11 The Design of Experiments

by **Patrick Boily** (inspired by **David Haziza**)

In the world of data-driven decision-making, it is not enough to simply possess vast datasets (which are often passively collected) and analytical prowess.

The manner in which experiments are **designed**, **conducted**, and **analyzed** can make a huge difference in the validity and reliability of the conclusions that analysts draw.

The design of experiment provides the foundation for sound experimental methodology, enabling scientists and data professionals to meticulously control variables, uncover hidden patterns, and discern causality amidst the complexity of real-world data.*

11.1 Basic Notions

At its core, statistics serves as the science of **collecting**, **analyzing**, and **deriving meaningful conclusions** from data.

Data can be obtained through several primary methods, each with its own unique characteristics.

One common approach to data collection involves conducting **sample surveys**. These surveys are often carried out by entities such as National Statistical Offices and polling market firms.†

The main objective of sample surveys is typically to estimate parameters for finite populations. For instance, they may aim to determine the average income within the Canadian population or calculate the unemployment rate.‡

Another method involves **observational studies**, where researchers gather data by observing and recording natural occurrences. These studies provide valuable insights into real-world phenomena but may not always allow for the establishment of causality between variables.

Experimentation represents a powerful way to investigate causal relationships. In experiments, researchers manipulate one or more variables and observe the effects on others. This controlled approach helps establish potential **causal networks**,[1] a crucial aspect of scientific inquiry.

These foundational concepts lay the groundwork for our exploration of experimental design.

* More details, examples, and exercises are available in [2, 5], among others.
† Such as *Statistics Canada* or *EKOS Research*, say.
‡ Survey sampling is explored in depth in Chapter 10.

1: Or **cause-and-effect** connections.

11.1.1 Experiments

The essence of **experimental studies** lies in the comparison of **treatments** and their respective **outcomes**. Researchers leverage experiments to address crucial questions, often revolving around topics such as:

- Is a drug a safe and effective cure for a disease? This could involve testing how AZT affects the progression of AIDS.
- What combination of protein and carbohydrate sources provides the optimal nutrition for growing lambs?
- How will long-distance telephone usage patterns change if our company introduces a different rate structure for our customers?
- Can an ice cream manufactured with a new kind of stabilizer match the palatability of our current ice cream?

A fundamental aspect of scientific reasoning involves drawing conclusions from experiments that have been meticulously designed, executed appropriately, and rigorously analyzed. Key elements include the **treatments** and **experimental units** to be employed, the **methodology** for assigning treatments to units, and the measured **responses**.

Note that the environment and observation conditions must be carefully **controlled** and **fixed**.[2]

2: Explanatory variables are under the direct control of the researchers; some are intentionally **altered**, while others are held **constant**.

Observational Studies against Experiments

Both observational studies and experiments are typically employed to establish relationships between two or more measured quantities. However, there is a fundamental distinction between observational studies and experiments.

In an observational study, researchers do not actively manipulate or create data; instead, they solely observe the characteristics of pre-existing data. Consequently, an observational study entails the observation of units/individuals and the measurement of variables of interest, **without any attempt to influence their responses**.

Conversely, an experiment involves the **deliberate imposition of specific treatments** on individuals/units to observe their responses. Causal inferences find justification in experiments, where the explanatory variables $x_1, \cdots, x_p$, often referred to as the "possible causes," are directly controlled by the researcher. Such experiments are known as **randomized trials** because the values of the explanatory variables are assigned to experimental units through some random mechanism.

In observational studies, the values of the explanatory variables are **observed** rather than assigned by the researcher, alongside the value of the response. In such studies, causal inferences are **not warranted** because, although efforts can be made to "control" for certain "confounding" factors, it is generally impossible to control for all relevant factors.

What constitutes a **relevant** (or confounding) **factor** in observational studies? It is a factor that both **influences the response variable(s)** and **relates to the explanatory variable(s)** on which the research focuses.

A drawback of observational studies is that the grouping of individuals into "treatments" is **beyond the experimenter's control**, and the mechanism underlying this grouping is often **unknown**.

Consequently, observed differences in responses between treatment groups may be attributable to **hidden mechanisms** rather than to the treatments.

Example Consider a dataset from Canada's *Health Care System* comparing the effectiveness of two procedures for treating prostate disease:

1. traditional surgery, or
2. a new method that does not require surgery.

The dataset includes many patients suffering from prostate disease, with their doctors choosing one of the two methods. Initially, the study found that patients treated with the new method were significantly more likely to die within 8 years. H

However, further data analysis revealed that this conclusion was incorrect. Why? What potential confounding variables might be at play?

Definitions

Some concepts will re-appear time and time again in this chapter, and so we take the time to define them properly.

- **Treatments** represent the different procedures under examination. These could encompass various types or amounts of fertilizer in agronomy or distinct long-distance rate structures in marketing.
- An **experimental unit** refers to the physical entity that can be randomly assigned to a treatment. This unit may be an individual, an animal, a plot of land receiving fertilizer, and so forth, upon which measurements are taken.[3]
- The **dependent** (or response) **variable**, often denoted by Y, represents the observed outcome after applying a treatment to an experimental unit.
- **Randomization** involves the use of a known and perfectly controlled probabilistic mechanism to assign treatments to units.
- A **factor** in an experiment is a controlled independent variable, a variable whose levels are determined by the experimenter. Factors combine to create treatments. For instance, the baking treatment for a cake may involve specific time and temperature settings, with each variable varied independently.
- A **level** denotes the intensity setting (or value) of a factor.
- The **effect** is the change in the response caused by a change in a factor.
- A **lurking** (or hidden) **variable** is an uncontrolled variable that falls outside the experimenter's awareness and control, which could influence the experiment's outcome.
- A **cell** refers to the subset of data occurring at the intersection of one level of every treatment.

3: It does not have to be a "physical" entity *per se*, as the data may arise in a simulation context (see Chapter 12).

Example In each of five different campuses across the country, we selected 10 students at random to assess their attitudes toward industrial pollution. Each student responded to a specific set of questions, and their responses were aggregated into a total interview score.

Campus	**I**	**II**	**III**	**IV**	**V**
Score	172	248	236	250	241

- Experimental unit: a student
- Response variable: total aggregated score
- Factor: campus, with 5 levels
- There are 5 cells in this experiment

Example We would like to compare the effects of three different insecticides on a particular variety of string beans. Four plots were prepared, with each plot subdivided into three rows. Each row was planted with 100 seeds and then maintained under the insecticide assigned to the row. The insecticides were randomly assigned to the rows within a plot so that each insecticide appeared in one row in all four plots. The response variable was the number of seedlings that emerged per row.

	Plot			
Row	**I**	**II**	**III**	**IV**
1	(A) 121	(A) 73	(B) 144	(B) 134
2	(B) 128	(B) 141	(C) 118	(A) 85
3	(C) 112	(C) 118	(A) 109	(C) 111

Of course, we do not need to physically refer to the rows in order; in fact, it might make more sense to represent the experiment using the treatments instead of the location.

	Plot			
Insecticide	**I**	**II**	**III**	**IV**
A	(1) 121	(1) 73	(3) 109	(2) 85
B	(2) 128	(2) 141	(1) 144	(1) 134
C	(3) 112	(3) 118	(2) 118	(3) 111

- Experimental unit: variety of string beans
- Response variable: number of seedlings
- Factors: plot and insecticide
- Levels of the factors:
  - Plot: four levels (I, II, III, IV)
  - Insecticide: three levels (A, B, C)
- There are $3 \cdot 4 = 12$ cells in this experiment

Example We aim to test whether a chemical agent can prevent symptomatic infection from a respiratory diseases. A clinical trial was conducted where patients received either the compound (C) or a placebo (P). The treatment was administered to both men (M) and women (F), each belonging to a specific age group. The information is summarized below.

	Gender			
	M		F	
	Drug			
Age	P	C	P	C
29−	(102) 0.31	(99) 0.29	(105) 0.28	(105) 0.30
30-59	(117) 0.35	(119) 0.31	(120) 0.31	(119) 0.27
60+	(89) 0.38	(85) 0.41	(91) 0.38	(90) 0.37

- Experimental unit: individual on which the infected/non-infected status is measured
- Response variable: 1 = infection, 0 = no infection.
- Factors: gender, drug, and age group
- Levels of the factors:
  - Drug: two levels (compound and placebo)
  - Gender: two levels (male and female)
  - Age group: three levels (29−, 30-60, 60+)
- There are $2 \cdot 2 \cdot 3 = 12$ cells in this experiment

11.1.2 Useful Distributions

We have encountered several probabilistic and statistical concepts that arise time and time again in applications.[4] We briefly mention those properties that will be useful in the analysis and design of experiments.

4: See Chapters 6, 7, 8, 9, and 10.

Sample Mean and Sample Variance Consider a **random sample**

$$\mathcal{Y} = \{y_1, \ldots, y_n\}$$

drawn from a population with mean μ and variance σ^2, where $\mathrm{E}(y_i) = \mu$ and $\mathrm{Var}(y_i) = \sigma^2$ for $i = 1, \ldots, n$.

We assume that the sample observations in $\mathcal{Y}$ are **independent and identically distributed** (i.i.d), indicating that they were generated from the same distribution (or from the same population $\mathcal{U}$).

The **sample mean** and **sample variance** of $\mathcal{Y}$ are given by:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y})^2.$$

As a reminder, both the sample mean and the sample variance are **unbiased estimators** of the population mean and the population variance, respectively:

$$\mathrm{E}(\overline{y}) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(y_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu,$$
$$\mathrm{Var}(\overline{y}) = \frac{1}{n^2}\sum_{i=1}^{N}\mathrm{Var}(y_i) = \frac{1}{n^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{n},$$

and

$$\begin{aligned}\mathrm{E}(s^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{N}\mathrm{E}(y_i^2) - n\mathrm{E}(\overline{y}^2)\right] = \frac{1}{n-1}\left[\sum_{i=1}^{N}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\ &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] = \sigma^2.\end{aligned}$$

Probability Distributions The distribution of sample observations is described by a probability distribution. For a continuous variable Y, the probability distribution is characterized by a **density function**, denoted as $f(y)$, with the following properties:

$$f(y) \geq 0, \quad P(a \leq Y \leq b) = \int_a^b f(y)dy, \quad \int_{-\infty}^{+\infty} f(y)dy = 1.$$

The mean of a probability distribution, denoted by μ, serves as a measure of **centrality location** and is defined as:

$$\mu = \mathrm{E}(Y) = \int_{-\infty}^{+\infty} yf(y)dy.$$

The variance σ^2 can be used to quantify the **dispersion** of a variable:

$$\sigma^2 = \mathrm{Var}(Y) = \mathrm{E}[(y-\mu)^2] = \int_{-\infty}^{+\infty}(y-\mu)^2 f(y)dy.$$

Normal Distributions If Y follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2, its probability density function is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad -\infty < y < \infty$$

.

If $y_1, \ldots, y_n$, is a random sample generated from a $\mathcal{N}(\mu, \sigma^2)$, then $\overline{y}$ and s^2 are **statistically independent**.

Normal distributions are entirely characterized by their expectation $\mathrm{E}(Y) = \mu$ and variance $\mathrm{Var}(y) = \sigma^2$; any other normal random variable with the same properties must in fact be exactly $Y \sim \mathcal{N}(\mu, \sigma^2)$. We can **standardize** any such random variable:

$$Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

The resulting random variable Z is said to be **standard normal**.

We have discussed normal distributions in detail in Section 6.3.3; the primacy of normal distributions in statistical applications is explained by the following oft-used result.

Central Limit Theorem: let $Y_1, \ldots, Y_n$, be n i.i.d. random variables with mean μ and variance σ^2. The random variable

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to $Z \sim \mathcal{N}(0,1)$, where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$.[5]

5: A sequence $\{X_n\}$ of random variables, with cumulative distribution functions $\{F_1\}$ converges in distribution to a random variable X with cumulative distribution function F if $F_n(x) \to F(x)$ for all x where F is continuous.

χ^2 Distributions If $Z_1, \ldots, Z_k \sim \mathcal{N}(0,1)$ are k i.i.d. random variables, then the random variable

$$Y = Z_1^2 + \cdots + Z_k^2$$

follows a χ^2_k distribution (with k degrees of freedom).[6]

6: We have also used the notation $\chi^2(k)$ in these notes.

The probability density function of such a random variable is

$$f(y) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} y^{k/2-1} e^{-y/2}, \quad y > 0,$$

where Γ is the Gamma function.

When $Y \sim \chi^2_k$, we have $\mathrm{E}(Y) = k$ and $\mathrm{Var}(Y) = 2k$.

As the degrees of freedom parameter k increases, the chi-square distribution converges in distribution to a normal distribution with a mean equal to k and a variance equal to $2k$. This convergence is a direct consequence of the Central Limit Theorem.

Now, if we have a random sample $y_1, \ldots, y_n$ generated from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, we can make the following observation:

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}.$$

This implies that we can obtain an unbiased estimator of σ^2 by dividing the sum of squares by the number of degrees of freedom, which is $n-1$. This unbiased estimator of the population variance will prove useful when introduce **ANOVA tables**.

Student's $T-$Distributions If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$ independent, then the distribution of the random variable

$$W = \frac{Z}{\sqrt{Y/k}}$$

is that of a Student $T-$distribution with k degrees of freedom, denoted by $W \sim t_k$.[7]

7: We have also used the notation $t(k)$ in these notes.

The probability density function of the $T-$distribution is :

$$f(w) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)\left(1+\frac{w^2}{k}\right)^{\frac{k+1}{2}}}, \quad -\infty < w < \infty.$$

The $T-$distribution is symmetric, and its expected value is $\mathrm{E}(W) = 0$, while its variance is $\mathrm{Var}(W) = \frac{k}{k-2}$ for $k > 2$. As the degrees of freedom parameter k increases, W converges in distribution to the standard normal distribution $\mathcal{N}(0,1)$.

Fisher's $F-$Distributions If $X \sim \chi^2_u$ and $Y \sim \chi^2_v$ are independent, then the distribution of the random variable

$$W = \frac{\frac{X}{u}}{\frac{Y}{v}}$$

is that of a Ficher $F-$distribution with (u, v) degrees of freedom, denoted by $W \sim F_{u,v}$.[8]

8: We have also used the notation $F(u,v)$ in these notes. The order of the degrees of freedom is important: if $W \sim F_{u,v}$, then $\frac{1}{W} \sim F_{v,u}$.

The probability density function of the $F-$distribution is given by:

$$f(w) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \frac{w^{\frac{u}{2}-1}}{\left(1+\frac{u}{v}w\right)^{\frac{u+v}{2}}}, \quad w > 0.$$

The expectation of $W \sim F_{u,v}$ is only defined if $v > 2$; its variance is only defined if $v > 4$. In those cases, we have

$$\mathrm{E}(W) = \frac{u}{v-2} \quad \text{and} \quad \mathrm{Var}(W) = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)};$$

if $v \leq 4$, W does not have a well-defined variance, if $v \leq 2$, it does not have a well-defined expectation. Moreover, if $X \sim t(k)$, then $X^2 \sim F_{1,k}$.

11.2 Review of Hypothesis Testing

We have discussed hypothesis testing in detail in Section 7.4 (and in the chapters on applications); we briefly review its important features as it relates to the design of experiment.

11.2.1 Inference on the Population Mean

The customary Student $T-$test relies on several key assumptions:

1. a random sample of size n is selected for analysis;
2. the individual observations in this sample are denoted by $y_1, y_2, \ldots, y_n$;
3. these observations are assumed to have been generated from a normal population with a mean parameter μ and variance σ^2, expressed as:
$$y_1, y_2, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2).$$

However, what if the underlying population does not follow a normal distribution? The Student $t-$test exhibits robustness in the sense that the distribution of the test statistic remains relatively stable even when the normality assumption is not strictly met. This robustness holds, provided that the sampled population exhibits an **approximately mound-shaped** distribution.

In the context of hypothesis testing: we typically formulate both **null** and **alternative hypotheses** as follows. We pit the

null hypothesis (H_0): $\mu = \mu_0$

against the **two-tailed**

alternative hypothesis (H_1): $\mu \neq \mu_0$,

or either of the **one-tailed**

alternative hypothesis (H_1): $\mu > \mu_0$ (one-tailed test), or
alternative hypothesis (H_1): $\mu < \mu_0$ (one-tailed test).

We define the following terms related to hypothesis testing (see Table 11.5 for a summary):

- a **type I error** occurs when we wrongly reject the null hypothesis H_0 but it is in fact valid:

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true});$$

- a **type II error** occurs when we do not reject the null hypothesis H_0 but it should in fact be rejected:

$$\beta = P(\text{Type II error}) = P(\text{do not reject } H_0 \mid H_0 \text{ is false});$$

- the **power of the test** is the probability of correctly rejecting the null hypothesis when it is in fact false:

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

We discuss other types of error in one of the sidenotes of Section 7.4.1.

		Reality	
		H_0 **is true**	H_0 **is false**
Decision	**Reject** H_0	type I error (α)	right decision ($1-\beta$)
	Do not reject H_0	right decision ($1-\alpha$)	type II error (β)

Table 11.5: The four possible outcomes for hypothesis testing.

We usually set the **significance level** α of the test, typically chosen as $\alpha = 0.01, 0.05, 0.1$, and aim to construct a test with **high power** $1-\beta$, typically for $\beta = 0.1, 0.2$.

The **test statistic** t_0 is calculated as follows:

$$t_0 = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}.$$

If H_0 is true,[9] the distribution of t_0 follows a $T-$distribution with $n-1$ degrees of freedom (t_{n-1}).

9: We also say "under H_0".

For a two-tailed test at the level α, we **reject** H_0 when $|t_0|$ is greater than the **critical value** $t_{\alpha/2;n-1}$.[10] For a one-tailed test, either $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ or $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, we reject H_0 based on the sign of t_0:

10: See Section 7.3.2 for more information.

- for $H_1 : \mu > \mu_0$, we reject H_0 when $t_0 > t_{\alpha;n-1}$;
- for $H_1 : \mu < \mu_0$, we reject H_0 when $t_0 < -t_{\alpha;n-1}$.

We can then build an $100(1-\alpha)\%$ **confidence interval** for μ according to:

$$\overline{y} \pm t_{\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}$$

The **margin of error** m (sometimes known as the **bound on the error of estimation**, see Chapter 10) is

$$m = t_{\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}$$

We reject H_0 if $\left|\frac{\overline{y}-\mu_0}{s/\sqrt{n}}\right| > t_{\alpha/2;n-1}$, that is, if $\overline{y}$ lies in the **rejection region**

$$\overline{y} \geq \mu_0 + m \quad \text{or} \quad \overline{y} \leq \mu_0 - m.$$

Inference about μ: Power The power of a test depends on various factors, including the **specific alternative hypothesis**, the **significance level** α, the **variance** σ^2, and the **sample size** n.

We can think of the power as a function

$$\pi(\theta) = P(\text{reject } H_0 : \theta = \theta_0 \mid \text{observed sample}).$$

The **power function** $\pi(\theta)$ obviously depends on the **true value** of the parameter θ, of course, but may also be influenced by the **sample size** and the **rejection rule** or **significance level** of the test. By construction, we must have $\pi(\theta_0) = \alpha$.

We can compute the power of the Student $T-$test with the help of the following random variable: if $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$ are independent, the distribution of

$$W = \frac{Z+\delta}{\sqrt{X/k}}$$

is a **non-central $T-$distribution with k degrees of freedom and non-centrality parameter** δ, denoted by $W \sim t_k(\delta)$.[11]

11: When $\delta = 0$, this clearly reduces to the standard Student $T-$distribution.

We take a detailed look at computing the power of the test for a one-tailed test with hypotheses $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

In this case, we reject H_0 if $t_0 > t_{\alpha;n-1}$, which is equivalent to

$$\frac{\overline{y}-\mu_0}{s/\sqrt{n}} > t_{\alpha;n-1}.$$

The power function of the test can then be expressed as:

$$\pi(\mu) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} > t_{\alpha;n-1} \,\middle|\, H_0 \text{ is false}\right) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} > t_{\alpha;n-1} \,\middle|\, \mu > \mu_0\right).$$

To compute this probability, we first note that

$$\frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{(\overline{y} - \mu) + (\mu - \mu_0)}{s/\sqrt{n}} = \frac{\sqrt{n}(\overline{y} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma}{s/\sigma}.$$

According to the central limit theorem, $Z = \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

Furthermore, $X = (n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. If $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$,[12] then

12: In practice, we use $\delta \approx \sqrt{n}(\mu - \mu_0)/s$.

$$\frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{Z + \delta}{X/(n-1)}.$$

Under H_1, then, we have:

$$W = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}(\sqrt{n}(\mu - \mu_0)/\sigma).$$

Example Let $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d., with $s = 10$. We want to test $H_0 : \mu = 60$ against $H_1 : \mu > 60$; assume that we reject H_0 if $\overline{y} \geq 62$.

