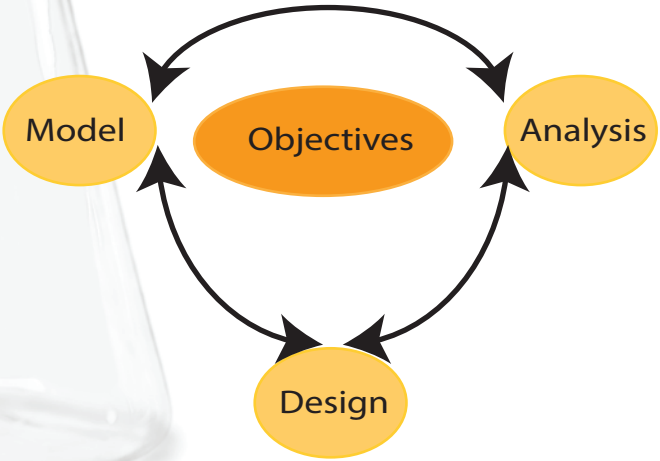
The Smart Design of Animal Experiments

Luc Wouters



The Smart Design of Animal Experiments

Luc Wouters

Contents

—More often than not, we are unable to reproduce findings published by researchers in journals.

Glenn Begley, Vice President Research Amgen (in Naik (2011))

—The way we do our research (with our animals) is stone-age.

Ulrich Dirnagl, Charité University Medicine Berlin (in Couzin-Frankel (2013))

1 Introduction

1.1. The reliability of biomedical research

Over the past decades, there have been growing concerns in the biomedical community about the reliability of research findings (Academy of Medical Science, 2015). These concerns are mostly supported by a survey conducted by *Nature* among 1,576 research scientists. It showed that 52% of the participants agreed that there was a significant crisis and 70% of them had tried and failed to reproduce another scientist's experiments (Baker, 2016). This "*replication crisis*" has serious consequences beyond the fact that valuable resources and money are wasted. First of all, studies that cannot be replicated create false hope for patients waiting for a life-saving cure. Furthermore, the high number of irreplicable studies makes scientists worry about the integrity of their enterprise, and may impede scientific progress. In this introductory chapter, we

shall consider the methodological flaws that are mainly responsible for this "*replication crisis*."

Reproducibility and replicability are two notions that are often conflated with each other, but it is a good idea to make a clear distinction between the two. In this text, we consider *replicability* as the replication of scientific findings using independent investigators, methods, data, equipment, and protocols. Replicability has long been and will continue to be the standard by which scientific claims are evaluated. By *reproducibility*, we mean that starting from the data gathered by the scientist, we can reproduce the same results, *p*-values, confidence intervals, tables, and figures as those reported by the scientist (Peng, 2009). Failure in the reproducibility of study results can be attributed to flaws in the operating procedures, such as the lack of good laboratory practices or failure to comply with them. The lack of replicability, on the other hand, can be ascribed to statistical fallacies, misconceptions, and other methodological issues (Begley and Ioannidis, 2015; Loscalzo, 2012; Peng, 2015; Prinz et al., 2011; Reinhart, 2015, pp. 97-103; van der Worp et al., 2010).

In contrast to reproducibility issues, investigators do not always recognise concerns about the replicability of their study. The following real-life examples illustrate the severeness of these reproducibility and replicability problems and the need to improve and possibly even transform the research process.

Example 1.1: Potti case A research team from Duke University led by Anil Potti, MD and Joseph Nevins MD published in 2006 a groundbreaking paper in *Nature Medicine* (Potti et al., 2006). They claimed that they had built an algorithm using genomic microarray data that allowed to predict which cancer patients would respond to chemotherapy. Such a development would spare patients the side effects of ineffective

treatments. Of course, this paper drew much attention and many independent investigators tried to reproduce the results.

Keith Baggerly and Kevin Coombes, two statisticians at MD Anderson Cancer Center, were also asked to have a look at the data. What they found was a mess of poorly conducted data analysis (Baggerly and Coombes, 2009). Some of the data was mislabelled, some samples were duplicated in the data, and some samples were even marked as both sensitive and resistant. Baggerly and Coombes concluded that they were unable to reproduce the analysis carried out by Potti et al. (2006), but the damage was done. Several clinical trials had started based on the erroneous results. In 2011, after several corrections, the original study by Potti et al. was retracted from *Nature Medicine*, stating: "*because we have been unable to reproduce certain crucial experiments*" (Potti et al., 2011).

Example 1.2: Scholl case In 2009, a group of researchers from Harvard Medical School published a study showing that cancer tumours could be destroyed by targeting the STK33 protein (Scholl et al., 2009). Scientists at *Amgen Inc.* pounced on the idea and assigned a team of 24 researchers to try to repeat the experiment with the objective of developing a new medicine. After six months of intensive lab work, it turned out that the project was a waste of time and money since it was impossible for the Amgen scientists to replicate the results (Babij et al., 2011; Naik, 2011).

Unfortunately, this was not the only problem of replicability the Amgen researchers encountered. During a decade Begley and Ellis (2012) identified a set of 53 "landmark" publications in preclinical cancer research, i.e. papers in top journals from reputable labs. A team of 100 scientists tried to replicate the results. To their surprise, in 47 of the 53 studies (i.e. 89%) the findings could not be replicated. This outcome was particularly disturbing since Begley and Ellis made every effort to work in close collaboration with the authors of the original papers and even tried to replicate the experiments in the laboratory of the original investigator. In some cases, 50 attempts were made to reproduce the original data, without obtaining the claimed result (Begley, 2012). What is even more troubling is that Amgen's findings were consistent with those of

others. In a similar setting, Bayer researchers found that only 25% of the original findings in target discovery could be validated (Prinz et al., 2011).

Example 1.3: Séralini case Séralini et al. (2012) published a 2-year feeding study in rats investigating the health effects of genetically modified (GM) maize NK603 with and without glyphosate-containing herbicides. The authors of the study concluded that GM maize NK603 and low levels of glyphosate herbicide formulations, at concentrations well below officially-set safe limits, induce severe adverse health effects, such as tumours, in rats. Apart from the publication, Séralini also presented his findings in a widely covered press conference, showing shocking photos of rats with enormous tumours. Consequently, this study had a severe impact on the general public and also on the interest of the industry. The paper was used in the debate over a referendum over labelling of GM food in California, and it led to bans on importation of certain GMOs in Russia and Kenya. However, shortly after its publication many scientists heavily criticised the study and expressed their concerns about the validity of the findings.

A polemic debate started with opponents of GMOs and also within the scientific community, which inspired media to refer to the controversy as *The Séralini affair* or *Séralini tumour-gate*. Subsequently, the European Food Safety Authority (2012) thoroughly scrutinised the study and found that it was of inadequate design, analysis, and reporting. Specifically, the number of animals was considered too small and not sufficient for reaching a solid conclusion. Eventually, the journal retracted Séralini's paper, claiming that it did not reach the journal's threshold of publication (Hayes, 2014)[1].

Example 1.4: Efficiency of the study Selwyn (1996, p. 2) describes a study where an investigator examined the effect of a test compound on hepatocyte diameters. The experimenter decided to study eight rats per treatment group, three different lobes of each rat's liver, five fields per lobe, and approximately 1,000 to 2,000 cells per field. At that time,

[1] Séralini managed to republish the study in Environmental Sciences Europe (Séralini et al., 2014), a journal with a considerably lower impact factor.

most of the work, i.e. measuring the cell diameters, was done manually, making the total amount of work, i.e. 15,000 - 30,000 measurements per rat, substantial. The experimenter complained about the overwhelming amount of work in this study and the tight deadlines that were set up. A sample size evaluation, conducted after completion of the study, indicated that sampling as few as 100 cells per lobe would have been without appreciable loss of information.

Doing good science and producing high-quality, reliable data should be the concern of every serious research scientist. Unfortunately, as shown in the first three examples, this is not always the case. As mentioned above, there is a genuine concern about the replicability of research findings. Moreover, it has even been argued that most research findings could be false (Ioannidis, 2005). In a recent paper, Begley and Ioannidis (2015) estimated that indeed 85% of biomedical research is wasted at large.

Pinpointing the replicability problem is one thing, but what are its underlying causes? Kilkenny et al. (2009), who surveyed 271 papers reporting laboratory animal experiments found that many studies had problems with the quality of reporting, quality of experimental design, and quality of statistical analysis. Most worrying was the fact that the quality of the experimental design in the majority of experiments was inappropriate or inefficient. Similar findings were obtained by Freedman et al. (2015) who tried to identify the root causes of the replicability problem and to estimate its economic impact. They estimated that in the United States alone approximately US$28B/year is spent on research that cannot be replicated. The main problems causing this lack of replicability are summarised in Figure 1.1. Issues in study design and data analysis accounted for more than 50% of the studies that could not be replicated. The value of a good experimental design and statistical analysis were also recognised by the researchers

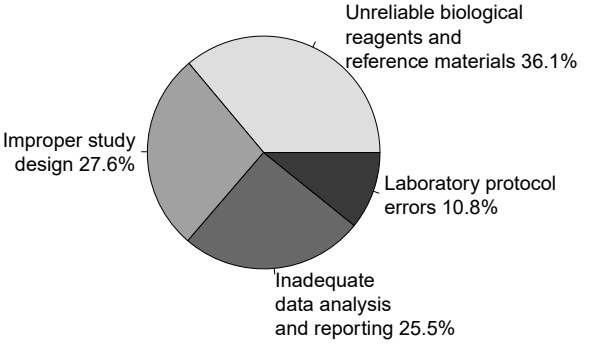


Figure 1.1. Categories of errors that contribute to the problem of replicability in life science research (source Freedman et al. 2015)

who took part in the *Nature* survey. Nearly 90% of the responders ticked "*More robust experimental design*", "*better statistics*" and "*better mentorship*" (Baker, 2016).

Not only scientists but also the journals have a great responsibility in guarding the quality of their publications. Peer reviewers and editors, who have little or no statistical training, let methodological errors pass undetected. Moreover, high-impact journals tend to focus on statistically significant results of unexpected findings, often without looking at the practical importance. Notably, in studies with insufficient sample size, this publication bias causes high numbers of false research claims. (Ioannidis, 2005; Reinhart, 2015, p. 25).

In addition to the problem of replicability of research findings, there has also been a dramatic rise in the number of journal retractions over the last decades (Cokol et al., 2008). In a review of all 2,047 biomedical and life-science research articles indexed

by PubMed and labelled as retracted on May 3, 2012, Fang et al. (2012) found that 21.3% of the retractions were due to error, while 67.4% of the retractions were attributable to misconduct, including fraud or suspected fraud (43.4%), duplicate publication (14.2%), and plagiarism (9.8%).

The lack of replicability, the increased number of retractions, and most notably, the false claims made by Potti et al. (2006), Scholl et al. (2009), and Séralini et al. (2012) caught the attention of mainstream media (Begley, 2012; Hotz, 2007; Lehrer, 2010; Naik, 2011; The Economist, 2013; Zimmer, 2012) and have put the integrity of science into question by the general public.

To summarise, a substantial part of the issues of replicability can be attributed to a lack of quality in the design and execution of the studies. When little or no thought is given to methodological issues, in particular to the statistical aspects of the study design, the studies are often seriously flawed and are not capable of meeting their intended purpose. In some cases, such as the Séralini study, the experiments were designed too small to enable an answer to the research question. Conversely, like in Example 1.4, there are also studies that waste valuable resources by using more experimental material than required.

To improve on these issues of credibility and efficiency, we need effective interventions and change the way how scientists look at the research process (Ioannidis, 2014; Reinhart, 2015, pp. 119-129. This transformation can be accomplished by introducing *statistical thinking* as a powerful and informed skill, based on the fundamentals of statistics, that enhances the quality of the research data (Vandenbroeck et al., 2006). While the science of statistics is

mostly involved with the complexities and techniques of statistical analysis, statistical thinking is a generalist skill that focuses on the application of nontechnical concepts and principles.

There are no precise, generally accepted definitions of statistical thinking. In our conceptualisation, we consider statistical thinking as a skill that helps to understand better how statistical methods can contribute to finding answers to specific research problems and what the implications are concerning data collection, experimental setup, data analysis, and reporting. Statistical thinking will provide us with a generic methodology to design insightful experiments.

Statistical thinking permeates the entire research process and, when adequately implemented, can lead to a highly successful productive research enterprise, as was demonstrated by the eminent scientist, the late Dr Paul Janssen. As pointed out by Lewi (2005), the success of Dr Paul Janssen could be attributed to a large extent on having a set of statistical precepts accepted by his collaborators. These formed the statistical foundation upon which his research was built and insured that research proceeded in an orderly and planned fashion, while at the same time having an open mind for unexpected opportunities. His approach was such a success that, when he retired in 1991, his laboratory had produced 77 original medicines over a period of fewer than 40 years, which is an absolute world record. Besides, at its peak, the Janssen laboratory produced more than 200 scientific publications per year (Lewi and Smith, 2007).

1.2. Structure of this text

Chapter 2 considers the architecture of experimental research, the phases of the scientific research process and, related to this, the different archetypes of scientists that can be distinguished. This chapter also introduces the concept of statistical thinking and the basic principles underlying smart research design. Chapter 3 is about the planning process and discusses the different types of experiment. Chapter 4 introduces the fundamental principles of statistical design. These principles are at the basis of the different experimental designs that are presented in Chapter 5. Several real-life examples from biomedical research are used to illustrate these designs.

Chapter 6 introduces the important concept of statistical power and shows ways to determine the required number of replicates. The relative importance of subsamples, as well as the problems with underpowered studies and effect size inflation, are also discussed here. The relation between statistical analysis and experimental design, as well as the true meaning of the concept p-value, are presented in Chapter 7. Chapter 8 is devoted to the finalisation of the design process in the study protocol.

Chapter 9 is about the interpretation and reporting of research findings. It shows which topics are to be included in the Methods section of a paper and how to summarise the data in the Results section. This chapter also gives indications on graphical displays and puts the relative importance of significance tests again into perspective. Relevant topics of the *ARRIVE* guidelines are also presented here. Finally, Chapter 10 discusses the role of the statistician, recapitulates the principles of statistical thinking and the problems of statistical significance.

1.3. Software

For statistical analysis, a researcher can choose from a multitude of software tools, such as *JMP*, *SAS*, *SPSS*, *GraphPad-Prism*, etc. However, for generating experimental designs and for sample size calculations the choice is limited, and most of the commercially available programs that have these features are quite expensive. In this text, several examples make use of **R** (R Core Team, 2017). The **R**-*system* is freely available and provides a versatile programming, statistical analysis, and graphical environment. The specific packages for experimental design developed in **R**, make up a powerful toolbox for randomisation, sample size calculations and for generating experimental designs, some of which can be fairly sophisticated. In the examples presented here, the code in **R** used to obtain a particular design, or the required sample size is shown in full detail. Further information on how to install **R** and the different packages can be found in Appendix C.

—Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write !

Samuel S. Wilks, (1951)

—Statistics is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics.

George Box, (1990)

2

Smart Research Design by Statistical Thinking

2.1. The architecture of experimental research

2.1.1. The controlled experiment

There are two basic approaches to implement a scientific research project. One strategy is to conduct an *observational study*[1] in which we examine the effect of naturally occurring variation, while the condition that we study is beyond our control. Although, there are often good and valid reasons for an investigator to conduct an observational study, the presence of concomitant, confounding variables can never be excluded, which weakens its conclusions.

An alternative to an observational study is an experimental or manipulative study in which we manipulate the experimental

[1] also called correlational study

Phase	⇒	**Deliverable**
Definition	⇒	Research Proposal
Design	⇒	Protocol
Data Collection	⇒	Data set
Analysis	⇒	Conclusions
Reporting	⇒	Report

Figure 2.1. Research is a phased process with each of the phases having a specific deliverable

system and measure the effect of our interventions on the experimental material. Since we have full control over the experimental environment, we call such studies *controlled experiments*. A well-designed controlled experiment eliminates the bias caused by confounding variables and allows to demonstrate causal relationships. We shall focus on controlled experiments and how statistical thinking can be of use to optimise their design and interpretation.

2.1.2. Scientific research as a phased process

From a systems analysis point of view, the scientific research process can be divided into five distinct stages:

1. definition of the research question
2. design of the experiment
3. conduct of the experiment and data collection
4. data analysis
5. reporting

Each of these phases results in a specific deliverable (Figure 2.1). The definition of the research question will usually end up in a research or grant proposal, stating the hypothesis related to the research (research hypothesis) and the implications or predictions

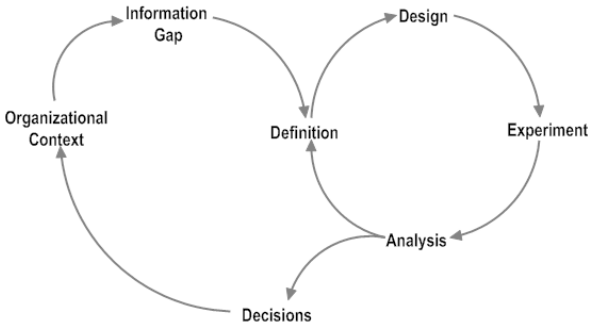


Figure 2.2. Scientific research as an iterative process

that follow from it. The design of the experiment needed for testing the research hypothesis is formalised in a written protocol. After the experiment has been carried out, the data will be collected and provide the experimental data set. Statistical analysis of this data set yields conclusions that answer the research question by accepting or rejecting the formalised hypothesis. Finally, a well carried out research project ends up in a report, thesis, or journal article.

2.1.3. Scientific research as an iterative, dynamic process

Scientific research is not a simple static activity, but as depicted in Figure 2.2, an iterative and highly dynamic process. A research project is carried out within some organisational or management framework which can be somewhat authoritative. In this academic, governmental or corporate context, the management objectives of the research project are put forward. The aim of our research project itself is to fill an existing information gap. Therefore, the research question is defined, the experiment is designed and carried out, and the data are analysed. The results of this analysis allow informed decisions to be made and provide a way of feedback to adjust the definition of the research question. On

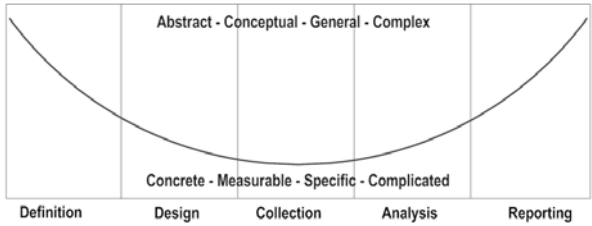


Figure 2.3. Modulating between the concrete and abstract world

the other hand, the experimental results can trigger research management to reconsider their objectives and, if necessary, request for more information.

2.2. Research styles - The smart researcher

The five phases that make up the research process modulate between the concrete and the abstract world (Figure 2.3). Definition and reporting are conceptual and complex tasks requiring a great deal of abstract reasoning. Conversely, experimental work and data collection are very concrete, measurable tasks handling with the practical details and complications of the specific research domain.

Scientists exhibit different styles in their research depending on the relative fraction of the available resources that they are willing to spend at each phase of the research process, which allows us to recognise different archetypes of researchers (Figure 2.4):

- the *novelist* who needs to spend a lot of time distilling a report from an ill-conceived experiment;
- the *data salvager* who believes that no matter how you collect the data or set up the experiment, there is always a statistical fix-up at analysis time;

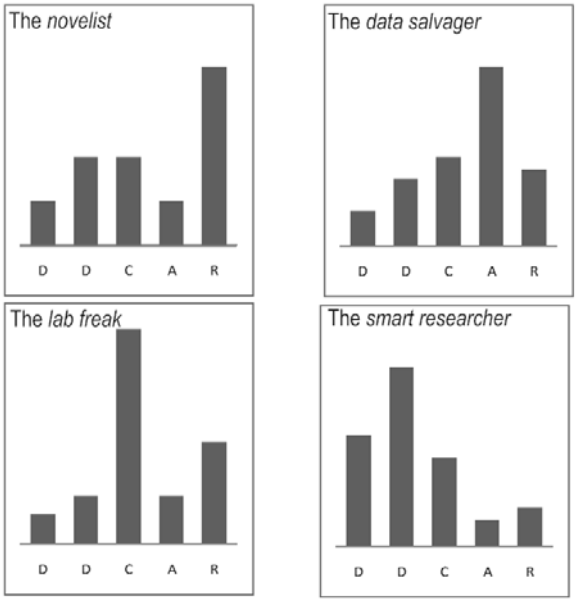


Figure 2.4. Archetypes of researchers based on the relative fraction of the available resources that they are willing to spend at each phase of the research process. D(1): definition phase, D(2): design phase, C: data collection, A: analysis, R: reporting

- the *lab freak* who firmly believes that if enough data are collected something interesting will always emerge;
- the *smart researcher* who is aware of the architecture of the experiment as a sequence of steps and allocates the major part of his time budget to the first two steps: definition and design.

The *smart researcher* is convinced that time spent planning and designing an experiment at the outset will save time and money in the long run. He opposes the *lab freak* by trying to reduce the number of measurements to be taken, thus effectively reducing the time spent in the lab.

Table 2.1. Statistical thinking versus statistics

Statistics	Statistical Thinking
Specialist skill	Generalist skill
Science	Informed practice
Technology	Principles, patterns
Closure, seclusion	Ambiguous, dialogue
Introvert	Extravert
Discrete interventions	Permeates the research process
Builds on good thinking	Valued skill itself

In contrast to the *data salvager*, the *smart researcher* recognises that the design of the experiment will govern how the data will be analysed, thereby reducing the time spent at the data analysis stage to a minimum. By carefully preparing and formalising the definition and design phase, the *smart researcher* can look ahead to the reporting phase with peace of mind, which is in contrast to the *novelist*.

2.3. Principles of statistical thinking

The *smart researcher* recognises the value of *statistical thinking* for his application area and he himself is skilled in statistical thinking, or he collaborates with a professional who masters this skill. As noted before, statistical thinking is related to but distinct from statistical science (Table 2.1). While statistics is a specialised technical skill based on mathematical statistics as a science on its own, statistical thinking is a *generalist skill* based on informed practice and focused on the applications of nontechnical concepts and principles.

The *statistical thinker* attempts to understand how statistical methods can contribute to finding answers to specific research

problems in terms of data collection, experimental setup, data analysis and reporting. He or she is able to postulate which statistical expertise is required to enhance the research project's success. In this capacity, the statistical thinker acts as a *diagnoser*.

In contrast to statistics, which operates in a closed and secluded mathematical context, statistical thinking is a practice that is fully integrated with the researcher's scientific field, not merely an autonomous science. Hence, the statistical thinker operates in a more ambiguous setting, where he is deeply involved in applied research, with a good working knowledge of the substantive science. In this role, the statistical thinker acts as an *intermediary* between scientists and statisticians and goes into dialogue with them. He attempts to integrate several potentially competing priorities that determine the success of a research project: resource economy, statistical power, and scientific relevance, into a coherent and statistically underpinned research strategy.

While the impact of the statistician on the research process is limited to discrete interventions, the *statistical thinker* truly permeates the research process. His combined skills lead to greater efficiency and increase the speed with which data, analyses, and conclusions become available. Moreover, these skills allow to enhance the quality and to reduce the associated cost. Statistical thinking then helps the scientist to build a case and negotiate it on fair and objective grounds with those in the organisation seeking to contribute to more business-oriented measures of performance. In that sense, the successful statistical thinker is a *persuasive communicator*. This comparison clearly shows that the power of statistics in research is actually founded upon good statistical thinking.

Smart research design is based on the seven basic principles of statistical thinking:

1. Time spent thinking about the conceptualisation and design of an experiment is time wisely spent.

2. The design of an experiment reflects the contributions from different sources of variability.

3. The design of an experiment balances between its internal validity (proper control of noise) and its external validity (the experiment's generalisability).

4. Good experimental practice provides the clue to bias minimisation.

5. Good experimental design is the clue to the control of variability.

6. Experimental design integrates various disciplines.

7. A priori consideration of statistical power is an indispensable pillar of an effective experiment.

—Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge.

Ronald A. Fisher (1935)

—The statistician who supposes that his main contribution to the planning of an experiment will involve statistical theory, finds repeatedly that he makes his most valuable contribution simply by persuading the investigator to explain why he wishes to do the experiment, by persuading him to justify the experimental treatments, and to explain why it is that the experiment, when completed, will assist him in his research.

Gertrude Cox (1950)

3

Planning the Experiment

The proper planning of an experiment is essential to its success and requires careful thought. Unfortunately, not all scientists realise its importance or are aware of the different steps and decisions that it involves. In this chapter, we shall consider in more detail the basics of the planning process.

3.1. The planning process

The first step in planning an experiment (Figure 3.1) is the specification of its objectives. The researcher should realise what the actual goal is of his experiment and how it integrates into the whole set of related studies on the subject. How does it connect with management or other objectives? How will the results from this particular study contribute to knowledge about the subject? Sometimes a preliminary exploratory experiment can be used to

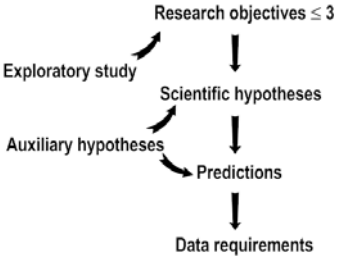


Figure 3.1. The planning process

generate clear questions that will be answered in a confirmatory experiment.

The study objectives should be well defined and written out as explicitly as possible. It is wise to *limit the objectives* of an experiment to a maximum of, say three (Selwyn, 1996, p. 9). Any more than that risks designing an overly complex experiment and could compromise the integrity of the study. Trying to accomplish each of many objectives in a single experiment stretches its resources too thin, and as a result, often none of the study objectives is satisfied. Objectives should also be reasonable and attainable and one should be realistic in what can be accomplished in a single experiment.

Example 3.1: Plaque disposition in brains of mice An experiment was planned to assess the effect of a pharmacological treatment on plaque disposition in the brains of a strain of transgenic mice (Bate and Clark, 2014, p. 8). The investigators initially hoped to compare the treated group to the control at 2, 3, 4, 6 and 12 months of age, which would result in 10 treatment groups (five time-points by two treatments). Unfortunately, with only 40 mice available, there would only be 4 animals per group per time-point, which would not allow them to detect any biologically relevant effect. With only 3 time-points selected (e.g. 2, 6 and 12 months), this number could be increased to 6 or 7 mice per group.

The study by Séralini et al. (2012) is also a typical example of a study where the research team tried to accomplish too many objectives. In this study, ten treatments were examined in both female and male rats. Since the research team apparently had a very limited amount of resources available, the investigators used only ten animals per treatment per sex. This was far below the 50 animals per treatment group that are standard in long-term carcinogenicity studies (Gart et al., 1986, p. 27; Haseman, 1984).

After having formulated the research objectives, the scientist will then transfer them into scientific hypotheses. Often it is impossible to study the research objective directly, but some surrogate experimental model is used instead. By doing so, an *auxiliary hypothesis* is assumed to hold, namely that the experimental model is adequate to the research objectives (Hempel, 1966, pp. 22-25).

Example 3.2: Séralini case (continued) Séralini et al. (2012) were not interested whether GMO's were toxic in *rats*. The real objective was to establish the toxicity in *humans*. As a surrogate for man, Séralini et al. (2012) chose the Sprague-Dawley strain of rat as the experimental model. The choice of the Sprague-Dawley rat strain by Séralini et al. (2012) received much criticism (European Food Safety Authority, 2012), since this strain is prone to the development of tumours.

Auxiliary hypotheses also play a role when it is difficult or even impossible to study the measure the variable of interest directly. In this case, an indirect measure might be available, and the investigator relies on the premise that this indirect measure is a valid surrogate for the actual target variable.

Based on both the scientific and auxiliary hypotheses, the researcher will then predict the test implications of what to expect if these hypotheses are true. Each of these predictions should be the strongest possible test of the scientific hypotheses. The de-

duction of these test implications also involves additional auxiliary hypotheses. As stated by Hempel (1966, p. 23), "*reliance on auxiliary hypotheses is the rule, rather than the exception, when testing scientific hypotheses*." Therefore, it is essential that the researcher is aware of the auxiliary assumptions he makes when predicting the test implications. Generating sensible predictions is one of the key factors of good experimental design. Good predictions follow logically from the hypotheses that we wish to test, not from other rival hypotheses, and lead us to insightful experiments in which we can test these predictions.

The next step in the planning process is then to decide which data are required to confirm or refute the predicted test implications. Throughout the sequence of formulating the question, the hypothesis, and the prediction, it is essential to assess each step critically with enough scepticism and even ask a colleague to play the *devil's advocate*. During the design and planning stage of the experiment, one should already have in mind how to communicate the study's proposal with external stakeholders. It is much better to identify problems at this early stage than to be confronted with them after the experiment has started. At the end of the experiment, the scientist should be able to determine whether the objectives have been met, i.e. whether the research questions were answered to satisfaction.

3.2. Types of experiments

We first distinguish between exploratory, pilot, and confirmatory studies. *Exploratory studies* or *learning* studies are used to explore a new research area. They tend to consist of a package of small and flexible experiments using different methodologies, with the

aim to discover as much as possible about the material or phenomenon under investigation. Replication, sample size, and formal hypothesis testing are less critical for this type of experiment. Currently, the vast majority of published research in the biomedical sciences originates from this sort of experiment (Kimmelman et al., 2014). The exploratory nature of these learning experiments is also reflected in the way the data are analysed. Exploratory data analysis, as opposed to confirmatory data analysis, is a flexible approach, based mainly on graphical displays, towards formulating new theories (Tukey, 1980).

Exploratory studies aim primarily at developing new research hypotheses and tend to generate more questions than they provide answers. However, they do not provide an unambiguous answer to the research question, since the data that generated the research hypothesis cannot be used for its support. Formulating a hypothesis based on a specific pattern in the data and using the same data to test this hypothesis is a kind of circular reasoning. Unfortunately, exploratory experiments tend to be reported as confirming experiments, where significant results are reported as if they were hypothesized before the experiment was conducted. This contributes to the false positive results in the literature (Lazic, 2016, p. 50). The study by Séralini et al. (2012) was, in fact, an exploratory study and much of the controversies around this paper would not have arisen if it would have been presented as such.

Pilot studies are designed to make sure the research question is sensible. They allow to refine the experimental procedures, to determine how variables should be measured, whether the experimental setup is feasible, etc. Pilot experiments are especially useful when the actual experiment is large, time-consuming or expen-

sive (Selwyn, 1996, pp. 45-48). Information obtained in the pilot experiment is of particular importance when writing the technical and study protocol of such studies. Pilot experiments are discussed in more detail in Section 3.3 (page 26).

Confirmatory studies are used to assess the test implications of a scientific hypothesis. In biomedical research, this assessment depends on statistical methodology. In contrast to exploratory studies, confirmatory experiments make use of rigid pre-specified designs and *a priori* stated hypotheses. Confirmatory studies are the main topic of this tutorial. Exploratory and confirmatory studies complement one another in the sense that the former generates the hypotheses that can be put to "crucial testing" in the latter. Besides, confirmatory and exploratory studies frequently overlap each other. For example, a confirmatory experiment can be set up to test whether a compound causes a specific effect on body weight. However, additional data may also be collected on food and water consumption, haematology, and clinical biochemistry. Even if these data are not incorporated in the formal analysis of the experiment, they may be used in an ancillary analysis and provide valuable information about the experiment. Furthermore, exploratory analysis of these concomitant data may suggest new hypotheses to be tested in future confirmatory experiments (Festing and Altman, 2002; Gaines Das, 2002).

A further distinction between different types of experiments is based on the type of objective of the experiment in question. The most straightforward experiment is an *estimation experiment* in which we try to estimate the value of a characteristic of a population. Next comes the *comparative experiment* in which two or more techniques, treatments, or levels of an explanatory variable are to be compared with one another. There are many examples

of comparative experiments in biomedical areas. For example in nutrition studies, different diets can be compared to one another in laboratory animals. In clinical studies, the efficacy of an experimental drug is assessed in a trial by comparing it to treatment with placebo. We shall focus primarily on designing comparative experiments for confirmation of research hypotheses.

The third type of experiment is the *optimisation experiment* which has the objective of finding conditions that give rise to a maximum or minimum response. Optimisation experiments are often used in product development, such as finding the optimum combination of concentration, temperature, and pressure that gives rise to the maximum yield in a chemical production plant. In animal experimentation optimisation experiments can be used to determine optimum conditions, such as age, gender, animal housing, etc for a response to treatment (Shaw et al., 2002). Dose-finding trials in animal research and clinical development are another example of optimisation experiments.

The fourth type of experiment is the *prediction experiment* in which the objective is to provide some statistical/mathematical model to predict new responses. Examples are dose-response experiments in pharmacology and immunoassay experiments.

The final experimental type is the *variation experiment*. This type of experiment has as objective to study the size and structure of bias and random variation. Variation experiments are implemented as *uniformity trials*, i.e. studies without different treatment conditions. For example, the assessment of sources of variation in *microtiter plate experiments*. These sources of variation can be plate effects, row effects, column effects, and the combination of row and column effects (Burrows et al., 1984). A variation

experiment can also tell us about the importance of cage location in animal experiments, where animals are kept in racks of 24 cages. Animals in cages close to the ventilation could respond differently from the rest (Young, 1989).

3.3. The pilot experiment

As researchers are often under considerable time pressure, there is the temptation to start as soon as possible with the actual experiment. However, a critical step in a new research project, which is often missed, is to spend a bit of time and resources at the beginning of the study collecting some pilot data. Preliminary experiments on a limited scale, or pilot experiments, are especially useful when we deal with time-consuming, critical, or expensive studies and are of great value for assessing the feasibility of the actual experiment. During the pilot stage, the researcher is allowed to make variations in experimental conditions such as the measurement method, the experimental set-up, etc.

The pilot experiment can also be of help to make sure that a sensible research question was asked. For instance, if our research question was about whether there is a difference in concentration of a particular protein between diseased and non-diseased tissue, it is of importance that this protein is present in a measurable amount. Carrying out a pilot experiment, in this case, can save considerable time, resources, and possible embarrassment. One could also wonder whether the effect of an intervention is large enough to warrant further study. A pilot experiment can then give a preliminary idea about the size of this effect and could be of help in making such a strategic decision.

A second crucial role for the pilot experiment is for the researcher to practice, validate and standardise the experimental techniques that will be used in the full experiment. When appropriate, trial runs of different types of assays allow fine-tuning them so that they will give optimal results. Provided the experimental techniques work well, carrying out a small-scale version of the actual experiment will yield some preliminary experimental data. These pilot data can be very valuable, as they allow to debug and fine-tune the experimental design, to calculate or adjust the required sample size, and to set up the data analysis environment.

The pilot experiment still belongs to the exploratory phase of the research project and is not part of the actual, final experiment. Therefore, to preserve the quality of the data and the validity of the statistical analysis, the pilot data cannot be included in the final data set.

—It is easy to conduct an experiment in such a way that no useful inferences can be made.

William Cochran and **Gertrude Cox** (1957).

—The development of the design of experiments is one of the greatest contributions of statistical science to science and technology. Yet, almost nobody knows about it!

