1

Statistics and medical science

1.1 Introduction

Many medical researchers have an ambiguous relationship with statistics. They know they need it to be able to publish their results in prestigious academic journals, as opposed to general public tabloids, but they also think that it unnecessarily complicates what should otherwise be straightforward interpretations. The most frustrated medical researchers can probably be found among those who actually do consult biostatisticians; they only too often experience criticism of the design of the experiment they want to do or, worse, have done – as if the design was the business of the statistician at all.

On the other hand, if you ask biostatisticians, they often consider medical science a contradiction in terms. Tradition, subjectivity and intuitive thinking seem to be such an integral part of the medical way of thinking, they say, that it cannot be called science. And biostatisticians feel fully vindicated by the hype that surrounded the term 'evidence-based medicine' during the 1990s. Evidence? Isn't that what research should be all about? Isn't it a bit late to realize that now?

This chapter attempts to explain what statistics actually contributes in clinical research. We will describe, from a bird's-eye perspective, the structure within which statistics operates, and the nature of its results. We will use most of the space to describe the true nature of one particular summary statistic, the *p*-value. Not because it necessarily is the right thing to compute, but because all workers in biostatistics have encountered it. How it is computed will be discussed in later chapters (though more emphasis will be put on its relative, the confidence interval).

Medicine is not a science *per se*. It is an engineering application of biology to human disease. Medicine is about diagnosing and treating individual patients in accordance with tradition and established knowledge. It is a highly subjective activity in which the physician uses his own and others' experiences to find a diagnostic fit to the signs and symptoms of a particular patient, in order to identify the appropriate treatment. For most of its history, medicine has been about individual patients, and about inductive reasoning. Inductive reasoning is when you go from the particular to the general, as in 'all crows I have seen have

Understanding Biostatistics, First Edition. Anders Källén.

Box 1.1 The philosophy of science

What is knowledge about reality and how is it acquired? The first great scholar of nature, Aristotle, divided knowledge into two categories, the original facts (axioms) and the deduced facts. Deduction is done by (deductive) logic in which propositions are derived from one or more premises, following certain rules. It often takes the shape of mathematics. When applied to natural phenomena, the problem are the premises. In a deductive science like mathematics there is a process to identify them, but in empirical sciences their nature is less obvious. So how do we identify them?

Early thinkers promoted the idea of induction. When repeated observations of nature fall into some pattern in the mind of the observer, they are said to induce a suggestion of a more general fact. This idea of induction was raised to an alternative form of logic, inductive logic, which forced a fact from multiple observations, a view which was vigorously criticized by David Hume in the mid-eighteenth century.

Hume's argument started with an analysis of causal relations, which he claimed were found exclusively by induction, never deduction, and contains an implicit assumption that unobserved objects resemble observed ones. The causal connection is by induction, not deduction, and the justification of the inductive process becomes a circular argument, Hume argues. This was referred to as 'Hume's dilemma', something that upset Immanuel Kant so much that he referred to the problem of induction as the 'scandal of philosophy'. This does not mean that if we have always observed something in a particular situation, we should not expect the same to happen next time. It means that it cannot be an absolute fact, and instead we are making a prediction, with some degree of confidence.

Two centuries later Karl Popper introduced refutationism. According to this there are no empirical, absolute facts and science does not rely on induction, but exclusively on deduction. We state working hypotheses about nature, the validity of which we test in experiments. Once refuted, a modified hypothesis is formulated and put to the test. And so on. This infinite cycle of conjecture and refutation is the true nature of science, according to Popper.

As an example, used by Hume, 'No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion'. It was a long-held belief in Europe that all swans were white, until Australia was discovered, and with it *Cygnus atratus*, the black swan.

Inductionism and refutationism both have their counterparts in the philosophy of statistics. In the Bayesian approach to statistics, which is inductive, we start with a summary of what we believe and update that according to experimental results. The frequentist approach, on the other hand, is one of refuting hypothesis. Each case is unique and the data of the particular experiment settle that case alone.

been black, therefore all crows are black'. It is the way we, as individuals, learn about reality when we grow up. However, as a foundation of science, induction has in most cases been replaced by the method of falsification, as discussed in Box 1.1. (It is of course not the case that medicine is exclusively about inductive reasoning: a diagnostic fit may well be put to the test in a process of falsification.)

Another peculiarity of medicine is ethics. Medical researchers are very careful not to put any patients at risk in obtaining the information they seek. This is often a complicating factor in clinical research when it interferes with the research objective of a clinical trial. For example, in drug development, at one important stage we need to show that a particular drug is effective. The scientific way to do this is by carrying out a clinical trial in which the response to the drug is compared to the response when no treatment is given. Everything else should be the same. However, in the presence of other effective drugs, it may not at all be ethical to withhold a useful drug for the sole reason that you want to demonstrate that a new drug is also effective.

Finally, there is the general problem of why it appears to be so hard for many physicians to understand basic statistical reasoning: what conclusions one may draw and why. To be honest, part of the reason why statistics is so hard to understand for non-statisticians is probably that statisticians have not figured it out for themselves. There is not one statistical philosophy that forms the basis for statistical reasoning, there are a number of them: frequentists versus Bayesians, Fisher's approach versus the Neyman–Pearson view. If statisticians cannot figure it out, how can they expect their customers to be able to do so?

These are some properties of medical researchers that statisticians should be aware of. Of course, they are not true statements about individual medics. They are statements about the group of medics, and statements about groups are what statistics is all about. This will be our starting point in Chapter 2 when we initiate a more serious discussion about the design of clinical trials. But before we do that we need to get a basic understanding of what it is statistics is trying to do. This journey will start with an attempt to describe the role of statistics within science.

1.2 On the nature of science

For almost all of the history of mankind the approach to health has been governed by faith, superstition and magic, often expressed as witchcraft. This has gradually changed since the period of the Enlightenment in the eighteenth century, so that doctors can no longer make empty assertions and quacks can no longer sell useless cures with impunity. The factor that has changed this is what we call science.

But what *is* science? We know what it does: it helps us understand and make sense of the world around us. But that does not define science; religion has served much the same purpose for most of mankind's history. Science is often divided into three subsets: natural sciences (the study of natural phenomena), social sciences (the study of human behavior and society), and mathematics (including statistics). The first two of these are empirical sciences, in which knowledge is based on observable phenomena, whereas mathematics is a deductive science in which new knowledge is deduced from previous knowledge. There is also applied science, engineering, which is the application of scientific research to specific human needs. The use of statistics in medical research is an example, as is medicine itself.

The science of mathematics has a specific structure. Starting from a basic set of definitions and assumptions (usually called axioms), theorems are formulated and proved. A theorem constitutes a mathematical statement, and its proof is a logical chain of applications of previously proved theorems. A collection of interlinked, proved, mathematical theorems makes up a mathematical theory of something. The empirical sciences are similar to this in many respects, but differ fundamentally in others. Corresponding to an unproved

mathematical theorem is a hypothesis about nature. The mathematical proof corresponds to an experiment that tests the hypothesis. A theory, in the context of empirical science, consists of a number of not yet refuted hypotheses which are bound together by some common theme.