- What is the power of the test when $n = 25$ and the true value of the mean is $\mu = 63$?

  In this case, we have $\delta \approx \sqrt{25}(63 - 60)/10 = 1.5$ and

  $$\pi(63) = P(\overline{y} \geq 62 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{25}} \,\middle|\, \mu = 63\right) = P(t_{24}(1.5) \geq 1) = 0.6933.$$

  We can compute this in R as follows:

  ```
  1 - pt(q=1, df=24, ncp=1.5)
  ```

  Thus, if $\mu = 63$, the probability of correctly rejecting H_0 is $\approx 70\%$.
- Repeat the calculation, but assuming that $n = 100$ instead. In this case, we have $\delta \approx \sqrt{100}(63 - 60)/10 = 3$ and

  $$\pi(63) = P(\text{reject } H_0 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{100}} \,\middle|\, \mu = 63\right) = P(t_{99}(3) \geq 2) = 0.8401,$$

  which can also be obtained in R as follows:

  ```
  1 - pt(q=2, df=99, ncp=3)
  ```

  We note that, for given values of μ and s, the power of the test increases as the sample size n increases.
- For an arbitrary n, we have $\delta \approx \sqrt{n}(63 - 60)/10 = 0.3\sqrt{n}$, and

  $$\begin{aligned}\pi(63) &= P(\text{reject } H_0 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{n}} \,\middle|\, \mu = 63\right) \\ &= P(t_{n-1}(0.3\sqrt{n}) \geq 0.2\sqrt{n}).\end{aligned}$$

- If the true parameter value is $\mu = 60$, then for an arbitrary sample size n, we have $\delta = \sqrt{n}(60 - 60)/10 = 0$ and

$$\pi(60) = P(\text{reject } H_0 \mid \mu = 60) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{n}} \middle| \mu = 60\right)$$
$$= P(t_{n-1} \geq 0.2\sqrt{n}) = \alpha.$$

  Note that $\pi(60)$ corresponds to the probability of a Type I error for a given decision rule and sample size. □

In general, the power of a test increases as:

- the effect $|\mu - \mu_0|$ increases for fixed values of n and s;
- the sample size increases for fixed values of μ and s;
- s decreases for fixed values of μ and n.

Sample Size When designing an experiment, it is crucial to determine an appropriate sample size. Typically, researchers aim to determine the sample size n that guarantees a high statistical power.[13] To achieve this, the they need to specify the following **key factors**.

13: Often set at $1 - \beta = 0.8$ or 0.9.

1. The desired **power**, which represents the probability of detecting a true effect if it exists;
2. the **significance level** α (the probability of making a Type I error);
3. the **effect size** $|\mu - \mu_0|$, which is chosen to represent a practically meaningful difference between groups or conditions, and
4. an estimate or range for the **population variance** σ^2.

We illustrate the process *via* a simple example.

Example Let $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d. We wish to test the following hypotheses:

$$H_0 : \mu = 100 \quad \text{against} \quad H_1 : \mu > 100.$$

We assume that 20 a plausible value for σ, and that the level of significance α is 0.05. If an effect $\mu - \mu_0 = 10$ is considered meaningful, what sample size is required to detect such a difference with a power of 0.9?

Given our assumption about σ^2, the distribution of the test statistic

$$Z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}$$

is standard normal, $\mathcal{N}(0, 1)$.

In order for $\mu - \mu_0$ to be 10, we must have $\mu = 110$; we can achieve a power of 0.9 as follows:

$$\pi(110) = P(\text{reject } H_0 \mid \mu = 110) = 0.9 \Leftrightarrow P\left(Z \geq z_{0.05} \mid \mu = 110\right) = 0.9$$
$$\Leftrightarrow P\left(\frac{\overline{y} - 100}{20/\sqrt{n}} \geq 1.645 \middle| \mu = 110\right) = 0.9 \Leftrightarrow P\left(\overline{y} \geq 1.645 \cdot \frac{20}{\sqrt{n}} + 100 \middle| \mu = 110\right) = 0$$
$$\Leftrightarrow P\left(\frac{\overline{y} - 110}{20/\sqrt{n}} \geq 1.645 - \frac{10\sqrt{n}}{20}\right) = 0.9.$$

What is the corresponding quantile of the standard normal distribution?

```
qnorm(p=0.9, mean=0, sd=1, lower.tail=FALSE)
```

```
[1] -1.281552
```

Then, we must have

$$1.645 - \frac{10\sqrt{n}}{20} = -1.29,$$

which is to say, $n \approx 35$. □

11.2.2 Inference on the Difference of Means

We start with an example borrowed from [4].

Motivational Example An experiment was conducted to compare the mean number of tapeworms in the stomachs of sheep that had been treated for worms against the mean number in those that were untreated.

A sample of 14 worms-infected lambs was randomly divided into two groups: 7 were injected with the drug and the remainder were left untreated. After a 6-month period, the lambs were slaughtered and the following worm counts were recorded.

Drug-treated sheep	18	43	28	50	16	32	13
Untreated sheep	40	54	26	63	21	37	39

How would we test the hypothesis that there is no difference in the mean number of worms between treated and untreated lambs? □

We will return to this example after some important notions.

To test for the **difference of means**, we assume two populations, denoted by I and II, in each of which the distribution of the response variable is taken to be normal.[14]

For Population 1, let μ_1 and σ_1^2 be the respective **population mean** and **variance**, and analogously, for Population II, μ_2, and σ_2^2.[15] A key assumption is that the population variances are equal: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Let $y_{1,1}, \ldots, y_{1,n_1}$ be a random sample of size n_1 drawn from Population I, with sample mean $\overline{y}_1$, and $y_{2,1}, \ldots, y_{2,n_2}$ be a random sample of size n_2 drawn from Population II, with sample mean $\overline{y}_2$. Crucially, these samples are assumed to be **independent**.

Expressed in distributional terms:

$$y_{1,1}, \ldots, y_{1,n_1} \sim \mathcal{N}(\mu_1, \sigma^2), \qquad y_{2,1}, \ldots, y_{2,n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$$

or equivalently:

$$\begin{aligned} y_{1,i} &= \mu_1 + \varepsilon_{1,i}, \quad \varepsilon_{1,i} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n_1, \quad \text{and} \\ y_{2,i} &= \mu_2 + \varepsilon_{2,i}, \quad \varepsilon_{2,i} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n_2. \end{aligned}$$

14: Note that in the motivational example, the response is the worm count, which cannot be normally distributed as negative and fractional values cannot arise. Nevertheless, that assumption may be a good approximation to reality (see Section 6.3.6, for instance).

15: Referring to the motivational example, μ_1 and μ_2 are the true worm count means in the populations of treated and untreated lambs, respectively.

The test's **null** and **alternative** hypotheses are:

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2;$$

the **treatment effect** is denoted by $\mu_1 - \mu_2$.[16]

We require a **test statistic** to determine whether to reject or accept the null hypothesis, H_0. Setting the level of the test as α,[17] we aim to formulate a test with a substantial power.

The customary $T-$statistic with significance level α is

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the **pooled estimate of the common variance** σ^2.

If the null hypothesis H_0 holds true, the test statistics t_0 follows a $T-$distribution with $n_1 + n_2 - 2$ degrees of freedom, $t_0 \sim t_{n_1+n_2-2}$. The decision to **reject the null hypothesis** at level α is made when

$$|t_0| > t_{\alpha/2;n_1+n_2-2}.$$

In practice, the decision often hinges on the p-value. The computation for the p-value (in the two-tailed case) is:

$$p-\text{value} = 2P(t_{n_1+n_2-2} > |t_0|);$$

that quantity is smaller then α if and only if the test rejects H_0 at level α.

Motivational Example (Cont.) We compute the required quantities.

```
y.1 <- c(18,43,28,50,16,32,13)
y.2 <- c(40,54,26,63,21,37,39)
(y.bar.1 <- mean(y.1))
(y.bar.2 <- mean(y.2))
(s.2.1 <- var(y.1))
(s.2.2 <- var(y.2))
```

```
[1] 28.57143
[1] 40
[1] 198.619
[1] 215.3333
```

The pooled estimate of the variance is easy to compute.

```
n.1 = length(y.1)
n.2 = length(y.2)
(n.1+n.2-2)
(s.2.p <- ((n.1-1)*s.2.1 + (n.2-1)*s.2.2)/(n.1 + n.2 - 2))
```

16: When the alternative hypothesis is in the form $H_1 : \mu_1 \neq \mu_2$, the test is a **two-tailed test**. If, however, the alternative hypothesis is either $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$, the test becomes a **one-tailed test**.

17: Common values: $\alpha = 0.01, 0.05, 0.1$.

```
[1] 12
[1] 206.9762
```

The test statistic is computed below.

```
(t_0 <- (y.bar.1 - y.bar.2)/sqrt(s.2.p*(1/n.1 + 1/n.2)))
```

```
[1] -1.486161
```

The $p-$value for the two-sided test is thus $2P(t_{12} > |-1.486161|)$.

```
2*pt(q=t_0, df=n.1 + n.2 - 2, lower.tail=TRUE)
```

```
[1] 0.1630303
```

Since the $p-$value is larger than $\alpha = 0.05$, we have insufficient evidence to reject H_0, which is to say that the observed data is compatible with the idea that the treatment has no effect. □

We have discussed this before (in Section 7.4, notably), but we will repeat it here for good measure: failure to reject the null hypothesis H_0 is not the same as accepting the null hypothesis H_0. We cannot **prove** H_0, we can only show that the observed data is at least compatible with it.[18]

18: We can **reject** H_0, however, which is equivalent to saying that the observed data is not compatible with it.

Power and Sample Size We now turn to the sample size determination n_1 and n_2. In a study, these are usually determined based on the need to offer **sufficient statistical power**.

When H_0 is true, the test statistic t_0 follows a Student $T-$distribution with $n_1 + n_2 - 2$ degrees of freedom. However, when H_0 is false, t_0 follows a non-central $T-$distribution with non-centrality parameter

$$\delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Suppose we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$.

The power function of the test is then given by

$$\begin{aligned}\pi(\mu_1 - \mu_2) &= P\left(\frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha;n_1+n_2-2} \,\middle|\, H_0 \text{ is false}\right) \\ &= P\left(\frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha;n_1+n_2-2} \,\middle|\, \mu_1 - \mu_2 > 0\right).\end{aligned}$$

The power function increases with δ. Thus, the power **increases** when:

1. $|\mu_1 - \mu_2|$ **increases** – a large difference between the means is easier to detect;

2. σ **decreases** – a given difference between μ_1 and μ_2 is easier to detect when the errors $\varepsilon_{\ell,j}$ are small, and/or
3. n_1 and/or n_2 **increases**.

Confidence Intervals We can construct an **approximate** $100(1-\alpha)\%$ **confidence interval for** $\mu_1 - \mu_2$**:**

$$\text{C.I.}(\mu_1 - \mu_2; 1-\alpha) \equiv \overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2;n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

In the previous example, we get a 95% confidence interval for $\mu_1 - \mu_2$ by computing

$$\text{C.I.}(\mu_1 - \mu_2; 0.95) \equiv (28.57 - 40) \pm 2.1788 \cdot 14.39 \sqrt{\frac{1}{7} + \frac{1}{7}} \iff (-28.18, 5.32).$$

Because the interval contains 0, we do not have enough evidence to reject H_0 – the data is not incompatible with the notion that $\mu_1 - \mu_2 = 0$.[19] This matches the $p-$test result from the previous section.

19: Which is not the same as saying that we accept $H_1 : \mu_1 - \mu_2 \neq 0$.

Paired-Difference Test When the samples are drawn independently from the two populations, we refer to the test as **unpaired**.[20] In a **paired** scenario, the units are not independent:[21] we could imagine selecting $n = 7$ sheep, testing them for tapeworm **before** treating them with a drug, then testing the same sheep for tapeworm **after** the treatment.

20: We often have $n_1 \neq n_2$.

21: In some sense, they are maximally dependent.

If a given specimen is somehow more likely to be afflicted by tapeworm due to genetics or farmer care, we wouldn't be surprised to find a link in its before/after measurements.

Motivational Example To compare the wear-and-tear qualities of two types of road paints, A and B, a sample of each is applied to a small area of five randomly selected roads. The roads operate as they normally do, with their specific usage patters, and the number of weeks to some "failure" threshold is recorded for each sample.

These measurements appear in the table below. Do the data present sufficient evidence to indicate a difference in the average wear for the two paint types?

Road	Paint A	Paint B
1	9.1	8.7
2	11.2	10.7
3	9.6	9.0
4	8.6	8.2
5	8.9	8.4

If we treated these samples as independent, we would be able be able to answer the question using the pooled variance s_p^2, computed with the help of $\overline{y}_A$, $\overline{y}_B$, s_A, s_B, and $n_A = n_B = 5$.

The **two-sample pooled $T-$test** would conclude that we cannot reject the null hypothesis $H_0 : \mu_A = \mu_B$, which is certainly thought-provoking given that the time to "failure" is systematically longer for Paint A than it is for Paint B. □

We have alluded to this problem at the start of the section: the two-sampled pooled $T-$test **is not the proper statistical test** to use in this case because the two samples are **not independent**.

Motivational Example (Cont.) Indeed, the (pair of) measurements Paint A and Paint B for a particular roadway are definitely **related**. The readings have approximately the same magnitude for a road but vary markedly from one road to another. Paint wear-and-tear is largely determined by **traffic volume** and **type**, the **weather**, and the **road surface**, say.

Since each road is likely to have different characteristics on that front, we expect a large amount of variability in the data from one road to another.

In designing the paint wear-and-tear experiment, the experimenters realized that the measurements would vary greatly from road to road. If the paint types (five of type A and five of type B) were randomly assigned to 10 roads, resulting in two independent random samples of size 5, this variability would result in a large standard error and make it difficult to detect a difference in the means.

Instead, they chose to "**pair**" the measurements, comparing the wear-and-tear for Paint A and Paint B on each of the five roads.

Road	Paint A	Paint B	Difference d
1	9.1	8.7	0.4
2	11.2	10.7	0.5
3	9.6	9.0	0.6
4	8.6	8.2	0.4
5	8.9	8.4	0.5

This experimental design, sometimes called a **paired-difference** or **matched pairs design**, allows us to eliminate the road-to-road variability by looking at only the five difference measurements shown above. These five differences form a **single random sample** of size $n = 5$. □

For a **paired-difference test** with n samples, we compute $d_i = y_{1,i} - y_{2,i}$ for $i = 1, \ldots, n$. The **null** and the **alternative hypotheses** are:

$$H_0 : \mu_d = 0$$

and

$$H_1 : \mu_d \neq 0 \quad \text{or} \quad H_1 : \mu_d > 0 \quad \text{or} \quad H_1 : \mu_d < 0,$$

while the **test statistic** is:

$$t_0 = \frac{\overline{d} - 0}{s_d/\sqrt{n}}, \tag{11.1}$$

where $\overline{d} = (d_1 + \cdots + d_n)/n$ and

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2.$$

For a two-tailed test at level α, we reject H_0 when

$$|t_0| > t_{\alpha/2;n-1}.$$

For a one-tailed test $H_0 : \mu_d = 0$ against $H_1 : \mu_d > 0$ (respectively, $H_1 : \mu_d < 0$), we reject H_0 when

$$t_0 > t_{\alpha;n-1}; \quad (\text{resp. } t_0 < -t_{\alpha;n-1}).$$

We can build an approximate $100(1-\alpha)\%$ confidence interval for μ_d using

$$\text{C.I.}(\mu_d; 1-\alpha) \equiv \overline{d} \pm t_{\alpha/2;n-1} \cdot \frac{s_d}{\sqrt{n}}.$$

Motivational Example (Cont.) We prepare the data.

```
d <- c(0.4, 0.5, 0.6, 0.4, 0.5)
n = length(d)
```

Simple calculations leads to $\overline{d}$ and s_d.

```
(d.bar <- mean(d))
(s.2.d <- var(d))
```

```
[1] 0.48
[1] 0.007
```

The test statistic t_0 can be computed easily.

```
(t_0 <- (d.bar - 0)/sqrt(s.2.d/n))
```

```
[1] 12.8285
```

At significance level α, the critical value of Student's T distribution with $n - 1 = 4$ degrees of freedom is $t_{\alpha/2;n-1}$, which can compute using either of the following ways in R.

```
alpha = 0.05
(t.crit = qt(p=1 - 0.05/2, df=n-1))
qt(p=0.05/2, df=n-1, lower.tail = FALSE)
```

```
[1] 2.776445
```

Since $12.829 = t_0 > t_{4;0.025} = 2.776$, we reject H_0 and we conclude that there is a difference in the mean wear-and-tear for paints A and B.[22]

We build an approximate 95% confidence interval for μ_d as follows.

22: Note that the observed value $t_0 = 12.829$ is quite large for the Student T distribution with 4 degrees of freedom, and the test result is highly significant.

```
c(d.bar - t.crit*sqrt(s.2.d/n),
  d.bar + t.crit*sqrt(s.2.d/n))
```

```
[1] 0.3761149 0.5838851
```

Note that this interval is much narrower than the interval that would have been obtained using the unpaired data, which indicates that the paired difference design increased the accuracy of the estimate – we have gained valuable information by using this design. □

The paired-difference test or matched pairs design used in the paint wear-and-tear experiment is a special case of an experimental design called a **randomized block design** (see Section 11.5). Importantly, the pairing (or blocking) must occur when the experiment is **planned**, and not after the data are collected.

11.2.3 Inference on the Population Variance

In some research situations, the primary interest lies in making inferences concerning **population variances** rather than focusing solely on population means. We begin by considering a test designed for a **single** population variance.

Imagine we have selected a random sample, represented as $y_1, \ldots, y_n$, from a population characterized by a mean of μ and a variance of σ^2. An important assumption is that the population from which this sample is drawn is **normally distributed**, i.e. $Y \sim \mathcal{N}(\mu, \sigma^2)$.

The hypothesis test pits

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{against} \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

The analysis uses the **test statistic**

$$\chi_0^2 = (n-1)s^2/\sigma_0^2.$$

Under the assumption that H_0 is indeed true, the distribution of χ_0^2 follows a χ_{n-1}^2 distribution.

We reject H_0 if $\chi_0^2 > \chi_{\alpha/2;n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2;n-1}^2$, with

$$P(W > \chi_{\alpha/2;n-1}^2) = P(W < \chi_{1-\alpha/2;n-1}^2) = \alpha/2, \quad \text{where } W \sim \chi_{n-1}^2.$$

We build an approximate $100(1-\alpha)\%$ **confidence interval for** σ^2 *via*:

$$\frac{(n-1)s^2}{\chi_{\alpha/2;n-1}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2;n-1}^2}.$$

Example An experimenter believes that the variability of her measuring apparatus has a standard deviation of $\sigma = 2.5$. During an experiment, the measurements recorded were 4.2, 5.3, and 10.3. The question arises: do these observations support or contradict her belief? We test this assertion using a significance level of $\alpha = 0.05$.

Firstly, we define our null and alternative hypotheses as:

$$H_0 : \sigma^2 = 6.25 \quad \text{against} \quad H_1 : \sigma^2 \neq 6.25.$$

We can find the test statistics χ_0^2 as follows.

```
x <- c(4.2, 5.3, 10.3)
n = length(x)
(s.2 = var(x))
```

```
[1] 10.57
```

```
sigma.2 = 6.25
chi.2.0 = (n-1)*s.2/sigma.2
```

```
[1] 3.3824
```

We can compute the critical χ_{n-1}^2 values at $\alpha = 0.05$.

```
alpha = 0.05
(crit.lv = qchisq(p=alpha/2, df=2))
(crit.uv = qchisq(p=1-alpha/2, df=2))
```

```
[1] 0.05063562
[1] 7.377759
```

We reject the null hypothesis H_0 if $\chi_0^2 > 7.38$ or $\chi_0^2 < 0.05$. Since the observed value of $\chi_0^2 = 3.3824$ lies between the critical values, we do not reject H_0.[23]

23: This indicates that the data does not provide sufficient evidence to dispute the experimenter's initial belief about the variability of her instrument.