John Nelder (1999)

4

Principles of Statistical Design

4.1. The importance of proper replication

In controlled experiments, scientists apply an intervention to an entity and observe its effect. The cornerstone of the scientific method is to provide evidence that the obtained results are reproducible by replicating the results. However, different aspects of the experiment can be replicated, and it is not always clear what needs to be repeated. For example, there can be multiple intervention-entity pairs, such as administering a drug to different animals. On the other hand, we could make multiple measurements of the outcome of the experiment, or we could administer the drug several times to the same animal and record each time the outcome. Alternatively, we could also conduct the entire experiment several times in different locations. The decision of what to repeat is crucial to the success of the experiment. Often the wrong type of repeat is chosen, and occasionally the

experiment is not repeated at all, or only repeated a few times. Therefore, the smart researcher will think, already at the onset of his research project, about which type of repeat he will use.

We first distinguish two types of replication. The first type consists of independently repeated data. It provides evidence of the reproducibility of our results and adds to the sample size. This type of replication is called *true*, *genuine*, or *absolute* repeat. The second type of replication does not provide evidence of the reproducibility of our results and does not add to the sample size. Terms like *pseudoreplication*, *subsampling*, or *repeated measurement* are used for this type of replication. As we shall see, erroneously taking pseudoreplicates as genuine repeats is a major problem in biomedical research.

Biologists often distinguish between biological repeats and technical repeats. We shall not follow this convention since biological replication is not always the same as a genuine repeat and technical replication is not necessarily the same as pseudoreplication. Instead, we will distinguish three levels of an experiment that can be replicated: the biological unit, the observational unit, and the experimental unit (Lazic, 2016, p. 50; Lazic et al., 2018).

4.1.1. Biological units

The biological unit is the entity about which we would like to make an inference. The purpose of the experiment is to test a hypothesis, estimate some property, or draw a conclusion about biological units. Biological units form the basis of the external validity in Section 4.4 (page 43). Examples of biological units are strains, litters, animals, cell lines, and cells.

4.1.2. Experimental units

The *experimental unit* is the smallest division of the experimental material to which a treatment can independently be assigned, such that any two units can receive different treatments. It is the entity that is randomly and independently assigned to one of the treatments. Experimental units should not influence each other and react independently to the treatment. Experimental units may correspond to (Lazic, 2016, p. 97; Lazic et al., 2018) :

- a biological unit of interest;
- groups of biological units;
- parts of a biological unit;
- a sequence of observations or measurements on a biological unit.

Correct identification of the experimental unit is of paramount importance for a solid design and analysis of the study.

4.1.3. Observational units

The last entity in our experiment is the *observational unit*, also called sampling unit or measurement unit. It is the entity on which observations or measurements are made. The observational unit may or may not correspond to the experimental unit or biological unit.

4.1.4. Choosing the right unit

In many experiments the biological unit, experimental unit, and observational unit coincide. However, in studies where replication is at multiple level, or when the biological units cannot be considered independent, investigators have often difficulties recognising the proper basic unit in their experimental material

(Lazic, 2016, pp. 94-122; Lazic et al., 2018;Vaux et al., 2012). We shall now consider some examples from biomedical research where the investigators had difficulties choosing the right unit, or made a wrong choice.

Example 4.1: Protective effect on cardiomyocytes Isolated rat cardiomyocytes provide an easy tool to assess the effect of drugs on calcium-overload (Ver Donck et al., 1986). Cardiomyocytes harvested from a single animal are isolated and seeded in plastic Petri dishes. The Petri dishes are randomly allocated to treatment with the experimental drug or with its vehicle. After a stabilisation period, the cells are exposed to a stimulating substance (i.e. veratridine), and the investigator counts the percentage viable, i.e. rod-shaped, cardiomyocytes in a dish.

The biological unit in this example is the single rat from which the cardiomyocytes were harvested. The Petri dishes are independently treated and constitute the experimental unit. The shortcoming of this experiment is that the conclusions are limited to the single animal from which the cardiomyocytes were obtained. For the conclusions to extend to "*rats*" in general, the investigators has to include more animals.

Example 4.2: Temme study - bile canaliculi diameters Temme et al. (2001) compared two genetic strains of mice, wild-type and connexin 32 (Cx32)-deficient. They measured the diameters of bile canaliculi in cryosections obtained from the livers of three wild-type and of three Cx32-deficient animals, thereby making several observations on each liver. Their results are shown in Figure 4.1.

It should be clear that Temme et al. (2001) mistakenly took the histological sections, which were the observational units, for experimental units and used them also as units of analysis. If we consider the genotype as the treatment, then it is evident that not the histological section, but the animal is the experimental unit. Moreover, histological sections obtained from the same animal will be more alike than sections from different animals. This interdependency invalidates the statistical analysis, as it was carried out by the investigators. Therefore, the correct experi-

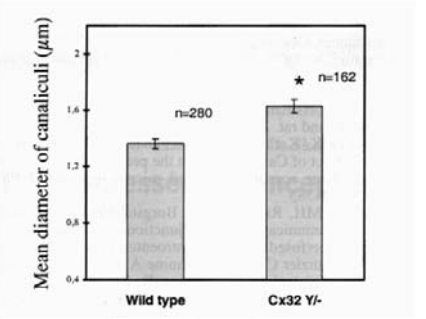


Figure 4.1. Morphometric analysis of the diameter of bile canaliculi in wild-type and Cx32-deficient liver. Means±SEM from three livers. *: P<0.005 (after Temme et al. (2001))

mental unit and unit of analysis is the animal, not the histological section. Hence, there were only three experimental units per treatment, certainly not 280 and 162 units[1]. The correct method of analysis calculates for each animal the average bile canaliculi diameter as a summary measure and takes this value as the response variable.

In studies involving microscopy, evaluation of treatment effects on individual cells or histological sections often leads to the wrong choice of experimental unit. One could wonder whether these are mistakes made out of ignorance or out of convenience, but with Hanlon's razor "*never attribute to malice that which is adequately described by incompetence*" in mind, we give these investigators the benefit of the doubt. Although such practices are always worrisome, more frightening is the fact that these studies get published in peer-reviewed high impact scientific journals.

The error made by treating multiple measurements on the same unit as if it were single measurements on multiple experimental

[1]If we recalculate the standard errors of the mean (SEM) using the appropriate number of experimental units, they are a factor 7-10 larger than the reported ones.

units is called pseudoreplication. Pseudoreplication results in an overly optimistic estimate of the precision of the experimental results and leads to false conclusions (Bate and Clark, 2014, pp. 106-110; Fry, 2014; Lazic, 2010; LeBlanc, 2004).

Independence of the experimental units can also be an issue in studies with animals housed together in cages, as is shown by the following example.

Example 4.3: Rivenson study - Toxicity of N-nitrosamines Rivenson et al. (1988) studied the toxicity of N-nitrosamines in rats and described their experimental set-up as:

> *The rats were housed in groups of 3 in solid-bottomed polycarbonate cages with hardwood bedding under standard conditions diet and tap water with or without N-nitrosamines were given ad libitum.*

Rats are undoubtedly the biological unit, but are they also the experimental unit? Since the treatment is supplied in the drinking water, it is impossible to provide different treatments to any two individual rats. Furthermore, the responses obtained within the different animals within a cage can be considered to be dependent upon one another in the sense that the occurrence of extreme values in one unit can affect the result of another unit. Therefore, the experimental unit here is not the single rat, but the cage.

An identical problem with the independence of the basic units is found in the study by Séralini et al. (2012). In their study, rats were housed in groups of two per cage, while the treatment was present in the food delivered to the cages.

Even when the animals are individually treated, e.g. by injection, group-housing can cause animals in the same cage to interact which would invalidate the assumption of independence of the experimental units. For instance, in studies with rats, a socially

dominant animal may prevent others from eating at certain times. Social dominance in rats and mice is also associated with "*barbering*", which is the trimming or plucking of fur or whiskers of cage-mates. Animals from barbered and unbarbered cages may differ in many ways (Lazic, 2016, p. 84). Another phenomenon occurs when mice are housed in a group, they usually lie together, thereby reducing their total surface area. A reduced heat loss per animal in the group is the result. Due to this *behavioural thermoregulation*, their metabolic rate is altered (Ritskes-Hoitinga and Strubbe, 2007).

In spite of the issues with group housing mentioned above, the single housing of gregarious animal species is considered detrimental to their welfare and regulations in Europe concerning animal welfare insist on group housing of such species (Council of Europe, 2006). However, when animals are housed together, the cage rather than the individual animal should be considered as the experimental unit (Fry, 2014; Gart et al., 1986, p. 28). Statistical analysis should take this into account by using appropriate techniques. Otherwise, wrongly treating the animals in a cage as independent units will yield a spurious indication of the reliability of the result. Fortunately, as is pointed out by (Fry, 2014), when the cage is the experimental unit, the total number of animals needed is not just a simple multiple of the number of animals per cage and the number of experimental units required. This is illustrated by the following example.

Example 4.4: Temme study (continued) Consider an experiment analogous to the one published by Temme et al. (2001). The investigator decides to restrict the number of histological sections for canaliculi diameter measurement to 25 per animal. Subsequently, for each animal, the average canaliculi diameter is calculated and taken as the response variable. The standard deviation (see page 57) of these average canaliculi

Table 4.1. Number of animals required per treatment group to achieve a power close to 80%, with various numbers housed per cage and with cage as the experimental unit

Number of animals per cage	Number of cages	Total number of animals	Standard deviation (of cage means)	Power
1	12	12	0.42	0.80
2	7	14	0.30	0.82
3	5	15	0.24	0.81
4	4	16	0.21	0.80
5	4	20	0.19	0.88

diameters is $0.42\ \mu m$. To detect a difference of $0.5\ \mu m$ in mean canaliculi diameter, with a power of 80% and a two-sided level of significance of 0.05, sample size calculations (see Chapter 6) show that 12 animals per treatment group are required.

When the animals are group-housed, and all animals in a cage receive the same treatment, the cage becomes the experimental unit. Accordingly, the investigator calculates for each cage the mean value of the canaliculi diameters of the individual animals. It can be shown that the standard deviation of these mean canaliculi diameters will be reduced by a factor $\sqrt{m}$, where m stands for the number of animals per cage[1]. Thus, with four animals per cage, the standard deviation of the mean diameter is half that of the standard deviation of the individual diameters.

Table 4.1 shows the effect of different scenarios of group housing for the current example. Housing the animals with two or three per cage, increases the total number of animals only marginally with two or three mice per treatment group. However, housing the animals five to a a cage would raise the total number substantially. In other words, with only two to three animals more per treatment group, we can accommodate both the animal welfare regulations and the statistical requirement of independence of experimental units.

[1]This follows from the central limit theorem

This example also highlights the three different levels of units in an experiment:

- observational unit: histological section
- biological unit: mouse
- experimental unit: cage

The former example does not take into account that, for some outcomes, group housing causes the animals to be more content, thereby reducing the variability, which would result in a more efficient experiment (Fry, 2014).

Standard reproductive studies, as they are used in teratology, involve exposure in utero of the dam. Consequently, the entire litter rather than the individual pup constitutes the experimental unit (Gart et al., 1986, p. 29; Lazic, 2016, pp. 108-109; Lazic et al., 2018).

Example 4.5: Effect on offspring A drug was tested for its capacity to reduce the effect of a mutation causing a common condition(Fry, 2014). Homozygous mutant female rats were randomly assigned to either a drug-treated or a control group. Then they were mated with homozygous mutant males, producing homozygous mutant offspring. Litters were weaned, the pups grouped five to a cage, and the effects on the offspring were observed. Here, although observations on the individual offspring were made, the biological and experimental units are the mutant dams that were randomly assigned to treatment. Therefore, the observations on the offspring should be averaged to give a single figure for each dam and these data are to be used for comparing the treatments.

A single biological unit can also relate to several experimental units, as is illustrated by the next example.

Example 4.6: Wound healing The efficacy of two agents at promoting regrowth of epithelium across a wound was evaluated by making 12 small wounds in a standardised way in a grid pattern on the back of a pig

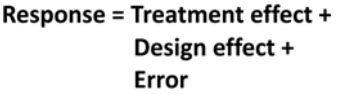


Figure 4.2. The response variable as the result of an additive model

(Fry, 2014). The wounds were far enough apart for effects on each to be independent. One of four treatments (negative control, positive control, agent 1, and agent 2) would then be applied at random to the wound in each square of the grid. In this case, the biological and experimental unit would be the wound and, as there are 12 of them, there would be three replicates for each treatment condition.

4.2. The structure of the response variable

Let us first consider some terminology. We refer to a *factor* as the condition or set of conditions that we manipulate in the experiment, e.g. the concentration of a drug. The *factor level* is the particular value of a factor, e.g. 15 $mg.kg^{-1}$, 30 $mg.kg^{-1}$, 60 $mg.kg^{-1}$. A *treatment* consists of a specific combination of factor levels, 15 $mg.kg^{-1}$ orally, 1.25 $mg.kg^{-1}$ intravenously. In *single-factor* studies, a treatment corresponds to a factor level. The characteristic that is measured and on which the effects of the different treatments are investigated, is referred to as the *response* or *dependent variable*. The definition of additional statistical terms can be found in Appendix B.

We assume that the response obtained for a particular experimental unit can be described by a simple *additive model* (Figure 4.2) consisting of the effect of the specific *treatment*, the effect of the *experimental design*, and an *error component* that describes the deviation of this particular experimental unit from the mean value of its treatment group.

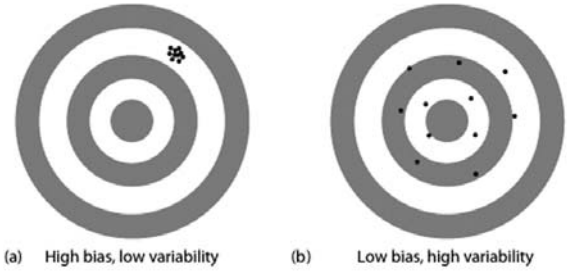


Figure 4.3. Bias and variability illustrated by a marksman shot at a bull's eye

Although the structure of this statistical model is simple, there are some strong assumptions associated with it:

- the treatment terms add rather than, for example, multiply;
- treatment effects are constant;
- the response in one unit is unaffected by the treatment applied to the other units.

These assumptions are particularly important in the statistical analysis. A statistical analysis is only correct when all of these assumptions are met.

4.3. Bias and variability

Bias and variability (Figure 4.3) are two important concepts when dealing with the design of experiments. By *bias*, we mean a systematic deviation in observed measurements from the true value, while by *variability*, we mean a random fluctuation about a central value. The terms bias and variability are also related to the concepts of *accuracy* and *precision* of a measurement process. The absence of bias means that our measurement is accurate, while little variability means that the measurement is precise.

A good experiment will eliminate or at best strongly minimise bias and will control for variability. Of the two, bias is the most important, since failure to minimise the bias of an experiment leads to erroneous conclusions. Conversely, the presence of substantial variability can sometimes be remediated by refinement of the experimental methods, increasing the sample size, or other techniques, but the study may still reach the correct conclusions.

One of the most important sources of bias in a study is the way experimental units are allocated to treatment groups, as is illustrated by the following example.

Example 4.7: Confounding bias A researcher plans to investigate the effect of an experimental treatment relative to a control treatment. She allocates all males to the control treatment and all females to the experimental treatment. At the end of the experiment, the investigator finds a substantial difference between the two treatment groups.

It is clear that, in the above example, the difference between the two treatment groups is a biased estimate of the true treatment effect, since it is intertwined or *confounded* with the difference between the males and the females. Gender is, in this case, a *confounding* factor and we refer to this type of bias as *confounding bias*.

Confounding bias can enter a study through less obvious routes. For instance, when all animals assigned to a specific treatment are kept in the same cage. Then, the treatment conditions are confounded with the influence of the different conditions of the cages. In the case where the experiment consists of a single cage per treatment, the comparisons between the treatments will be biased (Fry, 2014). The same reasoning applies to the position of the cages in a rack (Gart et al., 1986, p. 29) and the location

Table 4.2. Four types of bias affecting internal validity (after Bate and Clark, 2014, p. 123; van der Worp et al., 2010).

Type of bias	Definition	Example
Selection bias	Bias caused by a non-random allocation of animals in treatment groups.	Do we try to avoid the less healthy animals to the high dosage group?
Performance bias	Bias caused by differences, however subtle, in levels of husbandry care given to animals across treatment groups.	Are sick animals in the control group given the benefit of the doubt and kept alive longer than animals in the high dose group?
Detection bias	Bias caused when the researcher assessing the effect of the treatment knows which treatment the animal received.	When assessing animal behaviour, it is human nature to want to see a positive effect in your experiment.
Attrition bias	Bias caused by unequal occurrence and handling of deviations from the protocol and loss to follow-up between treatment groups.	If many animals are excluded from the high-dose group, should we take this into account?

of the rack itself (Gore and Stanley, 2005). Putting all the cages assigned to a particular treatment in the same rack or on the same shelf level of the rack can introduce confounding bias. In fact, the importance of rack location and shelf level on food consumption, body weight, body temperature (Gart et al., 1986, p. 29; Gore and Stanley, 2005; Greenman et al., 1983), and even on the occurrence of neoplasms (Greenman et al., 1984) have been demonstrated.

As shown in Table 4.2, four different types of confounding bias can be distinguished (Bate and Clark, 2014, p. 123; van der Worp

et al., 2010). It is essential that the researcher recognises these four sources of bias when planning the experiment and considers procedures that reduce their influence on the outcome of the study. We shall see that *randomisation* and *blinding* are effective strategies that adequately deal with the first three sources of bias.

Variability is everywhere in the natural world and is often substantial in the life sciences. While it is possible to eliminate bias from an experiment, variability can only be controlled and, despite a precise execution of the experiment, the measurements obtained in identically treated objects will yield different results. For example, cells grown in test tubes will vary in their growth rates and, in animal research, no two animals will behave the same. In general, the more complex the system that we study, the more factors will interact with each other, thereby increasing the variability between the experimental units. Consequently, experiments in whole animals will undoubtedly show more variability than in vitro studies on isolated organs.

When the variability is beyond our control or when its source cannot be measured, we shall refer to it as noise, random variability, or error. This uncontrollable variability masks the effects under investigation. It is the main reason why we have to replicate experimental units and why we use statistical methods to extract the necessary information. This large natural variability is not present in other scientific areas such as physics, chemistry, and engineering, where the studied effects are much larger than the variability.

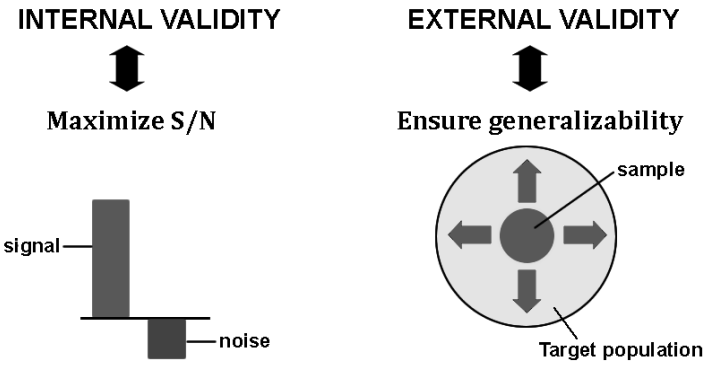


Figure 4.4. The fundamental dilemma: balancing between internal and external validity

4.4. Balancing internal and external validity

Internal validity refers to the fact that in a well-conceived experiment the effect of a given treatment can be attributed unequivocally to that treatment. Therefore, the influence of bias as mentioned above should be minimised as much as possible. However, even after complete elimination of the bias, the uncontrolled variability of the experimental material will still mask the treatment effect.

An experiment with a high level of internal validity should have a great chance to detect the effect of the treatment. If we consider the treatment effect as a signal and the inherent variability of our experimental material as noise, then a good experimental design will maximise the *signal-to-noise ratio* (Figure 4.4). Increasing the signal can be accomplished by choosing experimental material that is more sensitive to the treatment. Identification of factors that increase the sensitivity of the experimental material could be carried out in preliminary pilot experiments. Reducing

the noise is another way to increase the signal-to-noise ratio. This can be accomplished by including more experimental units, but this is not a very efficient approach. An alternative way for noise reduction is by using experimental material that is as much alike as possible, resulting in reduced natural variability. The use of cells harvested from a single animal is an example of noise reduction by using experimental material that is very similar.

External validity is related to the extent that our conclusions can be generalised to the target population (Figure 4.4). The choice of the target population, how a sample is selected from this population and the experimental procedures used in the study are all determinants of its external validity. An essential condition to ascertain external validity is that the experimental material should mimic the target population as close as possible. In animal experiments, specifying species and strain of the animal, the age and weight range and other characteristics determine the target population and make the study as realistic and informative as possible. The basis of the external validity of an experiment is formed by replication of the biological unit. External validity can become jeopardised when we work in a highly controlled environment with experimental material that is very uniform.

Thus there is a trade-off between internal and external validity, as one goes up, the other comes down. Fortunately, as we shall see, there are statistical strategies for designing a study such that the noise is reduced, while the external validity is maintained.

4.5. Requirements for a good experiment

Cox (1958, pp. 5-13) enunciated the following requirements for a *good experiment*:

1. treatment comparisons should as far as possible be free of systematic error (bias);
2. the comparisons should also be made sufficiently precise (low variability);
3. the conclusions should have a wide range of validity (external validity);
4. the experimental arrangement should be as simple as possible;
5. uncertainty in the conclusions should be assessable.

These five criteria determine the basic elements of the design of the study. We have discussed already the importance of the first three conditions in the preceding sections, the following section provides some basic strategies that can be used to fulfill these requirements.

4.6. Strategies for minimising bias and maximising signal-to-noise ratio

Maximising the signal-to-noise ratio is the fundamental principle of statistical design of experiments. (Figure 4.5). By doing so, the scientists enhances the internal validity of his study. The signal can be maximised by the proper choice of the measuring device and experimental domain. The noise is minimised by reducing bias and variability. Strategies for minimising the bias are based on good experimental practice, such as the use of controls, blinding, the presence of a protocol, calibration, randomisation, ran-

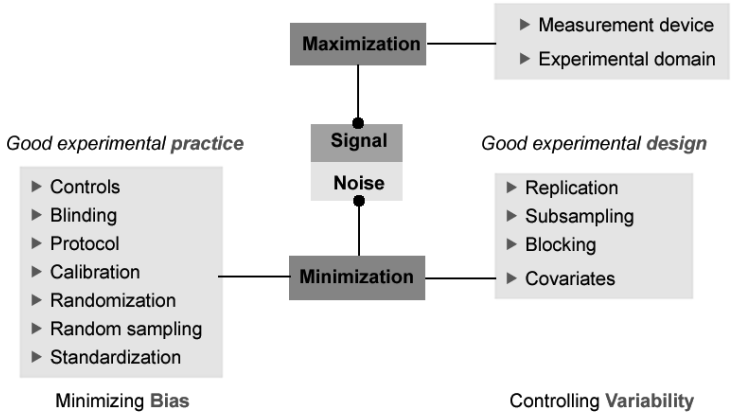


Figure 4.5. Overview of strategies for minimising the bias and maximising the signal-to-noise ratio

dom sampling, and standardisation. Variability can be minimised by elements of experimental design, such as replication, blocking, covariate measurement, and subsampling. Random sampling is also used to enhance the external validity. We shall now consider each of these strategies in more detail.

4.6.1. Strategies for minimising bias - good experimental practice

4.6.1.1. The use of controls

In biomedical studies, a control or reference standard is a standard treatment condition against which all others may be compared. The control can either be a *negative* or a *positive control*. The term *active control* is also used for the latter. Some studies, such as Example 4.6 (page 37), include both negative and positive con-

trols. In this case, the purpose of the positive control is mostly to provide an internal validation of the experiment[1].

When negative controls are used, subjects can sometimes act as their own control (*self-control*), in which case the subject is first evaluated under standard conditions (i.e. untreated). Subsequently, the treatment is applied, and the subject is re-evaluated. This design, also called *pre-post design*, has the property that all comparisons are made within the same subject. In general, variability within a subject is smaller than between subjects. Therefore, this is a more efficient design than comparing control and treatment in two separate groups. However, the use of self-control has the shortcoming that the effect of treatment is *confounded with the effect of time*, thus introducing a potential source of bias. Furthermore, blinding, which is another method to minimise bias, is impossible in this type of design.

Another type of negative control is where one treated group does not receive any treatment at all, i.e. the experimental units remain *untouched*. Just as in the previous case of self-control, *untreated controls* cannot be blinded. Moreover, applying the treatment (e.g. a drug) often requires additional manipulation of the subjects (e.g. injection). The effect of the treatment is then intertwined with that of the manipulation, and consequently, it is potentially biased.

Vehicle control (laboratory experiments) or *placebo control* (clinical trials) are terms that refer to a control group that receives a

[1] Active controls are used mainly in equivalence or non-inferiority studies, where the purpose is to show that a given therapy is equivalent or non-inferior to an existing standard.

matching treatment condition without the active ingredient. Another term for this type of control, in the context of experimental surgery, is *sham control*. In the sham control group subjects or animals undergo a faked operative intervention that omits the step thought to be therapeutically necessary. This type of vehicle control, placebo control, or sham control is the most desirable and effectively minimises bias. In clinical research, the placebo-controlled trial has become the gold standard. However, in the same context of clinical research ethical consideration may sometimes preclude its application.

4.6.1.2. Blinding

Researchers' expectations may influence the study outcome at many stages. For instance, the experimental material may unintentionally be handled differently based on the treatment group, or observations may be biased to confirm prior beliefs. *Blinding* is an effective strategy for minimising this subconscious *experimenter bias*.

In a recent survey of studies in evolutionary biology and the life sciences at large, Holman et al. (2015) found that in unblinded studies the mean reported effect size was inflated by 27%, and the number of statistically significant findings was substantially larger as compared to blinded studies. The importance of blinding in combination with randomisation in animal studies was also highlighted by Hirst et al. (2014). Despite its importance, blinding of experimenters is often neglected in biomedical research. For example, in a systematic review of studies on animals in nonclinical research, van Luijk et al. (2014) found that only 24% reported blinded assessment of the outcome, while only 15% considered blinding of the caretaker/investigator.

Two types of blinding must be distinguished. In *single blinding*, the investigators are uninformed regarding the treatment condition of the experimental subjects. Single blinding neutralises *investigator bias*. The term *double blinding* in laboratory experiments means that both the experimenter and the observer are uninformed about the treatment condition of the experimental units. In clinical trials double blinding means that both investigators and subjects are unaware of the treatment condition.

Two strategies for blinding have found their way to the laboratory: *group blinding* and *individual blinding*. Group blinding involves identical codes, say *A, B, C,* etc., for entire treatment groups. The major drawback of this approach is that, when results accumulate, the investigator will be able to break the code. A much better blinding strategy is to assign a code (e.g. sequence number) to each experimental unit individually and to maintain a list that indicates which code corresponds to which particular treatment. The sequence of the treatments in the list should be randomised. In practice, this individual blinding procedure often involves an independent person that maintains the list and prepares the treatment conditions (e.g. drugs).

Especially, when the outcome of the experiment is subjectively evaluated, blinding *must* be considered. However, there is one situation where blinding does not seem to be appropriate, namely in toxicologic histopathology. Here, the bias that would be reduced by blinding is a bias favouring the diagnosis of a toxicological hazard and therefore a conservative safety evaluation, which is appropriate in this context (Neef et al., 2012). In contrast, a blinded evaluation would result in a reduction in the sensitivity to detect anomalies. In this context, Holland and Holland (2011) suggested

that for toxicological work both an unblinded and blinded evaluation of histologic material should be carried out.

4.6.1.3. The presence of a technical protocol

The presence of a written technical protocol, describing in full detail the specific definitions of measurement and scoring methods is imperative to minimise potential bias. The technical protocol specifies practical actions and gives guidelines for lab technicians on how to manipulate the experimental units (animals, etc.), the materials involved in the experiment, the required logistics, etc. It also gives details on data collection and processing. Last but not least, the technical protocol lays down the personal responsibilities of the technical staff. The importance and contents of the other protocol, the study protocol, will be discussed further in Chapter 8.

4.6.1.4. Calibration

Calibration is an operation that compares the output of a measurement device to standards of known value, leading to correction of the values indicated by the measurement device. Calibration neutralises *instrument bias*, i.e. the bias in the investigator's measurement system.

4.6.1.5. Randomisation

Randomisation, together with blinding, is an essential tool for the elimination of confounding bias in experiments. In an overview of systematic reviews of animal studies, Hirst et al. (2014) found that failure to randomise is likely to result in overestimation of treatment effects across a range of disease areas and outcome measures.

Formal randomisation, in our context, is the process of allocating experimental units to treatment groups or conditions according to a well-defined stochastic law[1]. Randomisation is a critical element in proper study design. It is an objective and scientifically accepted method for the allocation of experimental units to treatment groups. Formal randomisation ensures that the effect of uncontrolled sources of variability has equal probability in all treatment groups. In the long run, randomisation balances treatment groups on unimportant or unobservable variables, of which we are often unaware. Any differences that exist in these variables after randomised treatment allocation can then be attributed to the *play of chance*.

Example 4.8: Plate location effect in 96-well plates Experiments in 96-well microtiter plates suffer from a substantial systematic bias that causes under- or over-estimation of biological activity. In addition to the systematic difference between plates, there is also a bias that depends on the row and column of the well. Burrows et al. (1984) described the presence of these plate location effects in ELISA assays and Levasseur et al. (1995) and Faessel et al. (1999) reported the presence of systematic parabolic patterns in cell growth experiments. These investigators were not able to show conclusively the underlying causes of this systematic error, which could be of considerable magnitude. Therefore, they concluded that only by random allocation of the treatments to the wells these systematic errors could be minimised.

To accomplish this, they developed an ingenious method for randomising treatments in 96-well microtiter plates. Drugs are serially diluted into tubes in a 96-tube rack. Next, a randomisation map is generated using a MS Excel™-macro. The randomisation map is then taped to the top of an empty tube rack. The original tubes are then transferred to the new rack by pushing the numbered tubes through the corresponding numbered rack positions. Using a multipipette system drug-containing medium is then transferred from the tubes to the wells of a 96-well mi-

[1] See glossary in Appendix B.

crotiter plate. At the end of the assay, the random data file generated by the plate reader is imported into the Excel spreadsheet and automatically resorted.

This procedure is one way of dealing with the presence of bias in microtiter plates. As we shall see, alternative methods use an appropriate experimental design such as a Latin square design (see Section 5.2.4). A more complicated method to deal with this bias is to construct a statistical model that corrects for the row-column effect (Schlain et al., 2001; Straetemans et al., 2005).

As shown in the above example, randomisation is an operation that effectively can turn lethal bias into more manageable random error (Vandenbroeck et al., 2006). The random allocation of experimental units to treatment conditions also provides an unbiased estimate of the standard error of the treatment effects, makes experimental units independent of one another and justifies the use of significance tests. In this sense, randomisation is a necessary condition for a rigorous statistical analysis (Cox, 1958, pp. 83-85; Fisher, 1935, pp. 17-21). Besides, randomisation is also of use as a device for blinding the experiment.

Example 4.9: Randomisation at different stages In neurological research, investigators randomly allocate animals to treatment groups. At the end of the experimental procedures, the animals are sacrificed, slides are made from specific target areas of the brain and investigated microscopically. At each of these stages, errors can arise leading to biased results if the original randomisation order is not maintained.

As shown in the above example, errors and bias can arise at various stages in the experiment. Therefore, to eliminate all possible bias, it is essential to cover all important sources of variability connected with the experimental units in the randomisation procedure. Furthermore, as far as practical, experimental units receiving the same treatment should be dealt with separately and independently at all stages at which errors may arise. If this is

not the case, additional randomisation procedures should be introduced (Cox, 1958, pp. 81-83).

To summarise, randomisation should apply to each stage of the experiment (Fry, 2014):

- allocation of independent experimental units to treatment groups
- order of exposure to test alteration within an environment
- order of measurement

Therefore, when the cage is the experimental unit, the investigators should consider randomising the arrangement of cages within the rack or room, the administration of substances, and the taking of samples, even though this adds an extra burden to the laboratory staff. Of course, all this can be accomplished by maintaining the original randomisation sequence throughout the experiment.

Formal randomisation requires the use of a *randomisation device*, such as the tossing of a coin, the use of randomisation tables (Cox, 1958, pp.295-300), or the use of computer software (Kilkenny et al., 2009). Methods of randomisation using MS Excel™ and **R** (R Core Team, 2017) are contained in Appendix D.

Some investigators are convinced that not randomisation, but a *systematic arrangement* is the preferred way to eliminate the influence of uncontrolled variables. For example, when one wants to compare two treatments A and B, one possibility is to set up pairs of experimental units and always assign treatment A to the first member of the pair and B to the remaining unit. However, if there

	Group 1	Group 2
	256	350
	248	290
	314	301
	316	302
	295	272
	309	330
Mean	289.7	307.5
St. Dev.	30.2	28.1

Group 1	Group 2
256	350
301	290
314	248
316	302
295	272
309	330
298.5	298.7
22.3	37.4

Figure 4.6. Trying to improve the random allocation by reducing the intergroup variability increases the intragroup variability

is a systematic effect, such that the first member of each pair consistently yields a higher or lower result than the second member, the estimated treatment effect will be biased.

To accommodate for this possible bias, some researchers devised rather smart arrangements, e.g. the alternating sequence *AB, BA, AB, BA,*.... However, here too it cannot be excluded that a particular pattern in the uncontrolled variability coincides with this arrangement. For instance, if the investigator tests eight experimental units per day, the first unit on a given day will always receive treatment *A*. Furthermore, in the case of such a systematic arrangement, the statistical analysis, which is now based on the false assumption of randomness, will underestimate the true standard error and yield misleading results.

Example 4.10: Improvement on randomisation A researcher had randomised rats over two treatment groups, but then realised that the mean body weights of the treatment groups were different. So, she exchanged two animals between the treatment groups (Figure 4.6). Now, the mean body weight of the two groups was almost the same. However,

in doing so, she did not realise that the within-group variability was altered and was now markedly different between the groups.

Like in the above example, researchers are sometimes tempted to improve on the initial random allocation of animals by rearranging individuals so that some baseline characteristics are the same. However, this improvement by subjective assessment (Cox, 1958, pp. 77-78) changes other aspects of the data, such as the variability. Furthermore, randomisation, a necessary condition for valid statistical analysis, is broken and the results will be wrong. Later, we shall see that blocking is the correct way of dealing with the heterogeneity of baseline data such as body weight.

Formal randomisation must be distinguished from *haphazard allocation* to treatment groups (Kilkenny et al., 2009). For example, an investigator wishes to compare the effect of two treatments (A and B) on the body weight of rats. All twelve animals are delivered in a single cage to the laboratory. The researcher then takes six animals out of the cage and assigns them to treatment A, while the remaining animals will receive treatment B. Although, many scientists would consider this as a random assignment, it is not. Indeed, one could imagine the following scenario: heavy animals react slower and are easier to catch than the smaller animals. Consequently, the first six animals will on average weigh more than the remaining six.