What we think we know about the world is very much the result of an inductive process, derived from experiences and learning. The difference between science and religion is not about content, but about the way knowledge is obtained. A statement can only be a scientific statement if it can be tested, and science is qualified by the extent to which its predictions are borne out; when a model fails a test it has to be modified. Science is therefore not static, it is dynamic. Old 'truths' are replaced by new 'truths'. It is like an enormous jigsaw puzzle in which pieces are constantly replaced and added. Sometimes replacement is with a set of new pieces that give a clearer picture of the overall puzzle, sometimes a piece turns out to be wrong and needs to be replaced by a new, fundamentally different, one. Sometimes we need to tear up an entire part of the jigsaw puzzle and rebuild it. The basic requirement of the individual pieces in this jigsaw puzzle is that each one addresses a question that can be tested for validity. Science is a humble practice; it tells us that we know nothing unless we have evidence and that our state of knowledge must always be open to scrutiny and challenge.

The fundamental difference between empirical sciences and mathematics is that a mathematical proof proves the hypothesis (i.e., theorem), whereas in empirical sciences experiments are designed to disprove the hypothesis. A particular hypothesis can be refuted by an observation that is inconsistent with the hypothesis. But the hypothesis cannot be proved by experiment – all we can say is that the outcome of the experiment is consistent with it.

Example 1.1 Like most people before modern times, the Greeks thought that the earth was the center of everything. They identified seven moving objects in heaven – five planets, the sun and the moon – and Ptolemy worked out a very elaborate model for how they move, using only circles and circles moving on circles (epicycles). The result was an explanation of the heavens (planets, at least) that fulfilled all the criteria of science. They made predictions that could be tested, and these never failed. When the idea of putting the sun at the center of this system emerged, it was not found to work better in any way; it did not produce better predictions than the Greek model. It was not until Johannes Kepler managed to identify his famous three laws that astronomers actually got a sun-centered description of the heavens that even matched the Greek version. This meant that there were two competing models with no one really ahead.

However, this changed with Isaac Newton. With his law of gravitation the science of the heavens took a gigantic leap forward. In one go, he reduced the complex behavior of the planets to a few fundamental and universal laws. When these laws were applied to the planets they not only predicted their movements to any precision measurable, they also allowed a new planet to be discovered (Neptune, in 1846). So many experiments were conducted over hundreds of years with outcomes consistent with Newton's theory, that it was very tempting to consider it a true fact. However, during the twentieth century some astronomical observations were made that were inconsistent with the mathematical predictions of the theory, and it is today superseded by Albert Einstein's theory of general relativity in cosmology. As a theory though, Newton's theory of gravitation is still good enough to be used for all everyday activities involving gravitation, such as sending people to the moon.

This example illustrates an important point about science which must be kept in mind, namely that 'all models are wrong, but some are useful', a quotation often attributed to the

English statistician George Box. Much of the success of Newton's physics was due to the fact that it was expressed in mathematical terms. As a general rule scientific theory seems to be least controversial when it can be expressed in the form of mathematical relationships. This is partly because this requires a rather well-defined logical foundation to build on, and partly because mathematics provides the logical tool to derive the correct predictions.

That one theory replaces another, sometimes with fundamental effects, is common in biology, not least in medicine. (On my bookshelf there are three books on immunology, published in 1976, 1994 and 2006, respectively. It is hard to see that they are about the same science. On the other hand, there is also a course in basic physics from 1950, which could serve well as present-day teaching material – in terms of content, if not style.) We must always consider a theory to be no more than a set of hypotheses that have not yet been falsified. In fact, mathematics also has an element of this, since a theorem that has been proved has been so only to the extent that no one has yet found a fault in the proof. There are quite a few examples of mathematical theorems that have been held to be true for a period of time until someone found a mistake in their proofs.

1.3 How the scientific method uses statistics

To produce objective knowledge is difficult, since our intuition has a tendency to see patterns where there is only random noise and to see causal relationships where there are none. When looking for evidence we also have a tendency, as a species, to overvalue information that confirms our hypothesis, and we seek out such confirmatory information. When we encounter new evidence, the quality of it is often assessed against the background of our working assumption, or prior belief, leading to bias in interpretation (and scientific disputes).

To overcome these human shortcomings the so-called scientific method evolved. This is a method which helps us obtain and assess knowledge from data in an objective way. The scientific method seeks to explain nature in a reproducible way, and to use these explanations to make useful predictions. It can be crudely described in the following steps:

1. Formulate a hypothesis.

2. Design and execute an experiment which tests the hypothesis.

3. Based on the outcome of the experiment, determine if we should reject the hypothesis.

To gain acceptance for one's conclusion it is critical that all the details of the research are made available for others to judge their validity, so-called *peer review*. Not only the results, but also the experimental setup and the data that drive the experimenter to his conclusions. If such details are not provided, others cannot judge to what extent they would agree with the conclusions, and it is not possible to independently repeat the experiment. As the physicist Richard Feynman wrote in a famous essay, condemning what he called 'cargo cult science',

> if you are doing an experiment, you should report everything that you think might make it invalid – not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked – to make sure that the other fellow can tell if they have been eliminated.

A key part of the scientific method is the design, execution and analysis of an experiment that tests the hypothesis. This may employ mathematical modeling in some way, as when one uses statistical methods. The first step in making a mathematical model related to the hypothesis is to quantify some entities that make it possible to do calculations on numbers. These quantities must reflect the hypothesis under investigation, because it is the analysis of them that will provide us with a conclusion. We call a quantity that is to be analyzed in an experiment an *outcome measure*, because it is a quantitative measure of the outcome of the experiment. After having decided on the outcome measure, we design our experiment so that we obtain appropriate data. The statistical analysis subsequently performed provides us with what is essentially only a summary presentation of the data, in a form that is appropriate to draw conclusions from.

So, for a hypothesis that is going to be tested by invoking statistics, the scientific method can be expanded into the following steps:

1. Formulate a hypothesis.
2. Define an outcome measure and reformulate the hypothesis in terms of it. This involves defining a statistical model for the data. This version of the hypothesis is called the *null hypothesis* and is formulated so that it describes what we want to reject.
3. Design and perform an experiment which collects data on this outcome measure.
4. Compute statistical summaries of the data.
5. Draw the appropriate conclusion from the statistical summaries.

When the results are written up as a publication, this should contain an appropriate description of the statistical methods used. Otherwise it may be impossible for peers to judge the validity of the conclusions reached.

The statistical part of the experiment starts with the data and a model for what those data represent. From there onwards it is like a machine that produces a set of summaries of the data that should be helpful in interpreting the outcome of the experiment. For confirmatory purposes, rightly or wrongly, the summary statistic most used is the p-value. It is one particular transformation of the data, with a particular interpretation under the model assumption and the null hypothesis. It measures the probability of the result we observed, or a more extreme one, given that the null hypothesis is true. Thus a p-value is an indirect measure of evidence against the null hypothesis, such that the smaller the value, the greater the evidence. (Often more than one model can be applied to any given set of data so we can derive different p-values for a given hypothesis and set of data – as in the case of parametric versus non-parametric tests.)

Note that, as a consequence of the discussion above, the conclusion from the experiment is either that we consider ourself as having proved the null hypothesis wrong, or we have failed to prove it wrong. Never is the null hypothesis proved to be true. To understand why, look at the hypothesis ‘there are no fish in this lake’ which we may want to test by going fishing. There are two possible outcomes of this test: either you get a fish or you do not. If you catch a fish you know there is (or was) fish in the lake and have disproved the hypothesis. If you do not get any fish, this does not prove anything: it may be because there were no fish in the lake, or it may be because you were unlucky. If you had fished for longer, you may

have had a catch and therefore rejected the null hypothesis. There is a saying that captures this and is worth keeping in mind: 'Absence of proof is not proof of absence.' Failure to reject a hypothesis does not prove anything, but it may, depending on the nature and quality of the experiment, increase one's confidence in the validity of the null hypothesis – that it to some degree reflects the truth. As such it may be part of a theory of nature, which is held true until data emerge that disprove it.