She can build an approximate 95% confidence interval for σ^2 by using the formula.

```
c((n-1)*s.2/crit.uv, (n-1)*s.2/crit.lv)
```

```
[1] 2.865369  417.4927
```

This wide range implies a high level of uncertainty about the true variance, which further underscores the need for more data (or a different testing approach). □

11.2.4 Inference on the Ratio of Variances

We now turn our attention to the case of comparing two population variances. Consider two **normal populations**, labeled I and II. Denote the population variances associated with each populations by σ_1^2 and σ_2^2.

We draw a random sample of size n_1 from Population I:

$$y_{1,1}, \ldots, y_{1,n_1} \sim \mathcal{N}(0, \sigma_1^2)$$

and similarly from Population II:

$$y_{2,1}, \ldots, y_{2,n_2} \sim \mathcal{N}(0, \sigma_2^2).$$

The samples are **unpaired**, and so assumed to be **independent** of one another.

The **hypothesis test** for the variances is framed as:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

The **test statistic** employed for the test is

$$F_0 = s_1^2 / s_2^2.$$

Under the assumption that H_0 is true, the distribution of F_0 follows an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The null hypothesis H_0 is **rejected** at significance level α if

$$F_0 > F_{\alpha/2;n_1-1,n_2-1} \quad \text{or} \quad F_0 < F_{1-\alpha/2;n_1-1,n_2-1},$$

with

$$P(W > F_{\alpha/2;n_1-1,n_2-1}) = P(W < F_{1-\alpha/2;n_1-1,n_2-1}) = \alpha/2, \quad \text{where } W \sim F_{n_1-1,n_2-1}.$$

Equivalently, we can express a $100(1-\alpha)\%$ **confidence interval for the ratio** σ_1^2/σ_2^2 *via*:

$$s_1^2/s_2^2 \cdot F_{1-\alpha/2;n_2-1,n_1-1} < \sigma_1^2/\sigma_2^2 < s_1^2/s_2^2 \cdot F_{\alpha/2;n_2-1,n_1-1}.$$

Note the order of the degrees of freedom.[24]

24: We may need to leverage the relationship

$$F_{1-\gamma,\nu_1,\nu_2} = \frac{1}{F_{\gamma,\nu_2,\nu_1}}$$

in the analysis.

Example The same experimenter is concerned that the variability of her responses may not be the same when she is using two different experimental procedures.

She conducts a preliminary study with random samples of $n_1 = 11$ and $n_2 = 9$ responses and obtains $s_1^2 = 8.25$ and $s_2^2 = 4.32$, respectively. Do the sample variances present sufficient evidence to indicate that the population variances are unequal?

The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

The test statistic is given by $F_0 = 8.25/4.32 = 1.91$. We reject H_0 at level $\alpha = 0.05$ if

$$F_0 > F_{0.025,10,8} = 4.29 \quad \text{or} \quad F_0 < F_{0.975,10,8} = 0.26;$$

therefore, we cannot reject H_0 based on the observed data: there is insufficient evidence to indicate a difference in the population variances.[25]

25: Perhaps if we increased the sample sizes?

We can build a 95% confidence interval for σ_1^2/σ_2^2 *via*:

$$\begin{aligned} \text{C.I.}(\sigma_1^2/\sigma_2^2; 0.95) &\equiv (8.25/4.32 \cdot F_{0.975,8,10}, 8.25/4.32 \cdot F_{0.025,8,10}) \\ &\equiv (8.25/4.32 \cdot 0.23, 8.25/4.32 \cdot 3.85) \\ &\equiv (0.44, 7.36). \end{aligned}$$

Because the confidence interval includes 1 (which corresponds to the situation of equal variances), we cannot reject H_0 at significance level $\alpha = 0.05$.

11.3 One-Way Classification

In the worm/sheep example of Section 11.2.1, we were primarily concerned with comparing the worm counts in treated versus untreated lambs, represented as $\mu_1 - \mu_2$. Within the context of experimental designs, the drug administered (or lack thereof) to the lambs is considered a **factor** with two levels: **treated**, **untreated**.

As we progress through this chapter, our focus shifts to a model where the factor encompasses a levels, thereby giving rise to a treatments. The primary objective is to examine hypothesis testing for **equality among more than two population means**. To achieve this, we leverage a method of data analysis known as the **analysis of variance** (ANOVA).[26]

26: In essence, ANOVA can be perceived as a generalization of the customary $T-$test.

11.3.1 Completely Randomized Designs

In experiments where we have a treatments to compare and N units available for the study, a **completely randomized design** offers an efficient approach. To implement such a design:

1. decide on sample sizes $n_1, n_2, \ldots, n_a$ such that $n_1+n_2+\ldots+n_a = N$;
2. randomly allocate n_1 units to Treatment 1, n_2 units to Treatment 2, and so forth, until n_a units are assigned to Treatment a.

In this design, the N experimental units are randomly divided into a groups. Taking the worm/sheep example of Section 11.2.1 as an illustration, the $N = 14$ lambs were divided **at random** into $a = 2$ groups: the treated group and the untreated group.

Alternatively, one could view the completely randomized design as drawing random samples from each of a distinct populations. Each population represents a unique **level** (or treatment) of the **factor** under consideration.

Regardless of the perspective – whether through **random selection** or **random assignment** – completely randomized designs are centered around a **single factor**, which is why they are often referred to as a **one-way classification**.

The next example (modified from [3]) illustrates the basic notation.

Example A horticulturist is investigating the phosphorus content of tree leaves from three different varieties of apple trees (A, B and C). Random samples of five leaves from each three varieties are analyzed for phosphorus content. The observations are shown below.

variety	sample size	phosphorus content	totals	means
1	5	0.45, 0.50, 0.68, 0.60, 0.57	2.80	0.560
2	5	0.65, 0.70, 0.90, 0.84, 0.79	3.88	0.776
3	5	0.50, 0.70, 0.65, 0.63, 0.56	3.04	0.608

The **response variable** is the phosphorus content, the **factor** (with three levels) is the tree variety.

Notation

- $y_{i,j}$ is the j**th observation for the** i**th factor level** (group, class), $i = 1, \ldots, a$; $j = 1, \ldots, n_i$;
- n_i is the **number of sample observations for the** i**th factor level**;
- the **total sample size** is

$$N = \sum_{i=1}^{a} n_i;$$

- $y_{i,\bullet}$ is the **total of the sample observations for the** i**th factor level**, so

$$y_{i,\bullet} = \sum_{j=1}^{n_i} y_{i,j};$$

- $y_{\bullet,\bullet}$ is the **grand total of the sample observations**, so

$$y_{\bullet,\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{i,j};$$

- $\overline{y}_{i,\bullet}$ is the **average of the sample observations for the** i**th factor level**, so

$$\overline{y}_{i,\bullet} = y_{i,\bullet}/n_i;$$

- $\overline{y}_{\bullet,\bullet}$ is the **average of all sample observations**, so $\overline{y}_{\bullet,\bullet} = y_{\bullet,\bullet}/N$.

In the example, we have:

- $y_{i,j}$ is the phosphorus content from leave j of variety i, $i = 1, 2, 3$; $j = 1, \ldots, 5$;
- $n_1 = n_2 = n_3 \equiv n = 5$;
- $N = n \cdot 3 = 5 \cdot 3 = 15$;
- $y_{1,\bullet} = 2.80$, $y_{2,\bullet} = 3.88$, $y_{3,\bullet} = 3.04$;
- $y_{\bullet,\bullet} = 9.72$;
- $\overline{y}_{1,\bullet} = 0.560$, $\overline{y}_{2,\bullet} = 0.776$, $\overline{y}_{3,\bullet} = 0.608$;
- $\overline{y}_{\bullet,\bullet} = 0.648$.

11.3.2 One-Way Classification Model

We consider a populations (**groups**, **treatments**). Initially, we address the scenario of **balanced data**, with $n_i = n = N/a$ observations for each treatment i.

The data can be summarized in the following manner:

- from Population 1, we gather the observations $y_{1,1}, \ldots, y_{1,n}$;
- from Population 2, we gather the observations $y_{2,1}, \ldots, y_{2,n}$;
- ...
- from Population a, we gather the observations $y_{a,1}, \ldots, y_{a,n}$.

For each treatment $i = 1, \ldots, a$, we assume that the observations

$$y_{i,1}, \ldots, y_{i,n} \sim \mathcal{N}(\mu_i, \sigma^2).$$

Equivalently, we can express the model as

$$y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \; j = 1, \ldots, n,$$

with the errors $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ being i.i.d. random variables. We also assume a **common variance** σ^2 for the a populations.[27] The parameters to be **estimated** include $\mu_1, \ldots, \mu_a$, and σ^2.

27: See **homoscedasticity**, Chapter 8.

We can deduce that:

$$\mathrm{E}(y_{i,j}) = \mu_i \text{ for the } j\text{th observation in treatment group } i$$

and the variance is given by:

$$\mathrm{Var}(y_{i,j}) = \sigma^2 \quad \text{for all } i, j.$$

Alternative Reparametrization We can also recast the problem in a different manner:

$$\mu_i = \mu + (\mu_i - \mu) \equiv \mu + \tau_i,$$

where $\tau_i = \mu_i - \mu$ for all $i = 1, \ldots, a$. Here, τ_i represents the ith **treatment effect** (or treatment effect).

Given this, the **one-way classification model** can be expressed as:

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \; j = 1, \ldots, n,$$

where μ stands for the global (or common) mean applicable to all observations, and the error term $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. This yields an expectation of:

$$\mathrm{E}(y_{i,j}) = \mu + \tau_i.$$

The original model has a parameters, specifically: $\mu_1, \ldots, \mu_a$. However, the new model presents $a + 1$ parameters: $\mu, \tau_1, \ldots, \tau_a$. This makes the model **over-parametrized**.

Addressing this, we set the constraint:

$$\sum_{i=1}^{a} \tau_i = 0.$$

It's clear that the both the **original** model and the **reparametrized** model are equivalent, provided we adhere to the constraint. This constraint enables us to express:

$$\begin{aligned} \mu_1 &= \mu + \tau_1, \\ &\vdots \\ \mu_{a-1} &= \mu + \tau_{a-1}, \\ \mu_a &= \mu - (\tau_1 + \cdots + \tau_{a-1}), \end{aligned}$$

reducing the parameter count to a parameters: $\mu, \tau_1, \ldots, \tau_{a-1}$.

Overview Most often, the main objective in ANOVA is to determine if there are differences between the a populations (or treatments). A pertinent question arises: why do we need a new procedure to compare population means when Student's $T-$test is available?

Consider an instance with $a = 3$ population means: μ_1, μ_2, and μ_3. We could hypothetically test each of the **three pairs** of hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_0 : \mu_1 = \mu_3, \quad \text{and} \quad H_0 : \mu_2 = \mu_3$$

against the appropriate alternatives to identify where the differences (if any) are located.

But each test we conduct is prone to **errors** – consequently, the more tests we perform, the greater the likelihood that at least one of our conclusions will be erroneous.[28]

28: We will delve deeper into this subject at a later date.

ANOVA offers a **singular, comprehensive test** to evaluate the equality of the a population means. Once we discern if a genuine difference exists among the means, we can then use a designated procedure to pinpoint the origins of these differences.

The hypothesis tests pits

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{against} \quad H_1 : \mu_i \neq \mu_j, \text{ for at least one pair } (i, j),$$

or, in an equivalent form:

$$H_0 : \tau_1 = \cdots = \tau_{a-1} = 0 \quad \text{against} \quad H_1 : \text{at least one } \tau_i \neq 0.$$

11.3.3 Analysis of Variance

In the analysis of variance, we focus on **partitioning the total sum of squares**, starting with the **basic decomposition**

$$y_{i,j} - \overline{y}_{\bullet,\bullet} = (\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}) + (y_{i,j} - \overline{y}_{i,\bullet}).$$

Each component of the decomposition is interpreted as follows:

- $y_{i,j} - \overline{y}_{\bullet,\bullet}$ is the **total deviation** component;
- $\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}$ is the **deviation of the estimated factor level mean around the overall mean**, and
- $y_{i,j} - \overline{y}_{i,\bullet}$ is the **deviation around the estimated factor level mean**.

We can show (see Exercises) that the sums of squares decomposition for this scenario is:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{\bullet,\bullet})^2 = n\sum_{i=1}^{a}(\overline{y}_{i,\bullet}-\overline{y}_{\bullet,\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2,$$

or

$$\text{SST} = \text{SSA} + \text{SSE},$$

where:

- SST is the **total sum of squares**;
- SSA is the **treatment (or factor) sum of squares**, and
- SSE is the **error sum of squares**.

Given that the total sum of squares SST is **fixed**, an **increase** in SSA corresponds to a **decrease** in SSE and *vice versa*.

If all the observations within a given factor level are identical across all factor levels, then SSE = 0 and SST = SSA. Conversely, if all the estimated factor levels $\overline{y}_{i,\bullet}$ are equal, then SSA = 0 and SST = SSE.

Each sum of squares in the decomposition is associated to a **degree of freedom** (df):

- SST $\rightsquigarrow N-1$
- SSA $\rightsquigarrow a-1$
- SSE $\rightsquigarrow a(n-1) = N-a$

The decomposition's "structure" applies to the degrees of freedome:

$$N-1 = (a-1) + a(n-1) = a-1+N-a.$$

Variance Considerations The ith treatment **sample variance** is:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2, \quad i=1,\ldots,a;$$

we know that $\text{E}(s_i^2) = \sigma^2$ and $(n-1)s_i^2/\sigma^2 \sim \chi^2_{n-1}$ for all $i = 1,\ldots,a$.

Thus, we can express SSE as

$$\text{SSE} = \sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2\right] = \sum_{i=1}^{a}(n-1)s_i^2 = (n-1)\sum_{i=1}^{a}s_i^2,$$

and using the typical argument related to the trace of quadratic forms, we can show that

$$\text{SSE}/\sigma^2 \sim \chi^2_{N-a}.$$

Theorem: The random variable

$$\text{MSE} = \frac{\text{SSE}}{N-a} = \frac{n-1}{N-a}\sum_{i=1}^{a}s_i^2 = \frac{1}{a}\sum_{i=1}^{a}s_i^2$$

is an unbiased estimator of σ^2.[29]

So, what exactly does SSA estimate?

29: This holds true regardless of whether the factor level means μ_i are equal or not. Intuitively, this is reasonable: the variability of observations within each factor level is not influenced by the magnitude of the estimated factor level means when the populations are normal.

Theorem: the expectation of SSA is:

$$\begin{aligned}\mathrm{E(SSA)} &= \frac{1}{N}\sum_{i=1}^{a}\left\{n\sigma^2 + [n(\mu+\tau_i)]^2\right\} - \frac{1}{an}[an\sigma^2 + (an\mu)^2] \\ &= (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2.\end{aligned}$$

If we denote the mean square due to the factor A (commonly known as the **treatment mean square**) as $\mathrm{MSA} = \mathrm{SSA}/(a-1)$, then:

$$\mathrm{E(MSA)} = \sigma^2 + \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2.$$

In situations where all the factor level means are the same ($\mu_i \equiv \mu$), then we have $\tau_i^2 = (\mu_i - \mu)^2 \equiv 0$ and $\mathrm{E(MSA)} = \sigma^2$. Consequently, both MSE and MSA offer unbiased estimates of σ^2. However, when the μ_i's differ, MSA tends to be larger than MSE on average.

It can be shown (although it is beyond the scope of these notes) that:

- SSA/σ^2 follows a **non-central χ^2 distribution**:

$$\mathrm{SSA}/\sigma^2 \sim \chi^2_{a-1}\left(n\sum_{i=1}^{a}\tau_i^2/\sigma^2\right);$$

- the random variables SSE and SSA are **independent**.

$F-$Test for the Equality of Treatment Means How can we tell if the treatment means are identical?

The $F-$test pits

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{against} \quad H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i,j).$$

The test statistic to be used is

$$F_0 = \frac{\mathrm{MSA}}{\mathrm{MSE}}.$$

Large values of F_0 support H_1 since MSA will tend to exceed MSE when H_1 holds.[30] On the other hand, values of F_0 near 1 tend to support H_0 since both MSE and MSA have the same expected value when H_0 holds.[31] Hence, the appropriate test is an **upper-tail one**.

When H_0 holds, SSE/σ^2 and SSA/σ^2 are independent χ^2 variables. Therefore, under H_0,

$$F_0 = \frac{\mathrm{SSA}/(a-1)}{\mathrm{SSE}/(N-a)} \sim F_{a-1,N-a}.$$

When H_1 holds, that is, when the μ_i's are not all equal, F_0 does not follow the customary F distribution.[32]

It is thus reasonable to reject H_0 if we observe large values of F_0. Formally, we reject H_0 at significance level α if

$$F_0 > F_{\alpha;a-1,N-a}.$$

30: We have seen above that the ratio of the expected values, $\frac{\mathrm{E(MSA)}}{\mathrm{E(MSE)}}$, is greater than 1 under H_1.

31: Indeed, under H_0,

$$\frac{\mathrm{E(MSA)}}{\mathrm{E(MSE)}} = 1.$$

32: It follows instead a more complicated **non-central F distribution**.

We can construct an **ANOVA table** for the $F-$test for equality of treatment means in the one-way classification scenario, based on the test statistic $F_0 = \text{MSA}/\text{MSE}$ (see Table 11.11).

Source	SS	df	MS	F_0
Treatment	SSA	$a-1$	MSA	$F_0 = \text{MSA}/\text{MSE}$
Error	SSE	$N-a$	MSE	
Total	SST	$N-1$		

Table 11.11: ANOVA table for the equality of the treatment means μ_i in the one-way classification scenario.

From a computational perspective, the following equivalent formulas are sometimes used, since they are easier to handle when we do not use software:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSA} = \frac{1}{N}\sum_{i=1}^{a} y_{i,\bullet}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSE} = \text{SST} - \text{SSA}.$$

Example The ANOVA table for the phosphorus dataset of the previous section can be obtained as follows in R.

First we load the data.

```
variety <- c(1,2,3)
sample.size <- c(5,5,5)
content.1 <- c(0.45, 0.50, 0.68, 0.60, 0.57)
content.2 <- c(0.65, 0.70, 0.90, 0.84, 0.79)
content.3 <- c(0.50, 0.70, 0.65, 0.63, 0.56)
content <- rbind(content.1, content.2, content.3)
data <- data.frame(cbind(sample.size, content))
rownames(data) <- variety
colnames(data) <- c("sample.size", "leaf.1", "leaf.2",
                    "leaf.3", "leaf.4", "leaf.5")
data$totals <- rowSums(content)
data$means <- data$totals/data$sample.size
data
```

```
  sample.size leaf.1 leaf.2 leaf.3 leaf.4 leaf.5 totals means
1           5   0.45    0.5   0.68   0.60   0.57   2.80 0.560
2           5   0.65    0.7   0.90   0.84   0.79   3.88 0.776
3           5   0.50    0.7   0.65   0.63   0.56   3.04 0.608
```

We compute the necessary quantities and place them in the ANOVA table.

```
a = nrow(data)
n = length(content.1)
N = a*n
grand.mean = mean(unlist(data[,c(2:(n+1))]))
SST = sum((data[,c(2:(n+1))]-grand.mean)^2)
SSA = n * sum((data$means-grand.mean)^2)
SSE = SST - SSA
```

```
ANOVA = as.data.frame(cbind(c(SSA,SSE,SST),
                        c(a-1, N-a, N-1),
                        c(SSA/(a-1),SSE/(N-a),0),
                        c((SSA/(a-1))/(SSE/(N-a)),0,0)))
rownames(ANOVA) = c("Treatment", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
              SS df      MS       F0
Treatment 0.12864  2 0.06432 7.892025
Error     0.09780 12 0.00815
Total     0.22644 14
```

At significance level $\alpha = 0.05$, the critical value of $F_{2,12}$ is:

```
alpha=0.05
qf(p=1-alpha, df1 = a-1, df2 = N-a)
```

```
[1] 3.885294
```

Since $7.89 = F_0 > F_{0.05;2,12} = 3.89$, we reject H_0 at $\alpha = 0.05$ and we conclude that the mean phosphorus content is unlikely to be the same for all $a = 3$ varieties of trees. □

11.3.4 Estimation of Model Parameters

Recall that in the one-way classification model, $a + 1$ parameters require estimation, namely $\mu, \tau_1, \ldots, \tau_a$. We use the **least square estimation principle** to find them based on the observed data.