Example 4.11: Moment of randomisation In an experiment, brain cells were taken from animals and placed in Petri dishes, such that one Petri dish corresponded to one particular animal. The Petri dishes were then randomly divided into two groups and placed in an incubator. After 72 hrs incubation, one group of Petri dishes was treated with the experimental drug, while the other group received solvent.

Although the investigators made a serious effort to introduce randomi-

sation in their experiment, they overlooked the fact the placement of the Petri dishes in the incubator introduced a systematic error. Instead of randomly dividing the Petri dishes into two groups at the start of the experiment, they should have made random treatment allocation *after* the incubation period.

As pointed out before, it is essential for the randomisation procedure to cover all substantial sources of variability connected with the experimental units. As a rule, randomisation should be performed immediately before treatment application. Furthermore, after the initial randomisation procedure, the randomised sequence of the experimental units must be maintained. Otherwise, a new randomisation procedure is required.

4.6.1.6. Random sampling

Using a random sample from the target population increases the *external validity* of the study and allows us to make a broad inference, based on a population model of inference (Lehmann, 1975, pp. 55-65). However, in practice, it is often difficult or impractical to conduct studies with actual random sampling. For instance, clinical trials are usually conducted using eligible patients from a small number of study sites; while animal experiments are based on the available animals. The failure to select a random sample from the target population certainly limits the external validity of the study. Accordingly, it is one of the main reasons that results are not always replicable.

In some cases, maximising the external validity of the study is of great importance, especially in studies that attempt to make a broad inference towards the target population (population model). In gene expression experiments, for instance, scientists try to relate a specific pathology to the differential expression

of some specific gene probes (Nadon and Shoemaker, 2002). To minimise possible bias in the results, the investigators should use a random sample from the target population.

4.6.1.7. Standardisation

Standardisation of the experimental conditions is an effective way for bias minimisation. Besides, it also can also be used to reduce the intrinsic variability of the results. Examples of standardisation of the experimental conditions are the use of genetically or phenotypically uniform animals, environmental and nutritional control, acclimatisation, and standardisation of the measurement system. However, as discussed before, too much standardisation of the experimental conditions can jeopardise the external validity of the results.

4.6.2. Strategies for controlling variability - good experimental design

4.6.2.1. Replication of the experimental unit

Ronald Fisher[1] noted in his pioneering book *The Design of Experiments* (Fisher, 1935, pp. 60-64) that replication at the level of the experimental unit serves two purposes. The first is to increase the precision of estimation and the second is to supply an estimate of error by which the significance of the comparisons is to be judged.

The *precision* of an experiment depends on the standard deviation[2] (σ) of the experimental material and inversely on the num-

[1]Sir Ronald Aylmer Fisher (Londen,1890 - Adelaide 1962) is considered a genius who almost single-handedly created the foundations of modern statistical science and experimental design.

[2]See Appendix B for the definitions of standard deviation and standard error

ber of experimental units n. In a comparative experiment with two treatment groups (X_1) and (X_2) of equal size n and with mean values μ_1 and μ_2 respectively, this precision is quantified by the standard error of the difference between the two averages ($\mu_1 - \mu_2$) as:

$$\sigma_{\mu_1-\mu_2} = \sigma \times \sqrt{2/n} \tag{4.1}$$

where σ is the common standard deviation and n is the number of experimental units in each treatment group.

The standard deviation is determined by the intrinsic variability of the experimental material and the precision of the experimental work. Reduction of the standard deviation is only possible to a limited extent by refining experimental procedures. However, one can, by increasing the number of experimental units, effectively enhance the experiment's precision. Unfortunately, due to the inverse square-root dependence of the standard error on the sample size, this is not an efficient way to control the precision. Indeed, the standard error is halved by a fourfold increase in the number of experimental units, but a hundredfold increase in the number of units is required to divide the standard error by ten. In other words, replication at the level of the experimental unit is an effective, but also an expensive strategy to improve the precision. As we will see later, choosing an appropriate experimental design that takes into account the different sources of variability that can be identified, is a more efficient way to increase the precision.

4.6.2.2. Subsampling

As mentioned before, the standard deviation can be reduced to a limited extent by standardisation of the experimental conditions, but this method jeopardises the external validity of the experiment. However, in some experiments, it is possible to manipulate

the physical size of the experimental units. In general, units of a larger size will show a smaller relative variability than units of a smaller size, thereby improving the precision of the estimates. In other experiments, there are multiple levels of sampling. The process of taking samples below the primary level of the experimental unit is known as *subsampling*[1] ((Selwyn, 1996, pp. 48-49); (Lazic, 2016, pp. 94-122; Lazic et al., 2018) .

The experiment reported by Temme et al. (2001) where the diameter of bile canaliculi was measured in 3 animals/experimental condition, is an example of subsampling with animals at the primary level and histological sections at the subsample level. Multiple observations or measurements made over time are also considered subsamples. In biological and chemical analyses, it is standard practice to duplicate or triplicate independent determinations on samples from the same experimental unit. In this case samples and determinations within samples constitute two distinct levels of subsampling.

When subsampling is present, the standard deviation σ used in the comparison of the treatment means can be considered as composed of the variability between the experimental units (between-unit variability) and the variability within the experimental units (within-unit variability) and is equal to (Cox, 1958, p. 180; Snedecor and Cochran, 1980, p. 529):

$$\sigma = \sqrt{\sigma_n^2 + \frac{\sigma_m^2}{m}} \tag{4.2}$$

[1]Biologists refer to subsamples often as *technical* repeats.

where n and m are the number of experimental units and subsamples and σ_n and σ_m the between sample and within sample standard deviation.

Equation 4.1, which defined the standard error of the difference between two treatment groups, now becomes:

$$\sigma_{\mu_1-\mu_2} = \sqrt{\frac{2}{n}(\sigma_n^2 + \frac{\sigma_m^2}{m})} \tag{4.3}$$

Thus, by increasing the number of experimental units n we reduce the total variability, while the subsample replication m only affects the within-unit variability. A large number of subsamples makes only sense when the variability of the measurement at the sublevel σ_m is substantial as compared to the between-unit variability σ_n. How to determine the required number of subsamples will be discussed in Section 6.5 (page 128). As a conclusion, we can say that subsample replication is not identical (i.e. pseudoreplication) and not as useful as replication on the level of the actual experimental unit.

4.6.2.3. Blocking

Example 4.12: Running speed of dogs Consider a (hypothetical) study to compare the effect of two diets on running speed of dogs (Ruxton and Colegrave, 2003, pp. 70-71). We can do this by taking six dogs of varying age and then randomly allocate three dogs to diet A and the three remaining to diet B. However, as shown in the left panel of Figure 4.7, the variability between dogs will mask to a great extent the effect of diet.

A more intelligent way to set up the experiment is to group the dogs by age and make all comparisons within the same age group, thus removing the effect of different ages (Figure 4.7, right panel). With the variability due to age removed, the effect of the diets within the age groups is much more apparent.

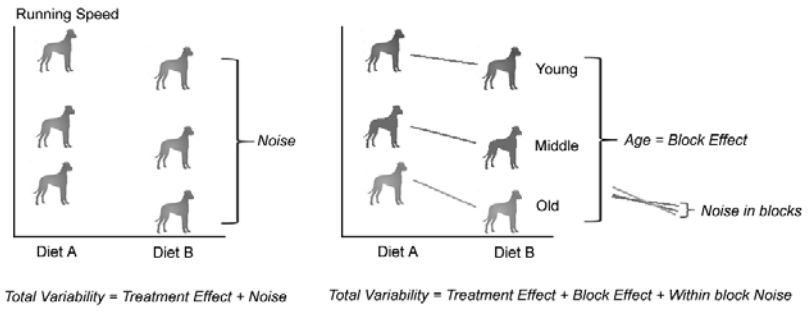


Figure 4.7. The effect of blocking illustrated by a study of the effect of diet on running speed of dogs. Not taking the age of the dog into account (left panel) masks most of the effect of the diet. In the right panel dogs are grouped (blocked) according to age and comparisons are made within each age group. The latter design is much more efficient (after Ruxton and Colegrave (2003, pp. 70-71)).

If we can identify one or more factors other than the treatment condition as potentially influencing the outcome of the experiment, then it may be appropriate to group the experimental units on these factors. We refer to such groupings as *blocks* or *strata*. Units within a block are then randomly assigned to the treatments. Examples of blocking factors are plate (in microtiter plate experiments), animal, cage, litter, date of experiment, or categorisations of continuous baseline data such as body weight, baseline measurement of the response, etc. What we effectively do by blocking, is to partition the variability between the individuals into variability between blocks and variability within blocks.

When the blocking factor has a substantial effect on the response, the between-block variability will be greater than the within block variability. We take this into account in the analysis of the data, typically by an analysis of variance (ANOVA) with blocks as an additional factor. Comparisons of treatments are

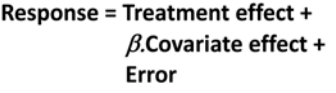


Figure 4.8. Additive model with a linear covariate adjustment

then carried out within blocks, where the variability is much smaller.

Blocking is an effective and efficient way to enhance the precision of the experiment. Furthermore, blocking allows reducing the bias due to an imbalance in baseline characteristics that are known to affect the outcome. However, blocking does not eliminate the need for randomisation. Within each block, one should randomly assign the experimental units to the treatments, thereby removing the effect of the remaining unknown sources of bias.

4.6.2.4. Covariates

Blocking on a baseline characteristic such as body weight is one possible strategy to eliminate the variability induced by the heterogeneity in weight of the animals or patients. Instead of grouping in blocks, or in addition to, one can also make use of the actual value of the measurement. Such a concomitant measurement is referred to as a covariate. It is an uncontrollable but measurable attribute of the experimental units (or their environment) that is *unaffected by the treatments* but may have an influence on the measured response. Examples of covariates are body weight, age, ambient temperature, measurement of the response variable before treatment, etc. The covariate filters out the effect of a particular source of variability. Rather than blocking, it represents a quantifiable attribute of the system studied.

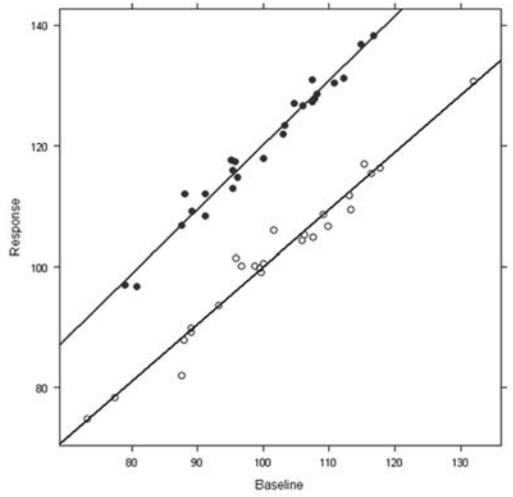


Figure 4.9. Results of an experiment with baseline as covariate. There is a linear relationship between the covariate and the response and this relationship is the same in both treatment groups.

The statistical model underlying the design of an experiment with covariate adjustment is conceptualised in Figure 4.8. The model implies that there is a linear relationship between the covariate and the response and that this relationship is the same in all treatment groups. In other words, there is a series of parallel curves, one per treatment group, relating the response to the covariate. This is exemplified in Figure 4.9, showing the results of an experiment with two treatment groups in which the baseline measurement of the response variable served as a covariate. The relationship between the covariate and the response variable is linear, and the slopes of the two lines are almost the same, i.e. the two lines are almost parallel.

4.7. Simplicity of design

In addition to sufficient precision, good external validity, and lack of bias, Cox (1958, p. 11) also stated that the design of our experiment should be as simple as possible. When the design of the experiment is too complex, it may be difficult to ensure adherence to a complicated schedule of alterations, especially if these are to be carried out by relatively unskilled people. An uncomplicated experimental design has the additional advantage that the statistical analysis will also be straightforward, without making unreasonable assumptions.

4.8. The calculation of uncertainty

This is the last of Cox's precepts (Cox, 1958, p. 12) for a *good experiment* (see Section 4.5, page 45). It is the only statistical requirement, but it is also the most important one. Unfortunately, it is also the requirement that researchers often neglect. Fisher (1935, p. 35) already lamented that:

> *It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available.*

Without the ability to estimate the error[1], there is no basis for statistical inference. Therefore, in a well-conceived experiment, we should always be able to calculate the uncertainty in the estimates of the treatment comparisons. This usually means esti-

[1] There is a big difference between the calculation of the standard error and its validity as an estimator of the true population standard error, which depends on some stringent criteria

Table 4.3. Multiplication factor to correct for the bias in estimates of the standard deviation based on small samples, after Bolch (1968).

n	Factor
2	1.253
3	1.128
4	1.085
5	1.064
6	1.051

mating the standard error of the difference between the treatment means.

To make this calculation in a rigorous manner, the set of experimental units must respond independently to a specific treatment and may only differ in a random way from the set of experimental units assigned to the other treatments. This requirement again stresses the importance of the independence of the experimental units and the randomness of the treatment allocation.

When the number of experimental units is small, the sample estimate s of the standard deviation is biased and underestimates the true standard deviation σ[1]. A multiplication factor to correct for this bias for normal distributions was proposed by Bolch (1968) and, for sample sizes smaller than seven, is given in Table 4.3. For a sample size of three, the estimate s should be increased with 13% to obtain an unbiased estimate of the population standard deviation σ.

Alternatively, one can also make use of the results of previous experiments to *guesstimate* the new experiment's standard devia-

[1] See page 121 for more details on the uncertainty in estimates of the standard deviation

tion. However, we then make the strong assumption that random variability is the same in the new experiment.

—And so it was ... borne in upon me that very often, when the most elaborate statistical refinements possible could increase the precision by only a few percent, yet a different design involving little or no additional experimental labour might increase the precision two-fold, or five-fold or even more

Ronald A. Fisher (1962).

5

Common Designs in Biological Experimentation

5.1. The three aspects of experimental design

When planning an experiment, one has to choose from a multitude of designs, some of which are employed more commonly than others in the area of biological research Unfortunately, the literature about experimental design is littered with technical jargon, which makes its understanding quite a challenge. There are completely randomised designs, randomised complete block designs, factorial designs, split plot designs, Latin square designs, Greco-Latin squares, Youden square designs, lattice designs, Placket-Burman designs, simplex designs, Box-Behrken designs, etc.

It helps to find our way through this jungle of designs by keeping in mind that the fundamental principle of experimental design

is to provide a synthetic approach to *minimise bias* and *control variability*. Furthermore, as shown in Figure 5.1, we can consider each of the specialised experimental designs as integrating three different aspects of the design (Hinkelmann and Kempthorne, 2008, pp. 33-34):

- the treatment design,
- the error-control design,
- the sampling & observation design.

The *treatment design* is concerned about which treatments are to be included in the experiment and is closely linked to the goals and aims of the study. Should a negative or positive control be incorporated in the experiment, or should both be present? How many doses or concentrations should be tested and at which level? Is the interaction of two treatment factors of interest or not? The *error-control design* implements the strategies that we learned in Section 4.6.2 (page 57) to filter out different sources of variability. The *sampling & observation* aspect of our experiment is about how experimental units are sampled from the population, how and how many subsamples should be drawn, etc.

These three aspects of experimental design determine the complexity of the study and the required resources. The number of treatments, the number of blocks, and the standard error govern the required resources, i.e. the number of experimental units, of a study. The more treatments or, the more blocks, the more experimental units are needed. The complexity of the experiment is determined by the underlying statistical model of Figure 4.2. In particular, the error-control design defines the study's complexity. The randomisation process is a major part of this error-control design. As argued before, a justified and rigorous estimation of

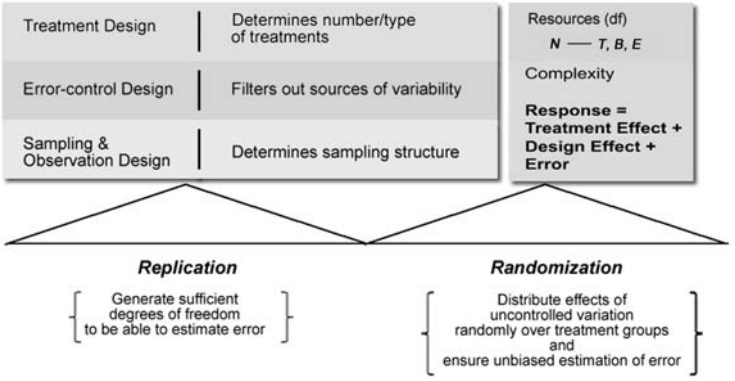


Figure 5.1. The three aspects of the design determine its complexity and the required resources

the standard error is only possible in a randomised experiment. Randomisation has the additional advantage that it distributes the effects of uncontrolled variability randomly over the treatment groups.

Replication of experimental units is a key factor for an effective experiment. The number of experimental units should be sufficient, such that an adequate number of *degrees of freedom* are available for estimating the experiment's precision (standard error). The degrees of freedom are the number of independent pieces of information to estimate the standard error. This parameter is related to the sampling & observation aspect of the design.

These three aspects of experimental design provide a framework for classifying and comparing the different types of experimental design that are used in the life sciences. As we shall see, each of these designs has its advantages and disadvantages.

5.2. Error-control designs

5.2.1. The completely randomised design

The completely randomised design (CRD) is the most common and simplest possible error-control design for comparative experiments. It is often the default design used by investigators who do not really think about design problems. The idea is simple, each experimental unit is randomly assigned to exactly one treatment condition.

In the following example of a CRD, the investigators used randomisation, blinding, and individual housing of animals to guarantee the absence of systematic error and independence of experimental units.

Example 5.1: Proliferation of gastric epithelial cells An experiment was set up to assess the effect of chronic treatment with two experimental drugs as compared to their vehicles on the proliferation of gastric epithelial cells in rats. A total of 40 rats were randomly divided into four groups of each ten animals, using the MS Excel randomisation procedure described in Appendix D. To guarantee the independence of the experimental units, the animals were kept in separate cages [1]. Cages were distributed over the racks according to their sequence number. Blinding was accomplished by letting the sequence number of each animal correspond to a given treatment. One laboratory worker was familiar with the codes and prepared the daily drug solutions. Treatment codes were concealed from the rest of the laboratory staff that was responsible for the daily treatment administration and final histological evaluation.

[1] The experiment dates from before the implementation of the guidelines regarding group housing of gregarious animals (Council of Europe, 2006). However, the design is easily adapted to designs with animals housed in groups of three or four. The total number of animals should then be raised to 48 or 60 and cages are the experimental units.

The advantage of the CRD is that it is simple to implement, as experimental units are simply randomised to the various treatments. Its disadvantage is the lack of precision in the comparisons among the treatments, which is based on the variability between the experimental units.

5.2.2. The randomised complete block design

The concept of blocking as a tool to increase efficiency by enhancing the signal-to-noise ratio has already been introduced in Section 4.6.2.3 (page 60). The basic idea behind blocking is to partition the total set of experimental units into subsets (blocks) that are as homogeneous as possible. This allows eliminating the effect of isolated extraneous factors that contribute to the variability of the response variable from the comparisons between treatment groups. (Hinkelmann and Kempthorne, 2008, p. 45).

In a randomised complete block design (RCBD), the randomisation procedure randomises treatments separately within each block. The RCBD is called *complete* since all treatments are applied within each block. Consequently, treatments can be compared with one another within the blocks, which makes it a very useful and reliable error-control design. When a study is designed such that the number of experimental units within each block and treatment is equal, the design is also called a balanced design. A few examples will illustrate the use of the RCBD in the laboratory.

Example 5.2: Proliferation of gastric epithelial cells (continued) In the CRD of Example 5.1 (page 70), the rats were individually housed in a rack consisting of five shelves of each eight cages. On different shelves, rats are likely to be exposed to multiple varieties of light intensity, temperature, humidity, sounds, views, etc. As argued in Section 4.3,

housing conditions can lead to biased outcomes. Also here, the investigators suspected shelf level to affect the results. Therefore, they decided to switch to an RCBD in which the blocks correspond to the five shelves of the rack. Within each block separately the animals were randomly allocated to the treatments, such that in each block each treatment condition occurred exactly twice. This example also illustrates that, although all treatments are present in each block in an RCBD, more than one experimental unit per block can be allocated to a treatment condition.

There are two main reasons for choosing an RCBD design above a CRD. Suppose there is an extraneous factor that is strongly related to the outcome of the experiment. It would be most unfortunate if our randomisation procedure yielded a design in which there was a great imbalance on this factor. If this were the case, the comparisons between treatment groups would be confounded with differences in this nuisance factor and be biased. The second main reason for an RCBD is its possibility to considerably reduce the error variation in our experiment, thereby making the comparisons more precise. The main objection to an RCBD is that it makes the strong assumption that there is no interaction between the treatment variable and the blocking characteristics, i.e. that the effect of the treatments is the same among all blocks.

Example 5.3: Proliferation of gastric epithelial cells (continued) In addition to the shelf height, the investigators also suspected that the body weight of the animals might affect the results. Therefore, the animals were numbered in order of increasing body weight. The first eight animals were placed on the top shelf and randomised to the four treatment conditions. Then, the next eight animals were placed on the second shelf and randomised, etc. The top row of the rack contained the animals with the lowest body weight, and the bottom row the heaviest animals. Within a shelf, the animals were as much alike as possible with

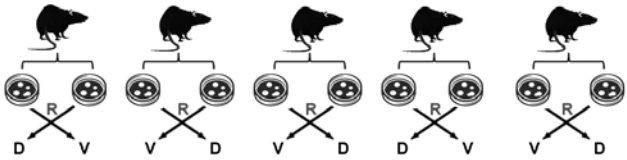


Figure 5.2. Outline of an experiment on isolated cardiomyocytes. Cardiomyocytes of a single animal were isolated and seeded in plastic Petri dishes. From the resulting five pairs of Petri dishes, one member was randomly assigned to drug treatment, while the remaining member received the vehicle.

regard to body weight and therefore, the resulting design simultaneously controlled for shelf height as well as body weight.

As in the above example, sometimes we want to include two or more blocking factors to reduce the unexplained variability in our experiment. In this case, we can treat the combinations of the blocking factor as a single new blocking factor and use it in an RCBD. The two (or more) blocking factors are now entirely confounded with each other, and it is impossible to determine the influence each blocking factor separately has on the response. However, we are not interested in the effects of these factors, but only want to control for them.

5.2.2.1. The paired design

Consider an RCBD with only two treatments and two experimental units per block. This is the simplest case of the RCBD and is called a *paired design*. We shall use the paired design to illustrate some properties of the RCBD.

Example 5.4: Protective effect on cardiomyocytes (continued) Example 4.1 (page 32) was about an experiment in which the protective effect of drugs on isolated cardiomyocytes was evaluated. Since comparisons were carried out using a single animal as biological unit, the con-

clusions of the experiment could not be generalised. Therefore, a paired experiment with myocytes from different animals and with the animal as blocking factor was carried out.

Figure 5.2 illustrates the experimental setting. Cardiomyocytes were harvested from a total of five animals. From each animal, two Petri dishes containing exactly 100 cardiomyocytes were prepared. From the resulting five pairs of Petri dishes, one member was randomly assigned to drug treatment, while the remaining member received the vehicle. After stabilisation and exposure to the stimulus, the number of viable cardiomyocytes in each Petri dish was counted. The resulting data are displayed in Figure 5.3.

The animals are the biological units, while the ten Petri dishes are the experimental units since we independently assigned the Petri dishes to vehicle or drug. The experimenter registers whether a myocyte is viable or not, so the cell constitutes the observational unit. The statistical analysis should take the particular structure of the experiment into account. More specifically, the blocking has imposed restrictions on the randomisation such that data obtained from one animal cannot be freely interchanged with that from another animal. This is illustrated in the right panel of Figure 5.3 by the lines that connect the data from the same animal. It is clear that for each pair the drug-treated Petri dish consistently yielded a higher result than its vehicle control counterpart. Since the different pairs (animals) are independent of one another, the mean difference and its standard error can be calculated. The mean difference is 7.0 with a standard error of 2.51.

We will now use the above example to have a look at the gain in efficiency obtained by blocking.

5.2.2.2. Efficiency of the RCBD

How much more efficient is the RCBD as compared to the CRD and does blocking always work? A few examples will help us to answer these questions.

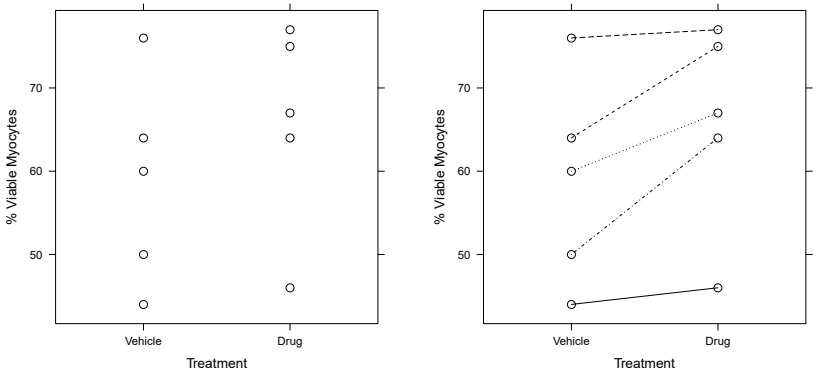


Figure 5.3. The gain in efficiency induced by blocking, illustrated in a paired design. In the left panel, we consider the myocyte experiment as a CRD and ignore the pairing. The two treatment groups largely overlap with each other. In this case, it is impossible to discern any difference due to treatment. In the right panel, we recognise the fact that the experimental units are in pairs. The lines connect the data of the same animal and show a marked effect of the treatment.

Example 5.5: Protective effect in cardiomyocytes (continued) Suppose that in Example 5.4 the experimenter would not have used blocking, i.e. consider it as if he had used myocytes originating from 10 completely different animals. In this case, the investigator would have distributed the 10 Petri dishes randomly over the two treatment groups, and it would be a CRD. The 10 Petri dishes would then be randomly distributed over the two treatment groups, and we would have been confronted with a CRD.

Let us assume that the results of this hypothetical experiment were identical to those obtained in the paired design. As is illustrated in the left panel of Figure 5.3, the two groups largely overlap one another. Since all experimental units are now independent of one another, the effect of the drug is evaluated by calculating the difference between the two mean values and comparing it with its standard error[1] Obviously, the mean difference is the same as in the paired experiment. However, the stan-

[1]As already mentioned in Section 4.6.2.1, page 57, the standard error on the difference between two means is equal to $\sigma\sqrt{2/n}$

dard error on the mean difference has risen considerably from a value of 2.51 to 7.83, i.e. the use of blocking induced a substantial increase in the precision of the experiment.

Examples 5.4 and 5.5 demonstrate that carrying out a paired experiment could enhance the precision of the experiment considerably, while the conclusions have the same validity as in a CRD. Neter et al. (1996, pp. 1089-1090) provide a method to compare designs on the basis of their relative efficiency. For the design in Example 5.4, the calculations show that this paired design is 7.7 times more efficient than the CRD in Example 5.5. In other words, about 8 times as many replications per treatment with a CRD are required to achieve the same results.

However, the forming of blocks of experimental units is only successful when the criterion upon which the pairing is based, is related to the outcome of the experiment. Using as a blocking factor, a characteristic that does not have a substantial effect on the response variables is worse than useless since the statistical analysis will lose power by taking the blocking into account. This can be of particular importance for small sized experiments. The following example illustrates such a case.

Example 5.6: Protective effect in cerebral ischaemia In this study (Haseldonckx et al., 1997), the neuronal protective effect of a drug was assessed in a rat model of brain ischemia. Global cerebral ischaemia was induced in rats by bilateral clamping of the carotid arteries and severe hypotension for 9 minutes. Five minutes after termination of ischaemia, treatment with an experimental drug or its vehicle was started. Seven days after the insult, the animals were sacrificed and the number of viable neurons/mm in the CA1 layer of the hippocampus was evaluated in a blinded manner. The investigators hypothesised that there would be a substantial variability connected to the particular day of the week an animal arrived in the study. Therefore, to eliminate this source of

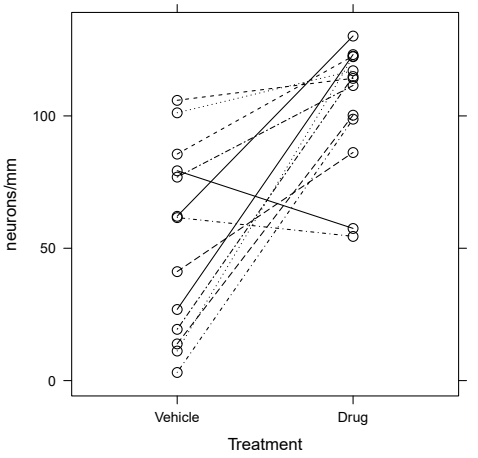


Figure 5.4. A case where the blocking criterium (animal pair) is not related to the response.

variability and possible bias, animals entered the study in pairs. From each pair, the investigators randomly selected one animal and treated it with the drug, while the remaining animal received the drug's vehicle.

The results of an experiment on 26 animals (13 pairs) are shown in Figure 5.4. Forming pairs based on the assumption of substantial daily variability was not very successful. There were animals treated with vehicle for which the outcome was low, while their drug-treated counterparts showed a large number of neurons and vice versa. Actually, there is no correlation at all (r = -0.05) between the outcome of the vehicle and of the corresponding drug-treated animals. The mean difference between the controls and treated animals is 51.2 neurons/mm, with a standard error of 12.3.

Next, let us consider the experiment as a CRD, i.e. without any restrictions on the randomisation, then the standard error on the difference between the two means is 12.0, which is almost the same. However, the standard error of the CRD has 24 degrees of freedom, while that for the paired design is based on only 12 degrees of freedom. Consequently, in this case, the paired experiment was less efficient than the CRD, since the

forming of pairs led to a serious loss in the degrees of freedom involved in the calculation of the standard error. Relative efficiency of the paired design in this experiment is reduced to 0.9, indicating that it is about 10% less efficient than a simple CRD.

As shown in the above example, blocking is only effective when the within-block variability of the response is substantially less than the between-block variability. Cox (1958, p. 168) balances the gain in precision against the loss in degrees of freedom and gives as a rule for blocking to be effective, that the ratio of the residual standard deviations must satisfy:

$$\frac{s_{crd}}{s_{rcb}} \geq \frac{1+\frac{1}{df_{crd}}}{1+\frac{1}{df_{rcb}}} \tag{5.1}$$

where s_{crd} and df_{crd} represent the standard deviation and corresponding degrees of freedom of the CRD , and s_{rcb} and df_{rcb} those of the RCBD. For the experiment in Example 5.4, we have $s_{crd} = 12.4$ and $s_{rcb} = 5.6$. Their ratio 2.21 is greater than $(1+\frac{1}{8})/(1+\frac{1}{4}) = 0.9$, indicating a substantial gain in efficiency of the RCBD. On the other hand, for the experiment in Example 5.6, we have $s_{crd} = 30.6$ and $s_{rcb} = 44.19$. Their ratio 0.69 is smaller than $(1+\frac{1}{24})/(1+\frac{1}{12}) = 0.96$, indicating a loss in efficiency of the RCBD

5.2.3. Incomplete block designs

In some circumstances, the block size is smaller than the number of treatments, and consequently, it is impossible to assign all treatments within each of the blocks. When a particular comparison is of specific interest, such as comparison with control, it is wise to include it in each of the blocks.

Table 5.1. Balanced incomplete block design for Example 5.7 with treatments A,B,C and D

	Sibling lamb pair					
Lamb	1	2	3	4	5	6
First	A	A	A	B	B	C
Second	B	C	D	C	D	D

Balanced incomplete block design (BIBD) allow all pairwise comparisons between treatments with equal precision, using a block size that is less than the number of treatments. To achieve this, the BIBD has to satisfy the following conditions (Bate and Clark, 2014, p. 55; Cox, 1958, p. 221):

- each block contains the same number of units;
- each treatment occurs the same number of times in the design;
- every pair of treatments occurs together in the same number of blocks.

BIBDs exist only for certain combinations of the number of treatments and number and size of blocks. When looking for a suitable BIBD, it can happen that if we add one or more additional treatments, a more appropriate design is found. Alternatively, omitting one or more treatments can yield a more efficient design (Cox, 1958). Software to construct BIBDs is provided by the *agricolae*-package in **R** (de Mendiburu, 2016).

Example 5.7: Lamb dietary study An experiment was carried out to assess the effect that vitamin A and a dietary protein supplement have on the weight gain of lambs over a 2-month period (Bate and Clark, 2014, pp. 55-56). There were four treatments, labelled A, B, C, and D in the study corresponding to a low dose and a high dose of vitamin A, combined with a low dose and a high dose of protein. A total of three replicates per treatment was considered as sufficient, and blocking was

carried using pairs of sibling lambs, so six pairs of siblings were used. With the number of treatments restricted to two per block, the BIBD shown in Table 5.1 was used.

A possible layout of the experiment is obtained in **R** using the *agricolae*-package (de Mendiburu, 2016) :

```
> library(agricolae)
> # Label 4 treatments A, B, C, D
> trt<-LETTERS[1:4]
> # Blocksize = 2
> # Change seed for other randomisation
> design.bib(trt,2,seed=543)$sketch
Parameters BIB
==============
Lambda     : 1
treatmeans : 4
Block size : 2
Blocks     : 6
Replication: 3

Efficiency factor 0.6666667

<<< Book >>>
     [,1] [,2]
[1,] "D"  "A"
[2,] "B"  "D"
[3,] "A"  "B"
[4,] "C"  "B"
[5,] "A"  "C"
[6,] "C"  "D"
```

The BIBD generated in the **R**-session is equivalent to the design shown in Table 5.1.

Example 5.8: Prostaglandin antagonists on fertility of mice Biggers et al. (1981) use a BIBD to compare the effects of intrauterine injection of six prostaglandin antagonists on the fertility of mice. Since the lumen of each uterine horn in a mouse is not connected with the other, each female can be used to compare the effect of two treatments.

There are six treatments (antagonists) and two experimental units (uterine horns) per animal. To compare the antagonists with each other

requires at least $\binom{6}{2}$ =15 animals. Each of the treatments is then replicated five times in the design.

The BIBD with the animal as the blocking factor is generated in **R** as:

```
> # Label 6 treatments A, B, C, D, E, F
> trt<-LETTERS[1:6]
> # Blocksize = 2
> # Change seed for other randomisation
> design.bib(trt,2,seed=4338)$sketch
Parameters BIB
==============
Lambda     : 1
treatmeans : 6
Block size : 2
Blocks     : 15
Replication: 5

Efficiency factor 0.6

<<< Book >>>
      [,1] [,2]
 [1,] "D"  "B"
 [2,] "B"  "C"
 [3,] "E"  "B"
 [4,] "A"  "F"
 [5,] "D"  "E"
 [6,] "C"  "E"
 [7,] "C"  "F"
 [8,] "C"  "D"
 [9,] "A"  "C"
[10,] "A"  "D"
[11,] "D"  "F"
[12,] "F"  "B"
[13,] "E"  "F"
[14,] "B"  "A"
[15,] "E"  "A"
```

As noted by Biggers et al. (1981), the major drawback of this design is that it assumes that the effect of a treatment in a uterine horn is local with no effects on the contralateral horn.