Failure to understand the difference between not being able to provide enough evidence to reject the null hypothesis and providing evidence for the null hypothesis is at the root of the most important misuse of statistics in medical research.

Example 1.2 In the report of a study on depression with three treatments – no treatment (placebo), a standard treatment, B, and a new treatment, A – the authors made the following claim: 'A is efficacious in depression and the effect occurs earlier than for B.' The data underlying the second part of this claim refer to comparisons of A and B individually versus placebo, using data obtained after one week. For A, the corresponding p-value was 0.023, whereas for B it was 0.16. Thus, the argument went, A was 'statistically significant', whereas B was not, so A must be better than B.

This is, however, a flawed argument. To make claims about the relative merits of A and B, these must be directly compared. In this case a crude analysis of the data tells us what the result should be. In fact, the first p-value was a result of a mean difference (versus placebo) of 1.27 with a standard error of 0.56, whereas the second p-value comes from a mean difference of 0.79 with the same standard error. The mean difference between A and B is therefore 0.48, and since we should probably have about the same standard error as above, this gives a p-value of about 0.40, which is far from evidence for a difference.

The mistake made in this example is a recurrent one in medical research. It occurs when a statistical test, accompanied by its declaration of 'significant' or 'not significant', is used to force a decision on the truth or not of the null hypothesis.

1.4 Finding an outcome variable to assess your hypothesis

The first step in the expanded version of the scientific method, to reformulate the hypothesis in terms of a specific outcome variable, may be simple, but need not to be. It is simple if your hypothesis is already formulated in terms of it, as when we want to claim that women on the average are shorter than men. The outcome variable then is individual height. It is more difficult if we want to prove that a certain drug improves asthma in patients with that disease. What do we mean by improvement in asthma? Improvement in the lung function? Fewer asthma symptoms? There are many ways we can assess improvement in asthma, and we need to be more specific so that we know what data to collect for the analysis. Assume that we want to focus on lung function. There are also many ways in which we can measure lung function: the simplest would be to ask the patients for a subjective assessment of their lung function, though usually more objective measures are used.

Suppose that we settle for one particular objective lung function measurement, the forced expiratory volume in one second, FEV_1. We may want to prove that a new drug improves the patient's asthma by formulating the null hypothesis to read that the drug does not affect

FEV_1. If we subsequently carry out an experiment and from the analysis of it conclude that there is an improvement in lung function as measured by FEV_1, we have disproved the null hypothesis.

The question is what we have proved. The statistical result relates to FEV_1. How much can we generalize from this and actually claim that the asthma has been improved? This is a non-trivial issue and one which must be addressed when we decide on which outcome measure to use to reflect our original hypothesis.

Quality of life is measured by having patients fill in a particular questionnaire with a list of questions. The end result we want from the analysis of such a questionnaire is a simple statement: 'The quality of life of the patients is improved'. In order to achieve that, the scores on individual questions in the questionnaire are typically reduced to a summary number, which is the outcome variable for the statistical analysis. The result may be that there is an increase in this outcome variable when the treatment is given. However, the term 'quality of life' has a meaning to most people, and the question is whether an increase in the summary variable corresponds to an increase in the quality of life of the patients, as perceived by the patients. This question necessitates an independent process, in which it is shown that an increase in the derived outcome variable can in fact be interpreted as an improvement of quality of life – a validation of the questionnaire.

The IQ test constitutes a well-known example. IQ is measured as the result of specific IQ tests. If we show that two groups have different outcomes on IQ tests, can we then deduce that one group is more intelligent than the other group? It depends on what we mean by intelligence. If we mean precisely what the IQ test measures, the answer is yes. If we have an independent opinion of what intelligence should mean, we first have to validate that this is captured correctly by the IQ test.

Returning to the measurement of FEV_1, for a claim of improvement in asthma, lung function is such an important aspect of asthma that it is reasonable to say that improved lung function means that the asthma has improved (though many would require additional support from data that measure asthma symptoms). However, if we fail to show an effect of FEV_1 it does not follow by logical necessity that no other aspect of the asthma has improved. So we deliberately choose one aspect of the disease to gamble on, and if we win we have succeeded. If we fail, we may not be any wiser.

1.5 How we draw medical conclusions from statistical results

Before we actually come to the subject of this section we need to consider the ultimate purpose of science, which is to make predictions about the future. What we see in a particular study is an observation. What we want from the study is more than that: we want statements that are helpful when we need to make decisions in the future. We want to use the study to predict what will be seen in a new, similar study. It is an observation that in a particular study 60% of males, but only 40% of females, responded to a treatment. Unless your sample is very large it is not reasonable to generalize this to a claim that 60% of males and 40% of females will respond to the drug in the target population. It may be the best predictor we have at this point in time, but that is not the same thing. What we actually can claim depends on the statistical summary of the data. A more cautious claim may be that in general males respond better to the treatment than females. To substantiate this claim we analyze the data under the null hypothesis that there is no difference in the response rates for males and females.

Suppose next that we want to show that some intervention prolongs life after a cancer diagnosis. Our null hypothesis is that it does not. We assume that we have conducted an appropriate experiment (clinical trial) and that the statistical analysis provides us with $p = 0.015$. This means that, if there is no effect at all of the intervention, a result as extreme as that found in the experiment is so unlikely that it should occur in only 1.5% of all such clinical trials. This is our confidence in the null hypothesis (not to be confused with the probability of the null hypothesis) after we have performed the experiment.

That does not prove that the intervention is effective. No statistical analysis proves that something is effective. The proper question is: does this p-value provide sufficient support to justify our starting to act as if it is effective? The answer to that question depends on what confidence is required from this particular experiment for a particular action. What are the consequences if I decide that it is effective? A few possibilities are:

- I get a license for a new drug, and can earn a lot of money;
- I get a paper published;
- I want to take this drug myself, since I have been diagnosed with the cancer in question.

In the first case it is really not for me to decide what confidence level is required. It is the licensing authority that needs to be assured. Their problem is on the one hand that they want new, effective drugs on the market, but on the other hand that they do not want useless drugs there. Since all statistics come with an uncertainty, their problem is one of error control. They must make a decision that safeguards the general public from useless drugs, but at the same time they must not make it impossible to get new drugs licensed. This is a balancing act, and they do it by setting a significance level α such that if your p-value is smaller than α, they agree that the drug is proved to be effective. The significance level defines the proportion of truly useless drugs that will accidentally be approved and therefore the level of risk the licensing agency is prepared to take (if we include almost useless drugs as well, the proportion is higher). Presently one may infer that the US licensing authority, the Food and Drug Administration (FDA), has set the significance level at $0.025^2 = 0.000625$ when it comes to proving efficacy for their market, for reasons we will come back to.

The picture is similar if you want to publish a paper. In general there is an agreed significance level of 5% (two-sided) for that process. If your p-value is less than 5% you can publish a paper and claim that the intervention works. But that does not prove that the intervention works, only that you can get a paper published that claims so. The significance level used by a particular journal is typically not explicitly spelt out, since a remark by the eminent statistician R.A. Fisher led to the introduction of the golden threshold at 5% a long time ago (see Box 1.2), making it unnecessary to argue about it. That is really its only virtue – there is no scientific reason why it should not be 6% or 0.1%. In relation to this particular threshold we now also have some jargon, the term ‘statistical significance’ , which is discussed in some detail in Box 1.3.

