmacopoeia, which requires six measurements be carried out (Armstrong 2006, p. 14).

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{3}$$

Disintegration Time of Hard-Shell Capsules Containing Two Formulations, A and B

	Formulation A	**Formulation B**
	11.1	9.2
	10.3	10.3
	13.0	11.2
	14.3	11.3
	11.2	10.5
	14.7	9.5
n	6	6
Mean	12.43	10.33
Variance	3.36	0.74
Standard deviation	1.83	0.86

Two-Sample t-test

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)}} \tag{4}$$

Summary Statistics

- $\bar{x}_A = 12.43, \quad s_A^2 = 3.36, \quad n_A = 6$
- $\bar{x}_B = 10.33, \quad s_B^2 = 0.74, \quad n_B = 6$

t-test Results

Test Statistic

Check critical values

$$t = \frac{12.43 - 10.33}{\sqrt{\left(\frac{3.36}{6} + \frac{0.74}{6}\right)}} = 2.54 \tag{5}$$

Decision Rule

- Degrees of freedom: $df = n_A + n_B - 2 = 10$
- At $\alpha = 0.05 \rightarrow$ critical value ≈ 2.228
- Compare:
  - $|t| = 2.54 > 2.228$

Reject H_0: significant difference exists

Analysis of Variance (ANOVA)

- Used to compare means across multiple groups (>3)
- It assumes that a random sample has been taken from each population, that each population has a normal distribution, and that all the populations have the same variance.
- The question that ANOVA seeks to answer is, are there significant differences among the means of the groups?

Crushing Strengths of Tablets (kg)

Batch A	Batch B	Batch C
5.2	5.5	3.8
5.9	4.5	4.8
6.0	6.6	5.1
4.4	4.2	4.2
7.0	5.6	3.3
5.4	4.5	3.5
4.4	4.4	4.0
5.6	4.8	1.7
5.6	5.3	5.9
5.1	3.8	4.8

Summary Statistics + ANOVA Equation

	Batch A	**Batch B**	**Batch C**
n	10	10	10
Mean	5.46	4.92	4.11
Variance	0.59	0.69	1.34
SD	0.77	0.83	1.16

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{6}$$

ANOVA Results

Source	SS	df	MS	F
Between groups	9.23	2	4.62	5.28
Within groups	23.61	27	0.87	–
Total	32.84	29	–	–

- $F_{\text{calculated}} = 5.28$
- Degrees of freedom: $df_1 = 2$, $df_2 = 27$
- Compare with critical value from the F table

Check F critical value

Nonparametric Methods

When do we use them?

- When increasing sample size is not feasible
- When data are non-numerical or ordinal (e.g., ranks, categories, severity scales)
- Nonparametric tests
  - Require no assumptions about population distribution
  - Suitable for flexible, real-world data

When do we use them?

- When increasing sample size is not feasible
- When data are non-numerical or ordinal (e.g., ranks, categories, severity scales)
- Nonparametric tests
  - Require no assumptions about population distribution
  - Suitable for flexible, real-world data
- Consider data structure
  - Independent samples vs paired observations

When do we use them?

- When increasing sample size is not feasible
- When data are non-numerical or ordinal (e.g., ranks, categories, severity scales)
- Nonparametric tests
  - Require no assumptions about population distribution
  - Suitable for flexible, real-world data
- Consider data structure
  - Independent samples vs paired observations
- *Experimental design and the method of evaluating the results are inextricably linked*

Linear Regression

- Many experiments consist in changing the value of a factor *(independent variable or the predictor)* and measuring the response (*dependent variable, the outcome*).
- This produces many pairs of data points.

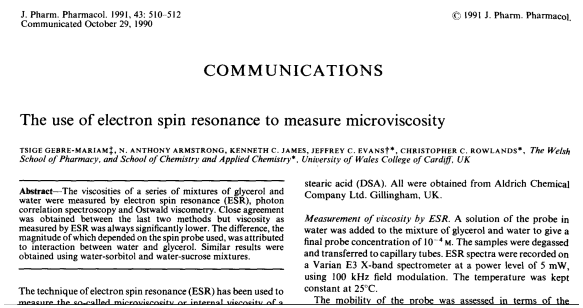
Linear Regression

- Many experiments consist in changing the value of a factor *(independent variable or the predictor)* and measuring the response (*dependent variable, the outcome*).
- This produces many pairs of data points.
- Convenient to present these in graphical form. Conventional to plot the factor on the X-axis (the abscissa) and the response on the Y-axis (the ordinate).

Linear Regression

- Many experiments consist in changing the value of a factor *(independent variable or the predictor)* and measuring the response (*dependent variable, the outcome*).
- This produces many pairs of data points.
- Convenient to present these in graphical form. Conventional to plot the factor on the X-axis (the abscissa) and the response on the Y-axis (the ordinate).
- In sum: Regression is the process of deriving a relationship between one or more factors and a response.

Gerbre-Mariam et al. 1991



J. Pharm. Pharmacol. 1991, 43: 510–512
Communicated October 29, 1990

© 1991 J. Pharm. Pharmacol.

COMMUNICATIONS

The use of electron spin resonance to measure microviscosity

TSIGE GEBRE-MARIAM‡, N. ANTHONY ARMSTRONG, KENNETH C. JAMES, JEFFREY C. EVANS†*, CHRISTOPHER C. ROWLANDS*, *The Welsh School of Pharmacy, and School of Chemistry and Applied Chemistry*, University of Wales College of Cardiff, UK*

Abstract—The viscosities of a series of mixtures of glycerol and water were measured by electron spin resonance (ESR), photon correlation spectroscopy and Ostwald viscometry. Close agreement was obtained between the last two methods but viscosity as measured by ESR was always significantly lower. The difference, the magnitude of which depended on the spin probe used, was attributed to interaction between water and glycerol. Similar results were obtained using water-sorbitol and water-sucrose mixtures.

The technique of electron spin resonance (ESR) has been used to measure the so-called microviscosity or internal viscosity of a

stearic acid (DSA). All were obtained from Aldrich Chemical Company Ltd. Gillingham, UK.