The designs presented in Example 5.7 and 5.8 consider only a single run of the BIBD. However, in most cases, more replicates

Table 5.2. Arrangement for a 4×4 Latin square design (LSD) controlling for column and row effects.

	Column			
Row	1	2	3	4
1	*A*	*B*	*C*	*D*
2	*D*	*A*	*B*	*C*
3	*C*	*D*	*A*	*B*
4	*B*	*C*	*D*	*A*

than given by the basic design are required. See Dean and Voss (1999) for more details on how to calculate the required number of replicates. BIBDs with more replicates can be generated by supplying an extra argument to the *design.bib*-function, specifying the number of replicates of the treatments. However, it is not always possible to obtain a balanced design, as is reported by the *design.bib*-function:

```
> # Label 6 treatments A, B, C, D, E, F
> trt<-LETTERS[1:6]
> # Block size 2,
> # 6 replicates of each treatment
> out<-design.bib(trt,2,6,seed=4338)$sketch

 Change r by  5, 10, 15, 20 ...

> # replicates must be multiple of 5, so 10 is OK
> out<-design.bib(trt,2,10,seed=4338)$sketch

Parameters BIB
==============
Lambda     : 2
treatmeans : 6
Block size : 2
Blocks     : 30
Replication: 10

Efficiency factor 0.6

<<< Book >>>
```

Table 5.3. Arrangement for the 5×5 LSD of Example 5.9 of placing animals in cage racks. The numbers indicate the five dose levels of the test compound (0 = vehicle control, - = empty)

	Column					
Row	1	2	3	4	5	6
1	*3*	*1*	*10*	*0*	*30*	-
2	*1*	*0*	*3*	*30*	*10*	-
3	*10*	*3*	*30*	*1*	*0*	-
4	*0*	*30*	*1*	*10*	*3*	-
5	*30*	*10*	*0*	*3*	*1*	-

5.2.4. Latin square designs

When the experimental units exhibit heterogeneity in two directions, such as the rows and the columns of cage racks, then we require a block design that accounts for both sources of variability. The LSD is an extension of the RCBD, with blocking done simultaneously on two characteristics that affect the response variable (Cox, 1958, pp. 35-44; Neter et al., 1996, pp. 1207-1225). In an LSD , the k treatments are arranged in a $k \times k$ square such as in Table 5.2. Each of the four treatments *A*, *B*, *C*, and *D* occurs exactly once in each row and exactly once in each column. The LSD is categorised as a *two-way block error-control design*.

Example 5.9: Weight gain study in mice Bate and Clark (2014, p. 58) illustrate the application of LSDs in animal research using an example provided by Gore and Stanley (2005). These authors considered a weight gain study in female CD-1 mice, which investigated the effect of a control vehicle and a test compound administered at four different doses.

The mice were housed singly in cages across three racks using an independent LSD in each rack, thereby ensuring that all five treatment groups were present in each row and column of every rack. The racks consisted of five shelves (rows) of each six cages. The last column of the rack was

left empty.

Table 5.3 illustrates a possible layout of the experiment. Using this design Gore and Stanley (2005) were able to show that the rack the animals were housed in influenced their water intake and that the body temperature of the mice depended on the row (shelf of the rack) that the cages were placed in.

The authors advocate the use of LSDs to allocate treatments to cages for future trials. Specifically, they warn against the use of more practically appealing designs such as:

- putting all replicates of treatments in the same row of cages in each of the racks;
- placing all replicates of a particular treatment in the same rack, perhaps in adjacent cages.

In either case, there is the risk of reaching false conclusions, due to this bias.

In the above example, 5×5 LSDs were used to control for the location of the cages in the racks. Another application of LSDs is in experiments where a pair of animals was tested each day, and the investigator expected a systematic difference not only between the pairs but also between the animal tested in the morning and the one tested in the afternoon.

In the case of 96-well microplates consisting of 8 rows and 12 columns, scientists often omit the edge rows and columns which are known to yield extreme results. The remaining 6 rows and 10 columns are then used for experimental purposes. However, using an LSD that appropriately corrects for row and column effects, the complete 96-well microplate can be used (Aoki et al., 2014; Burrows et al., 1984) , as is illustrated by the following example.

Example 5.10: Use of LSD in microtiter plates Aoki et al. (2014) describe their assay as follows: *Samples diluted 1:1,600 and serially-*

*diluted standards (*100 µL*) were placed in wells. Samples were arranged in triads, each containing samples from a case and two-matched controls. In order to minimize measurement error due to a spatial gradient in binding efficiency within plate, microplate wells were grouped into 6 blocks of 16 (*4×4*) wells and each set of 3 samples was placed in the same block using a version of Latin square design along with standards. Placement patterns were changed across blocks such that the influence of the spatial gradient on the signal-standard level relationship was minimised.*

In the above example, the authors do not explicitly state what they mean by *a version of Latin square design*. Presumably, an incomplete LSD (see Section 5.2.5), such as a balanced *lattice square* design was used. Such designs were proposed by Burrows et al. (1984) who investigated their use in quantitative ELISA. The interested reader is referred to the literature (Casella, 2008, pp. 262-265; Cox, 1958, p. 240) for this more advanced topic, but the idea is clear that this specific arrangement allows for unbiased comparisons. By *"placement patterns were changed across blocks"* the authors presumably mean randomisation of the rows and columns of the individual LSDs. The idea of using LSDs for the simultaneous control of row- and column-effects in microtiter plates also applies to the larger variants, such as the 384- and 1536-microwell plates.

The main advantage of the LSD is that it simultaneous balances out two sources of error. The disadvantage is the strong assumption that there are no interactions between the blocking variables or between the treatment variable and blocking variables. LSDs are also limited by the fact that the number of treatments, number of rows, and number of columns must all be equal. Fortunately, there are alternative arrangements that do not have this limitation (Cox, 1958, pp. 231-237; Hinkelmann and Kempthorne, 2008, pp. 394-395).

In a $k \times k$ LSD, only k experimental units are assigned to each treatment group. However, more experimental units may be required to obtain an adequate precision. The LSD can then be replicated, and several squares can be used to obtain the necessary sample size. In doing this, there are two possibilities to consider. Either one stacks the squares on top of each other (or next to each other) and keeps them as separate independent squares, or one completely randomises the order of the rows (or columns) of the design. For small experiments, such as in 2×2 LSDs, keeping the squares separate is not a good idea and leads to less precise estimation and loss of degrees of freedom[1]. However, when there is a reason to believe that the column (or row) effects are different between the squares, it does make sense to keep the squares separate.

The **R**-package *agricolae* (de Mendiburu, 2016) can generate random LSDs, e.g. a possible layout of the experiment in Example 5.9 is obtained by:

```
> library(agricolae) # load package agricolae
> trt<-c("0","1","3","10","30")
> # Latin square design
> # use seed for different randomization
> design.lsd(trt, seed=3489)[["sketch"]]
     [,1] [,2] [,3] [,4] [,5]
[1,] "1"  "30" "10" "3"  "0"
[2,] "0"  "10" "3"  "1"  "30"
[3,] "10" "1"  "0"  "30" "3"
[4,] "3"  "0"  "30" "10" "1"
[5,] "30" "3"  "1"  "0"  "10"
```

[1] The error degrees of freedom in an LSD with t treatments tested on t^2 experimental units are $(t-1)(t-2)$. When an LSD is replicated r times, the error degrees of freedom are $(t-1)(rt-t-1)$ when the squares are kept separate and $(rt-2)(t-1)$ when rows or columns of the squares are intermixed. The latter always results in a larger number, making it more precise.

5.2.5. Incomplete Latin square designs

In Section 5.2.3 (page 78), we considered the BIBD as block designs for which the number of treatments is larger than the block size. In LSDs the experimental units are classified in two directions analogous to the rows and columns of a Latin square. When the number of available units in one direction, or in both directions, is smaller than the number of treatments, then we need an incomplete LSD. We shall restrict our discussion to the case where only one direction has an incomplete character.

Youden Squares Youden squares are rectangles with $r = t$ rows (t = number of treatments) and $c < t$ columns. The designs combine the property of the LSD of eliminating heterogeneity in two directions with the property of the BIBD of comparing treatments with the same precision. The designs are called Youden squares after Youden (1937), who first introduced them. Youden squares have the property that every treatment occurs in every column, but not in every row.

Example 5.11: Gastrin assay in rats Colquhoun (1963) describes the use of Youden squares in an assay of gastrin in rats. Two doses of the standard preparation of gastrin and two doses of the preparation of unknown potency are tested in rats. Ideally, the four preparations would be tested in the same animal using a 4×4 LSD. However, it was found impractical to obtain responses from the animals to more than three treatment applications. Consequently, the fourth dose had to be given to another animal. The authors therefore resorted to a Youden square design, in which each row represents an animal, and the columns correspond to the order of administration.

Again, the **R**-package *agricolae* (de Mendiburu, 2016) is used to generate a Youden square design:

```
> library(agricolae) # load package
> trts<-c("A","B","C","D") # 4 doses of 2 drugs
> admins<-3 # administrations per animal
> outdesign <-design.youden(trts,admins,seed=3273)
> outdesign[["sketch"]]
     [,1] [,2] [,3]
[1,] "A"  "B"  "C"
[2,] "D"  "C"  "B"
[3,] "B"  "D"  "A"
[4,] "C"  "A"  "D"
```

Randomisation is achieved by changing the number in the *seed* argument of the function. Each treatment occurs exactly once in each column and since the columns represent the order of administration, the means of the columns can be used to judge whether there is a difference between the responses of the first, second, or third administration.

In microplate experiments, variations on LSDs such as Latin rectangles (Hinkelmann and Kempthorne, 2008, pp. 393-394) and incomplete Latin squares such as Youden squares and lattice squares (Cox, 1958, pp. 231-244; Hinkelmann and Kempthorne, 2008, pp. 394-395) can be used to remedy the spatial patterns present in these plates. Burrows et al. (1984) investigated the properties of a number of these designs for quantitative ELISA. One of the designs they considered (see Table 5.4) was a balanced lattice square[1] design that allows comparing, within one plate 16 treatments with one another, with five replicates for each treatment. Other arrangements, specifically for estimating relative potencies in the presence of microplate location effects , have also been proposed (Schlain et al., 2001).

[1]Balanced lattice square designs have the property that each pair of treatments appears once in each row and once in each column (Casella, 2008, pp. 262-263)

Table 5.4. A balanced lattice square arrangement of 16 treatments with 5 replicate squares on a single microtiter plate with eight lettered rows and twelve numbered columns (after Burrows et al. (1984))

	1	2	3	4	5	6	7	8	9	10	11	12
A	11	7	3	15	11	10	12	9	15	5	12	2
B	12	8	4	16	6	7	5	8	9	3	14	8
C	9	5	1	13	1	4	2	3	4	10	7	13
D	10	6	2	14	16	13	15	14	6	16	1	11
E	12	14	7	1	5	11	4	14				
F	3	5	16	10	1	15	8	10				
G	13	11	2	8	9	7	16	2				
H	6	4	9	15	13	3	12	6				

5.2.6. Randomised block designs and Laboratory Animal Experiments

The RCBD, BIBD, and LSD are all based on the concept of blocking as a means for the control of variability. These designs allow correcting for differences such as time and space effects, by arranging animals in homogeneous groups in such a way that only the variation within, but not between, groups contributes to the experimental error. As compared to the CRDs, these blocked designs are usually more powerful, are less subject to bias, have higher external validity and produce more repeatable results.

Unfortunately, in contrast to other fields of science and engineering, such as agricultural and industrial research, the concept of blocking has not found widespread application in studies involving laboratory animals. The failure to use these error-control designs has probably led to a substantial waste of animals, money, and scientific resources and slowed down the development of new treatments (Festing, 2014).

In experiments involving laboratory animals, Festing (2014) considers the use of randomised block experiments to:

- Spread the experiment over a period of time and/or space as a built-in repeatability check. For example, with four treatments and cage as the experimental unit, each block will consist of four cages. Block 1 starts at week 1, block 2 one week later, and so on. Each block may involve a different batch of animals, maybe of a slightly different age or weight or fed with a different diet. Cages may be placed at different levels in a rack. If the treatment effects remain unchanged, then this implies a good level of repeatability.
- Increase the power of the experiment by matching the experimental units in each block on age, weight, or location in the animal house. This is of particular importance with large experiments in which it is often difficult to obtain a sufficiently homogeneous group of animals.
- Take account of material which has a natural structure, such as the litter. Within-litter experiments in which each animal of the litter is used as an experimental unit and the litter is considered as a blocking variable.
- Split the experiment up into smaller bits (blocks) to make it more manageable. This strategy is particularly useful for large experiments and helps to minimise measurement errors and protocol violations because the work can be done under less time pressure
- Increase the external validity of an experiment because each block samples a different environment and/or time period

5.3. Treatment designs

5.3.1. One-way layout

The examples that we discussed up to now (apart from Example 5.7, page 79), all considered the treatment aspect of the design as consisting of a single factor. This factor can represent presence or absence of a single condition or several different related treatment conditions (e.g. Drug A, Drug B, Drug C). The treatment aspect of these designs is referred to as a *single factor* or *one-way* layout.

5.3.2. Factorial designs

In Example 5.7 (page 79), the investigators studied the joint effect of two factors, i.e. a high and a low dose of Vitamin A, combined with a high and a low dose of protein. Such an experimental set-up in which the joint effect of two or more factors is studied is called a *factorial design*. Example 5.7, in which only two factors were considered, each at two levels (high, low), is a typical case of a 2×2 *full factorial design*, the most straightforward and most frequently used factorial treatment design.

Full factorial designs, such as the above example, take all levels of the factors and all their combinations into consideration. These designs allow estimating the *main effects* of the individual treatments, as well as their *interaction effect*, i.e. the deviation from additivity of their joint effect. We shall use the 2×2 full factorial design to explore the basic concepts of factorial designs and statistical interaction.

Example 5.12: Assessing markers of atherosclerosis development Bate and Clark (2014, pp. 65-67) illustrate factorial experiments using a study conducted by Parkin et al. (2004) to assess whether the serum chemokines JE and KC could be used as markers of atherosclerosis

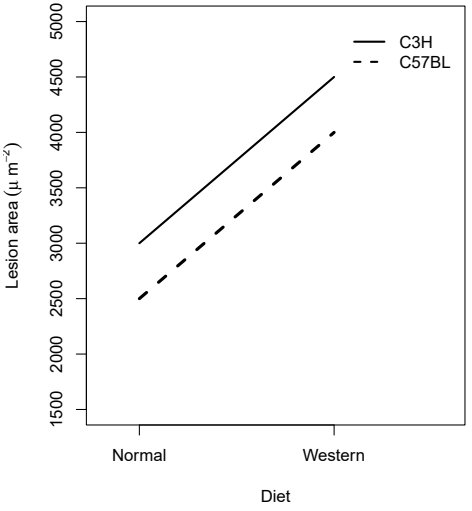


Figure 5.5. Plot of mean lesion area for the case where there is no interaction between *Strain* and *Diet*.

development in mice. Two strains of apolipoprotein-E-deficient (apoE$^{-/-}$) mice, C3H apoE$^{-/-}$ and C57BL apoE$^{-/-}$ were used in the study. These mice were fed either a normal diet or a diet containing cholesterol (the Western diet). After 12 weeks the animals were sacrificed, and their atherosclerotic lesion areas were determined.

The study design consisted of two categorical factors: *Strain* and *Diet*. The factor *Strain* contained two levels: C3H apoE$^{-/-}$ and C57BL apoE$^{-/-}$, as did the factor *Diet*: normal rodent diet and Western diet. In total there were four combinations of factor levels:

1. C3H apoE$^{-/-}$ + normal diet
2. C3H apoE$^{-/-}$ + Western diet
3. C57BL apoE$^{-/-}$ + normal diet
4. C57BL apoE$^{-/-}$ + Western diet

Let us consider some possible outcomes of this experiment.

No interaction When there is no interaction between *Strain* and *Diet*, the difference between the two diets is the same, irrespective of the mouse strain. As shown in Figure 5.5, the lines connecting the mean values of the two diets are parallel to one another. There is an overall effect of *Diet*, animals fed with the Western diet show larger lesions, and this effect is the same in both strains. There is also an overall effect of the *Strain*. Lesions are larger in the C3H apoE$^{-/-}$ than in the C57BL apoE$^{-/-}$ strain and this difference is the same for both diets.

Since the difference between the diets is the same, regardless which strain of mice they are fed to, it is appropriate to average the results from each diet across both strains and make a single comparison between the diets rather than making the comparison separately for each strain. In doing this, the external validity of the conclusions is broadened since they apply to both strains. Besides, the comparison between the two strains can be made, irrespective of the diet the animals are receiving. C3H apoE$^{-/-}$ have larger lesions than C57BL apoE$^{-/-}$ regardless of the diet. Both comparisons, use *all* the experimental units, which makes a factorial design a highly efficient design since all the animals are used to test simultaneously two hypotheses.

Moderate interaction When there is a *moderate* interaction, the direction of the effect of the first factor is the same regardless of the level of the second factor. However, the size of the effect of the first factor varies with the level of the second factor. This is exemplified by Figure 5.6, where the lines connecting both diets are not parallel anymore, though both indicate an increase in lesion size for the Western diet as compared to the normal diet. This increase is more pronounced in the C3H apoE$^{-/-}$ strain than in the C57BL

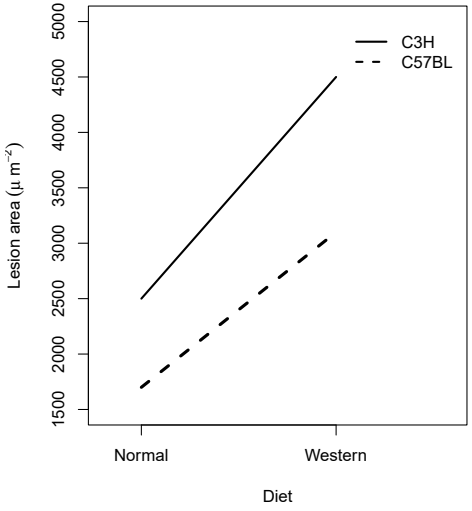


Figure 5.6. Plot of mean lesion area for the case where there is a moderate interaction between *Strain* and *Diet*.

apoE$^{-/-}$ animals. Hence, the C3H apoE$^{-/-}$ strain is more sensitive to changes in diet.

Strong interaction The effect of the first factor can also be entirely dependent on the level of the second factor. Figure 5.7 illustrates the presence of a *strong* interaction in our example. The Western diet always results in bigger lesions than the normal diet, but the effect in the C57BL apoE$^{-/-}$ strain is much more pronounced than in the C3H apoE$^{-/-}$ mice. Furthermore, when fed with normal diet the C3H apoE$^{-/-}$ mice show a larger lesion area than the C57BL apoE$^{-/-}$ strain. In contrast, when the animals receive Western diet, the reverse is true.

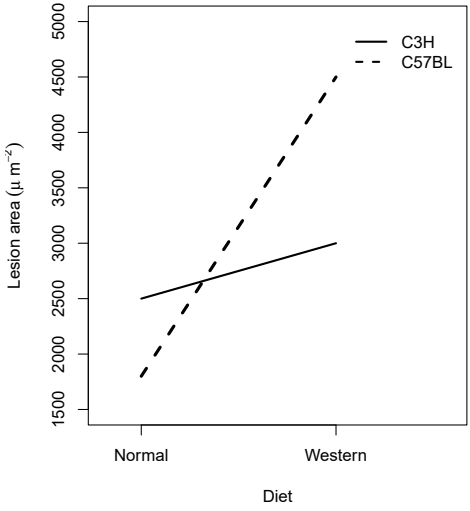


Figure 5.7. Plot of mean lesion area for the case where there is a strong interaction between *Strain* and *Diet*.

The factorial treatment design can be combined with any of the error control designs that we encountered in Section 5.2 (page 70). For instance, Example 5.7 illustrates a 2×2 factorial combined with a balanced incomplete block design.

It happens that researchers have planned a factorial experiment, but in the design and analysis phase failed to recognise it as such. In this case, they do not take full advantage of the factorial structure for interpreting treatment effects and often analyse and interpret the experiment with an incorrect procedure (Nieuwenhuis et al., 2011). We shall come back to that in Section 9.2.1.2 (page 166).

When the two factors consist of concentrations or dosages of drugs, researchers tend to conflate the statistical concept of interaction with the pharmacological concept of *synergism* (Selwyn, 1996, p. 56). However, the requirements for two drugs to be synergistic with each other are much more stringent than just the superadditivity associated with the statistical concept of interaction (Greco et al., 1995; Tallarida, 2000, pp. 157-171; Tallarida, 2001). It is easy to demonstrate that, due to the nonlinearity of the log-dose response relationship, superadditive effects will always be present for the combination, since the total drug dosage has increased, thus implying that a drug could be synergistic with itself. In connection with the 96-well plates and the presence of plate location effects, Straetemans et al. (2005) provide a statistical method for assessing synergism or antagonism.

Higher dimensional factorial designs Up to now we limited our discussion to 2×2 factorial designs. An obvious way to expand these designs is by considering more factors and more factor levels. For a small number of factors (up to 4) at two levels, full factorial designs can be used. However, for more factors or more factor levels, the number of possible combinations can become substantial. For instance, an experiment with five factors at two levels has to deal with 32 combinations. Also, the three-factor and higher interactions become difficult to interpret. One way of restricting the number of resources is by implementing a full factorial design with only a single replicate. But, these unreplicated factorial designs (UFDs) do not leave degrees of freedom for estimation of the error component and, consequently, their statistical analysis requires specific procedures (Montgomery, 2013, pp. 255-285; Neter et al., 1996, pp. 1241-1249). UFDs have only limited use when the number of factors becomes larger than six. For instance, when eight factors are thought to be of importance, each at two levels, a

UFD would require 256 experimental units and 1024 experimental units for ten factors.

When the number of factor combinations becomes large or the number of available animals is limited, fractional factorial designs (FrFDs) can be useful (Neter et al., 1996, pp. 1234-1269). These designs ignore the presence of higher-order effects at the design stage, by omitting certain treatments from the design, such that most of the degrees of freedom are devoted to the main effects and low-order interactions.

UFDs and FrFDs find application in exploratory studies that attempt to identify the possible sources of variation in an experiment. Therefore, significance tests are not applicable to these trials. Usually, these experiments are followed by more elaborate, well-designed studies in which the earlier findings are verified. In the following, we shall see the application of full factorial designs, UFDs, and FrFDs in the optimisation of animal experiments.

Optimising animal experiments by factorial designs. Optimisation of animal experiments, such that the maximum signal-to-noise ratio is obtained, leads to experiments in which the required number of animals is minimised. For this reason, investigators carry out experiments in which a vehicle control and a known positive control treatment are compared in connection with factors the investigator can control and that he thinks can be important in influencing the result. The researcher will then try to determine under which combination of factors the treatment effect, i.e. the mean difference between the positive control and the vehicle is maximised. The factors that are studied are animal-related characteristics such as sex, strain, age, diet, and health status, as well as aspects of the environment such as cage and group size, rack

location, bedding material. Other factors can be protocol specific, such as dose level, timing and route of administration, and the timing of observations.

The conventional approach for finding these optimum experimental conditions was to vary each factor of interest one at a time, keeping all other factors, which may influence the outcome, at a fixed level. However, as compared to factorial designs, this approach has certain disadvantages (Shaw et al., 2002) :

- each group of animals will contribute to understanding the effect of only a single factor, while in a factorial design each animal contributes to the understanding the effect of all the factors under exploration;
- the conventional approach overlooks the fact that the effect of one factor can depend on the level of another factor (i.e. interaction);
- in a factorial design, all potential factors are considered at the study outset, which avoids incremental changes to multiple studies over time.

Example 5.13: Development of an animal model for lung cancer Shaw et al. (2002) provide an example of the use of factorial designs for the development of an animal model for lung cancer. A worked-out version of this example can also be found in Bate and Clark (2014, pp. 71-75).

Multiple lung tumours can be induced in some strains of mice exposed to a carcinogen such as urethane. Animals that develop tumours can then be used as a model to test compounds that might prevent or reduce the incidence of cancer. In the study described by Shaw et al. (2002), a test compound (diallyl sulfide, an active ingredient of garlic) or vehicle was administered to mice before exposing them to the carcinogen, urethane. After a period, the animals were sacrificed and the number of lung tumours recorded.

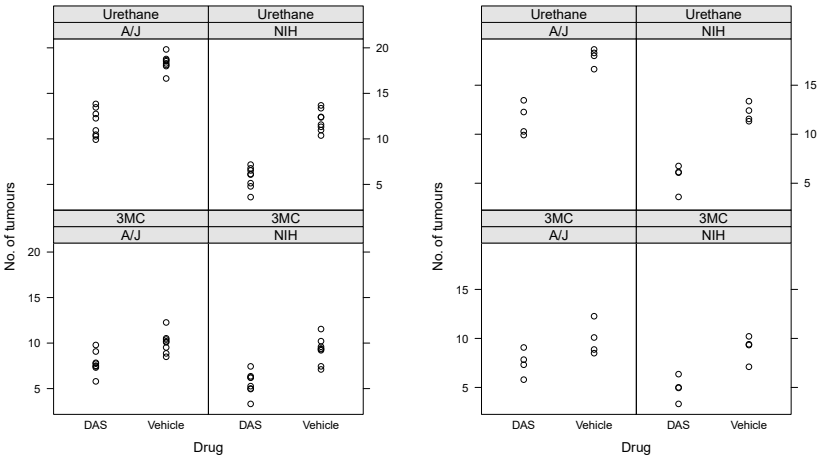


Figure 5.8. Plot of simulated data for the combinations of *Strain*, *Carcinogen*, and *Drug Treatment* for the complete dataset ($n = 64$) with two animals per factor combination and for a reduced dataset with only one animal per treatment ($n = 32$).

Several factors can influence the results, so the researchers decided to use a factorial design to investigate the importance of:

- *Strain*: two strains of mice were considered *A/J* and *NIH*.
- *Gender*: is there a difference between males and females?
- *Diet*: does the diet influence the results. Two diets were used *RM1* and *RM3*.
- *Carcinogen*: two carcinogens were tested, urethane and 3-methylcholanthrene (*3MC*).
- *Drug treatment*: diallyl sulfide (DAS) or vehicle.

If the investigators were to test each of the five possible factors separately on, say six animals for each group, then a total of 60 animals would be required. However, the possible interplay of the different $2^5 = 32$ combinations of the above factors would not be revealed in this manner. Therefore, the investigators decided to include all the combinations of factor levels in a full factorial design and to allocate two animals to each factor level combination, making a total of 64 animals.

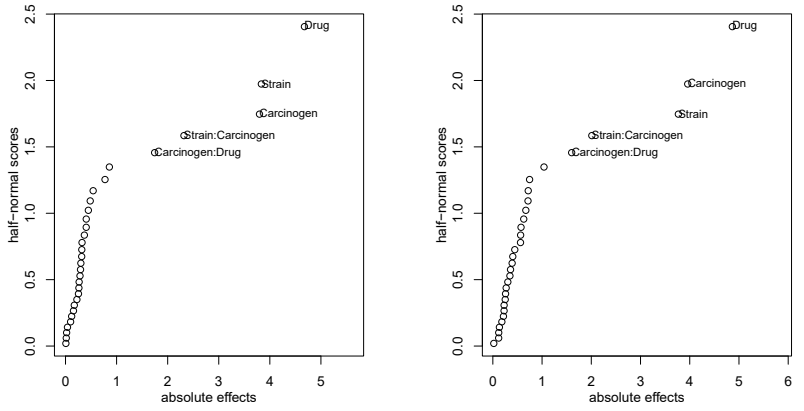


Figure 5.9. The half-normal probability plot (Daniel, 1959) allows identifying the important factors in a factorial design. Left panel full factorial with 2 replicates per treatment. Right panel unreplicated factorial.

The data were analysed by analysis of variance (ANOVA). After checking the model's assumptions, the investigators found that there were no significant third-, fourth-, or fifth-order interactions. Statistically significant two-way interactions were detected between carcinogen and strain, and between carcinogen and drug treatment. The main effects of drug treatment, strain, and carcinogen were also statistically significant.

Figure 5.8 shows the effects of these three significant factors (simulated data). Notably, the interaction of drug treatment with carcinogen is of value in optimising the experiment. We can see from Figure 5.8 that the difference between diallyl sulfide (DAS) and the vehicle was more pronounced when the carcinogen was urethane. Furthermore, the A/J strain appears to be more susceptible than the NIH strain to the carcinogenic action of urethane. Perhaps, we should use only this strain and urethane in future work, as this will maximise the window of opportunity for observing treatment effects.

The above full factorial approach allowed to investigate five factors and their interactions and enabled the investigators to eliminate *Gender*

Table 5.5. Half-fractional factorial design to investigate 6 factors in 16 runs

Run	Strain	Gender	Diet	Carcinogen	Drug
1	NIH	M	RM1	Urethane	Vehicle
2	NIH	M	RM2	Urethane	DAS
3	NIH	F	RM1	3MC	Vehicle
4	NIH	M	RM1	3MC	DAS
5	A/J	F	RM2	3MC	Vehicle
6	A/J	M	RM1	3MC	Vehicle
7	A/J	M	RM2	Urethane	Vehicle
8	A/J	M	RM2	3MC	DAS
9	NIH	F	RM2	Urethane	Vehicle
10	A/J	F	RM2	Urethane	DAS
11	A/J	F	RM1	Urethane	Vehicle
12	NIH	F	RM1	Urethane	DAS
13	A/J	F	RM1	3MC	DAS
14	A/J	M	RM1	Urethane	DAS
15	NIH	F	RM2	3MC	DAS
16	NIH	M	RM2	3MC	Vehicle

and *Diet* as being less critical. The problem is that, despite its effectiveness, 64 animals were needed to achieve this.

As an exercise, the analysis was repeated in a UFD using only 32 animals, i.e. one animal for each factor combination. Instead of a formal ANOVA, we now look at the data from an exploratory point of view. The normal and half-normal probability plots (Daniel, 1959) are graphical tools that help to identify important factors that influence the response. The normal probability plot is based on the idea that, when no factors significantly affect the outcome, the estimated effects would be like random samples drawn from a normal distribution. A plot of the ordered observed effects against their expected values under normality would then result in a straight line. Discernible deviations from this straight line indicate important effects. The half-normal probability plot is an extension of this idea, by plotting the absolute values of the effects, which should follow a half-normal distribution.

Figure 5.9 shows the half-normal probability plots of the full factorial with replicates (left panel) and the unreplicated factorial design (right panel). Interestingly, the results for the UFD, using only 32 animals

leads to the same conclusions as that using the complete dataset. However, power considerations (see Chapter 6) indicate that at least two replicates are required to detect a substantial treatment effect of 0.8 standard deviations. In general, two or three replications at each combination of factor levels are recommended (Bate and Clark, 2014, p. 72).

Next, consider the FrFD approach. FrFDs can be generated in **R** with the *FrF2*-package (Grömping, 2014) as:

```
> library(FrF2)
> FrF2(16,nfactors=5,factor.names=
+                 c("Strain","Gender","Diet","Carcin","Drug"))
   Strain Gender Diet Carcin Drug
1       1      1   -1     -1    1
2       1      1    1     -1   -1
3       1     -1   -1      1    1
4       1      1   -1      1   -1
5      -1     -1    1      1    1
6      -1      1   -1      1    1
7      -1      1    1     -1    1
8      -1      1    1      1   -1
9       1     -1    1     -1    1
10     -1     -1    1     -1   -1
11     -1     -1   -1     -1    1
12      1     -1   -1     -1   -1
13     -1     -1   -1      1   -1
14     -1      1   -1     -1   -1
15      1     -1    1      1   -1
16      1      1    1      1    1
class=design, type= FrF2
```

The result is a matrix with the five factors as columns and experimental units as rows. The two factor levels are indicated by -1 and 1. Table 5.5 shows the design in a more convenient form. The FrFD is half the size of a UFD. However, in FrFDs, not all effects, such as higher order interactions, can be estimated in an unbiased manner and some effects are confounded with others.

The half-normal probability plot (Daniel, 1959) for the FrFD of Table 5.5, based on the same (simulated) data as before, is shown in Figure 5.10. The critical factors were again *Carcinogen*, *Strain*, and *Drug* as main effects and the interaction between *Strain* and *Carcinogen*. In other words, the fractional factorial experiment with only 16 animals arrived

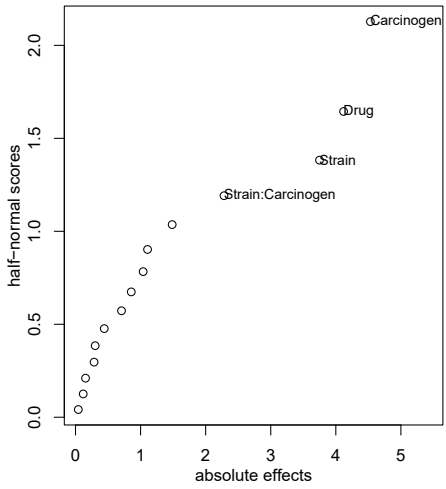


Figure 5.10. Half-normal probability plot of the effects in a half-fractional factorial experiment with only 16 animals

at essentially the same conclusion as the full factorial experiment that used 64 animals, namely that both carcinogen and strain, as well as their interaction, were important factors. Needless to say that the one-variable-at-a-time approach would have missed the interaction.

Two-stage procedure to reduce animal use. We have seen that UFDs and FrFDs can be used to reduce the total number of animals required when using a factorial design. However, this was at the cost of limiting our inferential possibilities. Not all interactions could be tested, and even main effect hypothesis testing became impossible.

An alternative to these procedures is a two-stage strategy, in which the researcher first runs a screening study to investigate all

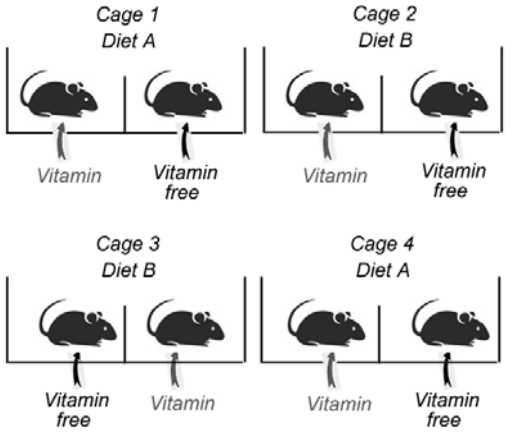


Figure 5.11. Outline of the split-plot experiment of Example 5.14. Cages, each containing two mice, were assigned at random to a number of dietary treatments and the colour-marked mice within the cage were then randomly selected to receive one of two vitamin treatments by injection.

the factors considered important (Bate and Clark, 2014, pp.75-77). The design at this stage should involve only limited replication at each combination of factor levels since here the researcher investigates a large number of factors. Once the most influential factors are identified, a second full factorial experiment is carried out considering only those factors that were declared important in the first screening experiment. In this second study, larger sample sizes are used and the interactions of the factors are investigated.