In the last situation in the bullet list above, the case where you had that particular cancer yourself, you really decide your own significance level. It may be very high, depending on how desperate you are. A significance level of 20% may be good enough for you. It may depend on side-effects and alternative options.

A situation where the interpretation of the p-value as a measure of confidence and its relation to what to do next becomes apparent, is in drug development. Clinical drug development

Box 1.2 The origin of the 5% rule

The 5% significance rule seems to be a consequence of the following passage in the book *Statistical Methods for Research Workers* by the inventor of the p-value, Ronald Aylmer Fisher:

> in practice we do not always want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. ... A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.

It is important that in Fisher's view a p-value below 0.05 does not force a decision, it only warrants a further investigation. Larger p-values are not worth investigating (note that he does not actually say anything about values between 0.05 and 0.1). On another occasion he wrote:

> This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained.

Nowadays we use the 5% rule in a different way. We use it to force decisions in single studies, referring to an error-rate control mechanism on the ensemble of studies, following a philosophy introduced by Jerzy Neumann and Egon Pearson (see Box 1.3).

is a staged process in which we sequentially try to answer more and more complex questions such as:

- Is the drug effective at all?
- What is the appropriate dose for this drug?
- Is the appropriate dose effective enough to get the drug licensed?

The monetary investment that needs to be made in order to answer these questions is usually very different. Moreover, the more confidence we want to have in the answer to a particular question, the more money it costs to get that confidence, because larger studies need to be performed. The decision on what confidence we need that a drug is effective at all before conducting a dose-finding study, could then depend on the cost of the latter. Or, rather, a balance between that cost and the loss in time to market, which in itself is a cost. The bottom line is that it may be strategically right for a pharmaceutical company to do a small study which only can produce limited confidence in efficacy, say a one-sided p-value at 10%, before gambling with a larger dose-range study, in order to save time.

In view of the present avalanche of statistical p-values pouring over us – by one estimate some 15 million medical articles have been published to date, with 5000 journals around the

Box 1.3 The meaning of the term 'statistical significance'

There are two alternative ways of looking at p-values and significance levels which are related to the philosophy of science. Here is a brief outline of these positions.

The *p*-value builds confidence. R.A. Fisher originally used p-values purely as a measure of inductive evidence against the null hypothesis. Once the experiment is done there is only one hypothesis, the null, and the p-value measures our confidence in it. There is no need for the significance level; all we need to do is to use the p-value as a measure of our confidence that it is correct to reject the null hypothesis. By presenting the p-value we allow any readers of our results to judge for themselves whether the test has provided enough confidence in the conclusion.

The significance level defines a decision rule. The Neyman–Pearson school instead emphasizes statistical hypothesis testing as a mechanism for making decisions and guiding behavior. To work properly this setup requires two hypotheses to choose between, so the Neyman–Pearson school introduces an alternative hypothesis, in addition to the null hypothesis. A decision between these is then forced, using the test and a predefined significance level α. The alternative is accepted if $p < \alpha$, otherwise the null hypothesis is accepted. Neyman–Pearson statistical testing is aimed at error minimization, and is not concerned with gathering evidence. Furthermore, this error minimization is of the long-run variety, which means that, unlike Fisher's approach, Neyman–Pearson theory does not apply to an individual study.

In a pure Neyman–Pearson decision approach the exact p-value is irrelevant, and should not be reported at all. When formulated as 'reject the null hypothesis when $p < \alpha$, accept it otherwise', only the Neyman–Pearson claim of $100\alpha\%$ false rejections of the null hypothesis with ongoing sampling is valid. This is because α is the probability of a set of potential outcomes that may fall anywhere in the tail area of the distribution of the null hypothesis, and we cannot know ahead of time which of these particular outcomes will occur. That is not the same as the tail area that defines the p-value, which is known only after the outcome is observed.

This dualism between Fisher's inductive approach to p-values and the error control of Neyman and Pearson is really about what p-values imply, not what they are. For Fisher it is about inductive learning, for Neyman and Pearson it is about decision making. For Fisher, the Neyman–Pearson view is not relevant to science, since one does not repeat the same experiment over and over again. What researchers actually do is one experiment, from which they should communicate information, not force a yes–no decision.

world constantly adding to that number – a strict adherence to a rule such as 'if $p < 5\%$ I can say I have an effect, otherwise not', is a bit primitive, to say the least. Assume (probably incorrectly) that all statistical analyses done are done in a correct manner. Then 5% of all cases investigated where there is no true effect or association, are out there as false effect or relationship claims. We cannot, using statistics, guarantee that there are no false 'truths' in circulation, and this level may be appropriate. But most hypotheses tested are part of a bigger context, a theory. If the result we present is a trivial modification of, or an add-on to, what is already known, we may need less assurance than if the result may set an earthquake in

motion and have a major impact on society. Ultimately the judgement about the correctness of the null hypothesis will depend on the existence of other data and the relative plausibility of the alternatives.

In fact, in a medical context it is probably a good idea to be a little relaxed about the first ground-breaking result. Let it be reproduced before you actually believe it. This only means that you work with a lower significance level when you draw your conclusion from such results, whereas for reports that more or less only confirm previous reports you may work on a higher significance level. In essence this means that you take a more inductive evidence approach in your use of p-values, as compared to a strict decision-theoretic one (see Box 1.3).

The very low significance level the FDA have set for proving efficacy, referred to earlier, is an example of this. In order to prove efficacy in the eyes of FDA you need to do so in two independent studies, each with a (two-sided) test at 5%. Since licensing efficacy only goes in one direction, this means that their significance level within a particular study is half of this, 2.5%, and they will only accept that the drug is an effective treatment if both studies succeed. That a treatment with no effect whatsoever should pass this hurdle then occurs with a probability as low as 0.000625. (Actually, this is a debatable point, because there is some lack of clarity about how many unsuccessful related studies are allowed. The presence of such studies obviously impacts on this probability calculation.)

The discussion in this section, about the separation between statistical results and the conclusion to be drawn, seems not to be clear to many statisticians in the pharmaceutical industry or health authorities. How else can we explain the rise in the late 1990s of the non-inferiority trial? This – to my mind peculiar – concept is discussed in Box 1.4. The mistake made with the non-inferiority trial concept is precisely a confusion about the relation between a statistical result, in this case the confidence interval for a particular parameter, and the conclusion we draw from that result. As discussed above, any conclusion should be drawn in a particular context. One such context can be that a health authority allows a particular result to mean that efficacy is demonstrated beyond any reasonable doubt, and grants you a license to sell the drug. In another situation the result may be part of a decision to switch standard treatment at a particular hospital. In a third example it may provide sufficient evidence to test the new drug on a particular patient. Each of these situations calls for a decision, and for each decision we need a standard of proof. Once that is decided, the action should be taken without reference to a statement like 'A is not inferior to B', only to the actual result, the confidence interval. The problem, in a nutshell, is that one tries to build the whole decision process into the study, so that the study result forces a definite decision, instead of viewing the result of the study as a step in this process.