The sum of squares is defined as

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{i,j} - \mu - \tau_i)^2.$$

We find $\hat{\mu}$ and $\hat{\tau}_i$ that minimize L by differentiating L with respect to μ and τ_i, $i = 1, \ldots, a$, and setting to 0. This yields the **normal equations**:

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{i,j} - \hat{\mu} - \hat{\tau}_i) = 0 \quad (\mu\text{-equation}),$$
$$\sum_{j=1}^{n} (y_{i,j} - \hat{\mu} - \hat{\tau}_i) = 0 \quad (\tau_i\text{-equation}, i = 1, \ldots, a).$$

The corresponding system of linear equations is:

$$\begin{aligned} N\hat{\mu} + n \sum_{i=1}^{a} \hat{\tau}_i &= y_{\bullet,\bullet}, \\ n\hat{\mu} + n\hat{\tau}_1 &= y_{1,\bullet}, \\ &\vdots \\ n\hat{\mu} + n\hat{\tau}_a &= y_{a,\bullet}. \end{aligned}$$

Given the constraint $\tau_1 + \cdots + \tau_a = 0$, the solution is:

$$\begin{aligned}\hat{\mu} &= \overline{y}_{\bullet,\bullet},\\ \hat{\tau}_i &= \overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet} \quad \text{for } i = 1, \ldots, a.\end{aligned}$$

Thus, the **estimated treatment effect** for the ith treatment is

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i,\bullet};$$

the **difference between treatments** i and j is given by

$$\hat{\mu}_i - \hat{\mu}_j = \overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}.$$

Using the **pooled estimate** MSE as an estimator of σ^2, we can exhibit a $100(1-\alpha)\%$ confidence interval for μ_i *via*:

$$\overline{y}_{i,\bullet} \pm t_{\alpha/2;N-a}\sqrt{\frac{\text{MSE}}{n}};$$

for $\mu_i - \mu_j$, we have instead:

$$(\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) \pm t_{\alpha/2;N-a}\sqrt{\frac{2\text{MSE}}{n}}.$$

11.3.5 Unbalanced Designs

We could also opt for an **unbalanced design**, in which the number of observations n_i we sample in each treatment group i is not necessarily the same from one group to the other. However, a balanced design has several advantages.[33]

33: First and foremost, the theoretical derivations are simpler to obtain in the balanced case.

In particular, the power of the $F-$test is **larger** with balanced data. Indeed for $a = 2$ (two treatments), we can show that the power of the $F-$test is maximized when $\frac{1}{n} + \frac{1}{N-n}$ is minimized (see Exercises); if N is even and fixed, the minimum is thus achieved when $n = N/2$.

Moreover, the F test is only **robust against unequal variances** when data is balanced. For the case of $a = 2$ treatments, the F test is equivalent to the Student's $T-$test, with $F_0 = t_0^2$.

If we define θ as the ratio σ_1^2/σ_2^2 and R as the ratio n_1/n_2, the Student's $T-$test can be expressed as:

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}\left(\frac{1}{s_p^2}\cdot\frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}{\frac{1}{n_1} + \frac{1}{n_2}}\right)^{1/2}.$$

When $n_1, n_2 \to \infty$,

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \to \sigma_2^2\frac{1+R\theta}{1+R}.$$

Consquently, $t_0 \to \mathcal{N}(0, (R+\theta)/(1+R\theta))$, and $(R+\theta)/(1+R\theta) = 1$ when $R = 1$, regardless of θ's value.

For unbalanced data, the sum of squares formulas must be modified:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSA} = \sum_{i=1}^{a}\frac{y_{i,\bullet}^2}{n_i} - \frac{y_{\bullet,\bullet}^2}{N}.$$

Finally, when we estimate the model parameters, we solve the **normal equations** subject to the constraint:

$$\sum_{i=1}^{a} n_i\hat{\tau}_i = 0$$

as opposed to the constraint $\sum_{i=1}^{a}\hat{\tau}_i = 0$.[34]

34: If we note that the latter can also be written as $n\tau_1 + \cdots + n\tau_a = 0$ in the balanced case, we see that it is simply a special instance of the unbalanced case. The same comment applies to the modified formula for SSA.

11.3.6 Contrasts

The analysis of variance can tell us an indication that not all the treatment groups have the same mean response, but an ANOVA does not, by itself, provide information about **which treatments** are different or in what ways they differ.

To get answers to these questions, we must examine the **treatment means**, or equivalently, the **treatment effects**. We can do so through **contrasts**, which enable us to focus in on **specific** (narrow) features of the data.[35]

35: In fact, a single contrast's focus is so narrow that it may obscure the overall picture.

By using several contrasts, we can move the focus around and explore more features. Intelligent use of contrasts involves choosing the contrasts so that they highlight interesting data features.[36]

36: But that's easier said than done without a solid understanding of the domain under study, which can be improved *via* data exploration, among others (see Chapter 18 for more information).

Linear Contrasts Linear combinations of the treatment effects μ_i

$$C = \sum_{i=1}^{a} c_i\mu_i, \quad \text{where } \sum_{i=1}^{a} c_i = 0 \text{ with } c_i \in \mathbb{R}.$$

are called **linear contrasts**; in general, we are interested in testing for

$$H_0 : C = 0 \quad \text{against} \quad H_1 : C \neq 0.$$

When there are a treatment effects, we sometimes identify the linear contrast C with its **signature vector** $(c_1, \ldots, c_a)$.

Examples

1. Suppose that we wish to test for

   $$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 \neq \mu_2;$$

   we must then work with the linear contrast $(1, -1, 0, \ldots, 0)$.[37]

   37: The linear contrast $(1, -1, 0, \ldots, 0)$ also does the trick, being equivalent to the one in the text when it comes to hypothesis testing.

2. Suppose that we wish to test for

   $$H_0 : \tfrac{1}{2}(\mu_1 + \mu_2) = \tfrac{1}{2}(\mu_3 + \mu_4) \quad \text{against} \quad H_1 : \tfrac{1}{2}(\mu_1 + \mu_2) \neq \tfrac{1}{2}(\mu_3 + \mu_4);$$

   we work with the linear contrast $(1/2, 1/2, -1/2, -1/2, 0, \ldots, 0)$.

To test a contrast hypothesis, we start by estimating C using

$$\widehat{C} = \sum_{i=1}^{a} c_i \overline{y}_{i,\bullet}.$$

Assume a balanced design; if the observations are obtained independently, we have $\text{Cov}(\overline{y}_{i_1,\bullet}, \overline{y}_{i_2,\bullet}) = 0$ if $i_1 = i_2$, and $\text{Var}(\overline{y}_{i,\bullet}) = \sigma^2/n$ for all i, so

$$\text{Var}(\widehat{C}) = \sum_{i=1}^{a} c_i^2 \text{Var}(\overline{y}_{i,\bullet}) = \frac{\sigma^2}{n} \sum_{i=1}^{a} c_i^2.$$

We obtain an **estimator** of $\text{Var}(\widehat{C})$ *via*:

$$\widehat{\text{Var}}(\widehat{C}) = \frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2.$$

It follows that the **test statistic** is given by

$$t_0 = \frac{\widehat{C}}{\sqrt{\frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2}}.$$

We can show that $t_0 \sim t_{N-a}$; therefore, we reject H_0 at significance level α if $|t_0| > t_{\alpha/2;N-a}$.

Instead of the $T-$test, however, we could use the **equivalent** $F-$test, with test statistic $F_0 = \frac{\text{SSC}}{\text{MSE}}$, which rejects H_0 at significance level α if $F_0 > F_{\alpha;1,N-a}$, where

$$\text{SSC} = \left(\sum_{i=1}^{a} c_i \overline{y}_{i,\bullet} \right)^2 \Bigg/ \sum_{i=1}^{a} c_i^2/n.$$

We can build a $100(1-\alpha)\%$ **confidence interval** for C is given by

$$\widehat{C} \pm t_{\alpha/2;N-a} \sqrt{\frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2}.$$

Example In the phosphorus dataset, suppose we want to test

$$H_0 : \mu_2 = \frac{\mu_1 + \mu_3}{2} \quad \text{against} \quad H_1 : \mu_2 \neq \frac{\mu_1 + \mu_3}{2}.$$

This is a contrast with $c_1 = -1/2$, $c_2 = 1$ and $c_3 = -1/2$.

The test statistics is given by

$$t_0 = \frac{(-1/2) \cdot 0.560 + 1 \cdot 0.776 + (-1/2) \cdot 0.608}{\sqrt{\left(\frac{0.00815/12}{5}\right) \{(-1/2)^2 + 1^2 + (-1/2)^2\}}} = \frac{0.192}{0.0142741} = 13.45094.$$

Since $|t_0| > t_{0.025,12} = 2.17881$, we reject H_0 and we conclude that there is enough evidence to conclude that μ_2 is different from the average of μ_1 and μ_3. □

Orthogonal Contrasts Two contrasts with coefficients $\{c_i\}$ and $\{d_i\}$ are **orthogonal** if $c_1d_1+\cdots+c_ad_a = 0$. For instance, the contrasts $-2\mu_1+\mu_2+\mu_3$ and $\mu_3 - \mu_2$ are orthogonal since

$$(-2)(0) + (1)(-1) + (1)(1) = 0.$$

If there are a treatments, we can find a set of $a - 1$ contrasts that are mutually orthogonal, that is, each one is orthogonal to all of the others. With 5 treatments (say), we can define 4 mutually orthogonal contrasts:

$$\begin{array}{lccccc} C_1 = & & & \mu_4 & -\mu_5 \\ C_2 = \mu_1 & & +\mu_3 & -\mu_4 & -\mu_5 \\ C_3 = \mu_1 & & -\mu_3 & & \\ C_4 = \mu_1 & -4\mu_2 & +\mu_3 & +\mu_4 & +\mu_5 \end{array}$$

The important feature of orthogonal contrasts is that they are **independent** (as random variables).[38]

38: An additional useful fact is that they **partition** the treatment sum of squares:

$$\text{SSA} = \sum_{i=1}^{a-1} \text{SSC}_i.$$

In other words, if we compute the sums of squares for a full set of orthogonal contrasts ($a-1$ contrasts for a groups), adding up those $a - 1$ sums of squares yields exactly the treatment sum of squares, which also has $a - 1$ degrees of freedom.

11.3.7 Multiple Comparisons

We have discussed **multiple hypothesis testing** in Section 8.2.3; how does it apply to design of experiments?

Example Suppose we want to compare four treatments, so $a = 4$. We may want to compare all the pairs

$$\begin{array}{lll} H_0 : \mu_1 = \mu_2, & H_0 : \mu_1 = \mu_3, & H_0 : \mu_1 = \mu_4, \\ H_0 : \mu_2 = \mu_3, & H_0 : \mu_2 = \mu_4, & H_0 : \mu_3 = \mu_4. \end{array}$$

Overall, there we have $k = 6$ possible tests of the form $H_0 : \mu_i = \mu_j$ against some fixed alternative type. □

In general, suppose that we wish to conduct k hypothesis tests. If the level of each individual test is α, then the overall error rate is likely to be **(much) larger** than α.

As an illustration, suppose that we conduct $k = 2$ tests, each one at significance level 5% .[39] Then, the probability of rejecting at least one of the null hypotheses when they are both true will be higher than 5%.

39: That is, the probability of a Type 1 error is 5% for each test separately.

Indeed, let E_j be the event that we reject the null hypothesis for the jth test, $j = 1, 2$. Then,

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= 0.05 + 0.05 - P(E_1 \cap E_2) = 0.1 - P(E_1 \cap E_2). \end{aligned}$$

As $0 \le P(E_1 \cap E_2) \le 0.5$, the probability of making at least one mistake is now between 5% and 10%.[40]

40: If the events are independent, then $0 < P(E_1 \cap E_2)$, and $P(E_1 \cup E_2) > 0.05$.

We can extend this argument to the general case of k tests. Suppose that the k null hypotheses H_0 are true. Once again, let's define E_j as the event that we reject the null hypothesis for the jth test, $j = 1, \cdots, k$.

Boole's inequality states that

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq \sum_{j=1}^{k} P(E_j) = \sum_{j=1}^{k} \alpha_j,$$

where α_j denotes the **probability of Type 1 error associated with the jth test**. If $\alpha_j = \alpha$ then

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq k\alpha.$$

For instance, for $k = 10$, and $\alpha_j = 0.05$ for all j, then the best that we can say is that

$$P\left(E_1 \cup \cdots \cup E_{10}\right) \leq 0.5.$$

Conclusion: the level of significance of a **family of tests** may differ from that of an **individual test**.

We use **multiple comparison procedures** to conduct multiple inference while **controlling the overall error rate**. The rationale behind these procedures is simple – we seek to ensure a **global significance level** below (or at) α. More specifically, we seek a procedure for which the probability of rejecting at least one the null hypotheses when they are all true is not larger than α.

Several procedures have been proposed in the literature, including:

- Bonferroni's (1936);
- Tukey's (1949);
- Scheffé's (1959).

Bonferroni's Procedure When investigating a particular set of k pairwise comparisons and/or contrasts, it is essential to specify the family of interest **in advance**. The **Bonferroni procedure** is versatile, and applicable whether the n_i's are equal or unequal and irrespective of whether the focus is on pairwise comparisons, contrasts, or a mix of both.

Instead of conducting each of the k tests at the usual α level, we conduct each test at the α/k level. With this adjustment, the probability of making at least one Type I error across all k tests is bounded by α:

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq \sum_{j=1}^{k} \frac{\alpha}{k} = k\left(\frac{\alpha}{k}\right) = \alpha.$$

For example, for an analysis involving 10 tests with an intended overall error rate of $\alpha = 0.05$, the Bonferroni correction would adjust the significance level for each test to $0.05/10 = 0.005$.

This method can also be extended to the construction of **simultaneous confidence intervals**. If we denote by $\text{C.I.}_1, \cdots, \text{C.I.}_k$ the associated confidence intervals, each constructed at a **coverage level** of $1 - \alpha$,[41] then the probability that all k intervals simultaneously contain their true parameter values is bounded above by $1 - \alpha$:

$$P\left(\bigcap_{j=1}^{k} E_j\right) \leq 1 - \alpha.$$

41: That is,

$$P(C_j \in \text{C.I.}_j) = 1 - \alpha, \quad j = 1, \cdots, k,$$

where C_j is the true value of the jth parameter or contrast of interest.

However, with Bonferroni's adjustment, if each interval is constructed to have a coverage probability of $1-\alpha/k$, then the **joint coverage probability** is at least $1-\alpha$:

$$P\left(\bigcap_{j=1}^{k} E_j\right) = 1 - P\left(\bigcup_{j=1}^{k} E_j^c\right) \geq 1 - \sum_{j=1}^{k} P(E_j^c) = 1 - \sum_{j=1}^{k} \frac{\alpha}{k} = 1 - \alpha.$$

An undoubted **advantage** of the Bonferroni method lies in its **generality**: it is applicable to a wide range of probability-based inferences across various distributions, not merely confidence intervals within a normal linear model.

But this method is not without its **drawbacks**. Chief among them being that for larger values of k, the individual significance level for each test can become **exceedingly stringent**.

With an overall error rate of $\alpha = 5\%$ and $k = 10$, say, the significance level for each test under Bonferroni's method is $1 - \alpha/k = 0.995$. This means each confidence interval might be so wide that its **practical utility** diminishes.[42]

42: In such scenarios, one might consider increasing the overall (joint) error rate, perhaps to 10%, to make the results **more easily interpretable**.

Tukey's Procedure The **Tukey multiple comparison procedure** is particularly valuable when our focus is on analyzing the set of **all pairwise comparisons** of **factor level means**. Specifically, when utilizing this method, the primary interest revolves around the tests defined by:

$$H_0 : \mu_i = \mu_j \quad \text{against} \quad H_1 : \mu_i \neq \mu_j.$$

When all sample sizes are balanced, the family confidence coefficient for the Tukey method aligns precisely with $1 - \alpha$, ensuring the family significance level is consistent with α. However, for unbalanced data, where sample sizes diverge, the Tukey procedure exhibits a **conservative behaviour**. This results in the family confidence coefficient surpassing $1 - \alpha$, and subsequently, the family significance level falling below α.

A key component of the Tukey procedure is the use of the **Studentized range distribution**. Given a set of i.i.d. random variables $y_1, \ldots, y_k \sim \mathcal{N}(\mu, \sigma^2)$, their **range** R is defined as:

$$R = \max\{y_1, \ldots, y_k\} - \min\{y_1, \ldots, y_k\}.$$

If s^2 be an estimator of σ^2 independent of R, and assume that $\frac{\upsilon s^2}{\sigma^2} \sim \chi^2_\upsilon$. Then the variable $\frac{R}{s}$ follows a **Studentized range distribution** $q_{k,\upsilon}$. Let $q_{\alpha;k,\upsilon}$ be the critical value for which

$$P\left(\frac{R}{s} > q_{\alpha;k,\upsilon}\right) = \alpha.$$

Theorem: suppose we have a means, $\overline{y}_{1,\bullet}, \ldots, \overline{y}_{a,\bullet}$, obtained from a independent normal samples, each of size n, with respective means $\mu_1, \ldots, \mu_a$ and a shared variance σ^2.[43]

43: That is, $\overline{y}_{i,\bullet} \sim \mathcal{N}(\mu_i, \sigma^2/n)$, for all i.

We know that MSE is an unbiased estimator of σ^2 independent of R and

$$\frac{(N-a)\text{MSE}}{\sigma^2} \sim \chi^2_{N-a}.$$

Under these conditions, the simultaneous probability for all pairwise comparisons is:

$$(\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) - q_{\alpha;a,N-a}\sqrt{\frac{\text{MSE}}{n}} < \mu_i - \mu_j < (\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) + q_{\alpha;a,N-a}\sqrt{\frac{\text{MSE}}{n}}.$$

The family confidence coefficient $1-\alpha$ pertaining to the multiple pairwise comparisons refers to the **proportion of correct families**, each consisting of all pairwise comparisons, when repeated sets of samples are selected and all pairwise confidence intervals are calculated each time.[44]

44: A family of pairwise comparisons is considered to be correct if **every pairwise comparison** in the family is correct.

This family confidence coefficient implies that, across repeated sampling, all pairwise comparisons in the family will be accurate in $100(1-\alpha)\%$ of the instances.

Transitioning our focus to **simultaneous testing**, the objective is to conduct a comprehensive set of tests that pit

$$H_0 : \mu_i = \mu_j \quad \text{against} \quad H_1 : \mu_i \neq \mu_j$$

for all potential pairwise comparisons. The pivotal test statistic in this context is:

$$q_0 = \frac{\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}}{\sqrt{\text{MSE}/n}}.$$

We reject H_0 at significance level α if $|q_0| \geq q_{\alpha;a,N-a}$.[45]

45: Selected percentiles for the Studentized range distribution can be found in tables, such as on this page. In R, we can use the functions `qtukey()` and `ptukey()`.

We illustrate the procedure with the help of a classical example.[46]

46: See here, for instance.

Example In a study of the effectiveness of different rust inhibitors, four brands (A, E, C, D) were tested. Altogether, 40 experimental units were randomly assigned to the four brands, with 10 units assigned to each brand. The results obtained after exposing the experimental units to severe weather conditions are given below.[47]

47: The higher the value, the more effective the rust inhibitor.

	Rust Inhibitor Brand			
	A	B	C	D
	$i=1$	$i=2$	$i=3$	$i=4$
$\overline{y}_{i,\bullet}$	43.14	89.44	67.95	40.47
$\overline{y}_{\bullet,\bullet} = 60.25$				
MSE = 6.14				

This study is a completely randomized design, where the levels of the single factor correspond to the four rust inhibitor brands. Suppose we are interested in all pairwise comparisons, which we evaluate *via* the Tukey procedure.

The important parameters are loaded below.

```
a = 4; N = 40; n = 10; alpha = 0.05
y.bar.1 = 43.14; y.bar.2 = 89.44
y.bar.3 = 67.95; y.bar.4 = 40.47
y.bar = 60.25; MSE = 6.14
(q.crit = qtukey(alpha, a, N-a, lower.tail = FALSE))
B = q.crit*sqrt(MSE/n)
```

```
[1] 3.808798
```

The 6 confidence intervals (with corresponding test statitsics) are computed as follows.