5.4. More complex designs

We shall now consider some specialised designs in which a factorial treatment design is combined with error-control designs of different complexity.

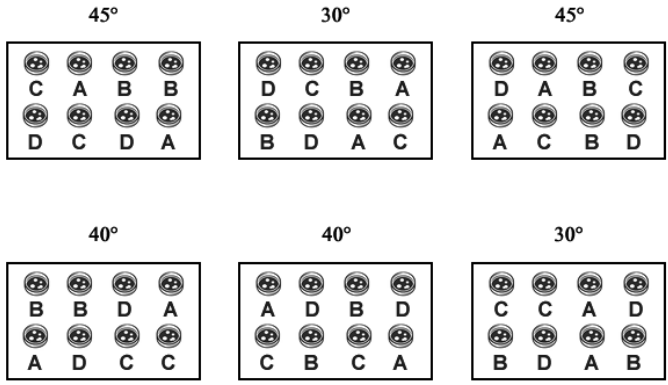


Figure 5.12. Outline of the split-plot experiment of Example 5.15. Six incubators were randomly assigned to three temperature levels in duplicate. In each incubator, eight Petri dishes were placed. Four growth media were randomly applied to the Petri dishes.

5.4.1. Split-plot designs

This type of design incorporates subsampling to make comparisons among the different treatments at two or more sampling levels. The split-plot design allows assessment of the effect of two independent factors using different experimental units as is illustrated by the following examples.

Example 5.14: Rat dietary study An example of a split-plot design is the experiment outlined in Figure 5.11 on diets and vitamins (Festing and Altman, 2002). Cages, each containing two mice, were assigned at random to a number of dietary treatments (i.e. cage was the experimental unit for comparing diets), and the colour-marked mice within the cage were randomly selected to receive one of two vitamin treatments by injection (i.e. mice were the experimental units for the vitamin effect).

Example 5.15: Yeast growth rate Another example is about the effects of temperature and growth medium on yeast growth rate (Ruxton and Colegrave, 2003, p. 78). In this experiment, Petri dishes are placed inside constant temperature incubators (see Figure 5.12). Within each

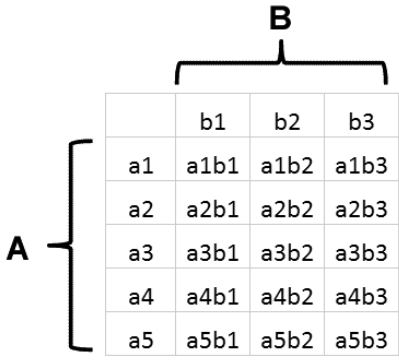


Figure 5.13. Schematic representation of a strip-plot experiment. Factor A is applied to whole plots in the horizontal direction at levels a1, a2, a3, a4, and a5. Factor B is also applied to whole plots at levels b1, b2 , and b3. Factors A and B are "orthogonal" to each other.

incubator, growth media are randomly assigned to the individual Petri dishes. Temperature constitutes the main-plot factor and growth medium the subplot factor. The experiment has to be repeated using several incubators for each temperature.

The term split-plot originates from agricultural research where fields are randomly assigned to different levels of a primary factor and smaller areas within the fields are randomly assigned to one level of another secondary factor. The split-plot design can be considered as two RCBDs superimposed upon one another (Hinkelmann and Kempthorne, 2008, p. 534). In our system of error-control and treatment designs, we regard it as a *two-way crossed (factorial) treatment design* and a *split-plot error-control design*.

The strip-plot or split-block design is a variant of the split-plot design in which a specific type of randomisation is used. In the strip-plot or strip-block design, both factors are applied to whole-

plots, which are "orthogonal" to each other. Schematically, this is represented as in Figure 5.13. There are two factors A and B which are applied to the two types of whole plots. As in Figure 5.13, both plots are orthogonal to one another. Although strip plot designs were originally developed for agricultural field trials, the next example shows that these designs have found a place in the modern laboratory.

Example 5.16: 96-well microtiter plate strip-plot layout Casella (2008, pp. 209-210) describes how Lansky (2002) applies strip-plot designs in 96-well microtiter plate assays. In these assays, samples or dilutions, are usually applied together with multichannel pipettes, whereby a solution is placed simultaneously across a row or down a column. This physical setup leads us naturally to the strip plot design.

The experiment (Casella, 2008, pp. 209-210; Lansky, 2002) consists of an assay in three 96-well plates (blocks). Table 5.6 shows the possible layout of a 96-well plate using a strip plot design. On each of the three assay plates, two rows are randomly assigned to each sample (reference and three test samples, labelled A,B,C,D), and one column is randomly assigned to each of the 12 serial dilutions. The serial dilutions are done using a multichannel pipette, separately on each plate, which applies the dilution to the entire column simultaneously. Random allocation of rows and columns to samples and dilutions separately in each plate allow protection against plate location effects.

5.4.2. The repeated measures design

The repeated measures design is a variant of the split-plot design. In a repeated measures design, we typically take multiple measurements on a subject over time. If any treatment is applied to the subjects or animals, they become the whole plots and *Time* is the subplot factor.

Figure 5.14 shows a typical experimental set-up in which two groups of animals are randomised over two (or more)

Table 5.6. Bioassay experiment of Example 5.16. Row and column indicators refer to the conventional coding of 96-well plates. The four samples (A, B, C, D) are in duplo applied to the rows, and the serial dilution level (dose) is applied to an entire column. *A1/1* in a cell means sample A, first replicate at dilution level 1, *B2/3* second replicate of sample B at dilution 3, etc.

	1	2	3	4	5	6	7	8	9	10	11	12
A	B1/2	B1/8	B1/10	B1/1	B1/11	B1/3	B1/12	B1/7	B1/4	B1/5	B1/6	B1/9
B	D2/2	D2/8	D2/10	D2/1	D2/11	D2/3	D2/12	D2/7	D2/4	D2/5	D2/6	D2/9
C	B2/2	B2/8	B2/10	B2/1	B2/11	B2/3	B2/12	B2/7	B2/4	B2/5	B2/6	B2/9
D	C1/2	C1/8	C1/10	C1/1	C1/11	C1/3	C1/12	C1/7	C1/4	C1/5	C1/6	C1/9
E	A1/2	A1/8	A1/10	A1/1	A1/11	A1/3	A1/12	A1/7	A1/4	A1/5	A1/6	A1/9
F	A2/2	A2/8	A2/10	A2/1	A2/11	A2/3	A2/12	A2/7	A2/4	A2/5	A2/6	A2/9
G	C2/2	C2/8	C2/10	C2/1	C2/11	C2/3	C2/12	C2/7	C2/4	C2/5	C2/6	C2/9
H	D1/2	D1/8	D1/10	D1/1	D1/11	D1/3	D1/12	D1/7	D1/4	D1/5	D1/6	D1/9

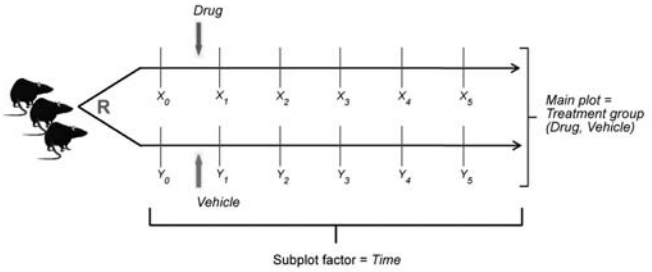


Figure 5.14. A typical repeated measures design. Animals are randomised to different treatment groups, the variable of interest (e.g. blood pressure) is measured at the start of the experiment and at different time points following treatment application.

treatment groups and the variable of interest is measured just before and at several time points following treatment application. When designing and analysing repeated measures designs, any *confounding* of the treatment effect with time, as was the case in Section 4.6.1.1 (page 46) when using self-controls, must be avoided. Therefore, like in the example of Figure 5.14, a *parallel control group* which protects against the time-related bias must always be included. Statistical analysis will then compare the changes from baseline between the two groups or use a more appropriate method, such as analysis of covariance.

5.4.3. The crossover design

Crossover designs or change-over designs are special types of repeated measures designs. While in a conventional repeated measures design, each animal or subject receives a single treatment and is then measured repeatedly, in a crossover design each subject receives different treatments over time. The crossover design considers each animal or subject as a block to which a sequence of treatments is applied over several test periods, one treatment per

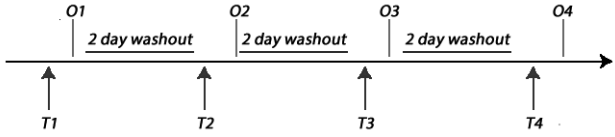


Figure 5.15. Outline of a four-period four-treatment crossover design

test period (Figure 5.15). Hence, the crossover design is a special case of the RCBD. Consequently, every pairwise treatment comparison will be carried out with the same level of precision. When applying these designs, it is of paramount importance that there is sufficient time between the test periods, the so-called *washout periods*, such that treatment effects do not influence future responses. Therefore, crossover designs cannot be used when the treatment affects the subjects permanently. Also, ethical concerns can prohibit the use of these designs.

The crossover design combines the randomised complete block with the repeated measures error-control designs. The following example illustrates the use in practice of the crossover design.

Example 5.17: Five-choice serial reaction time task Bate and Clark (2014, p. 60) describe the following experiment, originally published by Hille et al. (2008). 5-HT4 agonists are currently being developed as candidate treatments for Alzheimer's disease. The investigators evaluated two doses of a candidate drug for their effect on attention deficit in rats. The animals were trained over about 30 sessions to react to a visual stimulus (see Hille et al. (2008) or Bate and Clark (2014, p. 60) for more details). Since it takes a lot of time and effort to train the rats, they are considered a valuable resource. Therefore, it would be advantageous to treat the same animal more than once. Fortunately, the treatments in this experiment had only a short-term effect, so it was possible to administer a sequence of treatments over time. Two doses of the experimental drug (treatments A and B), nicotine as a positive control (treatment C) control and the vehicle (treatment V)

Table 5.7. Crossover design for Example 5.4.3 consisting of three stacked four-by-four Latin squares. Treatments are coded as A: 5-HT4 partial agonist at 0.1 mg/kg, B: 5-HT4 partial agonist at 1 mg/kg, C: nicotine at 0.2 mg/kg, V: vehicle

	Rat No.											
Test Period	1	2	3	4	5	6	7	8	9	10	11	12
1	V	A	B	C	V	A	B	C	V	A	B	C
2	A	V	C	B	C	B	A	V	B	C	V	A
3	B	C	V	A	A	V	C	B	C	B	A	V
4	C	B	A	V	B	C	V	A	A	V	C	B

were administered to 12 rats over four weeks. Between the test periods, a 2-day washout period was included. Rats were randomly assigned to the treatment sequences.

All treatments were administered to each rat, which makes it an RCBD. In each test period, three rats received the same treatment. The easiest way to construct this design is to make use of three four-by-four LSDs. Table 5.7 shows a possible design for this experiment. Using this design and with only 12 animals, the researchers were able to demonstrate that the 5-HT4 agonist augments attention in rats.

Crossover designs have many advantages, since:

- The experimental unit is not the animal or subject, but the animals or subjects within a test period. Therefore, each animal or subject generates more than one experimental unit. For example, in a three-period crossover study, the number of experimental units is three times the number of animals or subjects.
- All treatment comparisons are carried out within the animals or subjects. Therefore, differences between animals will not bias the treatment comparisons.
- All pairwise treatment comparisons are tested against the within-subject variability which is usually more precise than

the between-subjects variability. Therefore, fewer subjects or animals are required.

The major disadvantage of a crossover design is the presence of carry-over effects by which the results obtained for a specific treatment can be influenced by the previous treatment(s). In some cases, this can be dealt with by special types of crossover design that allow estimating the carry-over effect and correct for it (Jones and Kenward, 2003, pp. 117-149) Another important drawback is that crossover designs take longer to complete. From an ethical point of view, the discomfort placed upon the individual animal or subject by carrying out repeated treatments and procedures should also be considered.

Crossover designs are widely used in pharmacokinetics, in particular in studies showing the equivalence in bioavailability of pharmaceutical formulations, the so-called bioequivalence studies (Patterson and Jones, 2006). In-depth discussions of crossover designs can be found in Jones and Kenward (2003), and Senn (2002).

5.5. Dose-response designs

The designs that we have considered so far in this chapter provide a framework for comparing different treatments with one another. However, within the same framework we can also construct designs whose purpose is estimation and prediction. Dose-response experiments are carried out to estimate the shape of the dose-response relationship, to estimate the parameters of the functional model that describes this relationship, to determine a threshold dose (e.g. in toxicology), and to predict the response for intermediate doses.

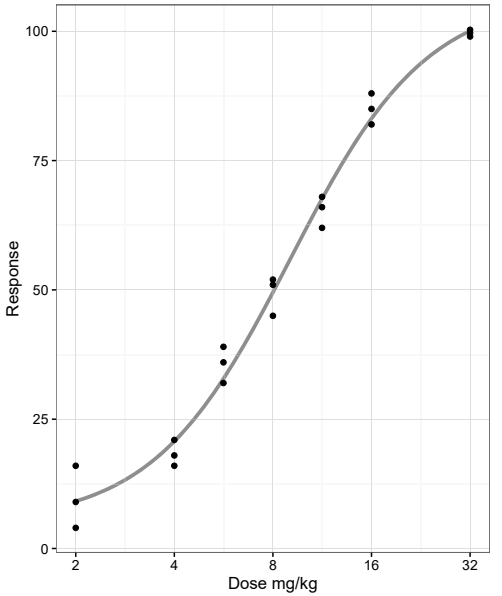


Figure 5.16. Analgesic action of morphine sulfate. Raw data and fitted 4-parameter log-logistic curve

These experiments play an important role in pharmacology and toxicology when developing new drugs (Tallarida and Jacob, 1979). A typical dose-response experiment is described in Tallarida (2000, p. 31) and presented, for illustrative purposes, in Bate and Clark (2014, p. 86).

Example 5.18: Analgesic action of morphine sulfate An experiment was conducted to assess the analgesic effect of morphine sulfate in rats. Rats were exposed to cold water, as a noiceptive stimulus, and the analgesic effect was assessed by measuring tail flick latency. There were seven doses of the drug and three rats were randomly assigned to each of the doses.

The outcome of the experiment, together with a curve describing the functional relationship between dose and response are displayed in Figure 5.16. When we look at the experiment from the point of view of

experimental design, we notice that the treatment aspect of the design is a dose-response design, while the error-control aspect is a CRD.

Dose-response designs are essentially one-way layout treatment designs in which the dose factor is a continuous rather than a discrete factor. This one-way layout treatment design can be combined with any of the error-control designs discussed earlier. However, at analysis time the error structure of the experiment must be taken into account (Pinheiro and Bates, 2001, pp. 273-414).

The primary purpose of the dose-response experiment is to investigate how the dose of a new compound influences the response. Whether a specific dose is statistically significant from the control or not is of less interest. The main difference with comparative experiments is that the choice of the dose levels is free and can be determined by objective criteria (Holland-Letz and Kopp-Schneider, 2015).

—Data, data, data. I cannot make bricks without clay.

Sherlock Holmes (in A.C. Doyle: *The Adventure of the Copper Beeches*)

—If you use p=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time. If, as is often the case, experiments are underpowered, you will be wrong most of the time.

David Colquhoun, (2014)

6

Sample Size and Power

6.1. The need for sample size determination

In most European countries and the USA, scientists are requested by the animal care committee to justify the number of animals that they want to include in their project, to ensure that it is appropriate (Dell et al., 2012). With too few animals, the experiment will lack sufficient statistical power to detect a real treatment effect and is a waste of animals, researcher's time and resources. On the other hand, with too many animals in the experiment, a biologically irrelevant effect could be declared statistically significant, and above all, some animals will suffer unnecessary harm (Bate and Clark, 2014, p. 7). Thus, the answer to how large an experiment should be is that it should be just big enough to give confidence that any biologically meaningful effect that exists can be detected.

6.2. The context of biomedical experiments

The estimation of the appropriate size of the experiment is straightforward and depends on the statistical context, the assumptions made, and the study specifications. Context and specifications on their turn depend on the study objectives and the design of the experiment.

In practice, the most frequently encountered contexts in statistical inference are point estimation , interval estimation, and hypothesis testing, of which hypothesis testing is the most important in biomedical studies.

6.3. The hypothesis testing context

In the hypothesis testing context[1], one defines a null hypothesis and, for the purpose of sample size estimation, an alternative hypothesis of interest. The null hypothesis will often be that the response variable does not depend on the treatment condition. For example, one may state as a null hypothesis that the population means of a particular measurement are equal under two or more different treatment conditions and that any differences found can be attributed to chance.

At the end of the study, when the data are analysed (see Section 7.3), we shall reject the null hypothesis in favour of the alternative hypothesis, or fail to reject the null hypothesis. As is indicated in Table 6.1, there are four possible outcomes at the end of the experiment.

[1]The hypothesis testing context, which in statistics is also known as the Neyman-Pearson approach, assumes the population model of statistical inference

Table 6.1. The decision process in hypothesis testing

	State of Nature	
Decision made	Null hypothesis true	Alternative hypothesis true
Do not reject null hypothesis	Correct decision $(1-\alpha)$	False negative β
Reject null hypothesis	False positive α	Correct decision $(1-\beta)$

When the null hypothesis is true, and we failed to reject it, we have made the correct decision. This is also the case when the null hypothesis is false, and we did reject it. However, two conclusions are erroneous. If the null hypothesis is true, and we incorrectly rejected it, then we made a *false positive decision*. Conversely, if the alternative hypothesis is true (i.e. the null hypothesis is false), and we failed to reject the null hypothesis we have made a *false negative decision*. In statistics, a false positive decision is also referred to as a *Type I error* and a false negative decision as a *Type II error*.

The basis of sample size calculation is formed by specifying an allowable rate of false positives and an allowable rate of false negatives for a particular alternative hypothesis and then to estimate a sample size just large enough so that these low error rates can be achieved. The allowable rate of false positives is called the *significance level* or *alpha level*, which we conventionally set at values of 0.01, 0.05, or 0.10. The false negative rate depends on the postulated alternative hypothesis and is usually described by its complement, i.e. the probability of rejecting the null hypothesis when the alternative hypothesis holds, which is called the *power* of the statistical hypothesis test. Power levels are usually expressed as percentages and values of 80%, or 90% are standard in sample size calculations.

Significance level and statistical power are already two of the four major determinants of the sample size required for hypothesis testing. The remaining two are the inherent variability in the study parameter of interest and the size of the difference to be detected in the postulated alternative hypothesis. Other key factors that determine the sample size are the number of treatments and the number of blocks used in the experimental design.

When the significance level decreases or the power increases, the required sample size will become larger. Similarly, when the variability is larger or the difference to be detected smaller, the required sample size will also become larger. Conversely, when the difference to be detected is large or the variability low, the required sample size will be small. It is convenient, for quantitative data, to express the difference in means as *effect size* by dividing it by the standard deviation[1]:

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{s} \tag{6.1}$$

The effect size Δ takes both the difference between the group means and the inherent variability into account. Cohen (1988, pp. 24-27) argues that effect sizes of 0.2, 0.5, and 0.8 can be regarded respectively as small, medium, and large. In basic biomedical research and, more specifically in animal research, effect sizes are likely to be large relative to other types of research because large doses of active compounds are often given to ensure that a response is detectable. Unfortunately, to date, no one has suggested small, medium, or large values for Δ in animal experi-

[1] When comparing mean values from two independent groups, the standard deviation for calculating the effect size, can be from either group when variances of the two groups are homogeneous. Alternatively, a pooled standard deviation can be calculated as $s_p = \sqrt{(s_1^2 + s_2^2)/2}$.

ments, but we shall follow (Shaw et al., 2002) and also consider values of $\Delta = 1.0$ and 1.2 as large effects.

6.4. Sample size estimation

6.4.1. Continuous data

Table 6.2. Values for the constant C used in sample size calculations

	Significance level α		
Power	0.1	0.05	0.01
60%	3.60	4.90	8.00
70%	4.71	6.17	9.61
80%	6.18	7.85	11.68
90%	8.56	10.51	14.88

Now that we are familiar with the concepts of hypothesis testing and the determinants of sample size, we can proceed with the actual calculations. The required sample size in each group for comparing two mean values is given by (Dell et al., 2012):

$$n = 1 + \frac{2C}{\Delta^2} \tag{6.2}$$

where the value of the constant C depends on the value of α and the power and is obtained from Table 6.2. For example, for an experiment to detect an effect $\Delta = 0.8$ at a significance level $\alpha = 0.05$ with a power of 80% ($\beta = 0.8$), requires:

$$n = 1 + \frac{2 \times 7.85}{0.8^2} \approx 26$$

animals in each treatment group. For a large effect size of $\Delta = 1.2$, with the same settings for α and β as before, the required number

of animals in each treatment group drops to:

$$n = 1 + \frac{2 \times 7.85}{1.2^2} \approx 12$$

Lehr (1992) simplified Equation 6.2 as:

$$n \approx {}^{16}/_{\Delta^2} \tag{6.3}$$

where Δ represents the effect size and n stands for the required sample size in each treatment group for a two-group comparison against a two-sided alternative with a power of 80% and a value of 0.05 as Type I error. The numerator of Lehr's equation relates to Table 6.2 and depends on the desired power and significance level. Alternative values for the numerator are 8 and 21 for powers of 50% and 90%, respectively.

One can also make use of a software package to obtain the required sample size. There is free software available to make the necessary calculations, and also some websites can be of help. In particular, there is the **R**-package *pwr* (Champely, 2017).

Example 6.1: Protective effect in cardiomyocytes (continued) Consider once again the CRD about cardiomyocytes discussed in Example 5.5. The pooled standard deviation of the two groups is 12.4. A large effect of 1.2 in this case, corresponds to a difference between both groups of $1.2 \times 12.4 \approx 15$ myocytes. Let's assume that we wish to plan a new experiment to detect such a difference with a power of 80% and we want to reject the null hypothesis of no difference at a significance level $\alpha = 0.05$, whatever the direction of the difference between the two samples (i.e. a two-sided test[1]).

The computations are carried out in **R** in a single line of code and show the same result as obtained above, namely that 12 experimental units are

[1] See Section 7.3 for a discussion of one-sided and two-sided tests.

required in each of the two treatment groups:

```
> require(pwr) # load the pwr package
> pwr.t.test(d=1.2,power=0.8,sig.level=0.05,type="two.sample",
+                 alternative="two.sided")

     Two-sample t test power calculation

              n = 11.94226
              d = 1.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Conversely, we can also use the software to determine the power of an experiment. Consider a proposal for an experiment with only five animals per treatment group, then the power to detect a difference $\Delta = 1.2$ at a two-sided level of significance $\alpha = 0.05$ is:

```
> pwr.t.test(d=1.2,n=5,sig.level=0.05,type="two.sample",
+                 alternative="two.sided")

     Two-sample t test power calculation

              n = 5
              d = 1.2
      sig.level = 0.05
          power = 0.3864373
    alternative = two.sided

NOTE: n is number in *each* group
```

In this case, the power to detect a difference of 15 myocytes (i.e. $\Delta = {}^{15}/_{12.4} = 1.2$) between treatment groups is only 39%.

Uncertainty in estimating the standard deviation. When we use previous studies or pilot experiments to estimate the standard deviation, we must realise that this estimate itself is also subject to variability. This is illustrated in Figure 6.1 where the distribution of the standard deviation in a sample of size $n = 5$ is highly

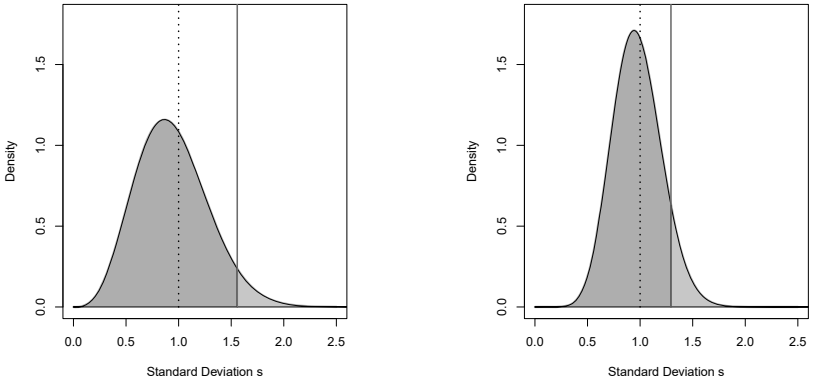


Figure 6.1. Distribution of the sample standard deviation based on $n = 5$ (left panel) and $n = 10$ replicates (right panel). The dotted vertical line indicates the true value σ, the solid line the upper 80% confidence limit

skewed, indicating that for this sample size the sample standard deviation tends to underestimate the true population standard deviation.

For a sample size of $n = 10$, the situation is markedly improved, and now estimates of the standard deviation can more reliably be used in the computation of Δ. To accommodate for the imprecision involved in estimating the standard deviation, some authors (Browne, 1995; Kieser and Wassmer, 1996) recommend basing the sample size calculations on the upper 80% confidence limit of the standard deviation. The sample size thus obtained guarantees that its corresponding power is at least equal to the planned power with a probability of 0.8. Analogously, the 90% confidence interval guarantees a probability of 0.9 that the corresponding power is at least the planned power. Table 6.3 contains values for the 80% and 90% upper confidence interval of the standard deviation. These can be used as a multiplier to adjust the stan-

dard deviation in the computation of the effect size. Alternatively, one can inflate the obtained sample size by multiplying it by the inflation factor.

Table 6.3. Upper 80% and 90% confidence limit of the standard deviation ($\sigma = 1$) and related inflation factor for the sample size n for different degrees of freedom used in estimating the standard deviation

	Standard Deviation		Inflation factor for n	
df	80%	90%	80%	90%
4	1.558	1.939	2.426	3.761
5	1.461	1.762	2.134	3.105
6	1.398	1.650	1.954	2.722
7	1.353	1.572	1.831	2.471
8	1.320	1.514	1.742	2.293
9	1.293	1.469	1.673	2.159
10	1.272	1.434	1.618	2.055
12	1.240	1.380	1.537	1.904
14	1.216	1.341	1.479	1.797
16	1.198	1.311	1.435	1.718
18	1.183	1.287	1.400	1.657
20	1.171	1.268	1.372	1.607
22	1.161	1.252	1.349	1.567
24	1.153	1.238	1.329	1.533
26	1.145	1.226	1.312	1.504

Example 6.2: Protective effect in cardiomyocytes (continued) In Example 6.1, the sample size calculation was based on an estimate of the standard deviation of 12.4 in a two-group comparison of each $n = 5$ replicates. The degrees of freedom involved in estimating the standard deviation are therefore $2 \times (n - 1) = 8$. From Table 6.3 we obtain for 8 degrees of freedom and an upper confidence limit of 80%, for the standard deviation a multiplication factor of 1.32 and, as inflation factor for the sample size, a value of 1.742.

A large effect of $\Delta = 1.2$ now corresponds to a difference of $1.2 \times 12.4 \times 1.32 = 20$ myocytes instead of the 15 myocytes of Example 6.1.

Alternatively, one could also adjust the previously obtained required sample size by multiplying it by the inflation factor of 1.742 to obtain the

required sample size of 21 animals in each treatment group for a future experiment.

Sample size based on the coefficient of variation. Biologists tend to think in percentages and often situations arise where the investigator looks for a percentage change in mean and also thinks of the variability in terms of percentages. For example, a scientist wants to set up a two-group experiment to detect a difference in means of 20% and expects the variability to be about 30%. A convenient rule of thumb to calculate the required sample size for a two-sided test with a power of 80% and a level of significance of 0.05 is given by (Van Belle, 2008, p. 29-30):

$$n = 8\frac{c_v^2}{d_p^2}[1 + (1 - d_p^2)] \tag{6.4}$$

where c_v is the coefficient of variation σ/μ and $d_p = (\mu_1 - \mu_0)/\mu_0$ is the proportionate change in means. For the situation described above, this becomes

$$n = 8 \times \frac{0.30^2}{0.20^2} \times [1 + (1 - 0.20)^2] = 29.52 \approx 30$$

Hence, the researcher will need 30 animals in each treatment group.

Paired experiments. Up to now, we restricted our discussion to comparisons in a CRD with two independent groups. For paired designs Equation 6.2 becomes (Dell et al., 2012):

$$n = 2 + \frac{C}{\Delta^2} \tag{6.5}$$

where the constant C is again obtained from Table 6.2 and Δ is the standardised effect size. However, in the definition of Δ in

Equation 6.1, we now use the standard deviation of the change in outcome, which is much smaller than the standard deviation of the absolute values.

Example 6.3: Protective effect in cardiomyocytes (continued) Consider the cardiomyocyte experiment again, but now correctly as a paired design, as discussed in Example 5.4. The standard deviation of the changes between vehicle-treated and drug-treated dishes is 5.61, which is much smaller than 12.4, the pooled standard deviation of the distribution of cardiomyocytes. However, since this standard deviation is based on only 4 degrees of freedom, it is an underestimate of the true standard deviation. To be 80% sure that our sample size is enough to cover a power of 80%, we multiply the standard deviation with the value 1.558 from Table 6.3 and use 8.74 as a conservative estimate of the standard deviation. The standardised effect size that corresponds to a difference of 20 cardiomyocytes now is Δ =2.29. The number of paired replicates, i.e. animals, detect this difference with a power of 80% at a level of significance $\alpha = 0.05$ is:

$$n = 2 + 7.85/(2.29^2) = 3.5 \approx 4$$

.

The **R**-package *pwr* yields comparable results:

```
> pwr.t.test(d=2.29,power=0.8,sig.level=0.05,type="paired",
+                 alternative="two.sided")

     Paired t test power calculation

              n = 3.770236
              d = 2.29
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

This small sample size, however, does not provide enough degrees of freedom to estimate the standard deviation in the new experiment. Therefore, it is recommended to use some additional pairs (animals).

6.4.2. Binary data

Binary data consider the occurrence of an event (alive/death, present/absent) and are usually expressed as proportions. In contrast to continuous data, sample size calculations for binary data (also called dichotomous data) do not require any knowledge of the standard deviation.

Let r_1 and r_2 represent the number of occurrences of the event in two treatment groups of size n_1 and n_2 respectively. The respective proportions p_1 and p_2 as estimates of the population proportions π_1 and π_2 are given by:

$$p_1 = \frac{r_1}{n_1}; p_2 = \frac{r_2}{n_2} \tag{6.6}$$

The null hypothesis that we test in our experiment is that $\pi_1 = \pi_2$, or $\Delta_\pi = \pi_1 - \pi_2 = 0$. The necessary sample size to test this hypothesis is obtained as (Dell et al., 2012):

$$n = C\frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{\Delta_\pi^2} + \frac{2}{\Delta_\pi} + 2 \tag{6.7}$$

As before, the constant C depends on the significance level α and desired power and is obtained from Table 6.2.

Example 6.4: Life span of cardiomyopathic hamsters Genetically inbred cardiomyopathic Syrian hamsters constitute a suitable animal model of congestive heart failure. Death, as a consequence of congestive heart failure, is usually observed within a year. Ver Donck et al. (1991) investigated the effect of a compound on the survival of cardiomyopathic hamsters after 300 days. The investigators expected 15% survival in the control group and 50% in the drug-treated group. What is the required sample size to detect this difference at a two-sided significance level of $\alpha = 0.05$ and power of 80%?