It is somewhat ironic that the non-inferiority study was modeled on so-called bioequivalence studies. A bioequivalence study is a particular type of pharmacokinetic study which drug makers run when they want to change some aspect of how a tablet is manufactured. Such studies follow rather precise rules in terms of how they should be analyzed: the 90% confidence interval of a particular mean ratio should lie between 0.8 and 1.25. If that is the case, the new formulation can replace the old one. The key difference between this type of study and the non-inferiority study is that for the bioequivalency study the result has a very specific follow-up action: you can switch to the new formulation. The bioequivalency result in itself is of no independent interest.

Box 1.4 The non-inferiority trial

The non-inferiority trial originally addressed the following specific problem. In order to prove efficacy, we need to prove that the new drug is better than taking no treatment. However, in many disease areas giving no treatment may be unethical; cancer treatments for which there are available alternative and established treatments may serve as an example. One way to approach this would be to take the new drug, A, and compare it with a standard treatment, B, which we agree is effective. If the difference in response between A and B is not too large, the argument goes, then A must also be effective. Such a trial was called a non-inferiority trial and its logic went like this: prespecify how much inferior A can be to B without casting doubts on A being effective. If our study achieves this objective we can claim that A is effective.

Unfortunately, that is not exactly true. Instead of using the argument to claim that A is effective, one claims that A is not inferior (in efficacy) to B. So the result becomes a statement about the relative merits of A and B, instead of the original intent to use B as a tool to declare A effective. The criterion that is typically used is that a confidence interval of a mean difference must stay within certain bounds. The study designers construct those limits, and the study logic dictates that if they succeed in getting the confidence interval within those limits, they are allowed to draw the conclusion that A is not inferior to B.

The problem here is that everyone needs to agree that the prespecified limits imply that A is not inferior to B. If the limits are widely agreed, there is no need to prespecify them – they would be universally accepted anyway. If they are not, it may be that the conclusion differs depending on its consequences. For some purposes it may be good enough, for others it may not.

Apart from the logical problem, there is an executional problem that is as important: how do we know that the trial could have picked up a difference? This is referred to as *assay sensitivity* and is a distinguishing feature between this type of trial and the superiority trial. If there is no assay sensitivity in a superiority trial, the trial will be unsuccessful, whereas for a non-inferiority trial it may be successful. This means that with a non-inferiority trial we also need to provide evidence that this particular trial was sufficiently sensitive; that the control behaved also in this trial as it had done in previous trials where it had shown efficacy. This is very much the same as referring to historical controls.

1.6 A few words about probabilities

Before we proceed we need to say a few words about probabilities. To set the scene, consider the following example.

Example 1.3 You meet a woman in the street who you know has two children, one of whom is a boy playing in your son's soccer team. What is the probability that her other child is a girl? The chances are that you will say 50%. The argument is deductive: there are two choices,

a boy or a girl, and there are the same number of boys and girls in the community. Is this a correct way of arguing?

The answer is no. The probability required should refer to an empirical statement: out of all two-children families with *at least* one boy, in what percentage is the other one a girl? With the appropriate model assumption, such as that a child in any family has the same probability of being a boy as being a girl, we can design an experiment to test the claim. Take two unbiased coins (with each coin representing a child so that heads (H) corresponds to a girl and tails (T) to a boy) and toss them, say, 100 times. Each time there is at least one H, note on paper if the other is a T. Out of your 100 experiments there will be some, say N, with at least one H, and out of these in a certain number of cases, say n, the other is a T. The number n/N is then an estimate of the probability that the other child is a girl. If you do this experiment, you will probably end up with a number closer to 2/3 than to 1/2. In fact if you do it on a computer instead, using a random number generator, with a very large number of experiments, you will get rather close to 2/3.

So you are advised to reject your hypothesis that the probability is 1/2. We will discuss why in a short while.

The type of probabilities we discuss here are relative frequencies, not observed relative frequencies but theoretical ones – entities that in principle can be estimated by observed frequencies. The concept of probability is actually non-trivial, and we will return to it at the end of this chapter. For now we assume that it is simple to define.

Probabilities are computed for events. If we denote an event (like that the other child is a girl) by A, we denote the probability that it occurs in a particular experiment by $P(A)$. In the previous example this is 2/3, which is the frequency if we do the experiment an infinite number of times. If A denotes an event, we denote by A^c the complement of that event (i.e., that it does not occur), and $P(A^c)$ is then the probability that A does not occur. It is computed as $P(A^c) = 1 - P(A)$, since we are dealing with relative frequencies.

Example 1.4 You are participating in a game show, in which the host has placed a car behind one of three doors and a goat behind each of the other two doors. The game host instructs you to choose one door by pointing at it. When you have done so, he opens one of the other two doors to reveal a goat. After you have seen that goat, you are given the opportunity to switch doors. You win whatever is behind the door you select.

The problem is simple: should you switch doors, or does it matter at all? The chances are that you think it does not matter. You have two doors to choose between, so there should be a 50% chance to find the car behind whichever door you selected first. Actually the probability is only 1/3 that it is behind the door you selected first, so the correct strategy is to switch doors. This particular problem is called the Monty Hall problem, and some of its history can be found in Box 1.5.

We now have two examples of what may well be counterintuitive probabilities. Intuition is perhaps nothing but a reflection of personal experience, and the reason why these examples appear counterintuitive may be a lack of the appropriate experience. In the first example we have that a family with precisely two children has one of the following structures: (B, B), (B, G), (G, B), (G, G), where B denotes boy, G denotes girl and the pair is written as (oldest, youngest). Moreover, if boys and girls are equally likely, we have the same number of

Box 1.5 The Monty Hall problem

The game discussed in Example 1.4 appeared in the 1990s in TV shows all over the world, and was loosely based on an American game show called *Let's Make a Deal*, hosted by Monty Hall. This game show epidemic had its origin in a letter to the column *Ask Marilyn* in the American journal *Parade* in February 1990. The columnist, Marilyn vos Savant, received the following question:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then asks you 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice?

Marilyn offered the correct solution, thereby provoking a debate involving some 10 000 readers, 92% of whom, including (legend has it) several hundred mathematics professors, said she was wrong. In fact, many harsh statements about the level of education in the country were made.

The original game show was, however, fundamentally different: Monty Hall did not let the participant switch door. The door was opened only to build excitement.

these different family types, so each of them constitute 25% of all families. It was part of the conditions of the problem that there was one boy in the family, but no more information than that. That means that the structure of the family in question is one of (B, B), (B, G), (G, B), and each of these have the same probability. In two of these we have a girl, so the probability that the other child is a girl is 2/3. There is an important subtle point here: if we instead know that *the oldest child* is a boy, there is 50% chance the other child is a girl. So the assumption must be spelt out in detail.

Before we leave this, let us repeat the discussion in a slightly different way. Let A be the event that a randomly chosen child from a two-child family is a girl, and let C be the event that the child chosen comes from a family with at least one boy. We then have that $P(A) = 1/2$ and $P(C) = 3/4$, and the probability we are interested in is the conditional probability that A occurs when we know that C has occurred, a probability we denote by $P(A|C)$. This is the frequency of A events among the C events, for which we have

$$P(A|C) = \frac{P(AC)}{P(C)} = \frac{1/2}{3/4} = \frac{2}{3}.$$

Here AC is the event that both A and C occur (i.e., the event that we have one boy and one girl in the family), an event which has probability 1/2. (The formula above, written as $P(AC) = P(A|C)P(C)$, implies a very basic probabilistic statement called Bayes' theorem, which relates the transposed conditional probabilities $P(A|C)$ and $P(C|A)$ to each other. To derive it we utilize the symmetry $P(AC) = P(CA)$; see Box 4.2.)