Measurement of viscosity by ESR. A solution of the probe in water was added to the mixture of glycerol and water to give a final probe concentration of 10^{-4} M. The samples were degassed and transferred to capillary tubes. ESR spectra were recorded on a Varian E3 X-band spectrometer at a power level of 5 mW, using 100 kHz field modulation. The temperature was kept constant at 25°C.

The mobility of the probe was assessed in terms of the

The use of electron spin resonance to measure micro viscosity

- Purpose of the study was to predict the viscosity of the contents of the pores in a glycerogelatin matrix.
  - Diffusion of solutes through glycerogelatin gels such as the shells of soft capsules.
  - In such systems, the gelatin forms a porous matrix, and the pores are filled with a mixture of glycerol and water.
  - As with all diffusion processes, the viscosity of the liquid medium was thought to play an important role.
  - With this in mind, the viscosities of a series of mixtures of glycerol and water were determined.

Gerbre-Mariam et al. 1991

Table 1: Viscosities of mixtures of glycerol and water at 23°C

	x	y	x^2	y^2	xy
	12.3	4.83	151.3	23.32	59.41
	18.5	6.32	342.3	39.94	116.92
	24.6	7.50	605.2	56.25	184.50
	30.8	9.66	948.6	93.32	297.53
	36.9	11.90	1361.6	141.61	439.11
Sum	123.1	40.21	3409.0	354.44	1097.47
Mean	24.6	8.04			

Linear Regression Model

The best-fitting straight line through a set of data points is called the **regression line**.

Linear Regression Model

The best-fitting straight line through a set of data points is called the **regression line**.

It can be estimated using least squares and takes the form:

Linear Regression Model

The best-fitting straight line through a set of data points is called the **regression line**.

It can be estimated using least squares and takes the form:

$$y = b_0 + b_1 x \tag{7}$$

Where:

- b_1: slope of the line
- b_0: intercept (value of y when $x = 0$)

Gerbre-Mariam et al. 1991

- The regression line is the line for which the sum of the vertical distances between it and the experimental points is less than the sum obtained with any other straight line.
- $Y_{\text{obs}} - Y_{\text{pred}}$

Gerbre-Mariam et al. 1991

- The regression line is the line for which the sum of the vertical distances between it and the experimental points is less than the sum obtained with any other straight line.
- $Y_{\text{obs}} - Y_{\text{pred}}$
- The slope (b_1) of the regression line, known as the regression coefficient, is calculated as:

$$b_1 = \frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sum(x^2) - \frac{(\sum x)^2}{n}} \tag{8}$$

Gerbre-Mariam et al. 1991

$$b_1 = \frac{1097.47 - 989.97}{3408.95 - 3030.72} = 0.284 \tag{9}$$

Gerbre-Mariam et al. 1991

$$b_1 = \frac{1097.47 - 989.97}{3408.95 - 3030.72} = 0.284 \tag{9}$$

$$y = 1.045 + 0.284x \tag{10}$$

R code: viscosity regression

```r
# Data
glycerol <- c(12.3, 18.5, 24.6, 30.8, 36.9)
viscosity <- c(4.83, 6.32, 7.50, 9.66, 11.90)

# Linear regression
model <- lm(viscosity ~ glycerol)

# Model summary
summary(model)

# Plot
plot(glycerol, viscosity)
abline(model)
```

Degrees of Freedom

- The reliability of a regression model depends on the number of data pairs used. More observations $\Rightarrow$ more reliable predictions.
- The degrees of freedom are given by:

$$\text{degrees of freedom} = n - (k + 1) \tag{11}$$

Where:

- $n =$ number of data pairs
- $k =$ number of explanatory variables

Degrees of Freedom

- The reliability of a regression model depends on the number of data pairs used. More observations $\Rightarrow$ more reliable predictions.
- The degrees of freedom are given by:

$$\text{degrees of freedom} = n - (k + 1) \tag{11}$$

Where:

- $n =$ number of data pairs
- $k =$ number of explanatory variables

Example:

- $n = 5$, $k = 1$

Degrees of Freedom

- The reliability of a regression model depends on the number of data pairs used. More observations $\Rightarrow$ more reliable predictions.
- The degrees of freedom are given by:

$$\text{degrees of freedom} = n - (k + 1) \tag{11}$$

Where:

- $n =$ number of data pairs
- $k =$ number of explanatory variables

Example:

- $n = 5$, $k = 1$
- $\Rightarrow$ degrees of freedom $= 5 - (1 + 1) = 3$

Degrees of Freedom

- The reliability of a regression model depends on the number of data pairs used. More observations $\Rightarrow$ more reliable predictions.
- The degrees of freedom are given by:

$$\text{degrees of freedom} = n - (k + 1) \tag{11}$$

Where:

- $n =$ number of data pairs
- $k =$ number of explanatory variables

Example:

- $n = 5$, $k = 1$
- $\Rightarrow$ degrees of freedom $= 5 - (1 + 1) = 3$

Regression Output

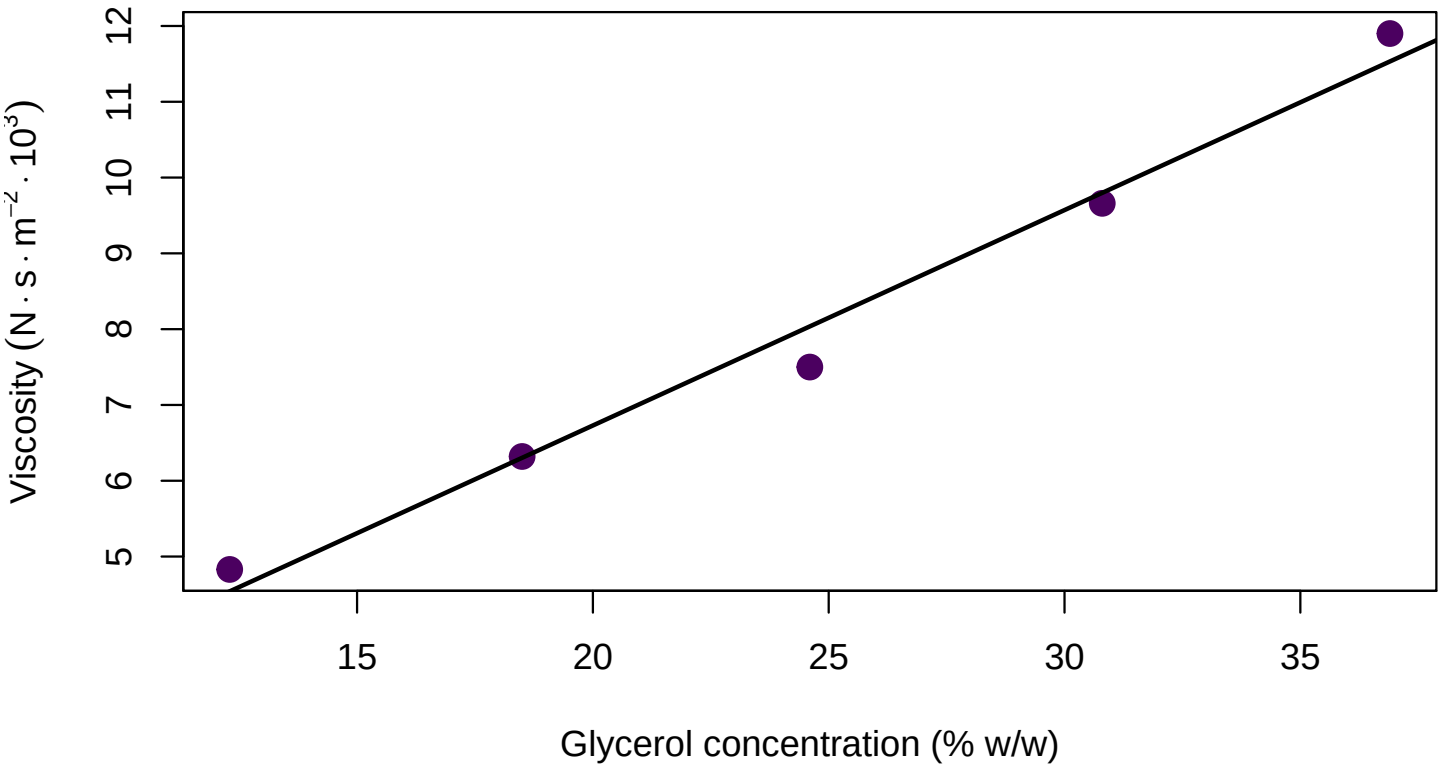
Table 2: Linear regression: viscosity on glycerol concentration

Term	Estimate	Std. Error	t-value	p-value
(Intercept)	1.0447	0.5623	1.8579	0.1602
glycerol	0.2842	0.0215	13.1974	0.0009

Table 3: Model fit statistics

Statistic	Value
R-squared	0.9831
Adjusted R-squared	0.9774
Residual SE	0.4188
F-statistic	174.1712
p-value	0.0009

Regression plot



F-test (Variance Ratio)

- The F-value tests whether the regression represents a real relationship (not random variation).
- It compares explained variation to unexplained variation (ANOVA framework).
- Decision rule:

$$F_{\text{calculated}} \quad \text{vs} \quad F_{\text{critical}}$$

- Example:

$$F = 174.17 \; > \; 34.1 \; \Rightarrow \; \text{model is statistically significant}$$

View F Table

Standard Errors of Coefficients

- The standard error indicates that if the experiment were to be repeated, the value of the coefficient $\beta 1$ should lie between 0.284 ± 0.022.

Standard Errors of Coefficients

- The standard error indicates that if the experiment were to be repeated, the value of the coefficient $\beta 1$ should lie between 0.284 ± 0.022.
- Larger standard error $\Rightarrow$ less reliable coefficient and weaker model representation.

Standard Errors of Coefficients

- The standard error indicates that if the experiment were to be repeated, the value of the coefficient $\beta 1$ should lie between 0.284 ± 0.022.
- Larger standard error $\Rightarrow$ less reliable coefficient and weaker model representation.
- Statistical significance can be tested using the t-ratio:

$$t = \frac{\text{coefficient}}{\text{standard error}}$$

Standard Errors of Coefficients

- The standard error indicates that if the experiment were to be repeated, the value of the coefficient $\beta 1$ should lie between 0.284 ± 0.022.
- Larger standard error $\Rightarrow$ less reliable coefficient and weaker model representation.
- Statistical significance can be tested using the t-ratio:

$$t = \frac{\text{coefficient}}{\text{standard error}}$$

- Example:

$$t = \frac{0.284}{0.022} = 12.91 > 2.35$$

$\Rightarrow$ coefficient is statistically significant (glycerol predicts viscosity).

Standard Errors of Coefficients

- The standard error indicates that if the experiment were to be repeated, the value of the coefficient $\beta 1$ should lie between 0.284 ± 0.022.
- Larger standard error $\Rightarrow$ less reliable coefficient and weaker model representation.
- Statistical significance can be tested using the t-ratio:

$$t = \frac{\text{coefficient}}{\text{standard error}}$$

- Example:

$$t = \frac{0.284}{0.022} = 12.91 > 2.35$$

$\Rightarrow$ coefficient is statistically significant (glycerol predicts viscosity).

Correlation Coefficient

- The correlation coefficient (r):

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} \tag{12}$$

Correlation Coefficient

- The correlation coefficient (r):

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} \tag{12}$$

- The value of the correlation coefficient ranges from -1 through 0 to $+1$.
- The higher the value of r, the greater the likelihood that x and y are correlated.

Multiple Regression Model

- The general form of a multiple regression model is:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{13}$$

- where:
  - Y is the response variable
  - $X_1, \ldots, X_n$ are the explanatory variables (factors)
  - $b_0, \ldots, b_n$ are the coefficients
- If there is only one factor ($n = 1$), the model simplifies to a simple linear regression.