From Table 6.2, for a power of 80% and $\alpha = 0.05$, we obtain a value

of 7.85 for the constant C. We also have $\Delta_\pi = 0.50 - 0.15 = 0.35$. Consequently, the required number of animals per treatment group is

$$n = 7.85 \times \frac{0.15 \times 0.85 + 0.50 \times 0.50}{0.45^2} + \frac{2}{0.35} + 2 \approx 32$$

The above method is very approximate and will tend to overestimate the required sample size , especially for π_1 or π_2 close to one or zero. A better approach uses the *pwr*-package:

```
> pwr.2p.test(h = ES.h(0.15,0.5),  sig.level = 0.05, power = .80,
+                 alternative = "two.sided")

     Difference of proportion power calculation
 for binomial distribution (arcsine transformation)

              h = 0.7753975
              n = 26.10885
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

This results in a required sample size of 27 animals in each treatment group.

6.5. How many subsamples

In Section 4.6.2.2 we defined the standard error of a two-group comparison when subsamples are present as:

$$\sqrt{\frac{2}{n}(\sigma_n^2 + \frac{\sigma_m^2}{m})} \tag{6.8}$$

where n and m are the number of experimental units and subsamples and σ_n and σ_m the between-sample, and within-sample standard deviation. Using this expression, we can establish the power for different configurations of an experiment.

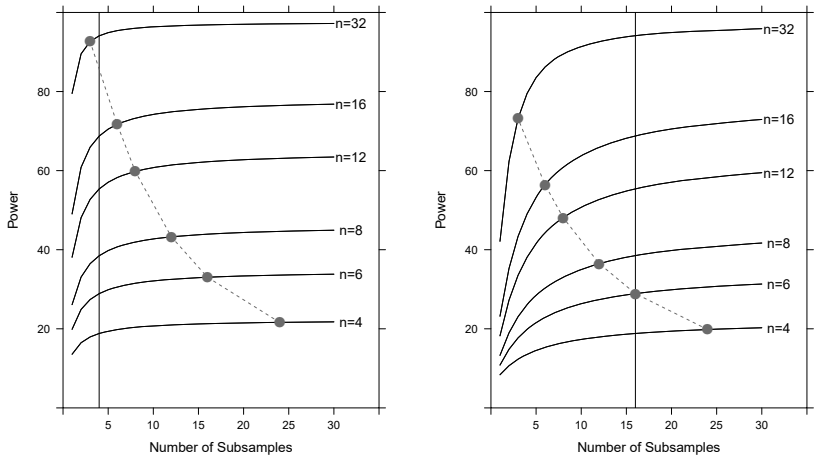


Figure 6.2. Power curves for a two-group comparison to detect a difference of $\Delta = 1$, with a two-sided t-test with significance level $\alpha = 0.05$ as a function of the number of subsamples m. Lines are drawn for different numbers of experimental units n in each group. For both left and right panel, the between-sample standard deviation (σ_n) is 1, while the within-sample standard deviation (σ_m) is 1 in the left panel and 2 in the right panel. The dots connected by the dashed line indicate where the total number of subsamples $2 \times n \times m$ equals 192. The vertical line indicates an upper bound to the useful number of subsamples of $m = 4(\sigma_m^2/\sigma_n^2)$.

Figure 6.2 shows the influence of the number of experimental units (n) and the number of subsamples (m) per experimental unit on the power of two experiments to detect a difference between two mean values of size 1, i.e. $\mu_1 - \mu_0 = 1$. For both experiments $\sigma_n = 1$. The left panel shows the case where $\sigma_m = 1$, while in the right panel $\sigma_m = 2$. The dots connected by a dashed line represent the power for experiments where the total number of subsamples equals 192 (96 per treatment group).

As is illustrated in the left panel of Figure 6.2, subsampling has only a limited effect on the power of the experiment when the within sample variability σ_m is the same size (or smaller) as the

between sample variability σ_n. In this case, it makes no sense taking more than say four subsamples per experimental unit, as is indicated by the vertical line in Figure 6.2. Furthermore, the sharp decline of the dashed line connecting the points with the same total number of subsamples indicates that subsampling, in this case, is not very efficient, at least when the cost of subsamples and experimental units is not taken into consideration. An experiment with 32 experimental units and three subsamples has a power of more than 90%, while for an experiment with the same total number of subsamples but with four experimental units and 24 subsamples per unit, the power is only about 20%.

The right panel of Figure 6.2 shows the case where the within sample standard deviation σ_m is twice the standard deviation between samples σ_n. In this example, taking more subsamples does make sense. The power curves keep increasing until the number of subsamples is about 16. The loss in efficiency by taking subsamples is also more moderate, as is indicated by the less sharp decline of the dotted line.

In both situations, the power curves have flattened after crossing the vertical line where $m = 4(\sigma_m^2/\sigma_n^2)$. This is known as Cox's rule of thumb (Cox, 1958, p. 181) about subsamples, which states that for the *CRD* there is not much increase in power when the number of subsamples m is greater than $4(\sigma_m^2/\sigma_n^2)$. Cox's ratio provides an upper limit for a useful number of subsamples. However, this rule of thumb does not take the different costs involved with experimental units and subsamples into account. In many cases, especially in animal research, the cost of the experimental unit is substantially larger than that of the subunit.

With differential costs taken into consideration, the optimum number of subsamples can be derived as (Snedecor and Cochran, 1980, p. 532):

$$m = \sqrt{\frac{c_n}{c_m} \times \frac{\sigma_m^2}{\sigma_n^2}} \tag{6.9}$$

Equation 6.9 shows that taking subsamples is of interest when the cost of experimental units c_n is large relative to the cost of subsamples c_m, or when the variation among subsamples σ_m is large relative to the variation among experimental units σ_n.

Example 6.5: Morphologic study in sheep In a morphologic study (Verheyen et al., 2014), the diameter of cardiomyocytes was examined in 7 sheep that underwent surgery and 6 sheep that were used as a control. For each animal, the investigators measured the diameters of about 100 epicardial cells.

An advanced statistical method, known as mixed model analysis of variance (Pinheiro and Bates, 2001) allowed to estimate σ_n^2 and σ_m^2 as 4.58 and 13.7 respectively. Surprisingly, the variability within an animal was larger than between the animals. If we were to set up a new experiment, we could limit the number of measurements to $4 \times 13.7/4.58 \approx 12$ per animal.

Alternatively, we can take the differential costs of experimental units and subsamples into account. It makes sense to assume that the cost of 1 animal is about 100 times the cost of one diameter measurement. Making this assumption, the optimum number of subsamples per animal would be $\sqrt{100 \times 13.7/4.58} \approx 18$. Thus the total number of diameter measurements could be reduced from 1300 to 234. Even if animals would cost 1000 times more than a diameter measurement, the optimum number of subsamples per animal would be 55, which is still a reduction of about 50% of the original workload. As a conclusion, this is a typical example of a study in which statistical input at the onset would have improved research efficiency considerably.

6.6. Multiplicity and sample size

When more than one statistical test is carried out on the data, the overall rate of false positive findings is higher than the false positive rate for each test separately (see Section 7.6, page 150). One way to circumvent this inflation of the false positive error rate is by setting the critical value of each test individually at a more stringent level.

The most straightforward adjustment, Bonferroni's adjustment (Neter et al., 1996, pp. 736-738), consists of just dividing the significance level of each test by the total number of comparisons. Bonferroni's adjustment maintains the error rate α of the totality of tests that are carried out in the same context at its original level. But, as we already noted on page 118, when the significance level is set at a lower value, the required sample size will necessarily increase. Fortunately, the increase in the required number of replicates is surprisingly small.

Figure 6.3 shows for a two-sided Student t-test with a significance level α of 0.05 and a power of 80% (left panel) and 90% (right panel), how the required sample size increases with an increasing number of comparisons. The percentage increase in sample size due to adding an extra comparison corresponds to the slope of the line segment connecting adjacent points in Figure 6.3.

For all values of Δ and power $(100 \times (1 - \beta))$, the evolution of the relative sample size is comparable. For powers of 80% and 90 %, carrying out two independent statistical tests instead of one involves a 20% larger sample size to maintain the overall error rate at its level of $\alpha = 0.05$. Similarly, when three or four independent tests are involved, the required sample size increases with 30%

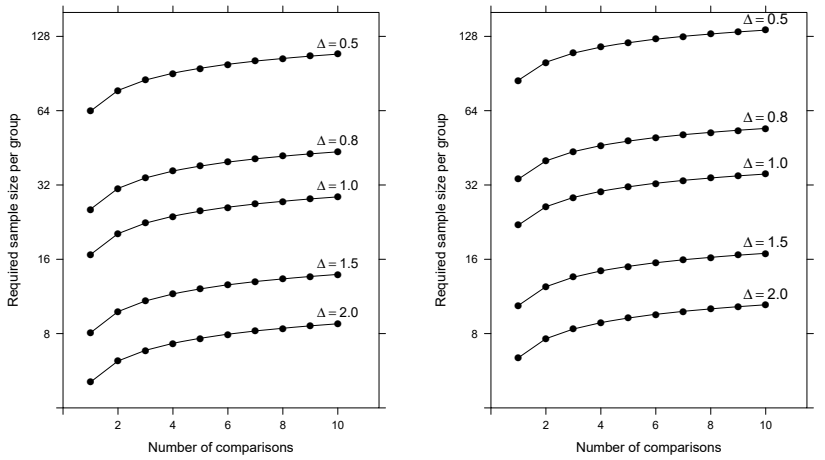


Figure 6.3. Required sample size of a two-sided test with a significance level α of 0.05 and a power of 80% (left panel) and 90% (right panel) as a function of the number of comparisons that are carried out. Lines are drawn for different values of the effect size (Δ). Note that the y-axis is logarithmic.

or 40% respectively. After four comparisons, the effect tapers off, and all curves approach linearity. Adding an extra comparison in the range of 4 - 10 comparisons, will increase the required sample size with about 2.7%, leading to a total increase in sample size for ten comparisons of about 70%. For a more substantial number of comparisons, Witte et al. (2000) noted that the relative sample size increases linearly with the logarithm of the number of comparisons.

Figure 6.3 also illustrates how sample size depends on the effect size. Large sample sizes are indeed required for detecting moderate-small differences. However, for the substantial differences, which we usually want to detect in early research, the required sample size reduces to an attainable level.

6.7. The problem with underpowered studies

A survey of articles that were published in 2011 (Tressoldi et al., 2013) showed that in prestigious journals such as *Science* and *Nature*, fewer than 3% of the publications calculate the statistical power before starting their study. More specifically in the field of neuroscience, published studies have a power between 8 and 32% to detect a genuine effect (Button et al., 2013).

Low statistical power might lead the researcher to wrongly conclude there is no effect from an experimental treatment when in fact an effect does exist. Also, in research involving animals, underpowered studies raise a significant ethical concern. If each individual study is underpowered, the true effect will only likely be discovered after many studies using many animals have been completed and analysed, using far more animal subjects than if the study had been done the first time properly (Button et al., 2013). Another consequence of low statistical power is that effect sizes are overestimated, and results become less reproducible (Button et al., 2013). The following example best illustrates this.

Example 6.6: Consider again the cardiomyocyte example. A sample size calculation, treating the experiment as if it was a CRD (which in reality it was not) was carried out in Example 6.1 (page 120) and yielded a required sample size of 12 animals in each group to detect a large treatment effect $\Delta = 1.2$ at a two-sided significant level $\alpha = 0.05$ with a power of 80%.

Imagine running several copies of this experiment, say 10,000. The effect sizes that are obtained from these experiments follow a distribution as displayed in the left-hand panel of Figure 6.4. The dark shaded area corresponds to experiments that produced a statistically significant increase (two-sided alternative, $\alpha = 0.05$). This subset yields a slightly increased estimate of the effect size of 1.26, which corresponds to an *effect size inflation* of 5%. This effect size inflation is to be expected when an

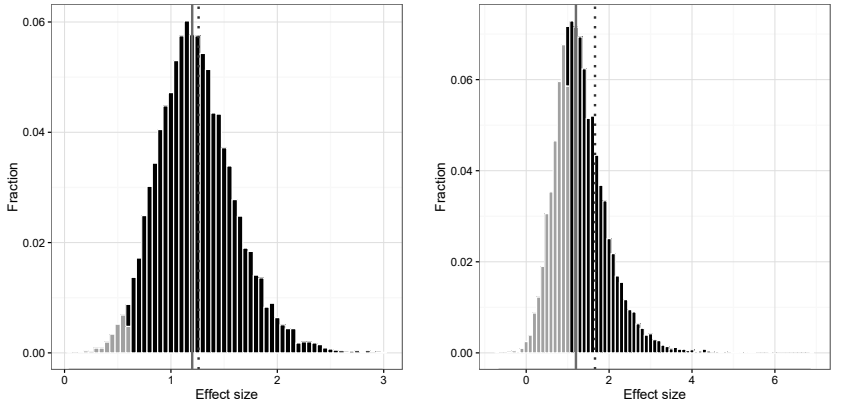


Figure 6.4. Running the cardiomyocyte experiment a large number of times, the measured effect sizes follow a broad distribution. In both plots the true effect size is $\Delta = 1.2$. The dark area represents statistically significant results (two-sided $p \leq 0.05$). The solid vertical line indicates the true effect size ($\Delta = 1.2$), while the dotted line represents the mean effect size as estimated from the statistically significant experiments. Left panel: 12 animals are used per treatment group which corresponds to a power of 80%. Right panel: only five animals per treatment group are used which results in an underpowered experiment with a power of 39%

effect has to pass a certain threshold such as statistical significance. A relative inflation of 5% as in this case is acceptable.

The situation is entirely different in the right-hand panel of Figure 6.4 where the experiments now consist of only five animals per treatment group and the power has dropped to 39%. The variability of the results is substantially larger as displayed by the larger scale of the x-axis. While the standard deviation of the mean effect size in the larger experiment was 0.37, this has now increased to 0.65, an increase by a factor 1.76 which corresponds to $\sqrt{12/5}$.

The significant experiments now constitute a much smaller part of the distribution. The mean effect size in this subset has now increased to 1.66, an inflation of 38%. In the context of the cardiomyocyte study, this would mean that the significant part of the underpowered studies on average will report an increase of 21 instead of the true value of 15 viable cardiomyocytes.

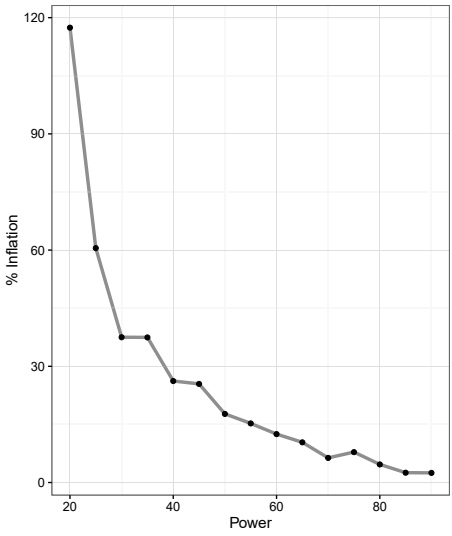


Figure 6.5. The winner's curse: effect size inflation as a function of statistical power.

The overestimation of the effect size in small, statistically significant studies is known as *truth inflation*, *Type M error* (M stands for magnitude) or the *winner's curse* (Button et al., 2013; Reinhart, 2015, pp. 23-26). As shown in Figure 6.5, effect size inflation is worst for small low-powered studies which can only detect treatment effects that happen to be large. Consequently, significant research findings of small studies will be biased in favour of inflated effects.

Inflation of the reported effect size in small published studies has severe consequences when we try to replicate a published finding and base our sample size calculations on the treatment effect as it was published. When this is an inflated estimate, the calculated sample size of the confirmatory experiment will be too low, and consequently, the new trial will most probably fail.

In the case of the cardiomyocyte experiment, planning of a confirmatory study based on the inflated effect size of 21 viable myocytes would result in two groups of 7 animals each and would have a power of only 54% to detect the true effect of an increase of 15 viable myocytes. To summarise, effect size inflation due to small, underpowered experiments is one of the major reasons for the lack of replicability in biomedical research.

6.8. Sequential plans

Sequential plans allow investigators to save on the experimental material, by testing at different stages, as data accumulate. These procedures have been used in clinical research and are now advocated for use in animal experiments (Fitts, 2010, 2011). Sequential plans are sometimes referred to as "*sequential designs*," but strictly speaking all types of designs that we discussed before can be implemented in a sequential manner.

Sequential procedures are entirely based on the Neyman-Pearson hypothesis decision-making approach that we saw in Section 6.3 and do not consider the accuracy or precision of the treatment effect estimation. Therefore, in the case of early termination for a significant result, sequential plans are prone to exaggerate the treatment effect. There is certainly a place for these procedures in exploratory research such as early drug screening, but a fixed sample size confirmatory experiment is needed to provide an unbiased and precise estimate of the effect size.

Example 6.7: Protection against traumatic brain injury In a search for compounds that offer protection against traumatic brain injury, a rat model was used as a screening test. Preliminary power calculations showed that at least 25 animals per treatment group were required to

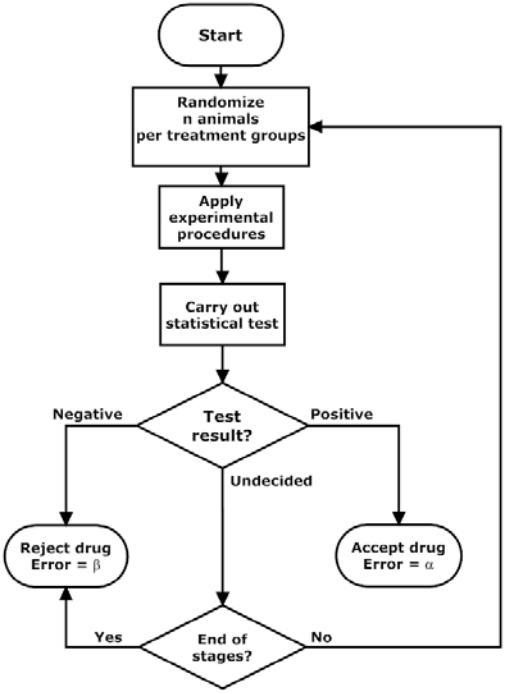


Figure 6.6. Outline of a sequential experiment

detect a protective effect with a power of 80% against a one-sided alternative with a Type I error of 0.05. Taking into consideration that a large number of test compounds would be inactive, a fixed sample size approach was regarded as unethical and inefficient. Therefore, a one-sided sequential procedure (Wilcoxon et al., 1963) was considered as more appropriate.

The procedure operated in different stages (Figure 6.6). At each stage, animals were selected, such that the group was as homogeneous as possible. The animals were then randomly allocated to the different treatment groups, as three per group. At a given stage the treatments consisted of several experimental compounds and their control vehicle. After measuring the response, the procedure allowed the investigator to make the decision to reject the drug as uninteresting, to accept it as active, or to continue with a new group of animals in a next stage.

After having tested about 50 treatment conditions, a candidate compound was selected for further development. An advantage of this screening procedure was that, given the biologically relevant level of activity that must be detected, the expected fractions of false positive and false negative results were known and fixed. A disadvantage of the method was that a dedicated computer program was required for the follow-up of the results.

—"How absurdly simple !", I cried.
"Quite so !", said he, a little nettled. "Every problem becomes very childish when once it is explained to you."

Dr Watson and Sherlock Holmes (in A.C. Doyle: *The Adventure of the Dancing Men*)

—We teach it because it's what we do; we do it because it's what we teach. (on the use of $p < 0.05$)

George Cobb, (2014) (in Wasserstein and Lazar (2016))

7

The Statistical Analysis

7.1. The statistical triangle

There is a one-to-one correspondence between the study objectives, the study design, and the statistical analysis (Figure 7.1). The objectives determine which of the experimental designs is the most appropriate and, on its turn, the chosen design governs how the data are analysed. Fisher (1935) enunciated this fundamental principle as follows:

> *All that we need to emphasise immediately is that, if an experiment does allow us to calculate a valid estimate of error, its structure must completely determine the statistical procedure by which this estimate is to be calculated. If this were not so, no interpretation of the data could ever be unambiguous; for we could never be sure that some other equally valid method of interpretation would not lead to a different result.*

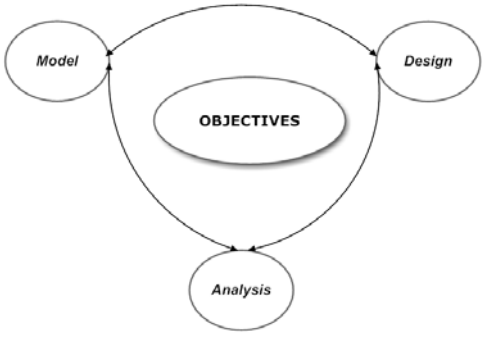


Figure 7.1. The statistical triangle: a conceptual framework for the statistical analysis

In other words, the choice of the statistical methods follows directly from the objectives and design of the study. With this in mind, many of the complexities of the statistical analysis have now almost become trivial.

7.2. The statistical model revisited

We already stated in Section 5.1 (page 67) that a statistical model underpins every experimental design and that we should consider the experimental results as being generated by this statistical model. This conceptual framework simplifies the statistical analysis to just fitting the statistical model to the data and comparing the model components related to the treatment effect with the error component of the model (Dowdy et al., 2004, pp. 265-272; Montgomery, 2013, pp. 65-80; Neter et al., 1996, pp. 78-80). Hence, the choice of the appropriate statistical analysis is straightforward. However, some statistical technicalities remain, such as the type of data and the assumptions we make about the distribution of the data.

7.3. Significance tests

The famous French scholar Pierre-Simon Laplace investigated in 1827 the tidal effect of the moon on the earth's atmosphere (Laplace, 1827, pp. 3-18). His readings of atmospheric pressure showed a modest effect of the moon's phases. To decide whether his findings were convincing or not, he formulated the null hypothesis that the moon did not influence atmospheric pressure and calculated the probability of obtaining by chance an even greater effect than the one he had observed. Like in many other aspects of his work, Laplace was again several years ahead of his time and conceived the scientific method that we still use today (Stigler, 1986, pp. 153-157; Stigler, 2016, Chapter 3). However, it was only after Ronald Fisher published his seminal work "*Statistical Methods for Research Workers*" (Fisher, 1925) that the use of significance tests and p-values for judging scientific evidence became widespread.

In contrast to the Neyman-Pearson hypothesis testing approach (Section 6.3, page 116), Fisher's significance test does not involve an alternative hypothesis, nor does it consider the concept of Type I and Type II errors. It only requires the definition of the null hypothesis, whose precise formulation is a crucial step in Fisher's methodology. The null hypothesis, says Fisher, *must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the "problem of distribution," of which the test of significance is the solution* (Fisher, 1935, p. 16).

Next to the formulation of the null hypothesis, a significance test involves the definition of a quantity called the *test statistic* and, based on the statistical model, its calculation from the obtained data. The following step is to establish the *sampling distribution* of

this test statistic under the assumption that the null hypothesis is true. This null distribution is then used to determine the probability of obtaining a value for the test statistic that is as extreme or more extreme than the one observed, assuming the null hypothesis holds. This probability is referred to as the *p*-value and is a measure for the "extremeness" of the data according to a hypothetical data distribution.

When Neyman and Pearson introduced their decision-making approach, Fisher most strongly disagreed with them and called their proposal a *misapprehension*. He objected to the use of Neyman-Pearson hypothesis testing in the natural sciences and saw its application limited to acceptance procedures in quality control (Fisher, 1973). For Fisher, the *p*-value was merely an informal measure to see how surprising the data were and whether they deserved a second look (Nuzzo, 2014; Reinhart, 2015, p. 11).

Nevertheless, it has become standard practice for researchers, to compare the *p*-value to a preset significance level α (usually 0.05). When the *p*-value is smaller than α, the null hypothesis is rejected. Otherwise, we fail to reject it, and the result is inconclusive. Actually, this way of working conflates the significance testing of Fisher with the formal decision-making approach of Neyman and Pearson into a hybrid form, the null hypothesis significance test (Pernet, 2018). Again this approach is criticised, as it forces researchers to turn a continuum of uncertainty into a dichotomous "reject/do-not-reject" decision (Amrhein et al., 2017). Instead, it is good practice to follow Fisher and report the actual *p*-values rather than $p \leq 0.05$ (see Section 9.2.1.2, page 166) since it puts equal weight on all *p*-values and allows the reader to decide for himself how much evidence is given. In this context, the *p*-value

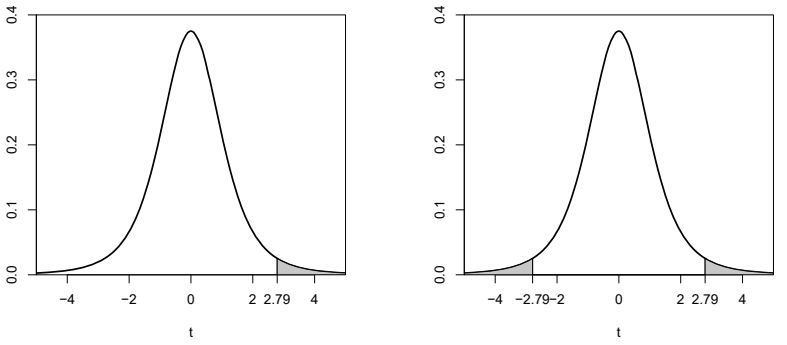


Figure 7.2. Distribution of the test statistic t for the cardiomyocyte example, under the assumption that the null hypothesis of no difference between the samples is true. Left panel: one-sided test. Right panel: two-sided test

is treated as a continuous measure of evidence against the null hypothesis (Amrhein et al., 2017).

Example 7.1: Protective effect on cardiomyocytes (continued) The cardiomyocytes experiment of Example 5.4 will help us to illustrate the concept of the significance test. The experiment was set up to test the null hypothesis of no difference between vehicle and drug. The (paired) design of this experiment is a special case of the RCBD with only two treatments, and the response is a continuously distributed variable. In this design, calculations can be simplified by evaluating for each pair separately the treatment effect, thus removing the block effect. We now must make some assumptions about the statistical model that generated the data. Specifically, we assume that the differences are independent of one another and originate from a normal distribution.

Next, we define a relevant test statistic, which, in this case, is the mean value ($\bar{x}$) of the differences, divided by its standard error[1]. For our example, we obtain a value of $7.0/2.51 = 2.79$ for this statistic. Under the

[1] the standard error of the mean of a sample is obtained by dividing the sample standard deviation by the square root of the sample size, i.e. $s_{\bar{x}} = SD/\sqrt{n} = 5.61/\sqrt{5} = 2.51$

assumptions made above and provided the null hypothesis of no difference between the two treatment conditions holds, the distribution of this statistic is known[1] and is depicted in Figure 7.2.

On the left panel of Figure 7.2, the value of the test statistic of 2.79, which was obtained from the experimental data is indicated and the area under the curve, to the right of this value, is shaded in grey. This area corresponds to the one-sided p-value, i.e. the probability of obtaining a greater value for the test statistic than the one obtained in the experiment. By definition, the total area under the curve equals one. Consequently, we can calculate the value of the shaded area. For our example, this results in a value of 0.024, which is the probability of obtaining a value of the test statistic that is as extreme or more extreme than 2.79, the value attained in the experiment, provided the null hypothesis holds.

Before the experiment was carried out, we were also interested in looking at the opposite result, i.e. we were also interested in a decrease in viable myocytes. Therefore, when we consider more *extreme* results, we should also look at values that are less than -2.79. This is done in the right panel of Figure 7.2. The sum of the two areas is called the *two-sided* p-value and corresponds to the probability of obtaining under the null hypothesis, a more *extreme* result than ± 2.79. In our example, the obtained two-sided p-value is 0.049.

An important caveat on the use of significance tests is that they merely reflect how much the obtained results can be attributed to chance alone, but do not tell us anything about their scientific relevance. With a large enough sample, we can detect the most uninteresting minimal difference (Lindsey, 1999, p. 170).

7.4. Verifying the statistical assumptions

It is always wise, before carrying out formal tests of significance, to make graphical displays of the data. Plotting individual data

[1] Under the null hypothesis and when the assumptions are true, the test statistic is distributed as a *Student t-distribution* with $n-1$ degrees of freedom.

allows to identify outliers and can give indications on whether the statistical model is appropriate or not. Such exploratory work is also a tool for gaining insight into the research project and can lead to new hypotheses.

The inferential results can be sensitive to distributional and other assumptions of the statistical analysis. It is, therefore, essential that we verify these assumptions. Informal methods such as diagnostic plotting are preferably used to assess the appropriateness of the statistical model (Neter et al., 1996, pp. 756-768; Montgomery, 2013, pp. 80-89). When planning the experiment, historical data or the results of exploratory or pilot experiments can already be used for a preliminary verification of the model assumptions. Another option is to use statistical methods that are robust against departures from the assumptions (Lehmann, 1975).

7.5. The meaning of statistical significance

As a measure of statistical evidence, the *p*-value is the most widely used, yet also the most misunderstood, misinterpreted, and sometimes even miscalculated index in biomedical research (Goodman, 2008). Maybe, the most widespread misconception is that an insignificant result implies that there is no difference between the treatment groups. However, in this reasoning, the scientist makes the logical mistake of *affirming the consequent*. Fisher (1935, p. 16) explicitly warned against this fallacy:

> *it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation.*

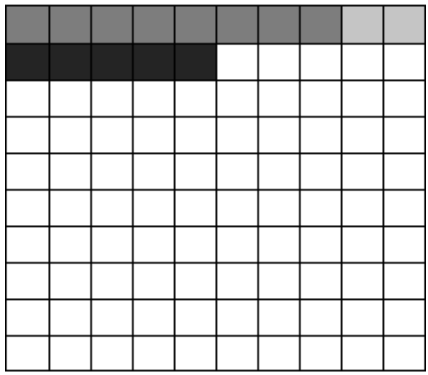


Figure 7.3. One hundred drugs are tested for activity against a biological target. Each drug occupies a square in the grid, the top row contains the drugs that are truly active. Statistically significant results are obtained only for the darker-grey drugs. The black cells are false positives (after Reinhart (2015, pp. 39-40)).

A plausible reason for an insignificant result is that the sample size is just too small. Inability to prove the null hypothesis is also in accordance with Popper (2002) who claimed that hypotheses can never be confirmed, only falsified.

Now, consider the opposite where, following the null hypothesis significance test, we obtain a significant result, what does it mean? An example will give us a better understanding of the value of a statistically significant finding.

Example 7.2: Screening of drugs In a laboratory, 100 experimental compounds are tested against a particular biological target. Figure 7.3 illustrates the situation. Each square in the grid represents a tested compound. In reality, only 10 drugs, which are located in the top row, are active against the target. We call this value of 10% the *prevalence* or *base-rate*. Let us assume that our statistical test has a power of 80%, which means that of the ten active drugs, eight are correctly identified (shown in darker grey). The threshold for the p-value to declare a drug

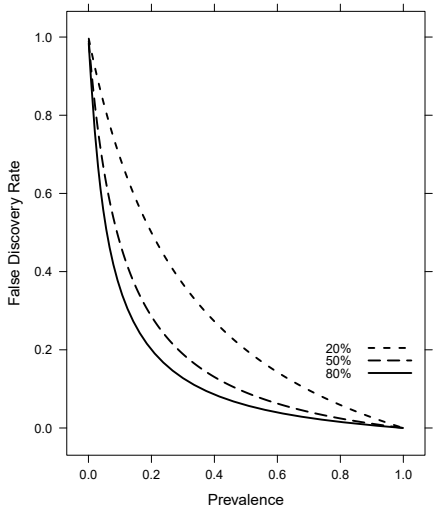


Figure 7.4. False discovery rate as a function of the prevalence π) and the power $100 \times (1-\beta)$, ($\alpha = 0.05$), lines are drawn for power $100 \times (1-\beta)$ of 80%, 50% and 20%.

statistically significant is set to 0.05, meaning that there is a 5% chance to declare an inactive compound incorrectly as active. Ninety drugs are in reality inactive, so about five of them will yield a significant effect. These correspond to the black squares in the second row of Figure 7.3. Hence, in the study 13 drugs are declared active, of which only eight are truly effective, i.e. the *positive predictive value* is about $8/13 = 62\%$, or its complement the *False discovery rate (FDR)* is about 38%.

From the above reasoning, it follows (Colquhoun, 2014; Wacholder et al., 2004) that the FDR depends on the threshold α, the power $(1-\beta)$ and the prevalence or base-rate π as:

$$\begin{aligned} FDR &= \alpha(1-\pi)/[\alpha(1-\pi)+\pi(1-\beta)] \\ &= 1/\{1+[\pi/(1-\pi)][(1-\beta)/\alpha]\} \end{aligned} \quad (7.1)$$

Table 7.1. Minimum false discovery rate *MFDR* for some commonly used critical valuescritical value of p

	p-value				
	0.1	0.05	0.01	0.005	0.001
MFDR	0.385	0.289	0.111	0.067	0.0184

For our example, the above derivation of the FDR yields a value of 0.36 when the prevalence of active drugs π is 10%, significance threshold α of 0.05 and a power $1-\beta$ of 80%. The FDR rises to 0.69 when the power is reduced to 20%, which means that under these conditions, 69% of the drugs (or other research questions) that were declared active, are in fact false positives.

The FDR depends highly on the prevalence rate as is illustrated in Figure 7.4, leading to the conclusion that when working in new areas where the a priori probability of a finding is low, say 1/100, a significant result does not necessarily imply a genuine activity. In fact, under these circumstances, even in a well-powered experiment (80% power) with a significance level of 0.05, 69% of the positive findings are false. To make things worse, such surprise, groundbreaking findings, often combined with exaggerated effect sizes due to a small sample size (Section 6.7) are the ones that are likely to be published in prestigious journals.

What is the value of $p \approx 0.05$ Consider an experiment whose results yield a p-value close to 0.05, say between 0.045 and 0.05. In how many instances does this result reflect a true difference? We already deduced that, when the power or prevalence rate is low, the FDR can easily reach 70%. However, what is the most optimistic scenario? In other words, what is the lowest value of the FDR? Irrespective of statistical power, sample size, and prior probability, Sellke et al. (2001) derived an expression for what they

call the *conditional error probability*, which is equivalent to the *minimum FDR* (*MFDR*). The *MFDR* gives the *minimum* probability that, when a test is declared *significant*, the null hypothesis is in fact true. Some values of the *MFDR* are presented in Table 7.1. For $p = 0.05$ the $MFDR = 0.289$, which means that a researcher who claims a discovery when $p \approx 0.05$ is observed will make a fool of him-/herself in about 30% of the cases. Even for a p-value of 0.01, the null hypothesis can still be true in 11% of the cases (Colquhoun, 2014).

The *FDR* is undoubtedly one of the key factors responsible for the lack of replicability in research and puts the Neyman-Pearson decision-theoretic approach with its irrational dichotomisation of the p-value into *significant* and *non-significant* certainly into question.

As it was noted already in the introductory chapter, the issues of reproducibility and replicability of research are worrying the scientific, but indeed also the statistical community, deeply. These concerns led the board of the *American Statistical Association* (*ASA*) to issue a statement on March 6, 2016, in which the organisation warns against the misuse of p-values (Wasserstein and Lazar, 2016). It was the first time in its 177-year-old history that the *ASA* made explicit recommendations on a fundamental matter in statistics. In summary, the *ASA* advises in its statement researchers to avoid drawing scientific conclusions or making decisions based on p-values alone. *P*-values should certainly not be interpreted as measuring the probability that the studied hypothesis is true or the probability that the data were produced by chance alone. Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations.

7.6. Multiplicity

We already touched the problem of multiplicity in Section 6.6 (page 131) and considered Bonferroni's adjustment and its impact on sample size. Multiplicity arises in many situations in research. For instance, a study might consider too many objectives, or it includes a large number of variables or measurements are made at many time points. All these inflate the Type I error, the probability of incorrectly rejecting the null hypothesis.

Example 7.3: Testing several doses of a drug Suppose a scientist tests 20 different doses of a drug on a specific outcome, and that she rejects the null hypothesis of no treatment effect for each dose separately when the p-value is less than or equal to 0.05.

In this case, the overall probability of falsely declaring the existence of a treatment effect, when all underlying null hypotheses are in fact true is $1-(1-0.05)^{20} = 0.64$, which means that in the long run in 64% of the cases she will reject the null hypothesis of no drug effect falsely. In other words, she will falsely claim a drug effect in 64% of the cases. The same multiplicity problem arises when an investigator tests a single dose of a drug on 20 mutually independent variables.

The problem of multiplicity is of particular importance and magnitude in gene expression microarray experiments (Bretz et al., 2005). For example, a microarray experiment examines the differential expression of 30,000 genes in a wild-type animal and a mutant. Assume that for each gene an appropriate two-sided two-sample test is performed at the 5% significance level. Then we expect to obtain roughly 1,500 false positives. Strategies for dealing with, what is often called the *curse of multiplicity*, in microarrays are provided by Amaratunga and Cabrera (2004, pp. 109-115), and Bretz et al. (2005).

The multiplicity problem must at least be recognised at the planning stage. Ways to deal with it (Bretz et al., 2010; Curran-Everett, 2000) should be investigated and specified in the protocol.

—Nothing clears up a case so much as stating it to another person.

Sherlock Holmes (in A.C. Doyle: *The Memoirs of Sherlock Holmes. Silver Blaze.*)

8 The Study Protocol

The writing of the study protocol finalises the end of the planning and design process. The complete study protocol consists of the technical protocol, which we already discussed in Section 4.6.1.3, and a more conceptual part, the research protocol. The research protocol contains sufficient scientific background, including relevant references to previous work, to understand the motivation and context for the study. In it, we describe the study's purpose and specify its primary and secondary objectives, the related hypotheses and working hypotheses which we shall test, as well as their consequential predictions. We also explain the experimental approach and rationale.

The protocol must contain a section on experimental design, with the definition of the experimental units and how they are assigned to treatments, as well as information and justification of the planned sample sizes. For long-term studies, the study protocol should provide a way that allows monitoring of the progress

of the study and means of preventing deviations from the protocol.

The study protocol also describes the appropriate statistical models that will be used to detect patterns of interest in the data and the selection strategies for choosing among them (Lindsey, 1999, pp. 5-9). Defining the statistical methods in the protocol is of particular importance since it allows preparation of the data analytic procedures beforehand and ensures against the misleading practice of *data dredging* or *data snooping*. Writing down the *statistical analysis* plan beforehand also prevents from trying several methods of analysis and report only those results that suit the investigator, which is a deceptive practice, known as *p*-hacking (Head et al., 2015). Such a practice is, of course, inappropriate, unscientific, and unethical. In this context, the study protocol is a safeguard for the replicability of research findings.

Many investigators consider writing a detailed protocol a waste of time. However, the *smart researcher* understands that by writing a good protocol, she is actually preparing her final study report or research paper. A well-written study protocol is even more essential when the design is complex or the study is collaborative. In this case, all people or organisations who collaborate on the study should be invited to provide input to the protocol. As a consequence, in long-term or complex studies that involve many departments and collaborators, several draft versions of the protocol may be required before a final version is obtained that is acceptable to all parties involved. Depending on the complexity of the study, a protocol may vary from a few to as much as 50 pages. Once we have formalised the final version of the protocol, it is es-

sential that it is followed as close as possible and every deviation of it should be documented.

It should be clear by now that the study protocol plays a crucial role in the conduct of the experiment. Lindsey (1999, p. 8) considers, the finalised protocol to serve, among other things, as:

- a specification of the scientific design, including motivation and aims;
- an operations manual by which all investigators know what is expected of them;
- a prior record of assumptions and hypotheses so that you cannot be accused of drawing *post hoc* conclusions.