What about the Monty Hall problem? The situation at the start of the game can be described as one of the triplets (C, G, G), (G, C, G) and (G, G, C). Here the position denotes a particular door; G denotes a goat and C denotes the car. Each of these are equally likely,

and therefore each has probability 1/3. This means that the probability is 1/3 that you picked the correct door from the start, and therefore 2/3 that you did not. Since it is more likely you picked the wrong door, you should switch if you are given the opportunity. Of course this may not win you the car in an individual game. But if you play it many times, with this strategy you will win it in about 67% of the games, as opposed to only in 33% if your strategy is not to switch door.

The reason why this was initially considered counterintuitive is that we often assume that if we have n choices, each choice has probability $1/n$ of being the correct one. But we may have information that invalidates this, just as picking a horse to bet on at random at the trot is a worse strategy than getting some knowledge about the fitness and qualities of the different horses before you make your bet.

There is one fundamental difference between the two examples we discussed above, both of which gave the probability 2/3. In the game show, if the rules are adhered to, our argument provides a correct probability and therefore the correct game strategy. In the example with the children, however, there are assumptions that are made in the computations that may not hold true in real life. The assumption is that the three pairs (B, G), (G, B) and (B, B) all occur with the same frequency in the relevant population. This may not be true, not only because the ratio boys/girls may not be precisely one, but also because family planning strategies may lead to unequal probabilities for the different pairs. So what we have in this case is not necessarily a true description of the world, only a model of it.

The reason for bringing up these examples is to point out how important it is that you understand the context in which you compute probabilities. Statistics is about probabilities, and ignorance around the context can not only produce bias in the results, but also lead to misleading or erroneous p-values. It may well be that a conditional probability that is involved, but which one may be less apparent. When probabilities are computed by the uninformed, disaster may strike, as the sad case of Sally Clark, outlined in Box 4.3, shows. Another interesting, but not disastrous, example might be the discovery of the basic genetic laws (see Box 1.6).

1.7 The need for honesty: the multiplicity issue

The story about p-values may appear rather simple: we start with a hypothesis, collect data and compute the p-value. However, there are a few important assumptions in this process that need to be understood in order for the analysis to provide credible conclusions. The key assumption is that you compute one p-value and that you have clearly identified *a priori* when and how you do that. This is because it is important to make sure that your choice is not data-driven. The reason for this can be summarized in the following sentence: 'The value of the p-value is influenced by the history behind its computation.' This section and the next will illustrate the importance of bearing this in mind.

The particular issue to be discussed here is called the multiplicity problem. Recall that when the p-value is below a certain, prespecified, significance level, we reject the null hypothesis. The multiplicity problem refers to the simultaneous application of this rule to a set of null hypotheses. It is one of the problems that many medical workers consider an unnecessary complication invented by statisticians in order to make it more difficult for physicians to draw the 'appropriate' conclusions.

Box 1.6 Did Mendel cheat?

In one of his experiments, the monk Gregor Mendel, the father of genetics, crossed two species of pea which, when cultivated, had shown themselves to be constant in color. One species was red, the other was white. The locus for color had two alleles: A for red and a for white, of which A is dominant (so that both AA and Aa become red and only aa white). In one experiment Mendel had 600 red colored peas in what is called an F2-generation, which means that the proportion of homozygotes (genotype AA) should be 1/3 (Aa is twice as common as AA). Thus Mendel expected 200 homozygotes, and counted to 201. A very good result!

Or was it? How did Mendel determine that a particular red pea is a homozygote? His method was to investigate the color of 10 offspring, obtained by self-fertilization. If all were red, he declared the parent to be a homozygote, otherwise to be a heterozygote (genotype Aa). The problem with this decision rule is that by chance alone a heterozygote pea can produce 10 red offspring! In fact, the probability for this is $(3/4)^{10}$, so the total probability of declaring a particular pea a homozygote (call that event B) is

$$P(B) = P(B|AA)P(AA) + P(B|Aa)P(Aa) = 1 \cdot \frac{1}{3} + \left(\frac{3}{4}\right)^{10} \cdot \frac{2}{3} = 0.371.$$

This means that in 600 red colored peas we expect, using Mendel's method, to declare 222.6 to be homozygotes, including 22.6 misclassifications. To obtain 201 is therefore rather unlikely!

However, this is not really a statistical problem. The same problem occurs in medicine, for example with screening activities. Consider the situation where a physician is carrying out a routine health check-up on a patient. As a part of this he takes a 'lab status': he draws blood which he sends to a laboratory. In return he gets measurements of a number of chemicals in various blood compartments. In order to assess the clinical implications of these numbers, to understand their relation to health, the laboratory also provides reference ranges for each of the measurements. These reference ranges define (we assume) an interval within which 95% of measurements from healthy individuals will fall.

Suppose we have on the list N different values. What is the probability that a healthy individual will be considered healthy after the physician has read through the list of test results? In other words, what is the probability that all N measurements will lie within their reference limits? We simplify the discussion by making the unrealistic assumption that these N measurements are independent of each other. (Two events are independent if the probability of both occurring equals the product of the individual probabilities.) Because of this assumption, the probability is 0.95^N that all values lie within their respective reference limits for a healthy individual. This means that the probability of at least one value lying outside its normal reference limit is given by

$$P_N(\alpha) = 1 - (1 - \alpha)^N, \quad \text{where } \alpha = 0.05.$$

This function $P_N(\alpha)$ is plotted in Figure 1.1 for a few choices of N.

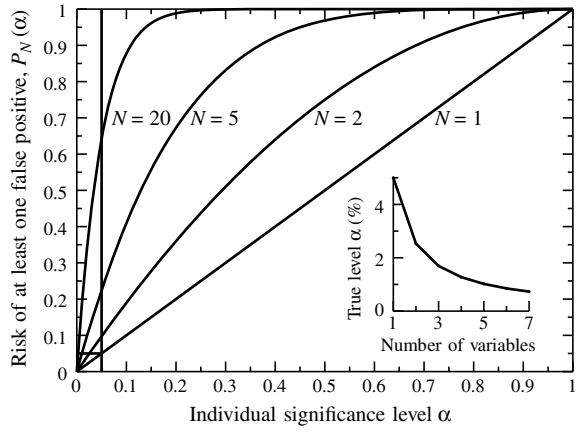

Figure 1.1 The graph of $P_N(\alpha)$ for a few values of N. In the small inset graph we see the significance level at which we need to do individual tests, in order to preserve the overall significance level at 5%.

If we declare a person healthy precisely when all values fall within their respective reference range, and we do this in a nationwide screening program, the number $P_N(0.05)$ will provide us with the percentage of healthy subjects who will wrongfully be found to be not healthy. This number is given by the intersection between the vertical line in Figure 1.1 and the curve describing the relevant function $P_N(\alpha)$. If the whole population is healthy, this is the fraction that will be declared sick. This number grows fast with N, which explains why we should not adhere strictly to a decision rule like this. When used for general health test purposes, the interpretation of data must be made with much more common sense. One looks for patterns, or uses the observation for follow-up testing to confirm or reject a hypothesis generated by the screening data.