Curve Fitting of Nonliner Relationships

- Quadratic Relationships

$$y = b_0 + b_1 x + b_2 x^2 \tag{14}$$

Curve Fitting of Nonliner Relationships

- Quadratic Relationships

$$y = b_0 + b_1 x + b_2 x^2 \tag{14}$$

- Cubic Equations

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 \tag{15}$$

Curve Fitting of Nonliner Relationships

- Quadratic Relationships

$$y = b_0 + b_1 x + b_2 x^2 \tag{14}$$

- Cubic Equations

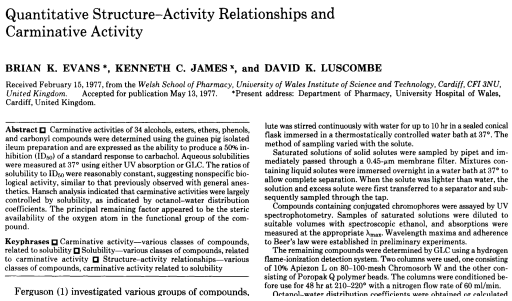
$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 \tag{15}$$

- Transformations

$$y = b_0 b_1^x \tag{16}$$

$$\log y = \log b_0 + x \log b_1 \tag{17}$$

Evans et al. 1977



Quantitative Structure–Activity Relationships and Carminative Activity

BRIAN K. EVANS *, KENNETH C. JAMES *, and DAVID K. LUSCOMBE

Received February 15, 1977, from the *Welsh School of Pharmacy, University of Wales Institute of Science and Technology, Cardiff, CF1 3NU, United Kingdom.* Accepted for publication May 13, 1977. *Present address: Department of Pharmacy, University Hospital of Wales, Cardiff, United Kingdom.

Abstract □ Carminative activities of 34 alcohols, esters, ethers, phenols, and carbonyl compounds were determined using the guinea pig isolated ileum preparation and are expressed as the ability to produce a 50% inhibition (ID_{50}) of a standard response to carbachol. Aqueous solubilities were measured at 37° using either UV absorption or GLC. The ratios of solubility to ID_{50} were reasonably constant, suggesting nonspecific biological activity, similar to that previously observed with general anesthetics. Hansch analysis indicated that carminative activities were largely controlled by solubility, as indicated by octanol–water distribution coefficients. The principal remaining factor appeared to be the steric availability of the oxygen atom in the functional group of the compound.

Keyphrases □ Carminative activity—various classes of compounds, related to solubility □ Solubility—various classes of compounds, related to carminative activity □ Structure–activity relationships—various classes of compounds, carminative activity related to solubility

Ferguson (1) investigated various groups of compounds,

lute was stirred continuously with water for up to 10 hr in a sealed conical flask immersed in a thermostatically controlled water bath at 37°. The method of sampling varied with the solute.

Saturated solutions of solid solutes were sampled by pipet and immediately passed through a 0.45-μm membrane filter. Mixtures containing liquid solutes were immersed overnight in a water bath at 37° to allow complete separation. When the solute was lighter than water, the solution and excess solute were first transferred to a separator and subsequently sampled through the tap.

Compounds containing conjugated chromophores were assayed by UV spectrophotometry. Samples of saturated solutions were diluted to suitable volumes with spectroscopic ethanol, and absorptions were measured at the appropriate λ_{max}. Wavelength maxima and adherence to Beer's law were established in preliminary experiments.

The remaining compounds were determined by GLC using a hydrogen flame-ionization detection system. Two columns were used, one consisting of 10% Apiezon L on 80–100-mesh Chromosorb W and the other consisting of Poropak Q polymer beads. The columns were conditioned before use for 48 hr at 210–220° with a nitrogen flow rate of 60 ml/min.

Octanol–water distribution coefficients were obtained or calculated

Quantitative Structure-Activity Relationships and Carminative Activity

- The objective of their experiment was to investigate the carminative properties of a series of 26 volatile compounds.

Key Variables

- The hypothesis is that carminative activity depends on two factors, with all compounds sharing an oxygen-containing substituent (linked to hydrogen, an alkyl, or an alkoxy group).

Key Variables

- The hypothesis is that carminative activity depends on two factors, with all compounds sharing an oxygen-containing substituent (linked to hydrogen, an alkyl, or an alkoxy group).
- Carminative activity is modeled using two factors:
  - Steric bulk of the substituent (van der Waals volume, V_w, nm^3)
  - Lipophilicity (octanol–water partition coefficient, P)

Key Variables

- The hypothesis is that carminative activity depends on two factors, with all compounds sharing an oxygen-containing substituent (linked to hydrogen, an alkyl, or an alkoxy group).
- Carminative activity is modeled using two factors:
  - Steric bulk of the substituent (van der Waals volume, V_w, nm^3)
  - Lipophilicity (octanol–water partition coefficient, P)
- Response variable:

Key Variables

- The hypothesis is that carminative activity depends on two factors, with all compounds sharing an oxygen-containing substituent (linked to hydrogen, an alkyl, or an alkoxy group).
- Carminative activity is modeled using two factors:
  - Steric bulk of the substituent (van der Waals volume, V_w, nm^3)
  - Lipophilicity (octanol–water partition coefficient, P)
- Response variable:
  - ID_{50} = concentration (M $\times 10^3$) required to reduce a standard carbachol response by 50%

Multiple Regression Analysis

Compound	**Substituent Group**	$\mathbf{V_w}$ (nm^3) (x_1)	$\log P$ (x_2)	$\log(1/\mathrm{ID}_{50})$ (M $\times 10^3$) (y)
Isobutanol	H	0.22	0.74	0.77
n-Butyl acetate	$CH_3C{=}O$	3.64	1.74	1.36
1,2-Dihydroxybenzene	H	0.22	0.95	1.02
1,3-Dihydroxybenzene	H	0.22	0.79	1.05
1,4-Dihydroxybenzene	H	0.22	0.55	0.91
1-Cresol	H	0.22	1.95	1.64
2-Cresol	H	0.22	1.99	1.54
3-Cresol	H	0.22	1.93	1.54
Dibutyl ether	$CH_3(CH_2)_3$	6.51	3.06	1.23
Diethyl ether	CH_3CH_2	3.41	0.80	0.59
3,4-Dimethylphenol	H	0.22	2.42	1.91
Di-isopropyl ether	$(CH_3)_2CH$	4.97	1.63	0.71
Di-n-propyl ether	$CH_3(CH_2)_2$	4.97	3.03	1.00
Ethyl acetate	$CH_3C{=}O$	3.64	0.70	0.59
Ethylvinyl ether	$CH_2{=}CH$	3.01	1.04	1.21
Eugenol	H	0.22	2.99	2.43
1-Hexanol	H	0.22	2.03	1.47
Menthol	H	0.22	3.31	2.13
2-Methoxyphenol	H	0.22	1.90	1.26
4-Methoxyphenol	H	0.22	1.34	1.32
1-Pentanol	H	0.22	1.16	1.11
2-Phenoxyethanol	H	0.22	1.16	0.90
Isopropyl acetate	$CH_3C{=}O$	3.64	1.02	0.96
n-Propyl acetate	$CH_3C{=}O$	3.64	1.50	0.94
Salicylaldehyde	H	0.22	1.76	1.70
Thymol	H	0.22	3.30	2.66

Table 4: Substituent groups with molar volumes, partition coefficients, and carminative activities of volatile compounds.