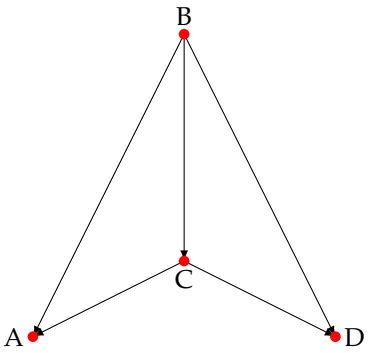
```
ci.2.1 = y.bar.2 - y.bar.1 +B*c(-1,1); q0.2.1 = (y.bar.2 - y.bar.1)/sqrt(MSE/n)
ci.3.1 = y.bar.3 - y.bar.1 +B*c(-1,1); q0.3.1 = (y.bar.3 - y.bar.1)/sqrt(MSE/n)
ci.4.1 = y.bar.4 - y.bar.1 +B*c(-1,1); q0.4.1 = (y.bar.4 - y.bar.1)/sqrt(MSE/n)
ci.3.2 = y.bar.3 - y.bar.2 +B*c(-1,1); q0.3.2 = (y.bar.3 - y.bar.2)/sqrt(MSE/n)
ci.4.2 = y.bar.4 - y.bar.2 +B*c(-1,1); q0.4.2 = (y.bar.4 - y.bar.2)/sqrt(MSE/n)
ci.4.3 = y.bar.4 - y.bar.3 +B*c(-1,1); q0.4.3 = (y.bar.4 - y.bar.3)/sqrt(MSE/n)
```

The simultaneous confidence intervals and tests for pairwise differences are shown in the table below.

Confidence Interval	**Test**		
	H_0	H_1	q_0
$43.3 < \mu_2 - \mu_1 < 49.3$	$\mu_2 = \mu_1$	$\mu_2 \neq \mu_1$	58.99
$21.8 < \mu_3 - \mu_1 < 27.8$	$\mu_3 = \mu_1$	$\mu_3 \neq \mu_1$	31.61
$-0.3 < \mu_4 - \mu_1 < 5.7$	$\mu_1 = \mu_4$	$\mu_1 \neq \mu_4$	3.40
$18.5 < \mu_2 - \mu_3 < 24.5$	$\mu_2 = \mu_3$	$\mu_2 \neq \mu_3$	27.37
$46.0 < \mu_2 - \mu_4 < 52.0$	$\mu_2 = \mu_4$	$\mu_2 \neq \mu_4$	62.39
$24.5 < \mu_3 - \mu_4 < 30.5$	$\mu_3 = \mu_4$	$\mu_3 \neq \mu_4$	35.01

Only in the comparison between A and D does the confidence interval include 0. Therefore, there is no clear evidence that either D or A is the better rust inhibitor. For the other pairs, we conclude that there is a difference in performance:

> B $\geq$ A, C $\geq$ A, B $\geq$ C, B $\geq$ D, C $\geq$ D (see the diagram format below).



We obtain the same conclusions if we look at the test statistics, and compare their absolute value to $q_{0.05;4,36} = 3.814$ – except for A and D, all differences are found to be statistically significant. □

Scheffé's Procedure The family of interest refers to the set of **all possible contrasts** among the factor level means:

$$C = \sum_{i=1}^{a} c_i \mu_i, \quad \text{where } \sum_{i=1}^{a} c_i = 0, \quad \text{with } c_i \in \mathbb{R}.$$

In essence, the family is comprised of estimates of all possible contrasts C or tests concerning all possible contrasts of the type:

$$H_0 : C = 0 \quad \text{versus} \quad H_1 : C \neq 0;$$

thus, the family consists of **infinitely many statements**.

The confidence level for the **Scheffé procedure** for the entire family is exactly $1 - \alpha$, regardless of whether the design is balanced or unbalanced.

Recall that

$$C = \sum_{i=1}^{a} c_i \mu_i$$

is estimated by

$$\widehat{C} = \sum_{i=1}^{a} c_i \overline{y}_{i,\bullet},$$

and that the variance of this estimate is

$$\widehat{\text{Var}}(\widehat{C}) = \frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2.$$

For simultaneous estimation through confidence intervals, the **Scheffé confidence intervals for the family of contrasts** C take the form:

$$\widehat{C} - W\sqrt{\widehat{\text{Var}}(\widehat{C})} < C < \widehat{C} + W\sqrt{\widehat{\text{Var}}(\widehat{C})},$$

where $W^2 = (a-1)F_{\alpha;a-1,N-a}$.[48]

48: See the justification for the Working-Hostelling test in Section 8.2.3 for an indication of how to prove this statement.

If we were to compute the confidence intervals for every conceivable contrast, then we would expect that the entire set of confidence intervals in the family would be accurate in roughly $100(1-\alpha)\%$ of the experimental repetitions. Note that the simultaneous confidence limits differ from those for a single confidence limit solely in terms of the **estimated standard deviation multiple** in front of the square root.

Considering the problem of **simultaneous testing**, we are interested in tests of the form:

$$H_0^C : C = 0 \quad \text{versus} \quad H_1^C : C \neq 0.$$

The corresponding test statistics are

$$F_0 = \frac{\widehat{C}^2}{(a-1)\widehat{\text{Var}}(\widehat{C})},$$

and we reject the specific test H_0^C if $F_0 > F_{\alpha;a-1,N-a}$.[49]

49: Given that applications of the Scheffé procedure never involve all conceivable contrasts, the confidence coefficient for the finite family of statements under consideration will exceed $1 - \alpha$. Thus, $1 - \alpha$ acts as a **guaranteed lower bound**. In a similar vein, the significance level for the finite family of tests will be below α.

The following example is found in [1].

Example The Kenton Food Company tested four different package designs for a new breakfast cereal. Twenty stores were selected as the experimental units. Each store was randomly assigned one of the package designs, with each package design assigned to five stores. A fire occurred in one store during the study period, so this store was dropped from the study. Hence, one of the designs was tested in only four stores.

The stores were chosen to be comparable in location and sales volume. Other relevant conditions that could affect sales, such as price, amount and location of shelf space, and special promotional efforts, were kept the same for all of the stores in the experiment.

Sales were observed for the study period; the results are recorded below.

	Package Design (i)				
	1	2	3	4	**Total**
n_i	5	5	4	5	**19**
$y_{i,\bullet}$	73	67	78	136	**354**
$\overline{y}_{i,\bullet}$	14.6	13.4	19.5	27.2	**18.63**

This study is a completely randomized unbalanced design with package type as the single, four-level factor.

For what it is worth, the package types had the following characteristics

- Package 1: 3-colour design, with a cartoon character;
- Package 2: 3-colour design, without a cartoon character;
- Package 3: 5-colour design, with a cartoon character;
- Package 4: 5-colour design, without a cartoon character.

The one-way classification ANOVA table for the observed data is:

Source	**SS**	**df**	**MS**	**F_0**
Treatment	588.2	3	196.07	18.585
Error	158.2	15	10.55	
Total	746.42	8		

We are interested in estimating the following 4 contrasts with family confidence coefficient 0.90:

$$
\begin{aligned}
C_1 &= \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} \\
C_2 &= \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2} \\
C_3 &= \mu_1 - \mu_2 \\
C_4 &= \mu_3 - \mu_4.
\end{aligned}
$$

We can compute the coefficient W for significance level $\alpha = 0.1$.

```
a=4; alpha=0.1; N=19;
(W = sqrt((a-1)*qf(alpha, df1=a-1, df2=N-a, lower.tail=FALSE)))
```

```
[1] 2.733014
```

We can easily compute the estimated contrasts.

```
y.bar.1 = 14.6; y.bar.2 = 13.4; y.bar.3 = 19.5; y.bar.4 = 27.2
C.hat.1 = (y.bar.1 + y.bar.2)/2 - (y.bar.3 + y.bar.4)/2
C.hat.2 = (y.bar.1 + y.bar.3)/2 - (y.bar.2 + y.bar.4)/2
C.hat.3 = y.bar.1 - y.bar.2
C.hat.4 = y.bar.3 - y.bar.4
```

The designed is unbalanced, so n is not constant. For the purposes of this exercise, we use the average value of the n_i for n. Moreover, we can read the value of MSE from the ANOVA table.

```
n = mean(c(5,5,4,5)); MSE = 10.55
```

We now compute the variance of the contrasts.

```
sum.c2.1 = 4*(1/2)^2; sum.c2.2 = 4*(1/2)^2
sum.c2.3 = 2*(1)^2; sum.c2.4 = 2*(1)^2
B.1 = sqrt(MSE/n*sum.c2.1); B.2 = sqrt(MSE/n*sum.c2.2)
B.3 = sqrt(MSE/n*sum.c2.3); B.4 = sqrt(MSE/n*sum.c2.4)
```

We are now able to obtain the joint 90% confidence intervals for the contrasts.

```
C.hat.1 + W*B.1*c(-1,1)
C.hat.2 + W*B.2*c(-1,1)
C.hat.3 + W*B.3*c(-1,1)
C.hat.4 + W*B.4*c(-1,1)
```

```
[1] -13.423064  -5.276936
[1] -7.3230638  0.8230638
[1] -4.560182  6.960182
[1] -13.460182  -1.939818
```

Note that the confidence interval for C_1 does not include 0. Hence, if we wished to test $H_0 : C_1 = 0$ versus $H_1 : C_1 \neq 0$ at 90% confidence (among 3 other contrasts), we would reject H_0 in favour of H_1, namely that the mean sales for the 3-colour and 5-colour designs differ.

The confidence interval provides additional information, however; the mean sales for the 5-colour designs exceed the mean sales for the 3-colour designs, by somewhere between 5.3 and 13.4 cases per store.

Using the other contrasts, the sales manager also concluded that no overall effect of cartoon characters in the package design is indicated by the data, although the use of a cartoon character in the 5-colour designs is associated with lower mean sales than when no cartoon character is used.[50] □

50: Is the link necessarily causal?

Bonferroni vs. Tukey vs. Scheffé If **all pairwise comparisons** are of interest, the Tukey procedure is superior to the Bonferroni and Scheffé procedures, leading to narrower confidence intervals. If **not all pairwise comparisons** are to be considered, the Bonferroni procedure may be prove to be a better choice (at times).

The Bonferroni procedure yields **tighter** confidence intervals than Scheffé's when the number of contrasts of interest is **about the same** as (or is **smaller than**) the **number of factor levels**. Indeed, the number of contrasts of interest must exceed the number of factor levels **by a considerable amount** before the Scheffé procedure becomes a better choice.

All three procedures are of the form

$$\text{Estimator} \pm \text{Multiplier} \cdot \text{SE}.$$

The only difference among the three procedures is the **multiplier**. In any given problem, one may then compute the Bonferroni and Scheffé multipliers (and, when appropriate, the Tukey multiplier), and select the smallest option.[51]

51: This is an appropriate choice because the multiplier does not depend on the observed data, only on the structure of the design and the desierd joint signficance level.

11.3.8 Model Validation

In our analysis of experimental results, we have primarily compared the average responses across various treatment groups. These comparisons have been conducted using an **overarall ANOVA test** or more targeted procedures based on **contrasts** and **pairwise comparisons**.

The foundation of these methods rests on the assumption that the data follows the model

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \dots, a; j = 1, \dots, n,$$

where μ symbolizes the global mean applicable to all observations, and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Note that the designed is assumed to be **balanced**.

The τ_i's are fixed but unknown coefficients, whereas the $\varepsilon_{i,j}$'s are independent normal random variables with constant but potentially unknown variance σ^2.

At its core, a model is essentially a **set of assumptions** – but we have done nothing so far to verify if (or ensure that) these assumptions are **reasonable**.

Specifically, we must verify three primary assumptions about the errors:

1. they are independent;
2. they are normally distributed, and
3. they have constant variance.

The model's analytical rigour and the consequent inferences largely depend on the extent to which the errors $\varepsilon_{i,j}$, adhere to these assumptions. Unfortunately, we **never observe the true errors** $\varepsilon_{i,j}$; the most accurate representation we possess for them are the **residuals** $e_{i,j}$, derived from the full model.

Validation must then be based on these **observable residuals** rather than the genuine errors. Due to the indirect nature of this process, diagnostics are sometimes **complicated**.

In any practical data set, it's almost inevitable that we encounter violations of **one or more** of these core assumptions. But there is reason for optimism: even in the face of **slight deviations**, the procedures can still yield **reasonable inferences**.

We now delve deeper into **diagnostics** and possible **remedial measures** for scenarios where the model assumptions are not met.

Residuals The (unobservable) errors are given by

$$\varepsilon_{i,j} = y_{i,j} - \mu - \tau_i.$$

After the model parameters have been estimated, we can compute the **residuals**

$$e_{i,j} = y_{i,j} - \hat{\mu} - \hat{\tau}_i = y_{i,j} - \hat{y}_{i,j} = y_{i,j} - \overline{y}_{i,\bullet}.$$

These residuals are often referred to as **raw residuals**.

The **error sum of squares** is simply

$$\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} e_i^2,$$

and the mean square error is

$$\text{MSE} = \frac{\text{SSE}}{N-a}, \quad \text{where } N = n_1 + \cdots + n_a.$$

At times, we may also use the **Studentized residuals**

$$d_{i,j} = \frac{e_{i,j}}{\sqrt{\text{MSE}}},$$

which we have discussed in Section 8.3.5.

Assessing Non-Normality The $qq-$plot, also known as the **normal probability plot**, is used to determine if the **errors align with a normal distribution**. The assessment is made by comparing the **observed quantiles** of the residuals with the **expected quantiles** from a normal distribution.

A **straight line** is indicative of errors following a normal distribution, albeit slight deviations at the tails are customary (and anticipated).[52] For **non-normal data**, the curvature of the plot provides insights into how the data varies from the normal distribution.

In the context of $qq-$plots, the choice between raw residuals and Studentized residuals is generally inconsequential.

52: See Section 8.3.5 for description and examples.

Assessing Non-Constant Variance We look for non-constant variance occurring when the responses within a treatment group all have the same variance σ_i^2, but the σ_i^2 differ between different groups.

This can be assessed visually by plotting the residuals, either $e_{i,j}$ or $d_{i,j}$, against the fitted values $\widehat{y}_{i,j}$. With constant variance, the **vertical dispersion** observed within the stripes of this plot remains **fairly consistent**; any **discernible pattern** in the residuals signals **non-constant variance**.

The most common deviations from constant variance are those where the residual variation depends on the mean. Usually we see variances increasing as the mean increases, but other patterns can occur.

Assessing Independence Serial dependence, also known as **autocorrelation**, is a common deviation from the assumption of independence in data analysis. This phenomenon emerges when consecutive data points, particularly those in **close temporal proximity**, exhibit excessive similarity (indicating **positive dependence**) or marked dissimilarity (suggesting **negative dependence**). Among these, positive dependence is the more prevalent form.

To visually discern the presence of serial dependence, analysts frequently use an **index plot**, which plots residuals on the vertical axis against their temporal sequence on the horizontal axis. By examining this plot, one can gauge the **degree of dependence**.

A **discernible drift** across the plot, for instance, is indicative of positive dependence. On the other hand, residuals **rapidly alternating** between positive and negative values, all the while centering around zero, typically suggest negative dependence.

Remedial Measures **Non-normality** and **non-constant variance** can sometimes be alleviated by transforming the response to a **different scale**:

- **skewness to the right** is often mitigated by employing a square root, logarithm, or other transformation to a power **smaller** than 1;
- in contrast, **skewness to the left** can be lessened by a square, cube, or other transformation to a power **greater** than 1;
- similarly, a prevalent method to address non-constant error variances is through the transformation of the **response variable**.

The **Box-Cox transformation** is particularly well-suited to such a situation, offering a suite of transformations indexed by a parameter λ:[53]

53: We also discuss it in Section 8.3.5.

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0. \end{cases}$$

The idea is to transform the data over a spectrum of λ values, perhaps between -3 and 3, and subsequently perform the ANOVA using $Y^(\lambda)$. We compute the sum of squared errors $\text{SSE}(\lambda)$ for every chosen λ.

Specifically, the optimal λ is the one that **maximizes the log-likelihood**

$$-\frac{N}{2}\log[\text{SSE}(\lambda)] + (\lambda - 1)\sum_{i=1}^{a}\sum_{j=1}^{n_i}\log(y_{i,j}).$$

And what can we do about the assumption of **data independence**? Unfortunately, straightforward methodologies to confront data dependence are limited. Advanced analytical techniques like **time series analysis** (see Chapter 9) and **spatial statistics** could be used to model such data, but these are beyond the scope of this chapter.

11.3.9 Power and Sample Size

So far, our focus has primarily been on analyzing experimental results. A new focus now emerges as we proceed: how do we determine an appropriate **sample size** for a completely randomized design?

Ideally, the sample size should be **as small as possible**, in order to optimize both time and costs, yet it must also be **sufficiently large** to fulfill the analytical requirements.[54]

54: Making an informed decision on the appropriate sample size requires the analysts to have some knowledge of the system being examined; we will discuss this further in Chapters 13 and 14.

We need two additional distributions to answer the original question:

- if $X_1, \ldots, X_a \sim \mathcal{N}(\mu_i, 1)$ are independent random variables, then $X_1^2 + \cdots + x_a^2$ follows a **non-central χ^2 distributions with a degrees of freedom and non-centrality parameter** $\delta = \mu_1^2 + \cdots + \mu_a^2$, denoted by $a\overline{X} \sim \chi_a^2(\delta)$;[55]
- if $X \sim \chi_n(\eta)$ and $Y \sim \chi_m$, then

$$F = \frac{X/n}{Y/m} \sim F_{n,m}(\eta),$$

  where $F_{n,m}(\eta)$ is the **non-central F distribution with n and m degrees of freedom and non-centrality parameter** η.

55: This definition is a generalization of the original definition of the (central) chi_a^2 distribution.

Recall that the statistic F_0 for testing

$$H_0 : \tau_1 = \cdots = \tau_a = 0 \quad \text{against} \quad H_1 : \tau_i \neq 0, \text{ for at least one } i$$

follows a distribution $F_{a-1,N-a}$ when H_0 is true. Under the alternative hypothesis H_1, this distribution assumption no longer holds.

Instead, the statistic F_0 follows a non-central $F_{a-1,N-a}(\delta^2)$, where

$$\delta^2 = n \sum_{i=1}^{a} \tau_i^2 / \sigma^2$$

is the non-centrality parameter.

This parameter essentially measures the extent to which the treatment means deviate from being equal, scaled relative to the variation of $\overline{y}_{i,\bullet}$, which is σ^2/n.

When computing the **power** for a specific sample size or determining the necessary **sample size** for a desired power, we have to use non-central F-distributions.

A potential complication arises from the fact that each value of the non-centrality parameter corresponds to a unique alternative distribution, meaning that there is a **distinct** non-central $F-$distribution for every possible non-centrality parameter value.

Example Suppose that $a = 5$ and that the treatment means are

$$\mu_1 = 11,\ \mu_2 = 12,\ \mu_3 = 15,\ \mu_4 = 18,\ \text{and}\ \mu_5 = 19.$$

From previous studies, we know that it is reasonable to expect that $\sigma^2 = 9$. What should N (or n) be in a balanced complete design if we use a test with $\alpha = 0.01$, assuming we want a power of at least $1 - \beta = 0.9$? □

In order to answer this question, we need to actually know **ahead of time** what the true individual values of $\mu_1, \ldots, \mu_5$ are, which may prove challenging; we also needed to specify a plausible value (or range of values) for σ^2.

An alternative approach is to determine the sample size N such that the **largest difference between** treatment means

$$\max\{\mu_i\} - \min\{\mu_i\}$$

is larger than a given value D.

If $D = \max\{\mu_i\} - \min\{\mu_i\}$, the non-centrality parameter is minimized when the other means are exactly in the middle of the interval

$$(\min\{\mu_i\}, \max\{\mu_i\}) = (\mu_{i_*}, \mu_{i^*}).$$

In that case, we would have

$$\tau_{i_*} = \mu_{i_*} - \mu = -\frac{D}{2} \quad \text{and} \quad \tau_{i^*} = \mu_{i^*} - \mu = \frac{D}{2},$$

and all other $\tau_i \equiv 0$, from which we conclude

$$\sum_{i=1}^{a} \tau_i^2 = 2(D/2)^2 = D^2/2.$$

It follows that

$$\delta^2_{\min} = nD^2/(2\sigma^2),$$

for a power equal to

$$P\left(F_{a-1,N-a}(\delta^2_{\min}) \geq F_{\alpha;a-1,N-a}\right).$$

Example With the data in the statement of the previous example, suppose that we have reason to believe that the largest difference between the treatment means is $D = 8$. Then

$$\delta^2_{\min} = n \cdot 8^2/(2 \cdot 9) = (32/9)n.$$

The power of the test is

$$P\left[F_{4,5(n-1)}((32/9)n) \geq F_{0.01;4,5(n-1)}\right].$$

We try different values of n, until we obtain a power which is at least 0.9.

```
for(n in c(2:9)){
delta.2.min = 32/9*n; df1 = 4; df2 = 5*(n-1); alpha = 0.01
crit = qf(0.01, df1=df1, df2=df2, lower.tail=FALSE)
print(c(n,
     pf(crit, df1=df1, df2=df2, ncp=delta.2.min,
     lower.tail=FALSE)))
}
```

```
[1] 2.0000000 0.0704121
[1] 3.0000000 0.2308392
[1] 4.0000000 0.4413316
[1] 5.0000000 0.6376441
[1] 6.0000000 0.7861772
[1] 7.0000000 0.8833954
[1] 8.0000000 0.9405001
[1] 9.0000000 0.9713123
```

With a value of $N = 40$ (i.e., with $n = 8$), the test's power is 94.1%. □

11.4 One-Way ANOVA with Random Effects

In the one-way ANOVA model from the previous section

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a;\ j = 1, \ldots, n,$$

where μ is the common mean to all observations and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$, the treatment effects $\tau_1, \ldots, \tau_a$ are viewed as **fixed**; this one-way ANOVA model is known as a **fixed-effect model**.