Multiple p-values work in exactly the same way. In fact, it is more than an analogy, it is the same math. The number $P_N(0.05)$ gives us the probability that we have at least one statistically significant test when performing N independent tests at the significance level $\alpha = 0.05$. This is therefore the true significance level for the procedure, if you make individual statistical tests at the 5% level. It follows that if we still want to do this panel of N independent tests, but that our overall risk of being wrong must not exceed α, we must find the number x that solves the equation $\alpha = 1 - (1 - x)^N$, namely $x = 1 - (1 - \alpha)^{1/N}$, which is approximately α/N. Graphically this is the same as finding the α-value that corresponds to the intersection of the curve with the line $y = 0.05$. These values are illustrated in the small picture in Figure 1.1. In particular, if we make two experiments we need to compare the p-value to the 2.5% level in order to draw our conclusion at the overall significance level 5%.

This correction (which does not require independence to hold true) is called the Bonferroni correction and is based on the assumption that we should compare all our tests to a common significance level. It can be improved upon by distributing the risk α at our disposal unevenly among the tests.

A natural follow-up question to the multiplicity problem is to ask how many tests one can make. This is addressed in Box 1.7.

Box 1.7 How many tests can we do?

A natural follow-up question to the discussion on multiplicity is how many tests we can do. There are two extreme answers. Either you can say that each test controls its own error rate, and that multiple testing therefore is not a problem. Alternatively, you can argue that the multiplicity issue is there as soon as more than one test, world-wide, has been done. So only one test is allowed, and that was done ages ago.

Ultimately, this is another instance of using p-values to guide behavior. What matters is what action the result triggers. If we use individual p-values in the inductive way of mainly measuring how extreme the signal–noise ratio is, we do not need to adjust significance level, because we do not really use it. When used in this way, p-values are sometimes called exploratory p-values.

The need to adjust for multiplicity arises when we make a claim. It is here the whole history must be accounted for in the computation. It is the validity of the claim that must be addressed based on data. This means that we somehow need to restrict how many claims we make per study, but statistics does not define the rules that should govern this.

A standard way of settling the history problem is to prespecify what tests to do and how the significance level is kept under control. Though a tool that solves one aspect of the problem, it must be exercised with care. It is not legitimate to specify 200 hypotheses for a study and then report the successful ones as having been 'prespecified'. To mitigate this problem various multiplicity control procedures have been invented in order to control the overall error rate. Though statistically sound, they are sometimes hard to understand in a non-biostatistical world. Effectively they mean that if we walk through the different p-values in a prespecified way, we are allowed to make a claim from the final p-value, but if we walk through the same p-values in a different, not prespecified way, that claim is not valid. The information about the particular variable in question is fixed, it is how well we predicted the outcome of other variables that defines the validity of the claim. Personally I have full sympathy with those who find it hard to understand how the appropriate medical action can hinge on what some statistician happened to prespecify.

1.8 Prespecification and p-value history

The multiplicity problem that we discussed in the previous section is one example that illustrates the need for a p-value to capture the full story. In the multiplicity case we need to say that we did all N tests in order to find one that is statistically significant. If we report only the significant p-values and omit to mention the others, we are cheating. The following example is related to this.

Example 1.5 It was noted at a particular workplace that there was an unusually high frequency of children with severe malformations born to women who had worked there during their pregnancy. When this observation was made, workforce records over the previous 5 years were collected, which showed that during that period 50 pregnant women had been employed there and as many as 5 of their babies were born with a severe malformation.

How likely is this result? It is known that in the general population about 2% of children are born with a malformation of (at least) such severity. We can then compute the probability of observing 5 or more such malformations among the 50 pregnancies. We simply have to list all possibilities. If we denote a malformed child by M and a healthy one by H, we list all sequences of Ms and Hs of length 50. A particular sequence with precisely k Ms has probability $p^k(1-p)^{n-k}$, where n = 50 and p = 0.02. If we therefore sum the probabilities for all such sequences, we find that the probability of getting precisely k malformed babies is given by

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

Here $\binom{n}{k}$, which is called a binomial coefficient, is the number of ways we can pick a set of k elements from a bigger set of n elements. The distribution given by these probabilities is called the binomial distribution and denoted by Bin(n, p). From this we can compute the probability of getting an observation that is at least 5 for a Bin(50, 0.02) distribution. It turns out to be very unlikely; it is 0.0032. Can we from this conclude that something in the environment is harmful to pregnancies?

The validity of such a conclusion depends on how the workplace that was investigated was chosen. Did we chose it at random, or did we first note the unusually high number and then start the investigation? In the former case, we are correct in drawing the conclusion above. However, as the situation is described it is most likely we have the second case. Also rare events occur by chance, and if we look for one and then use the fact that what we have found is rare to prove something, we are making a circular argument. In fact, for a Bin(50, 0.02) distribution, we have the following probabilities for different outcomes that can occur:

Outcome	0	1	2	3	4	≥ 5
Probability	0.364	0.372	0.186	0.0607	0.0145	0.00321

From this it follows that if we, for example, have 300 workplaces at which 50 women employees worked during their pregnancies, we expect 109 of these to have no malformed babies, 111 to have exactly one, 56, 18, 4 to have respectively 2, 3 and 4 and 1 to have at least 5 malformed babies, respectively. So, even though it is a rare event, we do expect this rare event to occur in about one in 300 such workplaces.

The key point here is that we need to carefully plan the experiment in order to guarantee that the statistical model we use is appropriate, so that we can draw valid conclusions. If, as in the case above, we observe a rare event, we must put this observation into proper perspective. By definition rare events do occur, though rarely, and we must ask ourselves why we came to observe this rare event – what mechanism underlies its detection? Would we react in the same way if it was another rare event that occurred? Is there a large number of rare events such that if any one of them occurred, we would have hit the alarm bell? Such considerations contribute to the appropriate probability model, and since we in general do not have control over this, the simple solution is to avoid making claims from unplanned observations. Instead the appropriate action in the case above would be to choose one or more similar workplaces and assess the outcome at these. We would then use our observation only as a trigger for a carefully planned experiment in which statistics can be used appropriately. If this is not a

possible way forward, one has to find other means than statistics to prove the point that there is some environmental hazard at that particular workplace.

How do we then make certain that a reported p-value really has followed the rules of the game? The standard answer is: prespecify. This means that we should write a protocol before we have collected and analyzed our data, outlining what we are going to do. If in this protocol we specify the hypothesis we want to test, and the data to use, we are in a good position to use statistics in a proper way.

We see the use of p-values in courts of law, albeit in a disguised form. A piece of evidence, which is an event, is presented by the prosecutor together with a more or less explicit calculation of the probability of this event occurring for an innocent person. If this probability is small, it is used as evidence against the accused. This is in complete analogy with how p-values are used in science, and therefore embeds the same problems. As an analogy to the discussion above we have the following example.

Example 1.6 Assume that a match in two DNA profiles occurs only once in 10 000 instances. Consider the following two situations.

1. A woman has been raped and foreign DNA has been obtained from her. Based on witness statements, a man has been identified and arrested. A DNA test shows a match to the sample from the woman and he is brought to trial.

2. A woman has been raped and foreign DNA has been obtained from her. The sample is compared against a database consisting of DNA from 20 000 men. A match is found and the man in question is brought to trial.

What is the value of the DNA test in these two situations? In the first case the probability of a match for an innocent subject is 1 in 10 000. In the second case we can compute the probability that at least one of the 20 000 in the database provides a match by chance alone. The result is $1-(1-\frac{1}{10000})^{20000} \approx 1-e^{-2} = 0.86$, which is considerably larger than the first probability.