—No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us

R. A. Fisher (1935)

—Statistical analysis allows us to put limits on our uncertainty, but not to prove anything.

Douglas G. Altman, (1991)

9 The Research Report

While in the previous chapters, we focused on the planning and design phase of the study, with the protocol as final deliverable, this chapter deals with some points to consider when we report the results of our research.

9.1. The ARRIVE Guidelines

In the introductory chapter, we noted that many studies show issues with the quality of reporting (Kilkenny et al., 2009). In an attempt to accommodate this quality issue, Kilkenny et al. (2010) published the ARRIVE guidelines for transparent reporting of research in animals. These guidelines (see Appendix E) provide a framework to help scientists report their research findings. They consist of a list of 20 items that should be included in scientific publications. We shall consider here only the issues that are of direct relevance to statistical thinking and smart experimental de-

sign, and discuss them as they appear in the different sections of a scientific publication, namely: the introduction, materials and methods, results, and discussion sections.

9.1.1. Introduction section

Items 3 and 4. The requirements of the Introduction section about the scientific background, the experimental approach and rationale, and the primary and secondary hypotheses were already discussed as an essential part of the study protocol (see Chapter 8). As mentioned before, the writing of a suitable study protocol should not be considered as time wasted, but as time and effort gained when the experiment reaches its end.

9.1.2. Methods section

Study design - Item 6. For each study, the size and number of experimental groups and control groups must be reported. Readers should be told about the weaknesses and strengths of the study design, e.g. whether randomisation and blinding were used as these add to the reliability of the data. A detailed description of the randomisation and blinding procedures, and how and when these were applied will allow the reader to judge the quality of the study. Reasons for blocking and the blocking factors should be given and how blocking was dealt with in the statistical analysis. When there is ambiguity about the experimental unit, the unit used in the statistical analysis, a single animal, a group (e.g. litter), or a cage of animals should be specified, and a justification for its choice should be provided. When animals are housed as a group, the cage, not the animal is the experimental unit (see Chapter 4).

Sample size - Item 10. Specify the total number of animals used in each experiment and each experimental group. Explain how the total number of animals was decided, provide details of any sample size calculation used (see Chapter 6). Consider a factorial design to increase the opportunity to observe drug effects allowing sample sizes to be reduced (see Section 5.3.2, page 91).

When making multiple statistical tests, there is always the risk of finding false positive results (see Section 7.6). One way to guard against this is to conduct multiple independent experiments. However, if a positive result was observed in only one of several experiments, then the reader should be made aware of this as it could indicate a false positive result (Bate and Clark, 2014).

Allocating animals to experimental groups - Item 11. Give full details of how animals were allocated to experimental groups, including randomisation or matching if done. Some form of randomisation should be used, but this must be done after a suitable experimental design (see Chapter 5) has been selected.

It is essential for the reader to know the order in which the animals were treated and assessed. If this was done in a non-random order, it is possible that confounding bias could have affected the outcome.

Statistical methods - Item 13. Statistical methods should be described in enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. The authors should state and justify which methods they used. A term like *tests of significance* is too vague and should be more detailed.

The level of significance and, when applicable, the direction of the statistical tests should be specified, e.g. "*two-sided p*-values less than or equal to 0.05 were considered to indicate statistical significance." Some procedures, e.g. analysis of variance, chi-square tests, etc. are by definition two-sided. Issues about multiplicity (Section 7.6) and a justification of the strategy that deals with them should also be addressed here. The unit of analysis in each dataset (e.g. single cell, single animal, group of animals, cage of animals) must be specified. The authors should also describe the methods they used to assess whether the data met the assumptions of the statistical analysis.

The software used in the statistical analysis and its version should also be specified. When the **R**-*system* is used (R Core Team, 2017), both **R** and the packages that were used should be referenced.

9.1.3. The Results section

Numbers Analysed - Item 15. The number of experimental units in each group included in each analysis should always be reported, such that the reader has an indication of the sensitivity of the results, in the sense that it allows the reader to decide whether the study was adequately powered, underpowered, or overpowered. Report absolute numbers instead of percentages.

If any animals were excluded from the analysis, their number must be indicated, and an explanation should be given how (statistically or otherwise) the exclusion criteria were defined. Any discrepancies with the number of units randomised to treatment conditions should be accounted for.

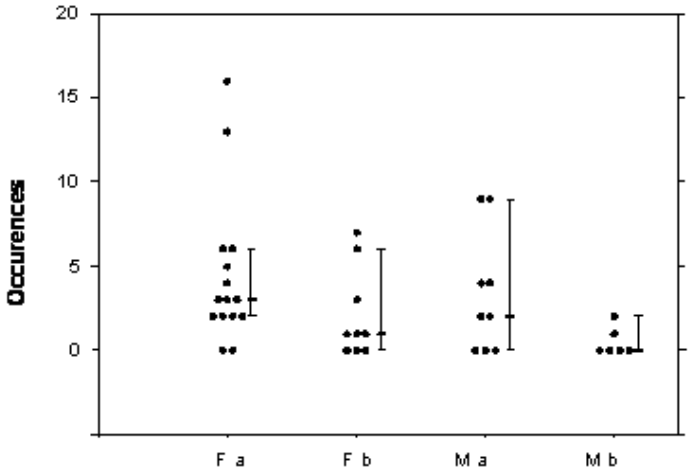


Figure 9.1. Scatter diagram with indication of median values and 95% distribution-free confidence intervals

Outcomes and estimation - Item 16. Findings should be quantified and presented with appropriate indicators of measurement error or uncertainty. As measures of spread and precision, standard deviations (SD) and standard errors (SEM) should not be conflated. Standard deviations are a measure of spread and as such a descriptive statistic, while standard errors are a measure of the precision of the mean.

Normally distributed data should preferably be summarised as *mean (SD)*, not as *mean* ± *SD*. For non-normally distributed data, medians and interquartile ranges are the most appropriate summary statistics. The practice of reporting the *mean* ± *SEM* should preferably be replaced by the reporting of confidence intervals, which are more informative. Extremely small datasets should not

be summarised at all, but instead be reported or displayed as raw data as in Figure 9.1.

When reporting SD (or SEM) one should realise that for positive variables (i.e. variables measured on a ratio scale, a scale with a true zero point) such as concentrations, durations, and counts, the mean minus $2 \times SD$ (or minus $2 \times SEM \times \sqrt{n}$) which indicates a lower 2.5% of the distribution, can lead to a ridiculous negative value. In this case, an appropriate 95% confidence interval based on the lognormal distribution or a distribution-free confidence interval will avoid such a pitfall.

Spurious precision detracts from a paper's readability and credibility. Therefore, unnecessary precision, particularly in tables, should be avoided. When presenting means and standard deviations, it is important to bear in mind the precision of the original data. Means should be given one additional decimal place more than the raw data. Standard deviations and standard errors usually require one more extra decimal place. Percentages should not be expressed to more than one decimal place, and with sample sizes smaller than 100, the use of decimal places should be avoided. Percentages should not be used at all for small samples. Note that the remarks about rounding apply only to the presentation of results, rounding should not be done at all before or during the statistical analysis.

9.2. Additional topics in reporting results

9.2.1. Graphical displays

Graphical displays complement tabular presentations of descriptive statistics. Graphs are better suited than tables for identifying

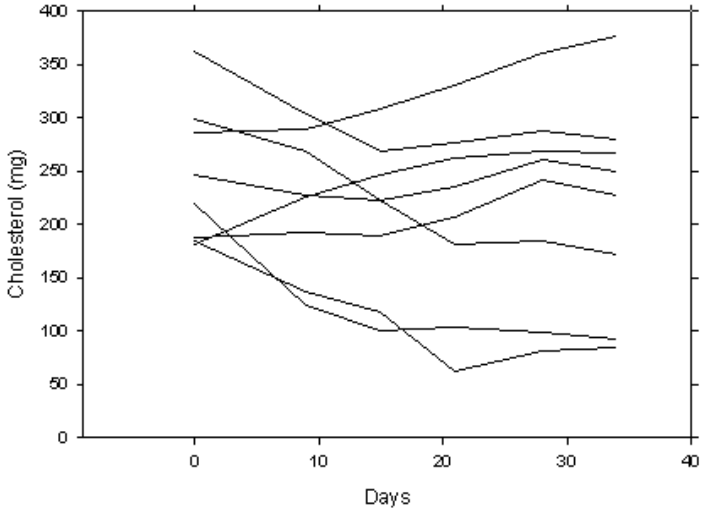


Figure 9.2. Graphical display of longitudinal data showing individual subject profiles

patterns in the data, whereas tables are better for providing large amounts of data with a high degree of numerical detail. Whenever possible, one should always attempt to graph individual data points, especially when treatment groups are small. Plots such as Figure 9.1 and Figure 9.2 are much more informative than the usual bar and line graphs showing mean values ± SEM (Weissgerber et al., 2015). These graphs are easily constructed in the **R**-language or using the Graphpad Prism software (GraphPad Software, 2016). Specifically, the **R**-package *beeswarm* (Eklund, 2010) can be of great help. Finally, avoid unnecessary 3D effects, as they distract from the content of your graph.

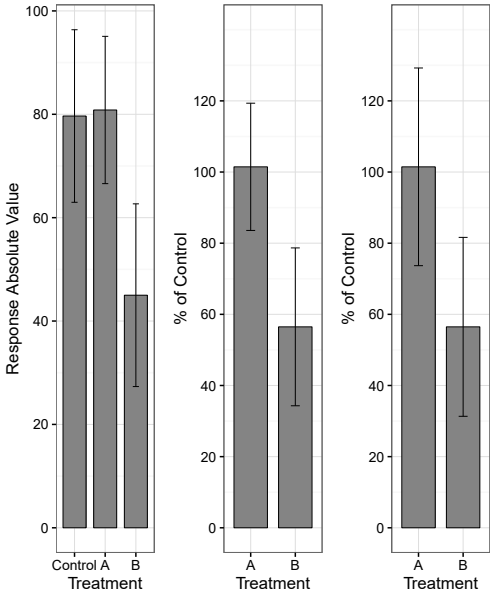


Figure 9.3. Misconception about the variability when computing per cent of control. Mean values and SD of the raw data are displayed in the left panel. The middle panel shows the data as percentage of control, but the researcher ignored the variability present in the control group. Therefore, the reported standard deviations are an underestimate of their true value. The panel on the right shows the percentages and the "*correct*" standard deviation calculated from equation 9.1

9.2.1.1. Percentage of control - A common misconception

Researchers often express the outcome of their experiments as a percentage of the control condition. However, by computing percentages, the units of measurement and the actual values are lost. Moreover, calculating percentages does not make sense when the data are not measured on a ratio scale, i.e. a measurement scale with a true zero point. For example, Celsius and Fahrenheit constitute interval scales rather than ratio scales. Consequently, a 5%

Table 9.1. A researcher calculates percentages based on the control group, but neglects the variability of this group. The standard deviations that he reports *SD %** considerably underestimate their value obtained from Equation 9.1, which is shown in the last column *SD %*

Treatment	n	Response	SD	% Response	SD %*	SD %
Control	6	79.7	16.69	100.0		
A	6	80.8	14.25	101.5	17.88	27.78
B	6	45.0	17.67	56.5	22.19	25.15

increase in body temperature is not the same for temperature measured in $^\circ C$ as when measured in $^\circ F$.

Whenever possible, a statistical model should be applied that incorporates the normalisation. For example, the four-parameter logistic model in dose-response experiments allows to estimate and adjust for the response of the negative and positive controls.

Example 9.1: Percentage of control group A scientist carries out an experiment in which two experimental groups A and B are compared with a control reference group C. He decides to express the results of the experimental groups as a percentage of the control mean. His results are summarised in Table 9.1 and depicted in the left panel of Figure 9.3.

The middle panel of Figure 9.3 shows the naive calculations in which the individual responses were divided by the mean of the control group and the variability of the controls is neglected. The standard deviations in the right panel of Figure 9.3 take the variability of the control group into account and are considerably larger than the former ones. It is obvious that the naive computation of percentages can be misleading.

As in the above example, the mean value of a control group can be used to calculate percentages, but in doing so the variability of the controls is often neglected. This way of normalising the response is typical in 96-well experiments such as MTT-assays.

However, the variability in the outcome is underestimated, and consequently, the results are misleading. The standard deviation of the ratio of two independent groups μ_X/μ_Y is given by[1]:

$$\sigma_{X/Y} = \sqrt{\frac{1}{\mu_Y^2}\sigma_X^2 + \frac{\mu_X^2}{\mu_Y^4}\sigma_Y^2} \tag{9.1}$$

For percentages, the standard deviation obtained from Equation 9.1 must be multiplied by 100.

9.2.1.2. Interpreting and reporting significance tests

When data are summarised in the Results section, the statistical methods that were used to analyse them should be specified in full detail. It is to the reader of little help to have in the Methods section a statement such as *"statistical methods included analysis of variance, regression analysis, as well as tests of significance"* without any reference to which specific procedure is reported in the Results part.

Tests of statistical significance should be *two-sided*. When comparing two means or two proportions, there is a choice between a two-sided or a one-sided test (see Section 7.3, page 141). In a one-sided test, the alternative hypothesis specifies the direction of the difference, e.g. experimental treatment greater than control. In a two-sided test, no such direction is specified. A one-sided test is rarely appropriate, and when one-sided tests are used, their use should be justified (Bland and Altman, 1994). For all two group

[1] The derivation of $\sigma_{X/Y}$ is based on the Delta method and assumes that the ratio X/Y is normally distributed, which it usually is not. For the interpretation of percentage of control, it is much better to make use of confidence intervals based on Fieller's theorem (Feuerstein et al., 1997; Fieller, 1940)

comparisons, the report should clearly state whether one-sided or two-sided p-values are reported.

Never present results simply as p-values, without the estimates. A statistically significant result does not mean that it is of any practical importance (Lindsey, 1999, p. 182). Exact p-values, rather than statements such as "$p < 0.05$" or even worse "NS" (not significant), should be reported where possible. Besides, the practice of dichotomising p-values into *significant* and *not significant* has no rational scientific basis at all and should be abandoned (Amrhein et al., 2017). This lack of rationality becomes apparent when one considers the situation where a study yielding a p-value of 0.049 would be flagged significant, while an almost equivalent result of 0.051 would be flagged as "NS". Reporting exact p-values, as Fisher recommended, would allow readers to compare the reported p-value with their own choice of significance levels.

One should also avoid reporting a p-value as $p = 0.000$, since a value with zero probability of occurrence is, by definition, an impossible value. No observed event can ever have a probability of zero. Therefore, such an extremely low p-value must be reported as $p < 0.001$. In rounding a p-value, it happens that a value that is technically larger than the significance level of 0.05, say 0.051, is rounded down to $p = 0.05$, which is incorrect and, to avoid this error, p-values should be reported to the third decimal. If a one-sided test is used and the result is in the wrong direction, then the report must state that $p > 0.05$ (Levine and Atkin, 2004), or even better report the complement of the p-value, i.e. $1 - p$.

Nonsignificant results. There is a common misconception among scientists that a nonsignificant finding implies that the null hypothesis can be accepted. We already pointed out in

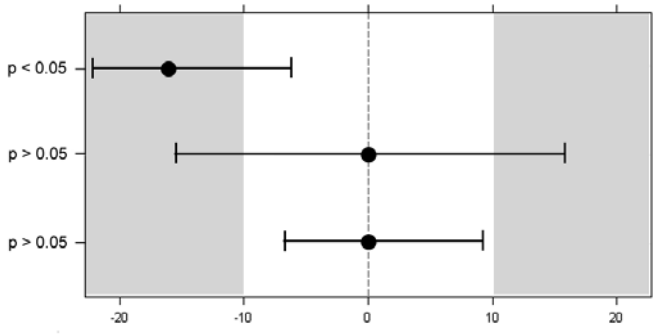


Figure 9.4. Use of confidence intervals for interpreting statistical results. Estimated treatment effects are displayed with their 95% confidence intervals. The shaded area indicates the zone of biological relevance.

Section 7.5 (page 145) that this is a logical fallacy. Accordingly, we must keep in mind that *lack of evidence is no evidence for lack of effect*. We also mentioned that a statistically significant effect is not necessarily of biomedical importance, nor that is it replicable. Therefore, we should avoid sole reliance on statistical hypothesis testing and preferably supplement our findings with confidence intervals which are more informative.

Confidence intervals for a difference of means or proportions provide information about the size of an effect and its uncertainty and are of particular value when the results of the test fail to reject the null hypothesis. This is illustrated in Figure 9.4 showing treatment effects and their 95% confidence intervals. The shaded area indicates the region in which results are important from a biological point of view. Three possible outcomes for treatment effects are shown here as mean values and 95% confidence intervals. The region encompassed by the confidence interval can be interpreted as the set of plausible values of the treatment effect.

The top chart shows a result that is statistically significant, and consequently, the 95% confidence interval does not encompass the zero effect line. Nevertheless, as indicated by the upper limit of the confidence interval, effect sizes that have no biological relevance are still plausible. The chart in the middle illustrates the result of an experiment that was not significant at the 0.05 level. However, the confidence interval reaches well within the area of biological relevance. Therefore, notwithstanding the nonsignificant outcome, this experiment is inconclusive. The third outcome concerns a result that was not significant, but the 95% confidence interval does not reach beyond the boundaries of scientific relevance. This nonsignificant outcome indicates that with 95% confidence, the treatment effect was also irrelevant from a biological point of view.

X is statistically significant, while Y is not. The sharp distinction scientists make between *significant* and *nonsignificant* findings often leads to comparisons of the sort *"X is statistically significant, while Y is not"*. A typical example of such a claim is a sentence like:

> *The percentage of neurons showing cue-related activity increased with training in the mutant mice ($P < 0.05$), but not in the control mice ($P > 0.05$).* (Nieuwenhuis et al., 2011)

Such comparisons are absurd, inappropriate and can be misleading. Indeed, *the difference between "significant" and "not significant" is not itself statistically significant* (Gelman and Stern, 2006). Unfortunately, such a practice is commonplace. A recent review by Nieuwenhuis et al. (2011) showed that in the area of cellular and molecular neuroscience the majority of authors erroneously claim an interaction effect when they obtained a significant result in one group and a nonsignificant result in the other. Given our discussion of *p*-values and nonsignificant findings, it is needless to say

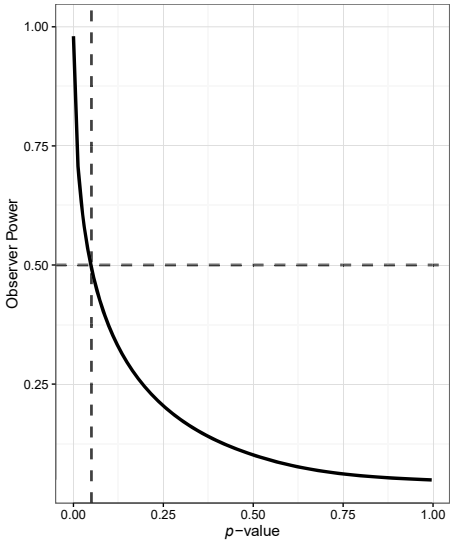


Figure 9.5. Observed power as a function of the p-value for a two-sided t-test with significance level $\alpha = 0.05$. When a test is marginally significant (i.e. $p = 0.05$), the estimated power is 50%

that this approach is completely wrong and misleading. The correct approach would be to design a factorial experiment and test the interaction effect of genotype on training.

In this context, it must also be noted that carrying out a statistical test to prove equivalence of baseline data is pointless. Tests of significance are not tests of equivalence. When baseline measurements are present, their value should be included in the statistical model.

Post hoc power analysis. Some software packages (e.g. SPSS) provide an estimate of the power to detect the observed treatment

effect in conjunction with the data analysis. Besides, several authors advocate a power analysis based on the observed treatment effect when the experiment yields a non-significant result. Moreover, some journal reviewers or editors go even further and require authors to carry out such calculations.

However, as pointed out by Hoenig and Heisey (2001) such a *post hoc* or *retrospective* power analysis based on the observed treatment effect is not only useless but also misleading since the observed power is directly related to the obtained p-value as shown in Figure 9.5. For a test at significance level $\alpha = 0.05$, when the result is marginally significant with a p-value of 0.05, the observed power is always 50%[1]. Higher p-values will necessarily correspond to powers less than 50%. Therefore, the observed power conveys no new information and only an *a priori* sample size or power calculation should be reported. As shown above, non-significant results can be better interpreted using confidence intervals.

Finally, when interpreting the results of the experiment, the smart researcher should bear in mind the topics covered in Section 6.7 about effect size inflation and Section 7.5 about the pitfalls of p-values.

[1]This equality holds for all levels of significance α

—You know my methods. Apply them.

Sherlock Holmes (in A.C. Doyle: *The Sign of the Four*)

—To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

R.A. Fisher (1938)

10

Concluding Remarks and Summary

10.1. Role of the statistician

What we did not touch yet, was the role of the statistician in the research project . The statistician is a professional particularly skilled in solving research problems. She should be considered as a team member and often even as a collaborator or partner in the research process in which she can have a critical role. Whenever possible, the statistician should be consulted, especially when there is doubt about the design, sample size, or statistical analysis. A statistician working closely together with a scientist can significantly improve the project's likelihood of success. Many applied statisticians become involved into the subject area and, by virtue of their statistical training, take on the role of statistical thinker, thereby permeating the research process. In a large number of

instances, this crucial role of the statistician is recognised and granted with a co-authorship.

The most effective way to work with a consulting statistician is to include her or him from the very beginning of the project, when the study objectives are formulated (Hinkelmann and Kempthorne, 2008). What always should be avoided, is contacting the statistical support group after the experiment has reached its completion, perhaps they can only say what the experiment *died of*.

10.2. Recommended reading

"*Statistics Done Wrong: The Woefully Complete Guide*" by Reinhart (2015) is essential reading material for all scientists working in biomedicine and the life sciences in general. This small book (152 pages) provides a well-written very accessible guide to the most popular statistical errors and slip-ups committed by scientists every day, in the lab and in peer-reviewed journals. Scientists working with laboratory animals should certainly read the article by Fry (2014) and the books "*The Design and Statistical Analysis of Animal Experiments*" by Bate and Clark (2014) and "*Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility*" by Lazic (2016). Although, the small book "*Statistical Principles Revealed*" by Jim Lindsey (Lindsey, 1999) is primarily intended for an audience of clinical researchers, most of the principles also apply to animal research and it is highly recommended.

For those interested in the history of statistics and the life of famous statisticians, "*The Lady Tasting Tea*" by Salsburg (2001) is a

lucidly written account of the history of statistics, and how statistical thinking revolutionised 20th Century science. A clear, comprehensive and highly recommended work on the design of experiments is the book by Selwyn (1996), while, on a more introductory level, there is the book by Ruxton and Colegrave (2003). A gentle introduction to statistics in general and hypothesis testing, confidence intervals and analysis of variance, in particular, can be found in the highly recommended book of the two Wonnacott brothers (Wonnacott and Wonnacott, 1990).

Comprehensive works at an advanced level on statistics and experimental design are the books by Neter et al. (1996), Hinkelmann and Kempthorne (2008), and Casella (2008). For those who want to carry out their analyses in the freely available **R**-language (R Core Team, 2017) is the book by Dalgaard (2002) a good starter, while the book by Everitt and Hothorn (2010) is at a more advanced level. Hints for efficient data visualisation can be found in the work of Tufte (1983) and in the two books by William Cleveland (Cleveland, 1993, 1994). Finally, there is the freely available e-book "*Speaking of Graphics*" (Lewi, 2006), which takes the reader on a fascinating journey through the history of statistical graphics[1].

10.3. Summary

We have looked at the complexities of the research process from the vantage point of a generalist. Statistical thinking was introduced as a non-specialist generalist skill that permeates the entire research process. The seven principles of statistical thinking were formulated as: 1) time spent thinking on the conceptualization

[1] http://www.datascope.be

and design of an experiment is time wisely spent; 2) the design of an experiment reflects the contributions from different sources of variability; 3) the design of an experiment balances between its internal validity (proper control of noise) and external validity (the experiment's generalizability); 4) good experimental practice provides the clue to bias minimisation; 5) good experimental design is the clue to the control of variability; 6) experimental design integrates various disciplines; 7) a priori consideration of statistical power is an indispensable pillar of an effective experiment.

We elaborated on each of these and finally discussed some points to consider in the interpretation and reporting of scientific results. In particular, the problems with a blind trust in statistical hypothesis tests and of exaggerated effect sizes in small significant studies were highlighted. Finally, we considered the reporting phase and had a look at the ARRIVE guidelines for the reporting of animal studies.

Appendices

A

List of Abbreviations and Mathematical Symbols

ANOVA	analysis of variance.
BIBD	balanced incomplete block design.
CRD	completely randomised design.
FDR	False discovery rate.
FrFD	fractional factorial design.
LSD	Latin square design.
RCBD	randomised complete block design.
UFD	unreplicated factorial design.
μ	population mean.
σ	population standard deviation.
m	size of subsample.
n	sample size.
c_v	coefficient of variation.
s	sample standard deviation.
$\bar{x}$	sample mean.

B

Glossary of Statistical Terms

accuracy: the degree to which a measurement process is free of bias.

additive model: a model in which the combined effect of several explanatory variables or factors is equal to the sum of their separate effects.

alternative hypothesis: in the Neyman-Pearson framework, a hypothesis which is presumed to hold if the null hypothesis does not; the alternative hypothesis is necessary for deciding upon the direction of the test and for estimating sample sizes.

analysis of variance (ANOVA): a statistical method of inference for making simultaneous comparisons between two or more means.

balanced design: a term usually applied to any experimental design in which the same number of observations is taken for each combination of the experimental factors.

bias: the long-run difference between the average of a measurement process and its true value.

binary data: also called dichotomous data, data that can take on only two distinct values.

biological unit: the entity about which we would like to make an inference; the biological unit may or may not be the same as the experimental unit.

blinding: the condition under which individuals are uninformed as to the treatment conditions of the experimental units.

blocking: arranging experimental units in groups (blocks), such that units within a block are expected to respond similarly to the treatment.

central limit theorem: a basic theorem from probability and statistics saying that for a random sample of size n from a population with mean μ and variance σ^2, the distribution of the sample means is approximately normal with mean μ and variance σ^2/n as the sample size increases.

coefficient of variation: the ratio of the standard deviation to the mean, only valid for data measured on a ratio scale.

completely randomised design (CRD): a design in which each experimental unit is randomised to a single treatment condition or set of treatments.

confidence interval: a random interval that depends on the data obtained in the study and is used to indicate the reliability of an estimate. For a given confidence level, if several confidence intervals are constructed based on independent repeats of the study, then on the long run, the proportion of such intervals that contain the true value of the parameter will correspond to the confidence level.

confounding: the phenomenon in which an extraneous variable, not under control of the investigator, influences both the factors under study and the response variable.

continuous data: data which, in theory, could take on any possible value in a given range, even an infinite range. In practice, the number and range of distinct values should be great enough so that the data would not be considered discrete.

covariate: a concomitant measurement that is related to the response but is not affected by the treatment.

critical value: the cutoff or decision value in hypothesis testing which separates the acceptance and rejection regions of a test.

data set: a general term for observations and measurements collected during any type of scientific investigation.

degrees of freedom: the number of values that are free to vary in the calculation of a statistic, e.g. for the standard deviation, the mean is already calculated and puts a restriction on the number of values that can vary; therefore the degrees of freedom of the standard deviation is the number of observations minus 1.

dichotomous data: also called binary data, data that can take on only two distinct values.

effect size: when comparing treatment differences, the effect size is the mean difference divided by the standard deviation (not standard error); the standard deviation can be from either group, or a pooled standard deviation can be used.

error degrees of freedom: degrees of freedom associated with the unexplained variation, i.e. the error component in a model.

estimation: an inferential process that uses the value of a statistic derived from a sample to estimate the value of a corresponding population parameter.

experimental unit: the smallest division of the experimental material to which different treatments or experimental conditions can be independently applied.

explanatory variable: a variable, also called predictor, which is used in a relationship to explain or to predict changes in the values of another variable; the latter is called the response or dependent variable.

external validity: extent to which the results of a study can be generalised to other situations.

factor: the condition or set of conditions that is manipulated by the investigator.

factorial design: an experimental design in which two or more series of treatments are tried in all combinations.

factor level: is the particular value that a factor can have. For example, a factor called "Treatment" might have "Control","Drug A","Drug B" as factor levels.

false discovery rate (FDR): the probability of making at least one false positive conclusion in a statistical analysis.

false negative: the error of accepting the null hypothesis when it is false, also referred to as Type II error.

false positive: the error of rejecting the null hypothesis when it is true, also referred to as Type I error.

hypothesis test: a formal statistical procedure where one rejects or does not reject a particular hypothesis on the basis of experimental data.

internal validity: extent to which a causal conclusion based on a study is warranted.

interquartile range: a measure of statistical dispersion equal to the difference between the 75th and 25th percentiles.

Latin square design (LSD): an experimental design used to control for the heterogeneity caused by two sources of variation.

median: the value that separates the higher half of a data set from the lower half.

normal probability plot: a graphical technique to identify substantive departures from normality; the data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line; departures from this straight line, indicate departures from normality.

null hypothesis: a hypothesis indicating "no difference"; in the Neyman-Pearson framework, the null hypothesis will either be accepted or rejected as a result of a statistical hypothesis test.

observational unit: the part of the experimental material on which the response is measured or observed; this is not necessarily identical to the experimental unit.

one-sided test: a statistical test for which the rejection region consists of either very large or very small values of the test statistic, but not of both.

parameter: a population quantity of interest, examples are the population mean and standard deviation.

pilot experiment: a preliminary study or experiment performed to gain initial information to be used in planning a subsequent, definitive study; pilot experiments are used to refine experimental procedures and provide information on sources of bias and variability.

population: the total collection of all units about which the researcher wishes to generalise the obtained results to.

population model: the assumption that the experimental units constitute a representative sample, obtained by random sampling from a well-defined larger population; the population model allows a broad inference, meaning that the conclusions can be extended to the population.

precision: the degree to which a measurement process is limited in terms of its variability about a central value.

protocol: a document describing the plan for a study; protocols typically contain information on the rationale for performing the study, the study objectives, experimental procedures to be followed, sample sizes and their justification, and the statistical analyses to be performed; the study protocol must be distinguished from the technical protocol which is more about lab instructions.

pseudoreplication: the error made by treating multiple measurements on the same unit as if it were single measurements on multiple experimental units. Pseudoreplication typically occurs when there are different levels of sampling. See also subsampling.

p-value: the probability of obtaining a value for a test statistic as extreme as, or more extreme than the observed one, provided the null hypothesis is true; small p-values are unlikely when the null hypothesis holds.

randomisation: a well-defined stochastic law for assigning experimental units to differing treatment conditions; randomisation may also be applied elsewhere in the experiment.

randomised complete block design (RCBD): a design in which all treatments are applied within each block and treatments are compared within the blocks.

sample: the collection of experimental units actually included in in a study; a sample from a population is considered a random sample when all units have an equal chance of inclusion in the sample and when subjects in the population can be considered as independent units.

sample size: the number of independent replicates of the experimental unit in each treatment group.

sampling distribution: the probability distribution of a given statistic based on a random sample; it is a theoretical rather than an empirical distribution that plays a major role in statistical inference.

significance level: the allowable rate of false positives, set prior to analysis of the data.

standard deviation: a measure of the variability of the data. It is defined as the square root of the variance; the standard deviation is a fundamental property of the underlying distribution and, unlike the standard error, is not altered by replication.

standard error: a measure of the precision of an estimate. As the sample size increases, the standard error decreases.

statistic: a mathematical function of the observed data.

statistical inference: the process of drawing conclusions from data that is subject to random variation.

statistical power: the probability of rejecting the null hypothesis when it is false and some specific alternative hypothesis holds.

stochastic: non-deterministic, chance dependent procedure, such as picking numbers out of a hat; a stochastic mechanism can be simulated in computer programs by pseudorandom number generators.

subsampling: the situation in which measurements are taken at several nested levels; the highest level is called the primary sampling unit; the next level is called the secondary sampling unit, etc.; when subsampling is present, it is of great importance to identify the correct experimental unit.

test statistic: a statistic used in significance and hypothesis testing; extreme values of the test statistic are unlikely to occur under the null hypothesis.

treatment: a specific combination of factor levels.

two-sided test: a statistical test for which the rejection region consists of both very large or very small values of the test statistic.

Type I error: error made by the incorrect rejection of a true null hypothesis.

Type II error: error made by not rejecting the null hypothesis when the alternative hypothesis is true.

variability: the random fluctuation of a measurement process about its central value.

variance: a measure of the spread of the data; the sample variance s^2 is defined as the sum of the squared deviations from the mean divided by one less than the number of observations:

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

C

Introduction to R

This Appendix describes the basic steps that are needed to install a working environment for experimental design in **R**, a free software platform that provides a very large number of statistical and graphical techniques. **R** runs on a wide variety of UNIX platforms, Windows, and MacOS.

C.1. Installation

The installation process in Windows, MacOS, and UNIX is pretty straight forward, assuming you are familiar with installing application software on your current platform. The software can be downloaded from the Comprehensive **R** Archive Network (CRAN),which can be reached at https://cran.r-project.org/mirrors.html. Here, you select a mirror close to you, e.g. https://lib.ugent.be/CRAN/ and choose the installation package corresponding to your platform. For Windows, the next window allows you to select from different options. Select ***base***, which opens a new window where you can download the current version of **R**. Alternatively, you can download **Architect** at https://www.openanalytics.eu/architect, or **RStudio** at https://www.rstudio.com/. Both are freely available and provide an integrated environment with a superb code editor. As an

example of their capabilities, the present document was entirely prepared in in LaTeX and **R** under the **Architect** environment.

C.2. Packages for experimental design

In this text, we made use of some specialized **R**-packages to help us in the design of experiments. The package *agricolae* (de Mendiburu, 2016) was used for generating most of the experimental designs of Chapter 5, *FrF2* for the fractional factorial designs (Grömping, 2014), and *pwr* (Champely, 2017) for sample size and power calculations (Chapter 6). These packages have to be installed before they can be used. This is done by:

```
> install.packages("agricolae") # Experimental Design
> install.packages("FrF2") # Fractional factorials
> install.packages("pwr") # Power and sample size
```

After installation, the packages become available after issuing the command *library(package)*. After the package is loaded, information about its usage is provided by typing *help(package)*, as shown below:

```
> library(pwr)
> help(pwr)
```

D

Randomisation in MS Excel™ and R

D.1. Completely randomised design

Suppose 21 experimental units have to be randomly assigned to three treatment groups, such that each treatment group contains exactly seven animals.

D.1.1. MS Excel™

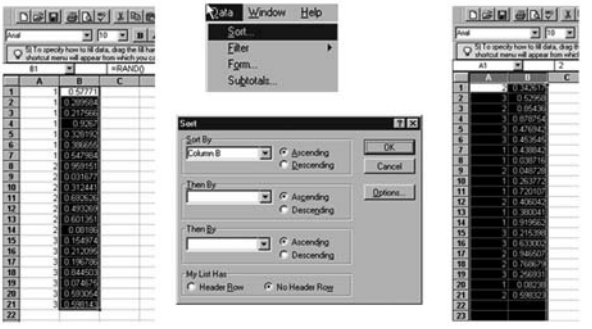


Figure D.1. Generating a completely randomised design in MS Excel™

A randomisation list is easily constructed using a spreadsheet program like MS Excel™. This is illustrated in Figure D.1. We enter in the first column of the spreadsheet the code for the treatment

(1, 2, 3). Using the *RAND()* function, we fill the second column with pseudo-random numbers between 0 and 1. Subsequently, we select the two columns and execute the *Sort* command from the *Data-menu*. In the *Sort-window* that appears now, we select column *B* as column to be sorted by. The treatment codes in column *A* are now in random order, i.e. the first animal will receive treatment 2, the second treatment 3, etc.

D.1.2. R

In the open source statistical language **R**, the same result is obtained by

```
> # make randomization process reproducible
> set.seed(14391)
> # sequence of treatment codes A,B,C repeated 7 times
> x<-rep(c("A","B","C"),7)
> x
 [1] "A" "B" "C" "A" "B" "C" "A" "B" "C" "A" "B" "C"
[13] "A" "B" "C" "A" "B" "C" "A" "B" "C"
> # randomise the sequence in x
> rx<-sample(x)
> rx
 [1] "B" "B" "B" "A" "A" "C" "C" "C" "B" "A" "C" "C"
[13] "A" "B" "C" "B" "A" "A" "C" "A" "B"
```

D.2. Randomised complete block design

Suppose 20 experimental units, organized in 5 blocks of size 4 have to be randomly assigned to 4 treatment groups A, B, C, D, such that each treatment occurs exactly once in each block.

D.2.1. MS Excel™

To generate the design in MS Excel™, we follow the procedure that is depicted in Figure D.2. We enter in the first column of

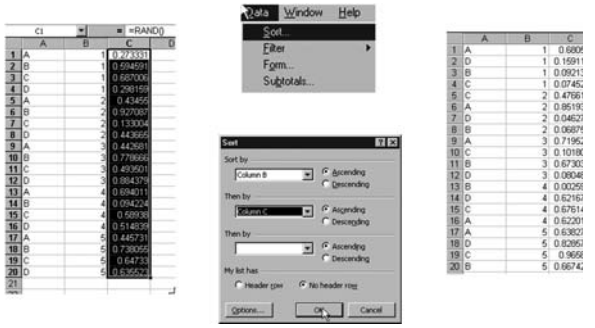


Figure D.2. Generating a randomised complete block design in MS Excel™

the spreadsheet the code for the treatment (A, C, B, D). The second column (Column B) is filled with an indication of the block (1:5). Using the *RAND()* function, we fill the third column with pseudo-random numbers between 0 and 1. Subsequently, we select the three columns and execute the Sort command from the Data-menu. In the Sort window that appears now, we select Column B as first sort criterion and Column C as second sort criterion. The treatment codes in column A are now for each block in random order, i.e. the first animal in block 1 will receive treatment A, the second treatment D, etc.

D.2.2. R

```
> set.seed(3223) # some number
> # treatments repeated 5 times
> treat<- rep(c("A","B","C","D"),5)
> # blocks numbered 1 to 5, each repeated 4 times
> blk<-rep(1:5,rep(4,5))
> # make design matrix of blocks and treatments
> design<-data.frame(block=blk,treat=treat)
> head(design,10) # first 10 exp units
   block treat
1      1     A
2      1     B
3      1     C
4      1     D
```

```
5      2     A
6      2     B
7      2     C
8      2     D
9      3     A
10     3     B
> # randomly distribute units
> rdesign<-design[sample(dim(design)[1]),]
> # order by blocks for convenience
> rdesign<-rdesign[order(rdesign[,"block"]),]
> # sequence of units within blocks
> # is randomly assigned to treatments
> head(rdesign,10)
   block treat
3      1     C
4      1     D
1      1     A
2      1     B
8      2     D
6      2     B
7      2     C
5      2     A
9      3     A
11     3     C
```

E

ARRIVE Guidelines

Table E.1. Selected items from the ARRIVE guidelines (Kilkenny et al., 2010)

Item	Point to consider	Recommendation
TITLE		
1	General	Provide as accurate and concise a description of the content of the article as possible
ABSTRACT		
2	General	Provide an accurate summary of the background, research objectives (including details of the species or strain of animal used), key methods, principal findings, and conclusions of the study
INTRODUCTION		
3	Background	a. Include sufficient scientific background (including relevant references to previous work) to understand the motivation and context for the study, and explain the experimental approach and rationale. b. Explain how and why the animal species and model being used can address the scientific objectives and, where appropriate, the study's relevance to human biology

Continued on next page

Table D.1 (continued)

Item	Point to consider	Recommendation
4	Objectives	Clearly describe the primary and any secondary objectives of the study, or specific hypotheses being tested
METHODS		
5	Ethical statement	Indicate the nature of the ethical review permissions, relevant licences (e.g. Animal [Scientific Procedures] Act 1986), and national or institutional guidelines for the care and use of animals, that cover the research.
6	Study design	For each experiment, give brief details of the study design, including: a. The number of experimental and control groups. b. Any steps taken to minimise the effects of subjective bias when allocating animals to treatment (e.g., randomisation procedure) and when assessing results (e.g., if done, describe who was blinded and when). c. The experimental unit (e.g. a single animal, group, or cage of animals). A time-line diagram or flow chart can be useful to illustrate how complex study designs were carried out.

Continued on next page

Table D.1 (continued)

Item	Point to consider	Recommendation
7	Experimental procedures	For each experiment and each experimental group, including controls, provide precise details of all procedures carried out. For example: a. How (e.g., drug formulation and dose, site and route of administration, anaesthesia and analgesia used [including monitoring], surgical procedure, method of euthanasia). Provide details of any specialist equipment used, including supplier(s). b. When (e.g., time of day). c. Where (e.g., home cage, laboratory, water maze). d. Why (e.g., rationale for choice of specific anaesthetic, route of administration, drug dose used).
8	Experimental animals	a. Provide details of the animals used, including species, strain, sex, developmental stage (e.g., mean or median age plus age range), and weight (e.g., mean or median weight plus weight range) b. Provide further relevant information such as the source of animals, international strain nomenclature, genetic modification status (e.g. knock-out or transgenic), genotype, health/immune status, drug- or testnaive, previous procedures, etc.

Continued on next page

Table D.1 (continued)

Item	Point to consider	Recommendation
9	Housing and husbandry	Provide details of: a. Housing (e.g., type of facility, e.g., specific pathogen free (SPF); type of cage or housing; bedding material; number of cage companions; tank shape and material etc. for fish). b. Husbandry conditions (e.g., breeding programme, light/dark cycle, temperature, quality of water etc. for fish, type of food, access to food and water, environmental enrichment). c. Welfare-related assessments and interventions that were carried out before, during, or after the experiment.
10	Sample size	a. Specify the total number of animals used in each experiment and the number of animals in each experimental group. b. Explain how the number of animals was decided. Provide details of any sample size calculation used. c. Indicate the number of independent replications of each experiment, if relevant.
11	Allocating animals to experimental groups	a. Give full details of how animals were allocated to experimental groups, including randomisation or matching if done. b. Describe the order in which the animals in the different experimental groups were treated and assessed.
12	Experimental outcomes	Clearly define the primary and secondary experimental outcomes assessed (e.g., cell death, molecular markers, behavioural changes).

Continued on next page

Table D.1 (continued)

Item	Point to consider	Recommendation
13	Statistical methods	a. Provide details of the statistical methods used for each analysis. b. Specify the unit of analysis for each dataset (e.g. single animal, group of animals, single neuron). c. Describe any methods used to assess whether the data met the assumptions of the statistical approach.
RESULTS		
14	Baseline data	For each experimental group, report relevant characteristics and health status of animals (e.g., weight, microbiological status, and drug- or test-naive) before treatment or testing (this information can often be tabulated).
15	Numbers analysed	a. Report the number of animals in each group included in each analysis. Report absolute numbers (e.g. 10/20, not 50%) b. If any animals or data were not included in the analysis, explain why.
16	Outcomes and estimation	Report the results for each analysis carried out, with a measure of precision (e.g., standard error or confidence interval).
17	Adverse events	a. Give details of all important adverse events in each experimental group. b. Describe any modifications to the experimental protocols made to reduce adverse events.

Continued on next page

Table D.1 (continued)

Item	Point to consider	Recommendation
DISCUSSION		
18	Interpretation Scientific implications	a. Interpret the results, taking into account the study objectives and hypotheses, current theory, and other relevant studies in the literature. b. Comment on the study limitations including any potential sources of bias, any limitations of the animal model, and the imprecision associated with the results. c. Describe any implications of your experimental methods or findings for the replacement, refinement, or reduction (the 3Rs) of the use of animals in research.
19	Generalizability Translation	Comment on whether, and how, the findings of this study are likely to translate to other species or systems, including any relevance to human biology.
20	Funding	List all funding sources (including grant number) and the role of the funder(s) in the study

Concluded

References

Academy of Medical Science (2015). Reproducibility and reliability of biomedical research: improving research practice. URL https://acmedsci.ac.uk/policy/policy-projects/reproducibility-and-reliability-of-biomedical-research

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, UK: Chapman & Hall.

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. New York, NY: J. Wiley.

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* **5**, e3544. doi:10.7717/peerj.3544.

Aoki, Y., Helzlsouer, K. J., and Strickland, P. T. (2014). Arylesterase phenotype-specific positive association between arylesterase activity and cholinesterase specific activity in human serum. *Int. J. Environ. Res. Public Health* **11**, 1422–1443. doi:doi:10.3390/ijerph110201422.

Babij, C. J., Zhang, R. J., Y. anf Kurzeja, Munzli, A., Shehabeldin, A., Fernando, M., Quon, K., Kassner, P. D., Ruefli-Brasse, A. A., Watson, V. J., Fajardo, F., Jackson, A., Zondlo, J., Sun, Y., Ellison, A. R., Plewa, C. A., T., S., Robinson, J., McCarter, J., Judd, T., Carnahan, J., and Dussault, I. (2011). STK33 kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer Research* **71**, 5818–5826. doi:10.1158/0008-5472.CAN-11-0778.

Baggerly, K. A. and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* **3**, 1309–1334. doi:10.1214/09-AOAS291.

Baker, M. (2016). Is there a reproducibility crisis? *Nature* **533**, 452–456.

Bate, S. T. and Clark, R. A. (2014). *The Design and Statistical Analysis of Animal Experiments*. Cambridge, UK: Cambridge University Press.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical research. *Nature* **483**, 531–533. doi:10.1038/483531a.

Begley, C. G. and Ioannidis, J. P. A. (2015). Reproducibility in science. *Circ. Res.* **116**, 116–126. doi:10.1161/CIRCRESAHA114.303819.

Begley, S. (2012). In cancer science, many "discoveries" don't hold up. *Reuters* **March 28**.
URL http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328

Biggers, J. D., Baskar, J. F., and Torchiana, D. F. (1981). Reduction of fertility of mice by the intrauterine injection of prostaglandin antagonists. *J. Reprod. Fert.* **63**, 365–372.

Bland, M. and Altman, D. (1994). One and two sided tests of significance. *BMJ* **309**, 248.

Bolch, B. (1968). More on unbiased estimation of the standard deviation. *The American Statician* **22**, 27.

Box, G. (1990). Commentary: [Communications between Statisticians and Engineers/Physical Scientists]. *Technometrics* **32**, 251–252. doi:10.2307/1269094.

Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple Comparisons Using R*. Boca Raton, FL: CRC Press.

Bretz, F., Landgrebe, J., and Brunner, E. (2005). Multiplicity issues in microarray experiments. *Methods Inf Med* **44**, 431–437.

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Med.* **14**, 1933–1940.

Burrows, P. M., Scott, S. W., Barnett, O., and McLaughlin, M. R. (1984). Use of experimental designs with quantitative ELISA. *J. Virol. Methods* **8**, 207–216.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 1–12. doi:10.1038/nrn3475.

Casella, G. (2008). *Statistical Design*. New Tork, NY: Springer.

Champely, S. (2017). *pwr: Basic Functions for Power Analysis*. R package version 1.2-1.
URL https://CRAN.R-project.org/package=pwr

Cleveland, W. S. (1993). *Visualizing Data.* Summit, NJ: Hobart Press.

Cleveland, W. S. (1994). *The Elements of Graphing Data.* Summit, NJ: Hobart Press.

Cochran, W. G. and Cox, G. (1957). *Experimental Designs*. New York, NY: John Wiley & Sons Inc., 2nd edition.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd edition.

Cokol, M., Ozbay, F., and Rodriguez-Esteban, R. (2008). Retraction rates are on the rise. *EMBO Rep.* **9**, 2. doi:10.1038/sj.embor.7401143.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2246630/

Colquhoun, D. (1963). Balanced incomplete block designs in biological assay illustrated by the assay of gastrin using a youden square. *Brit J Pharmacol* **21**, 67–77.

Colquhoun, D. (2014). An investigation of the discovery rate and the misinterpretation of p-values. *R. Soc. open sci.* **1**, 140216. doi:10.1098/rsos/140216.

Council of Europe (2006). Appendix A of the European Convention for the Protection of Vertebrate Animals used for Experimental and other Scientific Purposes (ETS No. 123. Guidelines for accomodation and care of animals (Article 5 of the Convention). Approved by the Multilateral Consultation.
URL https://www.aaalac.org/about/AppA-ETS123.pdf

Couzin-Frankel, J. (2013). When mice mislead. *Science* **342**, 922–925.
URL http://science.sciencemag.org/content/342/6161/922

Cox, D. (1958). *Planning of Experiments*. New York, NY: J. Wiley.

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *Am. J. Physiol. Regulatory Integrative Comp. Physiol.* **279**, R1–R8.

Dalgaard, P. (2002). *Introductory Statistics with R.* New York, NY: Springer.

Daniel, C. (1959). Use of half-normal plots in intepreting factorial two-level experiments. *Technometrics* **1**, 311–341.

de Mendiburu, F. (2016). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.2-4.
URL https://CRAN.R-project.org/package=agricolae

Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. New York, NY: Springer.

Dell, R. B., Holleran, S., and Ramakrishnan, R. (2012). Sample size determination. *ILAR J.* **43**, 207–213.

Dowdy, S., Wearden, S., and Chilko, D. (2004). *Statistics for Research*. Hoboken, NJ: J. Wiley & Sons, 3rd edition.

Eklund, A. (2010). beeswarm: The bee swarm plot, an alternative to stripchart. R package version 0.0.7.
URL http://CRAN.R-project.org/package=beeswarm

European Food Safety Authority (2012). Final review of the Séralini et al. (2012) publication on a 2-year rodent feeding study with glyphosate formulations and GM maize NK603 as published online on 19 September 2012 in Food and Chemical Toxicology. *EFSA Journal* **10**, 2986.
doi:10.2903/j.efsa.2012.2986.

Everitt, B. S. and Hothorn, T. (2010). *A Handbook of Statistical Analyses using R*. Boca Raton, FL: Chapman and Hall/CRC, 2nd edition.

Faessel, H., Levasseur, L., Slocum, H., and Greco, W. (1999). Parabolic growth patterns in 96-well plate cell growth experiments. *In Vitro Cell. Dev. Biol. Anim.* **35**, 270–278.

Fang, F. C., Steen, R. C., and Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17028–17033. doi:10.1073/pnas.1212247109.

Festing, M. F. W. (2014). Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *ILAR J* **55**, 472–476. doi:10.1093/ilar/ilu045.

Festing, M. F. W. and Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* **43**, 244–258.

Feuerstein, T. J., Roßner, R., and Schumacher, M. (1997). How to express an effect mean as percentage of a control mean? *J Pharmacol Toxicol Meth* **37**, 187–190. doi:10.12688/f1000research.6963.5.

Fieller, E. (1940). The biological standardization of insulin. *J Roy Statist Soc Suppl* **7**, 1–64. doi:10.2307/2983630.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1938). Presidential address: The first session of the Indian Statistical Conference, Calcutta. *Sankhya* **4**, 14–17.

Fisher, R. A. (1962). The place of the design of experiments in the logic of scientific inference. *Colloques Int. Centre Natl. Recherche Sci. Paris* **110**, 13–19.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. New York, NY: Hafner Press, 3rd edition.

Fitts, D. A. (2010). Improved stopping rules for the design of small-scale experiments in biomedical and biobehavioral research. *Behavior Research Methods* **42**, 3–22. doi:10.3758/BRM.42.1.3.

Fitts, D. A. (2011). Minimizing animal numbers: the variable-criteria stopping rule. *Comparative Medicine* **61**, 206–218.

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165. doi:10.1371/journal.pbio.1002165.

Fry, D. (2014). Experimental design: reduction and refinement in studies using animals. In K. Bayne and P. Turner, editors, *Laboratory Animal Welfare*, chapter 8, pages 95–112. London, UK: Academic Press.

Gaines Das, R. E. (2002). Role of ancillary variables in the design, analysis, and interpretation of animal experiments. *ILAR J* **43**, 214–222.

Gart, J. J., Krewski, D., N., L. P., Tarone, R. E., and Wahrendorf, J. (1986). *The design and analysis of long-term animal experiments*, volume 3 of *Statistical Methods in Cancer Research*. Lyon, France: International Agency for Research on Cancer.

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Semin Hematol* **45**, 135–140. doi:10.1053/j.seminhematol.2008.04.003.

Gore, K. H. and Stanley, P. J. (2005). An illustration that statistical design mitigates environmental variation and ensures unambiguous study conclusions. *Animal Welfare* **14**, 361–365.

GraphPad Software (2016). *GraphPad Prism version 7.00 for Windows*. La Jolla, California, USA.
URL www.graphpad.com

Greco, W. R., Bravo, G., and Parsons, J. C. (1995). The search for synergy: a critical review from a response surface perspective. *Pharmacol Rev* **47**, 331–385.

Greenman, D. L., Bryant, P., Kodell, R. L., and Sheldon, W. (1983). Relationship of mouse body weight and food consumption/wastage to cage shelf level. *Lab. Anim. Sci.* **33**, 555–558.

Greenman, D. L., Kodell, R. L., and Sheldon, W. G. (1984). Association between cage shelf level and spontaneous and induced neoplasms in mice. *J. natl Cancer Inst.* **73**, 107–113.

Grömping, U. (2014). R package FrF2 for creating and analyzing fractional factorial 2-level designs. *Journal of Statistical Software* **56**, 1–56.
URL http://www.jstatsoft.org/v56/i01/

Haseldonckx, M., Van Reempts, J., Van de Ven, M., Wouters, L., and Borgers, M. (1997). Protection with lubeluzole against delayed ischemic brain damage in rats. a quantitative histopathologic study. *Stroke* **28**, 428–432.

Haseman, J. K. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Enviromental Health Perspect.* **58**, 385–392.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569418/

Hayes, W. A. (2014). Retraction notice to "Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize" [Food Chem. Toxicol. 50 (2012): 4221-4231]. *Food Chem. Toxicol.* **52**, 244. doi:10.1016/j.fct.2013.11.047.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* **13**, e1002106. doi:10.1371/journal.pbio.1002106.

Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood-Cliffs, NJ: Prentice-Hall.

Hille, C., Bate, S., Davis, J., and Gonzalez, M. I. (2008). 5-HT4 receptor antagonism in the five-choice serial reaction time task. *Behavioural Brain Research* **195**, 180–186.

Hinkelmann, K. and Kempthorne, O. (2008). *Design and Analysis of Experiments. Volume 1. Introduction to Experimental Design*. Hoboken, NJ: J. Wiley, 2nd edition.

Hirst, J. A., Howick, J., Aronson, J. K., Roberts, N., Perera, R., Koshiaris, C., and Heneghan, C. (2014). The need for randomization in animal trials: An overview of systematic reviews. *PLOS ONE* **9**, e98856. doi:10.1371/journal.pone.0098856.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* **55**, 19–24. doi:10.1198/000313001300339897.

Holland, T. and Holland, C. (2011). Unbiased histological examinations in toxicological experiments (or, the informed leading the blinded examination). *Toxicol. Pathol.* **39**, 711–714. doi:10.1177/0192623311406288.

Holland-Letz, T. and Kopp-Schneider, A. (2015). Optimal experimental designs for dose-response studies with continuous endpoints. *Archiv Toxicol* **89**, 2059–2068. doi:10.1007/s00204-014-1335-2.

Holman, L., Head, M. L., Lanfear, R., and Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol.* **13**, e1002190. doi:10.1371/journal.pbio.1002190.

Hotz, R. L. (2007). Most science studies appear to be tainted by sloppy analysis. *The Wall Street Journal* **September 14**.
URL http://online.wsj.com/article/SB118972683557627104.html

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* **2**, e124. doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Med.* **11**, e1001747. doi:10.1371/journal.pmed.1001747.

Jones, B. and Kenward, M. G. (2003). *Design and Analysis of Cross-Over Trials*. Boca Raton, FL: Chapman & Hall/CRC, 2nd edition.

Kieser, M. and Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biom. J.* **8**, 941–949.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol.* **8**, e1000412. doi:doi:10.1371/journal.pbio.1000412.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F. W., Cuthill, I. C., Fry, D., Hutton, J., and Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLOS ONE* **4**, e7824. doi:10.1371/journal.pone.0007824.

Kimmelman, J., Mogil, J. S., and Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol.* **12**, e1001863. doi:10.1371/journal.pbio.1001863.

Lansky, D. (2002). Strip-plot designs, mixed models, and comparisons between linear and non-linear models for microtitre plate bioassays. In W. Brown and A. R. Mire-Sluis, editors, *The Design and Analysis of Potency Assays for Biotechnology Products. Dev. Biol.*, volume 107, pages 11–23. Basel: Karger.

Laplace, P. (1827). *Mémoire sur le flux et reflux lunaire atmosphérique*. Connaissance des Temps pour l'An 1830.

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis. *BMC Neuroscience* **11**. doi:10.1186/1471-2202-11-5.

Lazic, S. E. (2016). *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility*. Cambridge, UK: Cambridge University Press.

Lazic, S. E., Clarke-Williams, C. J., and Munafò, M. R. (2018). What exactly is 'N' in cell culture and animal experiments. *PLoS Biol.* **16**, e2005282. doi:10.1371/journal.pbio.2005282.

LeBlanc, D. C. (2004). *Statistics: Concepts and Applications for Science*. Sudbury, MA: Jones and Bartlett Publishers.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, CA: Holden-Day.

Lehr, R. (1992). Sixteen s squared over d squared: a relation for crude sample size estimates. *Stat. Med.* **11**, 1099–1102.

Lehrer, J. (2010). The truth wears off. *The New Yorker [online]* **December 13**. URL http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off

Levasseur, L., Faessel, H., Slocum, H., and Greco, W. . (1995). Precision and pattern in 96-well plate cell growth experiments. In *Proceedings of the American Statistical Association, Biopharmaceutical Section*, pages 227–232. Alexandria, Virginia: American Statistical Association.

Levine, T. R. and Atkin, C. (2004). The accurate reporting of software-generated p-values: a cautionary note. *Comm. Res. Rep.* **21**, 324–327. doi:10.1080/08824090409359995.

Lewi, P. J. (2005). The role of statistics in the success of a pharmaceutical research laboratory: a historical case description. *J Chemometr.* **19**, 282–287.

Lewi, P. J. (2006). Speaking of graphics. URL http://www.datascope.be

Lewi, P. J. and Smith, A. (2007). Successful pharmaceutical discovery: Paul Janssen's concept of drug research. *R&D Management* **37**, 355–361. doi:10.1111/j.1467-9310.2007.00481.x.

Lindsey, J. K. (1999). *Revealing Statistical Principles*. London, UK: Arnold.

Loscalzo, J. (2012). Irreproducible experimental results: causes,(mis)interpretations, and consequences. *Circulation* **125**, 1211–1214. doi:10.1161/CIRCULATIONAHA.112.098244.

Montgomery, D. C. (2013). *Design and Analysis of Experiments*. Hoboken, NJ: J. Wiley & Sons, 8th edition.

Nadon, R. and Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends in Genetics* **15**, 265–271.

Naik, G. (2011). Scientists' elusive goal: Reproducing study results. *The Wall Street Journal* **December 2**.
URL http://online.wsj.com/article/SB10001424052970203764804577059841672541590.html

Neef, N., Nikula, K. J., Francke-Carroll, S., and Boone, L. (2012). Regulatory forum opinion piece: blind reading of histopathology slides in general toxicology studies. *Toxicol. Pathol.* **40**, 697–699.

Nelder, J. A. (1999). From statistics to statistical science. *The Statistician* **48**, 257–269.

Neter, J., Kutner, M. H., Nachtsheim, C., and Wasserman, W. (1996). *Applied Linear Statistical Models*. Boston, MA: McGraw-Hill/Irwin, 4th edition.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E. (2011). Erroneous analysis of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* **14**, 1105–1107.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature* **506**, 150–152. doi:10.1038/506150a.

Parkin, S. L., Pritchett, J. P., Grimsdich, D. C., Bruckdorfer, K. R., Sahota, P. K., Lloyd, A., and Overend, P. (2004). Circulating levels of the chemokines JE and KC in female C3H apolipoprotein-E-deficient and C57BL apolipoprotein-E-deficient mice as potential markers of atherosclerosis development. *Biochemical Society Transactions* **32**, 128–130.

Patterson, S. and Jones, B. (2006). *Bioequivalence and Statistics in Clinical Pharmacology*. Boca Raton, FL: Chapman & Hall/CRC.

Peng, R. (2009). Reproducible research and biostatistics. *Biostatistics* **10**, 405–408. doi:10.1093/biostatistics/kxp014.

Peng, R. (2015). The reproducibility crisis in science. *Significance* **12**, 30–32. doi:10.1111/j.1740-9713.2015.00827.x.

Pernet, C. (2018). Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. *F1000Research* **4**, 621. doi:10.12688/f1000research.6963.5.

Pinheiro, J. C. and Bates, D. M. (2001). *Mixed effect models in S and S-Plus*. New York, NY: Springer Verlag.

Popper, K. (2002). *The Logic of Scientific Discovery*. London, UK: Routledge.

Potti, A., Dressman, H. K., Bild, A., Riedel, R., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J., and Nevins, J. R. (2006). Genomic signature to guide the use of chemotherapeutics. *Nature Medicine* **12**, 1294–1300. doi:10.1038/nm1491. (Retracted).

Potti, A., Dressman, H. K., Bild, A., Riedel, R., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J., and Nevins, J. R. (2011). Retracted: Genomic signature to guide the use of chemotherapeutics. *Nature Medicine* **17**, 135. doi:10.1038/nm0111-135. (Retracted).

Prinz, F., Schlange, A., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets. *Nature Rev. Drug Discov.* **10**, 712–713. doi:10.1038/nrd3439-c1.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reinhart, A. (2015). *Statistics Done Wrong: The Woefully Complete Guide*. San Francisco, CA: no starch press.

Ritskes-Hoitinga, M. and Strubbe, J. H. (2007). Nutrition and animal welfare. In E. Kaliste, editor, *The Welfare of Laboratory Animals*, chapter 5, pages 95–112. Dordrecht, The Netherlands: Springer.

Rivenson, A., Hoffmann, D., Prokopczyk, B., Amin, S., and Hecht, S. S. (1988). Induction of lung and exocrine pancreas tumors in F344 rats by tobacco-specific and Areca-derived N-nitrosamines. *Cancer Res.* **48**, 6912–6917.

Ruxton, G. D. and Colegrave, N. (2003). *Experimental Design for the Life Sciences.* Oxford, UK: Oxford University Press.

Salsburg, D. (2001). *The Lady Tasting Tea.* New York, NY.: Freeman.

Schlain, B., Jethwa, H., Subramanyam, M., Moulder, K., Bhatt, B., and Molley, M. (2001). Designs for bioassays with plate location effects. *BioPharm International* **14**, 40–44.

Scholl, C., Fröhling, S., Dunn, I., Schinzel, A. C., Barbie, D. A., Kim, S. Y., Silver, S. J., Tamayo, P., Wadlow, R. C., Ramaswamy, S., Döhner, K., Bullinger, L., Sandy, P., J.S., B., Root, D. E., Jacks, T., Hahn, W., and Gilliland, D. G. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137**, 821–834. doi:10.1016/j.cell.2009.03.017.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician* **55**, 62–71.

Selwyn, M. R. (1996). *Principles of Experimental Design for the Life Sciences*. Boca Raton, FL: CRC Press.

Senn, S. (2002). *Cross-over Trials in Clinical Research*. Chichester, UK: John Wiley & Sons Ltd., 2nd edition.

Séralini, G.-E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., Hennequin, D., and Vendômois, J. (2012). Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. *Food Chem. Toxicol.* **50**, 4221–4231. doi:10.1016/j.fct.2012.08.005.

Séralini, G.-E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., Hennequin, D., and Vendômois, J. (2014). Republished study: long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. *Environmental Sciences Europe* **26**, 14. doi:10.1186/s12302-014-0014-5.

Shaw, R., Festing, M. F. W., Peers, I., and Furlong, L. (2002). Use of factorial designs to optimize animal experiments and reduce animal use. *ILAR J* **43**, 223–232.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*. Ames, IA: Iowa State University Press, 7th edition.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard University Press.

Straetemans, R., O'Brien, T., Wouters, L., Van Dun, J., Janicot, M., Bijnens, L., Burzykowski, T., and M, A. (2005). Design and analysis of drug combination experiments. *Biometrical J* **47**, 299–308.

Tallarida, R. J. (2000). *Drug Synergism and Dose-Effect Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Tallarida, R. J. (2001). Drug synergism: its detection and applications. *J. Pharm. Exp. Ther.* **298**, 865–872.

Tallarida, R. J. and Jacob, L. S. (1979). *The Dose-Response Relation in Pharmacology*. New York, NY: Springer-Verlag.

Temme, A., Sümpel, F., Rieber, G. S. E. P., Willecke, K. J. K., and Ott, T. (2001). Dilated bile canaliculi and attenuated decrease of nerve-dependent bile secretion in connexin32-deficient mouse liver. *Eur. J. Physiol.* **442**, 961–966.

The Economist (2013). Trouble at the lab. *The Economist* .
URL http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correctingalarming-degree-it-not-trouble

Tressoldi, P. E., Giofré, D., Sella, F., and Cumming, G. (2013). High impact = high statistical standards? not necessarily so. *Nature Reviews Neuroscience* **8**, e56180. doi:10.1371/journal.pone.0056180.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT.: Graphics Press.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician* **34**, 23–25.

Van Belle, G. (2008). *Statistical Rules of Thumb.* Hoboken, NJ: J. Wiley, 2nd edition.

van der Worp, B., Howells, D. W., Sena, E. S., Porritt, M. J., Rewell, S., O'Collins, V., and Macleod, M. R. (2010). Can animal models of disease reliably inform human studies. *PLoS Med.* **7**, e1000245. doi:10.1371/journal.pmed.1000245.

van Luijk, J., Bakker, B., Rovers, M. M., Ritskes-Hoitinga, M., de Vries, R. B. M., and Leenaars, M. (2014). Systematic reviews of animal studies; missing link in translational research? *PLOS ONE* **9**, e89981. doi:10.1371/journal.pone.0089981.

Vandenbroeck, P., Wouters, L., Molenberghs, G., Van Gestel, J., and Bijnens, L. (2006). Teaching statistical thinking to life scientists: a case-based approach. *J. Biopharm. Stat.* **16**, 61–75.

Vaux, D. L., Fidler, G., and Cumming, G. (2012). Replicates and repeats - what is the difference and is it significant? *EMBO Rep.* **13**, 291–296. doi:10.1038/embor.2012.36.

Ver Donck, L., Pauwels, P. J., Vandeplassche, G., and Borgers, M. (1986). Isolated rat cardiac myocytes as an experimental model to study calcium overload: the effect of calcium-entry blockers. *Life Sci.* **38**, 765–772.

Ver Donck, L., Wouters, L., Olbrich, H. G., Mutschler, E., and Borgers, M. (1991). Effect of nebivolol on survival of cardiomyopathic hamsters with congestive heart failure. *J Cardiovasc Pharmacol* **18**, 1–3.

Verheyen, F., Racz, R., Borgers, M., Driesen, R. B., Lenders, M. H., and Flameng, W. J. (2014). Chronic hibernating myocardium in sheep can occur without degenerating events and is reversed after revascularization. *Cardiovasc Pathol.* **23**, 160–168. doi:10.1016/j.carpath.2014.01.003.

Wacholder, S., Chanoch, S., Garcia-Closas, M., El ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* **96**, 434–442. doi:10.1093/jnci/djh075.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* **70**, 129–133. doi:10.1080/00031305.2016.1154108.

Weissgerber, T. L., Milic, N. M., Wionham, S. J., and Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol.* **13**, e1002128. doi:10.1371/journal.pbio.1002128.

Wilcoxon, F., Rhodes, L. J., and Bradley, R. A. (1963). Two sequential two-sample grouped rank tests with applications to screening experiments. *Biometrics* **19**, 58–84.

Wilks, S. S. (1951). Undergraduate statistical education. *J. Amer. Statist. Assoc.* **46**, 1–18.

Witte, J. S., Elston, R. C., and Cardon, L. R. (2000). On the relative sample size required for multiple comparisons. *Statist. Med.* **19**, 369–372.

Wonnacott, T. H. and Wonnacott, R. J. (1990). *Introductory Statistics.* New York, NY.: J. Wiley, 5th edition.

Youden, W. J. (1937). Use of incomplete block replications in estimating tobacco-mosaic virus. *Contr. Boyce Thompson Inst.* **9**, 41–48.

Young, S. S. (1989). Are there location/cage/systematic nontreatment effects in long-term rodent studies? a question revisited. *Fundam Appl Toxicol.* **13**, 183–188.

Zimmer, C. (2012). A sharp rise in retractions prompts calls for reform. *The New York Times* **April 17**.
URL http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html

Author Index

Subject Index