But in some situations, the fixed-effect model is not appropriate; in this section, we consider treatments that are drawn randomly from a population of **potential treatments**, leading to a **random effects model**.

Examples: Fixe vs. Random Effects

- A business operates 50 machines that produce cardboard boxes for canned products. To analyze the consistency in the carton's durability, they randomly select ten machines out of the 50 and manufacture 40 boxes from each. They distribute 400 batches of feedstock cardboard randomly among these ten machines. Subsequently, the boxes' strength is assessed. This approach follows a completely randomized design, encompassing ten treatment groups and 400 units.

  In this context, a fixed-effect model is not suitable since the goal is to understand and draw conclusions **about the entire population of machines**, not merely the ten we tested in the experiment – we want to make assertions for the entire population, not just the random subset we examined. Moreover, if the experiment was repeated with a fresh batch of 10 machines, we would most likely end up with a completely distinct group of machines (and so with different observations).

- Imagine a home gardener conducting a small experiment using 24 tomato plants, divided into 4 varieties with 6 plants each. These varieties have piqued the gardener's interest after occasional use over recent summers. Now, the gardener plans to compare these varieties within a 12′ x 8′ garden patch. Each plant is randomly placed in one of the 2′ x 2′ sections. In this scenario, the gardener's focus is solely on these specific four varieties, with no consideration for any other types. The emphasis is strictly on the varieties being tested, and nothing else, so we can use fixed effects.

  Suppose, on the other hand, the 4 tomato varieties were chosen at random from a broader population of tomato types. In this scenario, we'd be dealing with random effects. If the experiment were repeated with a different batch of 4 varieties, it would likely result in a completely distinct group of tomato varieties.

- To determine how proficiently Ontario students can read by the conclusion of first grade, imagine we randomly select 6 schools within the province. From each chosen school, a group of students is randomly picked to undergo a reading assessment. Given that these schools are a random sample from a broader group of interest (all the schools in Ontario), we are operating under a random effect model.

  If our sole focus was on the performance of those specific 6 schools, then a fixed-effect model would have been appropriate. However, that is not the intention in this scenario.

11.4.1 Estimation of Model Parameters

The **one-way ANOVA model with random effects** is given by

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a;\ j = 1, \ldots, n,$$

where

- μ is the global (or common) mean to all observations;
- $\tau_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_T^2)$, $i = 1, \ldots, a$;
- $\varepsilon_{i,j} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$, $\quad i = 1, \ldots, a,\ j = 1, \ldots, n_i$;
- τ_i and $\varepsilon_{i,j}$ are independent.

It follows that

$$\begin{aligned}
\mathrm{E}(\tau_i) = 0, &\quad \mathrm{Var}(\tau_i) = \sigma_T^2, \quad \mathrm{Cov}(\tau_i, \tau_{i'}) = 0, i \neq i';\\
\mathrm{E}(\varepsilon_{i,j}) = 0, &\quad \mathrm{Var}(\varepsilon_{i,j}) = \sigma^2, \quad \text{and}\\
\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = 0, &\quad \text{except when } i = i' \text{ and } j = j';\\
\mathrm{Cov}(\tau_i, \varepsilon_{i',j'}) = 0, &\quad \text{for all } i \text{ and } i'.
\end{aligned}$$

Consequently, we have

$$\mathrm{E}(y_{i,j}) = \mathrm{E}(y_{i,j} \mid \tau_i) = \mathrm{E}(\mu + \tau_i + \varepsilon_{i,j} \mid \tau_i) = \mathrm{E}(\mu + \tau_i) = \mu$$

and

$$\mathrm{Var}(y_{i,j}) = \mathrm{Var}(y_{i,j} \mid \tau_i) + \mathrm{VarE}(y_{i,j} \mid \tau_i) = \sigma_T^2 + \sigma^2.$$

Although the τ_i's and the $\varepsilon_{i,j}$'s are **uncorrelated**, the $y_{i,j}$'s are **correlated**. Indeed, for those in the **same treatment class**, we have

$$\text{Cov}(y_{i,j}, y_{i,j'}) = \text{Cov}(\mu + \tau_i + \varepsilon_{i,j}, \mu + \tau_i + \varepsilon_{i,j'}) = \sigma_T^2, \text{ for } j \neq j',$$

whereas for those in **different treatment classes**, we have

$$\text{Cov}(y_{i,j}, y_{i',j'}) = \text{Cov}(\mu + \tau_i + \varepsilon_{i,j}, \mu + \tau_{i'} + \varepsilon_{i',j'}) = 0, \text{ for } i \neq i'.$$

Estimation of Parameters The **intra-class correlation coefficient** is defined as

$$\rho = \frac{\text{Cov}(y_{i,j}, y_{i,j'})}{\sqrt{\text{Var}(y_{i,j})}\sqrt{\text{Var}(y_{i,j'})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}, \quad \text{if } j \neq j'.$$

It is a measure of the correlation between two observations from the same factor level (or class); the parameters σ_T^2 and σ^2 are the **variance components**.

In practice, there are 4 family of parameters to be estimated and/or predicted: the common mean μ, the treatment effects τ_i, and the variance components σ_T^2 and σ^2.

The common mean and the variance components are **fixed parameters**; these we seek to **estimate**. The treatment effects are random variables, these we seek to **predict**.

11.4.2 Analysis of Variance

In the one-way fixed-effects ANOVA model, we considered the overall test of hypothesis $H_0 : \tau_1 = \cdots = \tau_a = 0$. In the context of a random- effects ANOVA model, this hypothesis is nonsensical as the τ_i's are **random variables**.

Instead, we look to test if the factor (treatment) has an impact on the **variability of the response** Y. The null hypothesis is then expressed as $H_0 : \sigma_T^2 = 0$. The alternative stipulates that the factor has an effect on the variability of the response Y, which we express as $H_1 : \sigma_T^2 > 0$.

In effect, if H_0 is valid, then all the τ_i's are equal, whereas if H_1 holds, then at least two of the τ_i's differ.

Despite the fact that the fixed-effects model is emphatically not equivalent to the random-effects model, their **analysis of variance** for a single-factor study (one-way classification) is conducted in similar fashions.

We can show (see Exercises) that

$$\text{E(MSE)} = \sigma^2 \quad \text{and} \quad \text{E(MSA)} = \sigma^2 + n\sigma_T^2.$$

It then follows that MSE and MSA have the same expectation σ^2 if $\sigma_T^2 = 0$. If $\sigma_T^2 > 0$, on the other hand, then E(MSA) > E(MSE) as $n > 0$.

Therefore, we would **reject** H_0 **at significance level** α if

$$F_0 = \frac{\text{MSA}}{\text{MSE}} > F_{\alpha;a-1,N-a}.$$

To understand why we compare the observed test statistic F_0 to critical values of the $F_{a-1,N-a}$ distribution, we first note that

$$\overline{y}_{i,\bullet} = \frac{1}{N}\sum_{j=1}^{n} y_{i,j} = \mu + \tau_i + \overline{\varepsilon}_{i,\bullet}, \quad \text{where}$$
$$\overline{\varepsilon}_{i,\bullet} = \sum_{j=1}^{n} \frac{\varepsilon_{i,j}}{n} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right),$$

from which it follows that

$$\overline{y}_{i,\bullet} \sim \mathcal{N}\left(\mu, \sigma_T^2 + \frac{\sigma^2}{n}\right), \quad i = 1, \ldots, a.$$

The random variables $\overline{y}_{i,\bullet}$ being i.i.d., we must then have

$$\frac{(a-1)\text{MSA}}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{a-1}.$$

In the context of a balanced design, SSA can be expressed as

$$\begin{aligned}\text{SSA} &= \sum_{i=1}^{a} n_i(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 = n\sum_{i=1}^{a}(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 \\ &= n\sum_{i=1}^{a}\left[(\tau_i - \overline{\tau}_\bullet) + (\overline{\varepsilon}_{i,\bullet} - \overline{\varepsilon}_{\bullet,\bullet})\right]^2,\end{aligned}$$

where

$$\overline{\tau}_\bullet = \sum_{i=1}^{a}\frac{\tau_i}{a} \quad \text{and} \quad \overline{\varepsilon}_{\bullet,\bullet} = \sum_{i=1}^{a}\frac{\overline{\varepsilon}_{i,\bullet}}{a}.$$

On the other hand, we have

$$(n-1)s_i^2 = \sum_{j=1}^{n}(y_{i,j} - \overline{y}_{i,\bullet})^2 = \sum_{j=1}^{n}(\varepsilon_{i,j} - \overline{\varepsilon}_{i,\bullet})^2.$$

According to the i.i.d. condition,

$$\frac{(n-1)s_i^2}{\sigma^2} \sim \chi^2_{n-1}$$

independently for all $i = 1, \ldots .a$. As a result, we then have

$$\frac{(N-a)\text{MSE}}{\sigma^2} = (N-a)\sum_{i=1}^{a}\frac{s_i^2}{\sigma^2} = \sum_{i=1}^{a}\frac{(n-1)s_i^2}{\sigma^2} \sim \chi^2_{N-a}.$$

Thus, MSA only depends on $\{\tau_1, \ldots, \tau_a\}$ and $\{\overline{\varepsilon}_{1,\bullet}, \ldots, \overline{\varepsilon}_{a,\bullet}\}$ and MSE only depends on $\left\{s_1^2, \ldots, s_a^2\right\}$. But the sets $\{\tau_1, \ldots, \tau_a\}$ and $\left\{s_1^2, \ldots, s_a^2\right\}$ are independent, as are the sets $\{\overline{\varepsilon}_{1,\bullet}, \ldots, \overline{\varepsilon}_{a,\bullet}\}$ and $\left\{s_1^2, \ldots, s_a^2\right\}$. Therefore, MSA and MSE are independent and so we have, by definition of the F distribution,

$$\frac{\sigma^2}{\sigma^2 + n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} \sim F_{a-1,N-a}.$$

Under $H_0 : \sigma_T^2 = 0$, this collapses to the decision protocol presented above.

11.4.3 Inference on σ^2, σ_T^2, and μ

As was the case with the fixed-effects model, we can conduct inference on the model parameters.[56]

56: Assuming, as before, a balanced model.

Confidence interval for σ^2 and σ_T^2 As before, MSE $= \widehat{\sigma^2}$ is an **unbiased estimator** of σ^2. Since $(N-a)\text{MSE}/\sigma^2 \sim \chi^2_{N-a}$, it follows from that we obtain a $100(1-\alpha)\%$ **confidence interval for** σ^2 *via*

$$\left[\frac{(N-a)\text{MSE}}{\chi^2_{\alpha/2;N-a}}, \frac{(N-a)\text{MSE}}{\chi^2_{1-\alpha/2;N-a}}\right].$$

But we also have

$$E\left(\frac{\text{MSA}-\text{MSE}}{n}\right) = \frac{\sigma^2}{n} - \frac{\sigma^2+n\sigma_T^2}{n} = \sigma_T^2;$$

consequently, $(\text{MSA}-\text{MSE})/n$ is an **unbiased estimator** of σ_T^2.

However, nothing precludes this estimator to take on **negative** values, which may occur when MSA $<$ MSE.[57] To overcome this issue, we use the **truncated estimator**

57: This can occur when we are evaluating MSE and MSA from actual data (not their expectations).

$$\hat{\sigma}_T^2 = \begin{cases} (\text{MSA}-\text{MSE})/n, & \text{if MSA} \geq \text{MSE}, \\ 0, & \text{otherwise.} \end{cases}$$

The distribution of $\hat{\sigma}_T^2$ is not simple since it is expressed as the linear combination of two chi-square distributions. As a result, we cannot derive an exact confidence interval for σ_T^2; we will have to settle for an **approximate confidence interval for** σ_T^2.

However, we can construct an **exact** confidence interval for the intra-class correlation coefficient $\rho = \sigma_T^2/(\sigma_T^2+\sigma^2)$. Indeed,

$$\begin{aligned}
1-\alpha &= P\left(F_{1-\alpha/2;a-1,N-a} \leq \frac{\sigma^2}{\sigma^2+n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} \leq F_{\alpha/2;a-1,N-a}\right) \\
&= P\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right) \leq \frac{\sigma_T^2}{\sigma^2} \leq \frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right) \\
&= P\left(g\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right)\right) \leq g\left(\frac{\sigma_T^2}{\sigma^2}\right) = \frac{\sigma_T^2}{\sigma_T^2+\sigma^2} \leq g\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right)\right),
\end{aligned}$$

where $g(t) = t/(t+1)$ is an **increasing** function. Therefore, a $100(1-\alpha)\%$ confidence interval for $\sigma_T^2/(\sigma_T^2+\sigma^2)$ is obtained *via*:

$$\left[\frac{\text{MSA}-F_{\alpha/2;a-1,N-a}\text{MSE}}{\text{MSA}+(n-1)F_{\alpha/2;a-1,N-a}\text{MSE}}, \frac{\text{MSA}-F_{1-\alpha/2;a-1,N-a}\text{MSE}}{\text{MSA}+(n-1)F_{1-\alpha/2;a-1,N-a}\text{MSE}}\right].$$

When $N-a$ is large, the estimator MSE of σ^2 becomes more precise and we can write $\sigma^2 \approx \text{MSE}$.

It follows that

$$1-\alpha \approx P\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right) \le \frac{\sigma_T^2}{\text{MSE}} \le \frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right)$$
$$= P\left(\frac{1}{n}\left(\frac{\text{MSA}}{F_{\alpha/2;a-1,N-a}}-\text{MSE}\right) \le \sigma_T^2 \le \frac{1}{n}\left(\frac{\text{MSA}}{F_{1-\alpha/2;a-1,N-a}}-\text{MSE}\right)\right),$$

which provides an approximate $100(1-\alpha)\%$ confidence interval for σ_T^2.

Confidence interval for μ Inferences about the global mean are simpler to obtain. The expression

$$\widehat{\mu} = \overline{y}_{\bullet,\bullet} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{i,j}$$

provides an unbiased estimator of μ. Its variance is given by

$$\text{Var}(\widehat{\mu}) = \frac{n\sigma_T^2+\sigma^2}{N};$$

an unbiased estimator of which is given by

$$\widehat{\text{Var}}(\widehat{\mu}) = \frac{\text{MSA}}{N}.$$

It follows that we can find a $100(1-\alpha)\%$ confidence interval for μ *via*

$$\overline{y}_{\bullet,\bullet} \pm t_{\alpha/2;a-1}\sqrt{\frac{\text{MSA}}{N}}.$$

11.4.4 Power of a Test

In the case of the $F-$test at significance level α for a one-way random-effects model, the power of the test

$$H_0 : \sigma_T^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_T^2 \neq 0$$

is:

$$P(\Delta) = P\left(\frac{\text{MSA}}{\text{MSE}} > F_{\alpha;a-1,N-a} \,\middle|\, \sigma_T^2 \neq 0\right)$$
$$= P\left(\frac{\sigma^2}{\sigma^2+n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} > \frac{\sigma^2}{\sigma^2+n\sigma_T^2}F_{\alpha;a-1,N-a}\right)$$
$$= P\left(F_{a-1,N-a} > \frac{1}{1+\Delta}F_{\alpha;a-1,N-a}\right),$$

where $\Delta = n\sigma_T^2/\sigma^2$.

11.5 Randomized Complete Block Designs

As the variance of the experimental error σ^2 **increases**, the corresponding confidence intervals get **longer** and the power of tests **decreases**.

All other things being equal, we would thus prefer to conduct experiments with units that are **homogeneous** so that σ^2 is as small as possible.

We can achieve this through **variance-reduction designs**, which almost exclusively use **blocking**. A **block of units** is a collection of units that are homogeneous in **some sense** – field plots located in the same general area, or units that came from a single supplier, say.

These similarities in the units themselves lead us to anticipate that units **within a block** may have **similar responses**.

When constructing blocks, the goal is to achieve homogeneity of the units within blocks, with the caveat that units in different blocks may be **dissimilar**.[58] The primary purpose of blocking is to remove or isolate the **block-to-block variability**. This helps ensure that this variability does not overshadow or **mask the treatment effects** under consideration.

58: Compare with the notion of stratified random sampling in Section 10.4.

A notable experimental design that makes use of this concept is the **Randomized Complete Block Design** (RCBD). This design is structured for comparing a treatments **across** b blocks. In this setup, treatments are **randomly assigned** to experimental units within a block – each treatment appears **exactly once** in every block. If a RCBD integrates a treatments within each of b blocks, then the **total number of observations** would be $N = ab$.

Randomized block designs can be seen as an **extension** of the paired-difference designs that were discussed in Section 11.2.

Examples: Randomized Complete Block Design

- A production supervisor is keen on comparing the mean assembly times of operators using three distinct methods: A, B, and C. Given the anticipated variation in assembly times across different operators, the supervisor employs an RCDB for the comparison.

  Specifically, six assembly-line operators are selected, each representing a block. Each operator is tasked with assembling the item three times, once for every method. The importance of the sequence in which the methods are applied is recognized, as factors like fatigue or heightened dexterity might influence the results. Therefore, every operator is assigned a randomized sequence of the three methods. For instance:

  - Operator 1 first uses method A, proceeds to B, and finishes with C.
  - Operator 2 first uses method A, proceeds to C, and finishes with B.
  - Operator 3 first uses method B, proceeds to A, and finishes with C.
  - Operator 4 first uses method B, proceeds to C, and finishes with A.
  - Operator 5 first uses method C, proceeds to A, and finishes with B.
  - Operator 6 first uses method C, proceeds to B, and finishes with A.

- The credit card industry is engaged in an intense competition for cardholders. Each company designs its unique, intricate reward and fee structure in an attempt to attract customers. Notably, the benefits or costs associated with a credit card can vary significantly depending on the cardholder's monthly spending.

  To investigate this, a consumer watchdog group set out to compare the average rewards or fees of four different credit card companies (A, B, C, D). They used three distinct spending levels as blocks:

  – low spending – \$500 per month,
  – middle spending – \$2,500 per month, and
  – high spending – \$10,000 per month.

  If the rewards are not monetary in nature, the watchdog group has first converted them to a monetary value. The average monthly rewards, as quoted by the credit card companies for cardholders across these spending levels, are presented in the table below.

Rewards	**Credit Card Company**			
Spending Level	**A** ($i = 1$)	**B** ($i = 2$)	**C** ($i = 3$)	**D** ($i = 4$)
Low ($j = 1$)	30	27	34	26
Middle ($j = 2$)	68	76	65	67
High ($j = 3$)	304	322	308	296

11.5.1 Analysis of Variance

In an RCBD, we consider two key factors: **treatments** and **blocks**, both of which play a significant role in influencing the response. Let $y_{i,j}$ represent the response when the ith treatment is applied within the jth block. The underlying RCBD is described *via*:

$$y_{i,j} = \mu + \tau_i + \beta_j + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, b,$$

where the error terms $\varepsilon_{i,j}$ are independent random variables from a $\mathcal{N}(0, \sigma^2)$ distribution.

In this model, the parameter μ represents the global effect, while τ_i denotes the treatment effect for the ith treatment level, and β_j indicates the effect associated with the jth block.[59]

Both treatments and blocks are regarded as **fixed factors**; the expected value of the response can thus be expressed as:

$$\mathrm{E}(y_{i,j}) = \mu + \tau_i + \beta_j.$$

Just as in the one-way (single-factor) fixed-effect experimental design discussed previously, the RCBD model is **over-parameterized**.[60] The primary aim is to test the **uniformity of the treatment means**, effectively examining the presence or absence of an effect for Factor A.

59: We also refer to the treatment as the **first factor** (or Factor A), and to blocking as the **second factor** (or Factor B).