Summary Statistics

	x_1	x_2	y
Total	41.17	44.79	33.95
Mean	1.58	1.72	1.31
SD	2.01	0.86	0.53

$$\sum y^2 = 51.42, \quad \sum x_1^2 = 166.29, \quad \sum x_2^2 = 95.55$$

$$\sum x_1 y = 41.75, \quad \sum x_2 y = 67.14, \quad \sum x_1 x_2 = 73.65$$

Table 5: Summary statistics and cross-products used in multiple regression estimation.

Equation

$$\log\left(\frac{1}{\mathrm{ID}_{50}}\right) = 0.670 - 0.132V_w + 0.490\log P \qquad (18)$$

Equation

$$\log\left(\frac{1}{\mathrm{ID}_{50}}\right) = 0.670 - 0.132V_w + 0.490\log P \qquad (18)$$

- Substitute V_w and $\log P$ into the regression equation

Equation

$$\log\left(\frac{1}{\mathrm{ID}_{50}}\right) = 0.670 - 0.132V_w + 0.490\log P \qquad (18)$$

- Substitute V_w and $\log P$ into the regression equation
  - Obtain predicted values of $\log\left(\frac{1}{ID_{50}}\right)$

Equation

$$\log\left(\frac{1}{\mathrm{ID}_{50}}\right) = 0.670 - 0.132V_w + 0.490\log P \qquad (18)$$

- Substitute V_w and $\log P$ into the regression equation
  - Obtain predicted values of $\log\left(\frac{1}{ID_{50}}\right)$
- Standard error of the coefficients and the Intercept

Equation

$$\log\left(\frac{1}{\mathrm{ID}_{50}}\right) = 0.670 - 0.132V_w + 0.490\log P \qquad (18)$$

- Substitute V_w and $\log P$ into the regression equation
  - Obtain predicted values of $\log\left(\frac{1}{ID_{50}}\right)$
- Standard error of the coefficients and the Intercept
- F Value

Rank Correlation (Spearman)

- Used when data are **ranked (ordinal)**, not continuous
- Does not assume linearity or equal spacing in the original data
- Measures association between two ranked variables using r_s

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \tag{19}$$

- $d =$ difference between the ranks for each observation
- $n =$ number of observations (e.g., number of medicines)

Rank Correlation (Spearman)

- Used when data are **ranked (ordinal)**, not continuous
- Does not assume linearity or equal spacing in the original data
- Measures association between two ranked variables using r_s

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \tag{19}$$

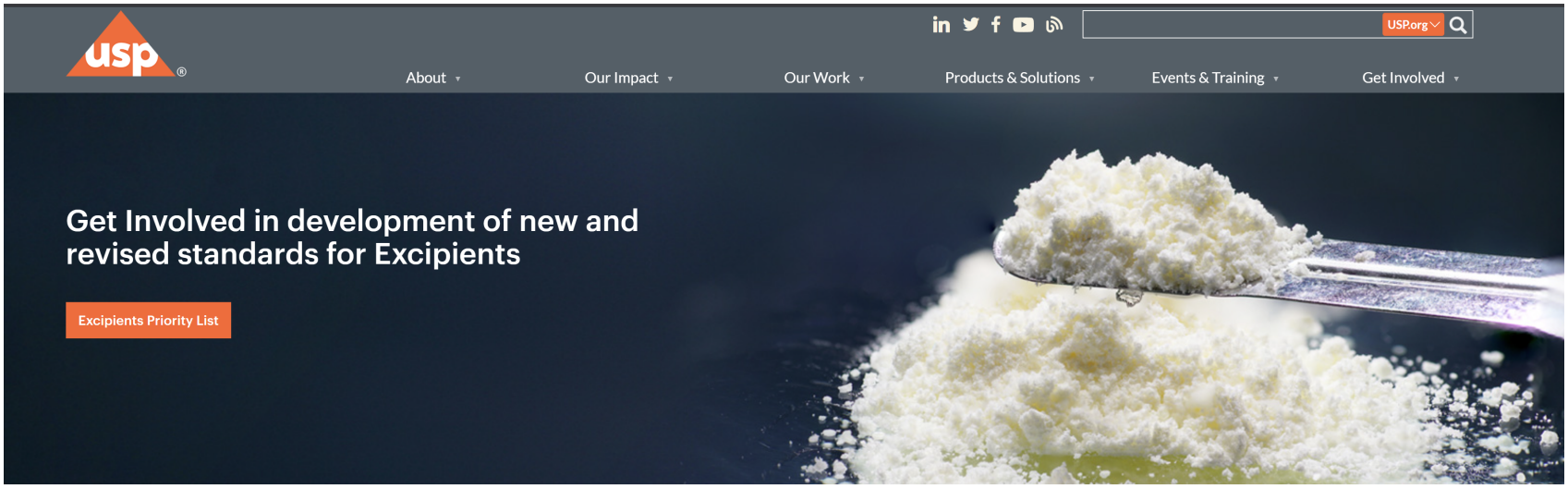
- $d =$ difference between the ranks for each observation
- $n =$ number of observations (e.g., number of medicines)

- For larger samples ($n > 10$), r_s can be tested using Student's t

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \tag{20}$$

Multivariate Methods

Excipients



USP / Our Work

Excipients

Because they comprise up to 90% of medications, the quality of inactive ingredients (excipients) is critical for a drug to be safe and effective. Our documentary standards provide the appropriate, validated test procedures to establish the identity, purity and quality of excipients, while our reference standards are authentic specimens that have been approved as suitable for use as comparison standards in USP or NF tests and assays. At every step, we're protecting the public's health by helping to prevent poor¬quality medication from entering the marketplace.