60: We can bypass this problem by enforcing constraints on the treatment and block effects:

$$\sum_{i=1}^{a} \tau_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0.$$

Formally, we test for

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0, \quad \text{against} \quad H_1 : \tau_i \neq 0 \quad \text{for at least one } i.$$

The **totals** for the ith treatment, the jth block, and the overall total of the $N = ab$ observations are given, respectively, by

$$y_{i,\bullet} = \sum_{j=1}^{b} y_{i,j}, \quad i = 1, \ldots, a, \quad y_{\bullet,j} = \sum_{i=1}^{a} y_{i,j}, \quad j = 1, \ldots, b$$

$$y_{\bullet,\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{i,j} = \sum_{i=1}^{a} y_{i,\bullet} = \sum_{j=1}^{b} y_{\bullet,j}.$$

Similarly, we define the various **means**

$$\overline{y}_{i,\bullet} = \frac{y_{i,\bullet}}{b}, \quad \overline{y}_{\bullet,j} = \frac{y_{\bullet,j}}{a}, \quad \text{and} \quad \overline{y}_{\bullet,\bullet} = \frac{y_{\bullet,\bullet}}{N}.$$

Example (cont.) In the credit card example from earlier in the section, the totals and means are given in the table below.

Rewards	**Credit Card Company**					
Spending Level	**A** $(i = 1)$	**B** $(i = 2)$	**C** $(i = 3)$	**D** $(i = 4)$	**Totals** $y_{\bullet,j}$	**Means** $\overline{y}_{\bullet,j}$
Low $(j = 1)$	30	27	34	26	117	29.25
Middle $(j = 2)$	68	76	65	67	276	69
High $(j = 3)$	304	322	308	296	1230	307.5
Totals $y_{i,\bullet}$	402	425	407	389	1623	
Means $\overline{y}_{i,\bullet}$	134	141.7	135.7	129.7		135.25

The **total sum of square** (SST) can be expressed as the sum of three sums of squares:

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \overline{y}_{\bullet,\bullet})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left[(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}) + (\overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet}) + (y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet})\right]^2 \\ &= b\sum_{i=1}^{a}(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 + a\sum_{j=1}^{b}(\overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet})^2,\end{aligned}$$

or, using the customary symbols (along with the corresponding degrees of freedom):

$$\begin{aligned}\text{SST} &= \text{SSA} + \text{SSB} + \text{SSE} \\ N - 1 &= (a-1) + (b-1) + (a-1)(b-1) = ab - 1\end{aligned}$$

There are equivalent formulas (which are slightly easier to use) for the sums of squares:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{a}\sum_{j=1}^{b} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \\
\text{SSA} &= \frac{1}{b}\sum_{i=1}^{a} y_{i,\bullet}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \\
\text{SSB} &= \frac{1}{a}\sum_{j=1}^{b} y_{\bullet,j}^2 - \frac{y_{\bullet,\bullet}^2}{N},
\end{aligned}$$

Finally, SSE is obtained as

$$\text{SSE} = \text{SST} - \text{SSA} - \text{SSB}.$$

It can be shown that

$$\frac{\text{SSA}}{\sigma^2} \sim \chi^2_{a-1}\left(b\sum_{i=1}^{a}\frac{\tau_i^2}{\sigma^2}\right), \quad \frac{\text{SSB}}{\sigma^2} \sim \chi^2_{b-1}\left(a\sum_{j=1}^{b}\frac{\beta_j^2}{\sigma^2}\right),$$

and

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)}.$$

As has been the case throughout, we can also show that the three sums of squares SSA, SSB, and SSE are **mutually independent**. The corresponding mean squares are obtained in the usual way:

$$\text{MSA} = \frac{\text{SSA}}{a-1}, \quad \text{MSB} = \frac{\text{SSB}}{b-1}, \quad \text{and} \quad \text{MSE} = \frac{\text{SSE}}{(a-1)(b-1)}.$$

We can show (see Exercises) that

$$\begin{aligned}
\text{E(MSA)} &= \sigma^2 + \frac{b}{a-1}\sum_{i=1}^{a}\tau_i^2, \\
\text{E(MSB)} &= \sigma^2 + \frac{a}{b-1}\sum_{j=1}^{b}\beta_j^2, \\
\text{E(MSE)} &= \sigma^2.
\end{aligned}$$

We can test for the absence of a treatment effect (Factor A) by pitting $H_0 : \tau_1 = \cdots = \tau_a = 0$ against $H_1 : \tau_i \neq 0$ for at least one i, using the test statistics

$$F_0 = \frac{\text{MSA}}{\text{MSE}},$$

which follows an $F_{a-1,(a-1)(b-1)}$ distribution under H_0.

All of this is summarized in Table 11.19.[61]

61: A "large" value of the ratio MSB/MSE implies that blocking was a good strategy.

Source	**SS**	**df**	**MS**	$\mathbf{F_0}$
Treatment	SSA	$a-1$	MSA	$F_0 = \text{MSA/MSE}$
Block	SSB	$b-1$	MSB	
Error	SSE	$(a-1)(b-1)$	MSE	
Total	SST	$N-1$		

Table 11.19: ANOVA table for the equality of the treatment means τ_i in a two-factor randomized complete block design.

Example (cont.) In the credit card example from earlier in the section, we have a randomized block design with $b = 3$ spending levels (blocks) and $a = 4$ companies (treatments), so there are $N = ab = 12$ observations.

We start by loading the data.

```
content.1 <- c(30, 27, 34, 26)
content.2 <- c(68, 76, 65, 67)
content.3 <- c(304, 322, 308, 296)
data <- data.frame(rbind(content.1, content.2, content.3))
rownames(data) <- c("Low", "Middle", "High")
colnames(data) <- c("A", "B", "C", "D")
row.totals <- rowSums(data)
row.means <- rowMeans(data)
data <- cbind(data, row.totals, row.means)
col.totals <- colSums(data)
col.means <- colMeans(data)
data <- rbind(data, col.totals, col.means)
rownames(data) <- c("Low", "Middle", "High", "col.totals",
                    "col.means")
data[4,6] <- NA; data[5,5] <- NA
```

```
             A        B        C        D row.totals row.means
Low         30  27.0000  34.0000  26.0000        117     29.25
Middle      68  76.0000  65.0000  67.0000        276     69.00
High       304 322.0000 308.0000 296.0000       1230    307.50
col.totals 402 425.0000 407.0000 389.0000       1623
col.means  134 141.6667 135.6667 129.6667                135.25
```

We compute the necessary quantities and place them in the ANOVA table.

```
a = ncol(content)
b = nrow(content)
N = a*b
grand.mean = data[b+2,a+2]
SST = sum((data[c(1:b),c(1:a)]-grand.mean)^2)
SSA = b * sum((data[b+2,c(1:a)]-grand.mean)^2)
SSB = a * sum((data[c(1:b),a+2]-grand.mean)^2)
SSE = SST - SSA - SSB
ANOVA = as.data.frame(cbind(c(SSA,SSB,SSE,SST),
            c(a-1, b-1, (a-1)*(b-1), N-1),
            c(SSA/(a-1),SSB/(b-1),SSE/((a-1)*(b-1)),0),
            c((SSA/(a-1))/(SSE/((a-1)*(b-1))),(SSB/(b-1))/(SSE/((a-1)*(b-1))),0,0)))
rownames(ANOVA) = c("Treatment", "Block", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
                SS df          MS         F0
Treatment    222.25  3    74.08333    1.84058
Block     181180.50  2 90590.25000 2250.68944
Error        241.50  6    40.25000
Total     181644.25 11
```

At significance level $\alpha = 0.05$, the critical value of $F_{4-1,(4-1)(3-1)} = F_{3,6}$ is given below.

```
qf(0.05, df1 = a-1, df2 = (a-1)*(b-1), lower.tail=FALSE)
```

```
[1] 4.757063
```

We see that $F_0 = \text{MSA/MSE} = 1.84 < F_{0.05;3,6} = 4.76$; therefore, the results do not show a significant difference in the treatment means. That is, there is insufficient evidence to indicate a difference in the credit card companies' monthly rewards.[62]

62: The ratio MSB/MSE is quite large, which suggests that blocking is effective, even if we cannot say that the treatment is so.

11.5.2 Estimation of Model Parameters

The RCBD model parameters are the grand mean μ, the treatment effects τ_i, and the blocking effect β_j, which can be estimated from the data as follows.

We seek to minimize the sum of squares errors:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\varepsilon_{i,j}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \mu - \tau_i - \beta_j)^2.$$

We determine the model values of μ, τ_i and β_j by differentiating the expression above, setting the gradient to 0, and solving for the parameters. In the RCBD context, this leads to:

$$\begin{aligned}
\mu &: \ -2\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \\
\tau_i &: \ -2\sum_{j=1}^{b}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \quad i = 1, \ldots, a, \\
\beta_j &: \ -2\sum_{i=1}^{a}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \quad j = 1, \ldots, b.
\end{aligned}$$

After some simplifications, we obtain the following system of linear equations:

$$\begin{aligned}
\mu &: \ N\hat{\mu} = y_{\bullet,\bullet}, \\
\tau_i &: \ b\hat{\mu} + b\hat{\tau}_i = y_{i,\bullet}, \quad i = 1, \ldots, a, \\
\beta_j &: \ a\hat{\mu} + a\hat{\beta}_j = y_{\bullet,j}, \quad j = 1, \ldots, b,
\end{aligned}$$

whose solution is

$$\hat{\mu} = \overline{y}_{\bullet,\bullet}, \quad \hat{\tau}_i = \overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}, \quad \hat{\beta}_j = \overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet}.$$

11.5.3 Multiple Comparisons

We can compare two treatments i and i', by looking at the difference of treatments $\tau_i - \tau_{i'}$, which we estimate *via* $\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}$.

The variance of $\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}$ is given by

$$\text{Var}(\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) = \sigma^2 \cdot \frac{2}{b}.$$

We obtain an $100(1-\alpha)\%$ confidence interval for $\tau_i - \tau_{i'}$ in the usual manner:

$$\tau_i - \tau_{i'} : \ (\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) \pm t_{\alpha/2;(a-1)(b-1)} \sqrt{\text{MSE}} \sqrt{\frac{2}{b}}.$$

For simultaneous confidence intervals, we must use a modification (as in Section 11.3.7). If we use **Tukey's method**, for instance, the confidence interval with family confidence $100(1-\alpha)\%$ becomes

$$\tau_i - \tau_{i'} : \ (\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) \pm q_{\alpha;a,(a-1)(b-1)} \sqrt{\text{MSE}} \sqrt{\frac{1}{b}}.$$

11.5.4 Power and Sample Size

Whether or not Factor A has an effect, the distribution of the test statistic F_0 is a non-central $F_{a-1,(a-1)(b-1)}(\delta^2)$, with non-centrality parameter

$$\delta^2 = b \sum_{i=1}^{a} \tau_i^2 / \sigma^2.$$

To determine the **sample size**, we can use an approach similar to the one described in Section 11.3.9.

The differences between the treatment effects are $\tau_i - \tau_{i'}$; the largest difference between the treatment averages is thus

$$D = \max\{\tau_i\} - \min\{\tau_i\}.$$

The **minimal** non-centrality parameter is thus

$$\delta^2_{\text{min}} = bD^2/(2\sigma^2),$$

which yields a **test power** of

$$P\left(F_{a-1,(a-1)(b-1)}(\delta^2_{\text{min}}) \geq F_{\alpha;a-1,(a-1)(b-1)}\right).$$

11.5.5 Model Validation

As in the previously studied design, three basic assumptions about errors must be checked: **independence**, **normality**, and **homoscedasticity**. As before, we use the residuals to verify whether the assumptions seem reasonable. In the RCBD **predicted responses** are

$$\hat{y}_{i,j} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j = \overline{y}_{i,\bullet} + \overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet};$$

their **residuals** are thus

$$e_{i,j} = y_{i,j} - \hat{y}_{i,j} = y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet}.$$

11.6 Factorial Designs

In our discussions up to this point, we primarily focused on the foundational problem of understanding how a **single independent factor** influences the **response**. However, it's not uncommon in research to encounter situations where the interest lies in studying the **combined effects** of **multiple independent variables** on a given response. We call such experimental setups, where two or more factors are simultaneously investigated, **factorial designs**.

Consider an example where researchers wish to determine the effect of sleep deprivation on student test performance. If the study only revolves around the sleep factor and the test performance, it is a **simple (one-way) experiment**. But we can add a twist: what if the researchers also wants to know whether the impacts of sleep deprivation vary between high school and university students? This introduces a second factor, school level,[63] into the study, turning it into a factorial design.

63: Which is presumably linked to age.

Factorial designs can vary in their complexity. A frequently encountered type is the 2×2 factorial design, where **two factors** are being analyzed, and each factor has **two distinct levels**. The numeric representation of a factorial design offers quick insights: the number of digits indicates the **number of factors**, while the value of each number shows **how many levels** the corresponding factor has. For instance, a 4×3 factorial design consists of two factors, with the first having four levels and the second comprising three levels. Extending this understanding, a $2 \times 2 \times 2$ factorial design would mean the experiment has three factors, each of which having two levels.

11.6.1 Two-Way Factorial Experiments

We start by looking into **two-factor designs**. The data from a two-way factorial design can be illustratively showcased using a table, as in Table 11.21.

	B_1	B_2	B_3
A_1	$y_{1,1,1}, \ldots, y_{1,1,n}$	$y_{1,2,1}, \ldots, y_{1,2,n}$	$y_{1,3,1}, \ldots, y_{1,3,n}$
A_2	$y_{2,1,1}, \ldots, y_{2,1,n}$	$y_{2,2,1}, \ldots, y_{2,2,n}$	$y_{2,3,1}, \ldots, y_{2,3,n}$
A_3	$y_{3,1,1}, \ldots, y_{3,1,n}$	$y_{3,2,1}, \ldots, y_{3,2,n}$	$y_{3,3,1}, \ldots, y_{3,3,n}$
A_4	$y_{4,1,1}, \ldots, y_{4,1,n}$	$y_{4,2,1}, \ldots, y_{4,2,n}$	$y_{4,3,1}, \ldots, y_{4,3,n}$

Table 11.21: 4×3 factorial design treatment structure, with n observations per cell.

In this representation, **rows** align with the levels of one specific factor (designated as Factor A), while columns represent the levels of the second factor (Factor B).

In that design, there are $4 \times 3 = 12$ total treatments. In **balanced** factorial designs, the number of observations n per unique combination of factor levels (which we also call a **cell**) is the same value across **all combinations**. For the current discussion, we assume that the collected data is balanced, n responses to a cell.

Assume that we are working with an $a \times b$ two-way design; there are $N = abn$ observations in total. We refer to the kth response in the $(i, j)-$cell by $y_{i,j,k}$.

By similarity to the one-way design, we adopt the following notation:

$$y_{i,\bullet,\bullet} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{i,\bullet,\bullet} = \frac{y_{i,\bullet,\bullet}}{bn};$$

$$y_{\bullet,j,\bullet} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{\bullet,j,\bullet} = \frac{y_{\bullet,j,\bullet}}{an};$$

$$y_{i,j,\bullet} = \sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{i,j,\bullet} = \frac{y_{i,j,\bullet}}{n};$$

$$y_{\bullet,\bullet,\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{\bullet,\bullet,\bullet} = \frac{y_{\bullet,\bullet,\bullet}}{N}.$$

Example We are interested in determining if a medicated agent can help reduce inflammation among athletes. 6000 college-level athletes are assigned to 30 lots of 200 athletes each.

The 30 lots are divided at random into ten groups of three lots each, with each group receiving a different treatment.

A treatment is factorial combination of the medication dosage (Factor A, with two levels), and when the medication is applied (Factor B, with five levels: 1 hour after a game, immediately after the game, during the game, immediately before the game, 1 hour before game).

In each lot, the response is the number of athletes who experience inflammation at some point within a 24-hour period after the game.

Cases	**Application Period**				
Dosage	**1**	**2**	**3**	**4**	**5**
Low	10	6	8	12	19
	7	18	36	29	46
	9	16	19	35	37
High	3	7	9	10	15
	4	4	10	10	26
	7	0	4	0	10

The data is summarized below.

Cases	**Application Period**					
Dosage	**1**	**2**	**3**	**4**	**5**	$y_{i,\bullet,\bullet}$
Low	26	40	63	76	102	307
High	14	11	23	20	51	119
$y_{\bullet,j,\bullet}$	40	51	86	96	153	426

We will discuss how to estimate the two-way factorial design model parameters shortly. □

Typically, we are interested in the **treatment effects** and **interaction effects**. The mathematical representation of a two-way factorial experiment is given by the model:

$$y_{i,j,k} = \mu_{i,j} + \varepsilon_{i,j,k}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n;$$

the subscripts i and j serve as indices for the treatment levels A and B, respectively.

We can re-write the treatment effects as follows:

$$\begin{aligned}\mu_{i,j} &= \overline{\mu}_{\bullet,\bullet} + (\overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,\bullet}) + (\overline{\mu}_{\bullet,j} - \overline{\mu}_{\bullet,\bullet}) + (\mu_{i,j} - \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,j} + \overline{\mu}_{\bullet,\bullet}) \\ &= \mu + \tau_i + \beta_j + (\tau\beta)_{i,j}\end{aligned}$$

By adopting this perspective, we can reformulate the model as:

$$y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + \varepsilon_{i,j,k}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n.$$

As always, we incorporate constraints to avoid an over-parametrized model:

$$\sum_{i=1}^{a} \tau_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} (\tau\beta)_{i,j} = 0, \quad \sum_{j=1}^{b} (\tau\beta)_{i,j} = 0.$$

The **main treatment effects** are represented by τ_i (Factor A) and β_j (Factor B); the **interaction effect** by $(\tau\beta)_{i,j}$. This interaction plays a pivotal role in understanding the experiment's nuances.

The **row effects** tells us how the response changes as we transition from one row to the next, averaged across all columns. In contrast, the **column effect** tells us how the response changes as we move from once column to the next, averaged across all rows.

The **interaction effects** tell us how the change in response depends on columns when moving between rows, or how the change in response depends on rows when moving between columns. An interaction term between Factor A and Factor B means that the change in mean response going from level i_1 of Factor A to level i_2 of Factor A depends on the level of Factor B under consideration.[64]

64: We cannot simply say that changing the level of Factor A changes the response by a given amount; we may need a different amount of change for each level of Factor B.

Advantages Factorial experiments present several advantages.

- When the factors do not interact, factorial experiments are more efficient than one-at-a-time experiments, as the units can be used to assess the (main) effects for both factors. Units in a one-at-a-time experiment can only be used to assess the effects of one factor.
- When the factors interact, factorial experiments can estimate the interaction. One-at-at-time experiments cannot estimate interaction. Use of one-at-a-time experiments in the presence of interaction can lead to serious misunderstanding of how the response varies as a function of the factors.

When there is no interaction, then the **main treatment effects alone** are suffcient to describe the means of the response – such a model is said to be **additive**.

Estimation of Model Parameters As before, we seek to minimize the sum of squared residuals

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{i,j,k}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\mu-\tau_i-\beta_j-\gamma_{i,j})^2,$$

where we write $\gamma_{i,j}$ for $(\tau\beta)_{i,j}$ to simplify the notation.

We compute the partial derivatives and set them to 0:

$$\begin{aligned}
\mu &: \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{\bullet,\bullet,\bullet}-N\hat\mu = 0;\\
\tau_i &: \sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{i,\bullet,\bullet}-bn\hat\mu-bn\hat\tau_i = 0;\\
\beta_j &: \sum_{i=1}^{a}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{\bullet,j,\bullet}-an\hat\mu-an\hat\beta_j = 0;\\
\gamma_{i,j} &: \sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{i,j,\bullet}-n\hat\mu-n\hat\tau_i-n\hat\beta_j-n\hat\gamma_{i,j} = 0.
\end{aligned}$$

The system's solution is

$$\begin{aligned}
\hat\mu &= \overline{y}_{\bullet,\bullet,\bullet},\\
\hat\tau_i &= \overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet}, \quad i=1,\dots,a,\\
\hat\beta_j &= \overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet}, \quad j=1,\dots,b,\\
\hat\gamma_{i,j} &= \overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet}, \quad i=1,\dots,a,\ j=1,\dots,b.
\end{aligned}$$

Analysis of Variance The total sum of squares can be decomposed as

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\overline{y}_{\bullet,\bullet,\bullet})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left[(\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})+(\overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})+(\overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet})+(y_{i,j,k}-\overline{y}_{i,j,\bullet})\right]^2\\
&= bn\sum_{i=1}^{a}(\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})^2+an\sum_{j=1}^{b}(\overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet})^2+\sum_{i,j,k}(y_{i,j,k}-\overline{y}_{i,j,\bullet})^2,
\end{aligned}$$

which we can re-write simply as

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

The corresponding ANOVA table is shown below.

Source	SS	df	MS	F
Treatment A	SSA	$a-1$	MSA	$F_A = \text{MSA}/\text{MSE}$
Treatment B	SSB	$b-1$	MSB	$F_B = \text{MSB}/\text{MSE}$
Interaction AB	SSAB	$(a-1)(b-1)$	MSAB	$F_{AB} = \text{MSAB}/\text{MSE}$
Error	SSE	$ab(n-1)$	MSE	
Total	SST	$N-1$		

Table 11.25: ANOVA table for equality of factorial effects and of interaction effects, in a two-way design.

As always, there are equivalent formulas for the sums of squares:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}^2 - \frac{y_{\bullet,\bullet,\bullet}^2}{N}; \quad \text{SSA} = \sum_{i=1}^{a} \frac{y_{i,\bullet,\bullet}^2}{bn} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSB} = \sum_{j=1}^{b} \frac{y_{\bullet,j,\bullet}^2}{an} - \frac{y_{\bullet,\bullet,\bullet}^2}{N}; \quad \text{SSTR} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{i,j,\bullet}^2}{n} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSAB} = \text{SSTR} - \text{SSA} - \text{SSB}; \quad \text{SSE} = \text{SST} - \text{SSTR}.$$

Example In a comprehensive study aimed at understanding the growth dynamics of plants, we use a 33 factorial design, resulting in a total of 9 distinct treatments. For each treatment combination, we collect information on $n = 4$ replicates, ensuring robustness in the observations.

The response variable of interest is the height of the plants (in cm), all of the same species, after a span of 30 days. We examine two critical factors: the amount of daily sunlight exposure,[65] and the type of fertilizer used.[66] The responses are shown below.

65: With three specific levels (4 hours, 8 hours, and 12 hours).

66: With three unique compositions (Type 1, Type 2, Type 3).

Height (cm)	**Daily Sunlight Exposure** (hours)					
Fertilizer	**12**		**8**		**4**	
Type 1	14.0	19.8	12.6	13.2	1.5	8.0
	14.9	13.6	9.6	12.5	4.8	5.5
Type 2	14.0	14.5	4.4	3.0	3.0	6.0
	8.4	17.0	9.0	6.5	9.2	4.8
Type 3	13.8	11.0	17.4	12.0	9.6	10.4
	16.8	16.0	15.0	13.9	8.2	6.0

The primary objective of the study is not only to tease out the individual and interactive effects of sunlight exposure and fertilizer composition on the plant's growth, but also to pinpoint whether a particular fertilizer type consistently supports optimal growth across sunlight conditions. The data is summarized below.

Height (cm)	**Exposure** (hrs)			
Dosage	**12**	**8**	**4**	$y_{i,\bullet,\bullet}$
Type 1	62.3	47.9	19.8	130.0
Type 2	53.9	22.9	23.0	99.8
Type 3	57.6	58.3	34.2	150.1
$y_{\bullet,j,\bullet}$	173.8	129.1	77.0	379.9

We can also create this table in R.[67]

67: We will use the `tidyverse` package this time around, just to show it can be done.

```
data = data.frame(
  Fertilizer = as.factor(c(rep("Type 1",4),rep("Type 2",4),rep("Type 3",4))),
  Height_12 = c(14.0, 14.9, 19.8, 13.6, 14.0, 8.4, 14.5, 17.0, 13.8, 16.8, 11.0, 16.0),
  Height_8 = c(12.6, 9.6, 13.2, 12.5, 4.4, 9.0, 3.0, 6.5, 17.4, 15.0, 12.0, 13.9),
  Height_4 = c(1.5, 4.8, 8.0, 5.5, 3.0, 9.2, 6.0, 4.8, 9.6, 8.2, 10.4, 6.0))
```

```
library(tidyverse)
summary.main <- data |> group_by(Fertilizer) |>
    summarise(h12 = sum(Height_12), h8 = sum(Height_8), h4 = sum(Height_4))
totals <- summary.main$h12 + summary.main$h8 + summary.main$h4
summary.big <- data.frame(cbind(summary.main[,c(2:4)], totals))
summary.end <- summary.big |>
    summarise(h12 = sum(h12), h8 = sum(h8), h4 = sum(h4), totals = sum(totals))
summary.data <- rbind(summary.big,summary.end)
rownames(summary.data) <- c("Type 1", "Type 2", "Type 3", "totals")
summary.data
```

```
         h12    h8   h4 totals
Type 1  62.3  47.9 19.8  130.0
Type 2  53.9  22.9 23.0   99.8
Type 3  57.6  58.3 34.2  150.1
totals 173.8 129.1 77.0  379.9
```

We can obtain the ANOVA table as follows.

```
a = nrow(summary.data) - 1
b = ncol(summary.data) - 1
n = nrow(data)/a
N = a*b*n

SST = sum(data[,c(2:(b+1))]^2)-summary.data[4,4]^2/N
SSA = sum(summary.data[b+1,c(1:a)]^2)/(b*n)-summary.data[4,4]^2/N
SSB = sum(summary.data[c(1:b),a+1]^2)/(a*n)-summary.data[4,4]^2/N
SSTR = sum(summary.data[c(1:b),c(1:a)]^2)/n-summary.data[4,4]^2/N
SSAB = SSTR - SSA - SSB
SSE = SST - SSTR
MSA = SSA/(a-1)
MSB = SSB/(b-1)
MSAB = SSAB/((a-1)*(b-1))
MSE = SSE/(a*b*(n-1))

ANOVA = as.data.frame(cbind(c(SSA,SSB,SSAB,SSE,SST),
                            c(a-1, b-1, (a-1)*(b-1), a*b*(n-1), N-1),
                            c(MSA,MSB,MSAB,MSE,0),
                            c(MSA/MSE,MSB/MSE,MSAB/MSE,0,0)))
rownames(ANOVA) = c("Treatment A", "Treatment B", "Interaction AB", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
                      SS df         MS        F0
Treatment A    391.18722  2 195.593611 29.159629
Treatment B    106.83722  2  53.418611  7.963792
Interaction AB  96.13778  4  24.034444  3.583121
Error          181.10750 27   6.707685
Total          775.26972 35
```

Hypothesis Testing Before discussing the different hypothesis tests, we need the following results (see Exercises):

$$\mathrm{E(MSA)} = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\tau_i^2; \quad \mathrm{E(MSB)} = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2;$$

$$\mathrm{E(MSAB)} = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2; \quad \mathrm{E(MSE)} = \sigma^2.$$

In general, we may be interested in the following tests:

- presence/absence of interactions between Factor A and Factor B;
- presence/absence of a Factor A effect;
- presence/absence of a Factor B effect.

The hypothesis of **absence of interaction** between Factors A and B can be formulated as

$$H_0^{AB} : \mu_{i,j} - \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,j} + \overline{\mu}_{\bullet,\bullet} = (\tau\beta)_{i,j} = 0, \quad i = 1,\ldots,a; j = 1,\ldots,b.$$

In the absence of interaction, the **difference between averages** obtained by varying either Factor A or Factor B is the same regardless of the level of the other factor:

$$\begin{aligned}&\mu_{i,j} - \mu_{i,j'} = \mu_{i',j} - \mu_{i',j'}\\ &\mu_{i,j} - \mu_{i',j} = \mu_{i,j'} - \mu_{i',j'}, \quad i,i' = 1,\ldots,a,\ j,j' = 1,\ldots,b.\end{aligned}$$

The **absence of effect for Factor A** can be formulated as

$$H_0^A : \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,\bullet} = \tau_i = 0, \quad i = 1,\ldots,a.$$

In the absence of interaction, we can rewrite the hypothesis as

$$H_0^A : \mu_{i,j} = \mu_{i',j}, \quad i,i' = 1,\ldots,a,\ j = 1,\ldots,b,$$

which corresponds to the intuitive notion of the absence of effect of Factor A.

Similarly, the **absence of effect for Factor B** can be formulated as

$$H_0^B : \overline{\mu}_{\bullet,j} - \overline{\mu}_{\bullet,\bullet} = \beta_j = 0, \quad j = 1,\ldots,b.$$

In the absence of interaction, we can rewrite the hypothesis as

$$H_0^B : \mu_{i,j} = \mu_{i,j'}, \quad i = 1,\ldots,a,\ j,j' = 1,\ldots,b,$$

which corresponds to the intuitive notion of the absence of effect of Factor B.

The hypotheses H_0^{AB}, H_0^A and H_0^B use, respectively, the following tests:

$$F_{AB} = \frac{\mathrm{MSAB}}{\mathrm{MSE}} \sim F_{(a-1)(b-1),N-ab}(\delta_{AB}^2), \quad \delta_{AB}^2 = \frac{n}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2;$$

$$F_A = \frac{\mathrm{MSA}}{\mathrm{MSE}} \sim F_{a-1,N-ab}(\delta_A^2), \quad \delta_A^2 = \frac{bn}{\sigma^2}\sum_{i=1}^{a}\tau_i^2;$$

$$F_B = \frac{\mathrm{MSB}}{\mathrm{MSE}} \sim F_{b-1,N-ab}(\delta_B^2), \quad \delta_B^2 = \frac{an}{\sigma^2}\sum_{j=1}^{b}\beta_j^2.$$

When H_0^{AB}, H_0^A, and/or H_0^B hold, we note that $E(\text{MSAB})$, $E(\text{MSA})$, and/or $E(\text{MSB})$ take on the value $\text{MSE} = \sigma^2$, respectively. Thus, **large values** of F_{AB}, F_A, and/or F_B imply that the observations **do not support** the corresponding null hypotheses.

More generally, When H_0^{AB}, H_0^A, and/or H_0^B hold, the corresponding test statistic F_{AB}, F_A, and/or F_B follow a **central F-distribution**. Thus, we **reject** H_0^{AB}, H_0^A, and/or H_0^B, respectively, at **significance** α if

$$\begin{aligned} AB &: F_0 > F_{\alpha;(a-1)(b-1),N-ab}; \\ A &: F_0 > F_{\alpha;a-1,N-ab}, \quad \text{and/or} \\ B &: F_0 > F_{\alpha;b-1,N-ab}. \end{aligned}$$

In practice, we start by testing the **absence/presence of interactions**. If the interaction is **not significant**, then we perform the tests corresponding to treatment effects for Factors A and B.[68]

68: In the latter case, the hypotheses H_0^A et H_0^B can easily be interpreted; when the interaction is statistically significant, the interpretation of the treatment effect may be more challenging.

Example In the plant growth example, we have $F_{AB} = 3.58$, $F_A = 29.16$, and $F_B = 7.96$. At significance level $\alpha = 0.05$, we find:

```
qf(0.05, df1=(a-1)*(b-1), df2=N-a*b, lower.tail=FALSE)
qf(0.05, df1=a-1, df2=N-a*b, lower.tail=FALSE)
qf(0.05, df1=b-1, df2=N-a*b, lower.tail=FALSE)
```

```
[1] 2.727765
[1] 3.354131
[1] 3.354131
```

Since $3.58 > F_{0.05,4,27} = 2.73$, we reject H_0^{AB} and conclude that the interaction is significant at $\alpha = 0.05$. Also, since $7.96 > F_{0.05,2,27} = 3.35$ and since $29.16 > F_{0.05,2,27} = 3.35$, we reject both H_0^A and H_0^B, but it is not as obvious what the means for the data. □

11.6.2 Model Validation

The three basic model assumptions are still that the errors are **independent**, **normally distributed**, and have **constant variance**. As we have done before, we would use the residuals in lieu of the errors to validate these assumptions.

In the two-way balanced factorial design, the **predicted values** are given by

$$\hat{y}_{i,j,k} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{i,j} = \overline{y}_{i,j,\bullet};$$

the **model residuals** are thus given by

$$e_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k} = y_{i,j,k} - \overline{y}_{i,j,\bullet}.$$

11.6.3 Model Without Interaction

In the **absence of interaction**, the model simplifies to

$$y_{i,j,k} = \mu + \tau_i + \beta_j + \varepsilon_{i,j,k},\ i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n.$$

In that case, the estimators of the model parameters are given by

$$\begin{aligned}
\hat{\mu} &= \overline{y}_{\bullet,\bullet,\bullet} \\
\hat{\tau}_i &= \overline{y}_{i,\bullet,\bullet} - \overline{y}_{\bullet,\bullet,\bullet},\ i = 1, \ldots, a \\
\hat{\beta}_j &= \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,\bullet,\bullet},\ j = 1, \ldots, b,
\end{aligned}$$

and the decomposition of the total sum of squares is

$$\begin{aligned}
\text{SST} &= \text{SSA} + \text{SSB} + \text{SSE} \\
N - 1 &= (a-1) + (b-1) + [(a-1)(b-1) + ab(n-1)] \\
&= (a-1) + (b-1) + (N - a - b + 1).
\end{aligned}$$

The treatment sums of squares SSA and SSB are identical to those in the ANOVA model with interaction. The simpler formulas collapse to:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}^2 - \frac{y_{\bullet,\bullet,\bullet}^2}{N};\quad \text{SSA} = \sum_{i=1}^{a}\frac{y_{i,\bullet,\bullet}^2}{bn} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSB} = \sum_{j=1}^{b}\frac{y_{\bullet,j,\bullet}^2}{an} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};\quad \text{SSE} = \text{SST} - \text{SSA} - \text{SSB}.$$

The corresponding ANOVA table is given below:

Source	SS	df	MS	F
Treatment A	SSA	$a-1$	MSA	$F_A = \text{MSA/MSE}$
Treatment B	SSB	$b-1$	MSB	$F_B = \text{MSB/MSE}$
Error	SSE	$N-a-b+1$	MSE	
Total	SST	$N-1$		

Table 11.29: ANOVA table for equality of factorial effects, with no interaction effects, in a two-way design.

We test for the null hypotheses

$$H_0^A : \mu_{i,j} = \mu_{i',j} \quad \text{and} \quad H_0^B : \mu_{i,j} = \mu_{i,j'}$$

using the test statistics

$$F_A = \frac{\text{MSA}}{\text{MSE}} \sim F_{a-1,N-a-b+1}(\delta_A^2),\quad \delta_A^2 = \frac{bn}{\sigma^2}\sum_{i=1}^{a}\tau_i^2,$$

$$F_B = \frac{\text{MSB}}{\text{MSE}} \sim F_{b-1,N-a-b+1}(\delta_B^2),\quad \delta_B^2 = \frac{an}{\sigma^2}\sum_{j=1}^{b}\beta_j^2.$$

The analysis of the residuals is based on the following residuals

$$e_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k} = y_{i,j,k} - (\hat{\mu} + \hat{\tau}_i + \hat{\beta}_j) = y_{i,j,k} - \overline{y}_{i,\bullet,\bullet} - \overline{y}_{\bullet,j,\bullet} + \overline{y}_{\bullet,\bullet,\bullet}.$$

The rest of the analysis proceeds as before.

11.6.4 Multiple Comparisons

As in previous sections, we may want to perform **multiple comparisons**. More often than not, we are interested in constructing **simultaneous confidence intervals** that compare the effects for each factor.

Throughout, recall that we estimate σ^2 by

$$s^2 = \text{MSE} = \frac{\text{SSE}}{N-ab} = \frac{\text{SSE}}{ab(n-1)}.$$

Suppose that we are interested in all possible pairwise comparisons for treatment A; in that case, there are $K = \binom{a}{2} = a(a-1)/2$ possible pairs to test. For treatment B, there are $L = b(b-1)/2$ possible pairs to test.

We could use the **Bonferroni procedure** to do so; the simultaneous confidence intervals corresponding to Factor A take the form

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm t_{\alpha/(2K),N-ab}\sqrt{\text{MSE}}\sqrt{\frac{2}{bn}},$$

and those for Factor B, the form

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm t_{\alpha/(2L),N-ab}\sqrt{\text{MSE}}\sqrt{\frac{2}{an}}.$$

If instead we use **Tukey's method**, the simultaneous confidence intervals corresponding to Factor A are given by

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm q_{\alpha;a,N-ab}\sqrt{\text{MSE}}\sqrt{\frac{1}{bn}},$$

and those for Factor B, by

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm q_{\alpha;b,N-ab}\sqrt{\text{MSE}}\sqrt{\frac{1}{an}}.$$

For **Scheffé's approach**, the simultaneous confidence intervals corresponding to Factor A are

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm \sqrt{(a-1)F_{\alpha;a-1,N-ab}}^{1/2}\sqrt{\text{MSE}}\sqrt{\frac{2}{bn}},$$

and those for Factor B,

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm \sqrt{(b-1)F_{\alpha;b-1,N-ab}}^{1/2}\sqrt{\text{MSE}}\sqrt{\frac{2}{an}}.$$

For the two-way balanced factorial model **without interaction**, the simultaneous confidence intervals are similar, except that the number of degrees of freedom in the residual sum of squares SSE is now $N-a-b+1$. In that case, the estimator of σ^2 is

$$\tilde{s}^2 = \tilde{\text{MSE}} = \frac{\text{SSE}}{N-a-b+1}.$$

For instance, the simultaneous confidence intervals for Factor A obtained

using **Tukey's method** are given by

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm q_{\alpha;a,N-a-b+1}\sqrt{\tilde{\text{MSE}}}\sqrt{\frac{1}{bn}},$$

whereas the simultaneous confidence intervals for Factor B obtained *via* **Scheffé's approach**, say, are given by

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm \sqrt{(b-1)F_{\alpha;b-1,N-a-b+1}}\sqrt{\tilde{\text{MSE}}}\sqrt{\frac{2}{an}}.$$

11.6.5 Factorial Designs with Multiple Factors

The two-way factorial design can be naturally extended to **multiple factors**. For instance, the **three-way factorial design** $a \times b \times c$ is:

$$y_{i,j,k,l} = \mu_{i,j,k} + \varepsilon_{i,j,k,l}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c;\ l = 1, \ldots, n$$

where

$$\mu_{i,j,k} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{i,j} + (\tau\gamma)_{i,k} + (\beta\gamma)_{j,k} + (\tau\beta\gamma)_{i,j,k}.$$

With three factors, we can explore **second-order interactions** AB, AC, and BC, or the **third-order interaction** ABC. Such designs are out of scope for these course notes,[69] more details are available in [2, 5].

69: The ANOVA table for the $a \times b \times c$ design has 9 rows, but is otherwise what one would expect to see.

11.7 Exercises

1. Conduct an analysis of the paint example of Section 11.2.1 assuming that the samples are independent (unpaired test). Compare with the results of the paired test on the same data.
2. Recreate the analysis of the apparatus example of Section 11.2.4 using R. What if the sample sizes were $n_1 = 25$ and $n_2 = 30$, instead?
3. Show directly that the decomposition SST = SSA + SSE of one-way classification holds.
4. In a one-way classification model with $a = 2$, show that the power of the $F-$test is maximized when $\frac{1}{n} + \frac{1}{N-n}$ is minimized.
5. Use the least square estimation principles to establish the normal equations, and estimate the parameters in the unbalanced one-way classification model. What are the estimated treatment effects and the estimated difference between treatments in that case? What about their confidence intervals?
6. Compute the ANOVA table for the completely randomized unbalanced design in the Kenton Food Company example.
7. In the one-way random-effects ANOVA model, show that E(MSE) = σ^2 and E(MSA) = $\sigma^2 + n\sigma_T^2$.
8. In the one-way random-effects ANOVA model, show that

$$\frac{(a-1)\text{MSA}}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{a-1}.$$

9. In a two-factor RCBD, show that

$$\text{E(MSA)} = \sigma^2 + \frac{b}{a-1}\sum_{i=1}^{a}\tau_i^2,$$

$$\text{E(MSB)} = \sigma^2 + \frac{a}{b-1}\sum_{j=1}^{b}\beta_j^2.$$

10. Verify if the RCBD model assumptions are met for the credit card example.
11. Show directly that the total sum of squares in a balanced two-way factorial design breaks down as

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

12. In the two-way balanced factorial design, show that

$$\text{E(MSA)} = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\tau_i^2;$$

$$\text{E(MSB)} = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2;$$

$$\text{E(MSAB)} = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2;$$

$$\text{E(MSE)} = \sigma^2.$$

13. In the medical agent example (two-way factorial design), is the interaction effect significant at $\alpha = 0.05$? What about the dosage effect? The application period effect?
14. Produce simultaneous confidence intervals at family significance α for treatment effects (Factors A and B) in the medical agent and plant growth examples.

Chapter References

[1] M.H. Kutner et al. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[2] D.C. Montgomery. *Design and Analysis of Experiments*. Wiley, 2012.

[3] L. Ott and M. Longnecker. *A First Course in Statistical Methods*. Thomson-Brooks/Cole, 2004.

[4] R.L. Ott and M.T. Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, 2015.

[5] H. Scheffé. *Analysis of Variance*. London: John Wiley and Sons Inc., 1959.