Figure 2: The United States Pharmacopeia (USP)

Distance Matrices

- Many multivariate methods consist of measuring "distances," either between observations or populations.
- In distance matrices, distances between individual observations are determined.
- Materials from all sources must meet analytical specifications, but it is obviously desirable that the ingredients from the alternative sources resemble the original material as closely as possible.

Distance Matrices

- The technique is to compare the three samples in two-dimensional space.
- The distance between samples (the Euclidean distance) is calculated by Pythagoras' theorem.

Distance Matrices

Sample	Acid Value	Iodine Value
A	0.1	79
B	0.5	82
C	0.2	88
Mean	0.267	83.000
Standard deviation	0.208	4.583

$$(AB)^2 = (\text{acid value difference})^2 + (\text{iodine value difference})^2 \quad (21)$$

Distance Illustration

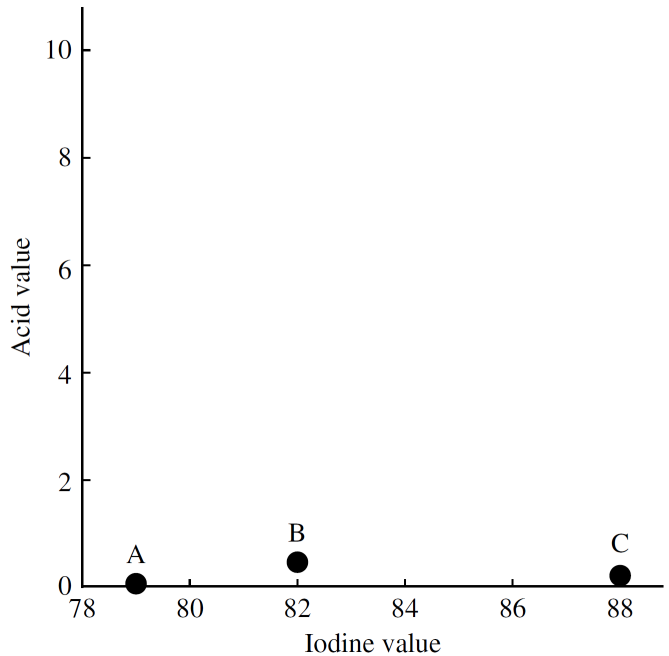


Figure 3: Graphical representation of analytical data from three samples of olive oil, using raw data

Standardization & Distance

Standardization:

$$z = \frac{x - \bar{x}}{s}$$

- Centers data at 0 and scales variance to 1
- Ensures all variables contribute equally

Example (Sample A, Acid value):

$$\frac{0.1 - 0.267}{0.208} = -0.803$$

Euclidean Distance (standardized space):

$$AB = 2.031 \qquad AC = 2.021$$

Conclusion:

$$AC < AB \Rightarrow \text{Sample C is closer to A}$$

Distance Illustration

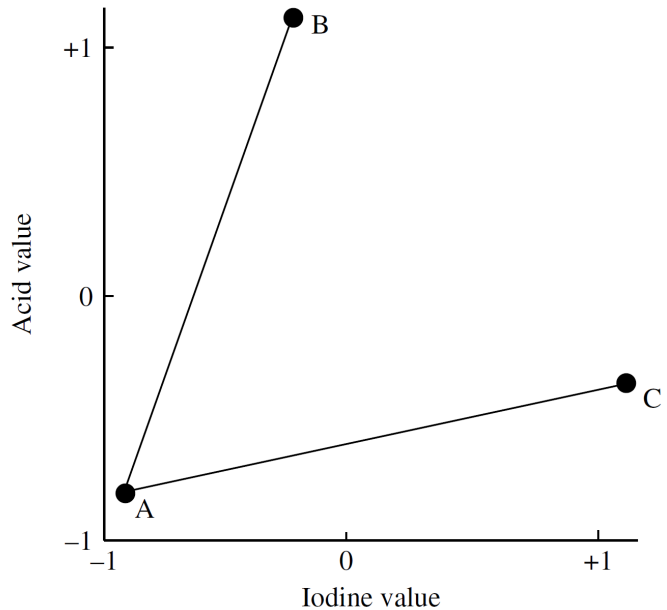


Figure 4: Graphical representation of analytical data from three samples of olive oil, using standardized data

Principal Component Analysis (PCA)

- Its objective is to reduce the number of variables needed to describe the data

Principal Component Analysis (PCA)

- Its objective is to reduce the number of variables needed to describe the data
- This is done using a small number of **linear combinations** (principal components)

Principal Component Analysis (PCA)

- Its objective is to reduce the number of variables needed to describe the data
- This is done using a small number of **linear combinations** (principal components)
- If there are n variables $(X_1, \ldots, X_n)$, there are n components $(Z_1, \ldots, Z_n)$

Principal Component Analysis (PCA)

- Its objective is to reduce the number of variables needed to describe the data
- This is done using a small number of **linear combinations** (principal components)
- If there are n variables $(X_1, \ldots, X_n)$, there are n components $(Z_1, \ldots, Z_n)$
- The principal components are **uncorrelated**

Principal Component Analysis (PCA)

- Its objective is to reduce the number of variables needed to describe the data
- This is done using a small number of **linear combinations** (principal components)
- If there are n variables $(X_1, \ldots, X_n)$, there are n components $(Z_1, \ldots, Z_n)$
- The principal components are **uncorrelated**
- **Why this matters:**
  - No correlation $\Rightarrow$ different aspects of the data
  - Correlation $\Rightarrow$ repeated information

PCA in R

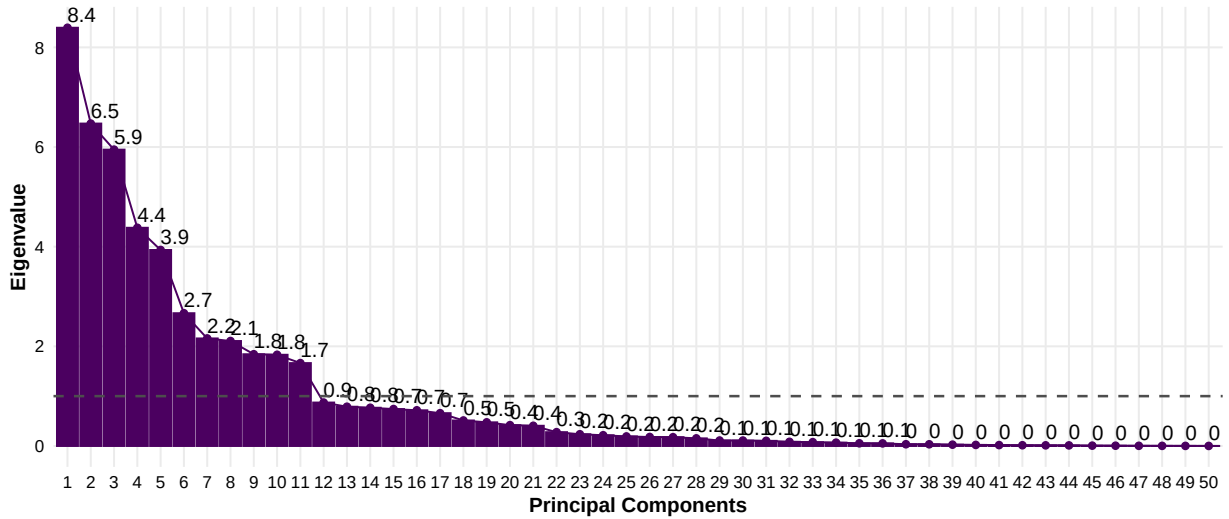
```r
# Load data
pca_df <- read.csv2()

# Keep only numeric predictor columns
# Here, y is excluded because PCA is unsupervised
x_data <- pca_df[, names(pca_df) != "y"]

# Run PCA
pca_model <- prcomp(
  x_data,
  center = TRUE,
  scale. = TRUE
)

# Check summary
summary(pca_model)
```

PCA Scree Plot



Total principal components: 50

Important principal components: 11 retained using the Kaiser criterion ($\lambda > 1$).

Factorial Design

Factorial Design

- Provides a means whereby the factors that may have an influence on a reaction or a process can be evaluated simultaneously and their relative importance assessed.

Factorial Design

- Provides a means whereby the factors that may have an influence on a reaction or a process can be evaluated simultaneously and their relative importance assessed.
  - The factors to be studied: Factors can be quantitative or they can be qualitative.
  - The levels of the factors: A commonly used starting point is to select the 25th and 75th percentile.
  - The response to be measured:usually defined in the experimental objectives. The response must be capable of being expressed numerically. Adjectival descriptions (big, bigger, and biggest) or ordinal numerals (designating the biggest response as 1, the next biggest as 2, and so on) are not permissible.

Factorial Design

- Simplest factorial design is one in which two factors are studied at two levels.

Factorial Design

- Simplest factorial design is one in which two factors are studied at two levels.
- The design consists of four experiments:
  - Experiment A: Both factors are at their lower levels.
  - Experiment B: The first factor is at its higher level and the second at its lower.
  - Experiment C: The first factor is at its lower level and the second factor at its higher.
  - Experiment D: Both factors are at their higher levels.

Two-Factor, Two-Level Factorial Design

Experiment	**Temperature (°C)**	**Catalyst (M)**	**Loss of E (%)**
A	20	0	10
B	40	0	25
C	20	0.1	30
D	40	0.1	45

Two-Factor, Two-Level Factorial Design

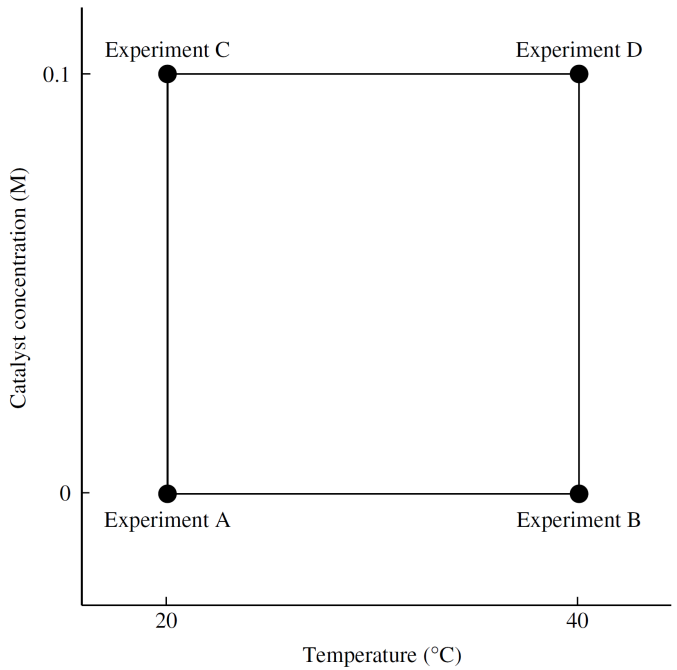


Figure 5: A two-factor, two-level experimental design.

Factorial Effects

$$\frac{1}{2}\left[(B+D)-(A+C)\right] \tag{22}$$

$$=\frac{1}{2}\left[(25+45)-(10+30)\right]=15 \tag{23}$$

Effect of catalyst concentration:

$$\frac{1}{2}\left[(C+D)-(A+B)\right] \tag{24}$$

$$=\frac{1}{2}\left[(30+45)-(10+25)\right]=20 \tag{25}$$

Interaction Between Factors

- So far we assumed that the factors act independently to produce their effects.

Interaction Between Factors

- So far we assumed that the factors act independently to produce their effects.
  - In some cases the level of one factor may govern the magnitude of the effect of another.
  - **This is termed factor interaction.**

Notation In Factorially Designed Experiments

- In factorial designs, raw factor values can differ greatly in scale (e.g., iodine values ~ 80 vs acid values < 1)
- This makes direct comparison and analysis difficult
- **Solution: Coding (Experimental Units, e.u.)**
  - Subtract the mean
  - Divide by the standard deviation
- **Key idea:** Coding brings all factors into the same range, so each has equal importance
- For two-level designs:
  - Lower level $\rightarrow -1$
  - Upper level $\rightarrow +1$
- Enables:
  - Fair comparison
  - Clear interpretation
  - Interaction terms

Factorial Design with Coded Values

Exp	X_1 (Temp)	X_2 (Cat.)	X_1X_2	**Loss (%)**
$(-1,-1)$	-1	-1	$+1$	10
$(+1,-1)$	$+1$	-1	-1	25
$(-1,+1)$	-1	$+1$	-1	30
$(+1,+1)$	$+1$	$+1$	$+1$	45

Note: Center $(0,0) \rightarrow 30^\circ$C, 0.05 M.

1 e.u. Temp $= 10^\circ$C $\quad|\quad$ 1 e.u. Cat. $= 0.05$ M

Three-Level Factorial Design

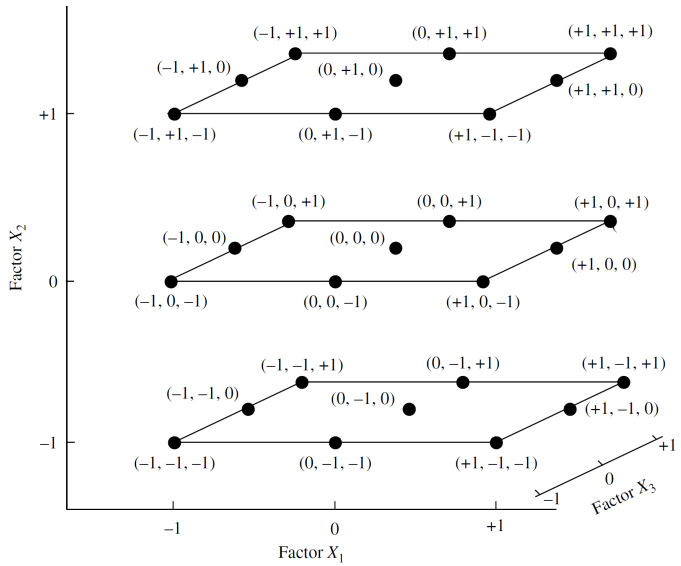


Figure 6: A three-factor, three-level experimental design.

$$L^F \tag{26}$$

Fractional Factorial Design

- As the number of factors in a design increases, the number of experiments needed to form a complete design can rapidly outgrow the resources available to the experimenter.
- Fractional designs are extremely useful in screening experiments where many factors are considered.
- Those factors that have large effects can be identified, and these can be more thoroughly investigated.

Artificial Neural Networks

Artificial Neural Networks

- Defined by Aleksander and Morton as "the study of networks of adaptable nodes which, through a process of learning from task examples, store experiential knowledge and make it available for use."
- The underlying function of a neural network is to identify patterns, that is, when presented with an input pattern, it produces an output pattern.

Artificial Neural Networks

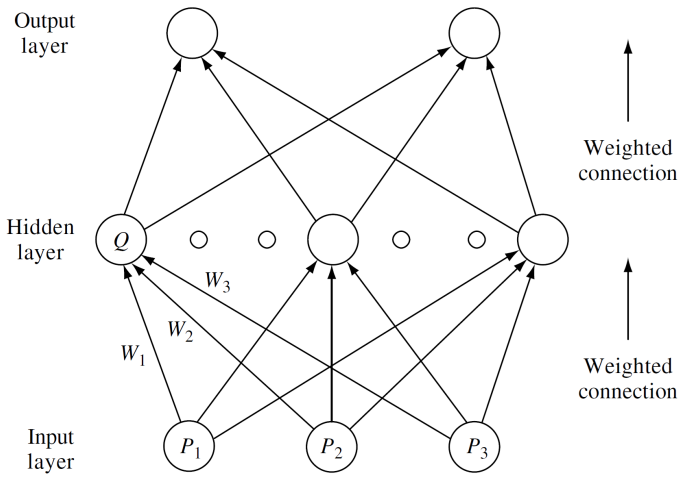


Figure 7: A fully interconnected artificial neural network with an input layer, one hidden layer and an output layer.

Neural Unit Computation

Weighted Sum (Input to Unit):

$$S_Q = O_1 W_1 + O_2 W_2 + O_3 W_3 \tag{27}$$

Sigmoid Activation Function:

$$F(S_Q) = \frac{1}{1 + e^{-S_Q}} \tag{28}$$

Min-Max Scaling (Normalization):

$$\text{converted value} = \frac{\text{value} - L}{U - L} \tag{29}$$

Thank You

- Merci!Thank You! ¡Gracias!
- Questions?

F Critical Values for ANOVA ($\alpha = 0.05$)

$df_2 \backslash df_1$	1	2	3	4
20	4.35	3.49	3.10	2.87
24	4.26	3.40	3.01	2.78
27	4.21	**3.35**	2.96	2.73
30	4.17	3.32	2.92	2.69
40	4.08	3.23	2.84	2.61

- Numerator df: $df_1 = 2$
- Denominator df: $df_2 = 27$
- Critical value: $F_{0.05,(2,27)} \approx 3.35$
- $F_{\text{calculated}} = 5.28 > 3.35$
- **Conclusion: reject** H_0

◂ Back to ANOVA results

t-Distribution Critical Values

Back to t-test

Table 6: t-critical values (two-tailed)

df	alpha = 0.10	alpha = 0.05	alpha = 0.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	**1.812**	**2.228**	**3.169**
12	1.782	2.179	3.055
15	1.753	2.131	2.947
20	1.725	2.086	2.845
30	1.697	2.042	2.750

F Distribution Table ($\alpha = 0.01$)

$d_1 \backslash d_2$	1	2	3	4
1	4052.2	499.5	**34.1**	21.2
2	98.5	99.0	30.8	18.0
3	34.1	30.8	29.5	16.7

- Highlighted value: $F_{1,3}^{0.01} = 34.1$
- Used in the regression test on the previous slide

Back to F-test

Critical Value Table

Table 7: Critical Values of the Correlation Coefficient ($\alpha = 0.05$)

n	**df** $(n-2)$	**Critical** r
3	1	0.997
4	2	0.950
5	3	0.878
6	4	0.811
7	5	0.754
8	6	0.707
9	7	0.666
10	8	0.632
15	13	0.514
20	18	0.444
30	28	0.361
50	48	0.279