1.9 Adaptive designs: controlling the risks in an experiment

A classical way of destroying your significance level is to repeatedly look at your data, and to decide to stop when you are ahead. Consider the following experiment: we want to compare two treatments that are truly equal in all respects. We do not know that, and decide to pick 40 subjects and randomly allocate half to one treatment and the other half to the other treatment. However, we decide to randomize the study in such a way that for each pair of patients included in the study, we assign one of them to each of the two treatments. In order to be more efficient, we also decide to analyze our data after each pair, and to stop the experiment if the p-value drops below 5%. What is now the probability that we falsely end up claiming that there is a difference between the two drugs?

Figure 1.2 shows what the true significance level $P(\alpha)$ would be, if we make each individual test at the significance level α. This corresponds to the $N = 20$ curve in Figure 1.1, except that in this case there is a dependence between the 20 tests done; each new test adds the observation from one new pair of subjects to the old data. In the graph two special observations are illustrated:

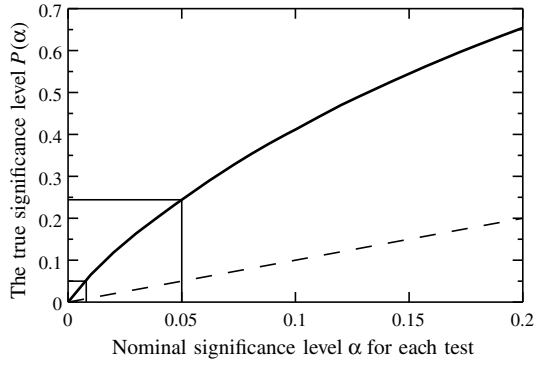

Figure 1.2 The true significance level $P(\alpha)$, as a function of the nominal significance level α of the individual tests, for the procedure discussed in the text, in which we analyze after each of the 40 pairs. The dashed line shows the corresponding function when we only look at data only after study completion.

- The value $P(0.05) = 0.25$, which gives the true significance level for an efficacy claim when individual tests are done at the conventional 5% level.
- The solution $\alpha = 0.007$ to the equation $P(\alpha) = 0.05$, which tells us at what level the individual tests should be done in order to protect the overall significance level at 5%.

The latter observation illustrates that we can in fact look repeatedly at data, if we so wish. But we need to be careful, so that we protect the overall significance level.

Adaptive designs for clinical trials are designs in which, on at least one occasion during the study execution, we take a look at the data and make a decision on the further conduct of the study. Depending on how this is done, it may or may not have consequences for how we distribute our α, the significance level. The example just given is an extreme type, called a *sequential design*. To be useful one must design decision rules in such a way that the total amount of α one can spend is properly distributed between the different looks, and in the process take into account the dependence between different tests.

A variation of this is called a *group sequential design* in which we only look a few times before the final readout. The problem is the same, we need to distribute our α, but more simply. However, we may not always like the end result, as the following real-life example illustrates.

Example 1.7 The TORCH study was a 3-year study in chronic obstructive pulmonary disease (COPD) patients with mortality as outcome. Even though there were four treatments in the study, one comparison was of primary interest: that of a certain combination product versus placebo. It had all the virtues of a good clinical design (double-blind, randomized, etc.) and it was planned to recruit 1510 patients per arm.

To carry out a 3-year study on mortality in which one treatment may be effective poses an ethical problem. If the effect is large, can we really wait for the study to come to completion? This ethical consideration forced the study designers to include two interim analyses, the first to be performed when 358 deaths had occurred in the study and the second when 680 had occurred. On each occasion a test for efficacy was performed, and the study was to be stopped

if a predefined significance level was reached, which depended on which interim analysis it was. In order to preserve the overall significance level it was decided (these are not the exact numbers used, but serve our purpose) to do the first test at significance level 0.0006 and the second at level 0.012, so that the final test would be done at level 0.038. This preserves the overall significance level, because

$$0.0006 + (1 - 0.0006)0.012 + (1 - 0.0006)(1 - 0.012)0.038 = 0.05.$$

As it happened, the study proceeded to completion, since neither of the interim analysis passed the test. Taking the data at face value, the primary comparison produced a p-value of 4.0%. However, it needed to be below 3.8% in order to preserve the overall significance level at 5%, so the study missed its objective: there was not sufficient evidence in the study to conclude that the mortality rate will be decreased if patients are treated with the combination product. In passing, we may note that the 'true' p-value, adjusted for the two interim analysis, was

$$p = 0.0006 + (1 - 0.0006)0.012 + (1 - 0.0006)(1 - 0.012)0.040 = 0.052,$$

very close to the conventional cut-off of 5%.

This example is really instructive on many levels. On the one hand, how can it be that the 5% level is so set in stone that a FDA Advisory Board can decide that 5.2% is not sufficient evidence for an effect on mortality, when the logical expectation from effects previously demonstrated by the combination treatment is that some benefit should be expected (the present medical paradigm is that COPD worsens as patients get exacerbations, and if the drug decreased the rate of these, it should really also prolong life). But accepting that, it takes some deep thinking to understand why we should be punished for introducing interim analyses that assure that if the effect is extremely obvious, we should not drag out the study unnecessarily but instead make the drug available to, among others, the patients in the placebo group. It is not hard to understand why some non-statisticians sometimes consider statistics to be more mysticism than science.

1.10 The elusive concept of probability

The word 'probability', along with some of its synonyms such as 'chance' and 'risk', is part of our everyday language. We have already had a first discussion around this concept, indicating that it is not as trouble-free a concept as it may appear. But the nature of the problem is wider than previously indicated, and lies at the heart of the difference between the two dominant schools in statistics, the frequentist school and the Bayesian school.

Serious thinking about probabilities started in connection with games, in particular in France in the seventeenth century, where playing games for money was one of the major occupations of the nobility. In this situation many problems can find a solution by the type of combinatorial argument that was used for the Monty Hall problem earlier. This means that we define the probability of an event A to be $P(A) = g/m$, where m is the number of possible outcomes of the game, and g the number of successful outcomes (satisfying the specific criteria that define A). The underlying assumption is that all possible outcomes are

equally likely, and what we do is make a list of all possible outcomes and count the proportion of successful ones.

However, this combinatorial definition is only useful at the gambling table, if what we compute is also what actually occurs. To determine if this is the case, we need to play a large number of games, say n, and count the number of times the event A occurs, call it n_A. We then expect that the proportion n_A/n, of occasions when the event A has occurred, approaches $P(A)$ as we increase the number of experiments. The frequentist school of statistics essentially derives the concept of probability from this property. The problem with this definition is how we determine $P(A)$ if we cannot do infinitely many experiments.

Probability theory comes to the rescue. One of its key statements is the law of large numbers which says that if you define how much error you can tolerate (defined by a small number $\epsilon > 0$) and you increase the number of experiments, your observed frequency will home in on the desired probability:

$$P\left(\left|\frac{n_A}{n} - P(A)\right| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Probability theory is a topic in mathematics. It does not concern itself with reality, but instead needs a starting point to build the logic from – the premises or axioms. Many important results in probability theory were derived without such a foundation, and it was not until the early 1930s that mathematicians decided how to define what a probability is. These were the axioms of Kolmogorov which define a probability as something that measures events in such a way that the total measures one and the probability of the sum (union) of exclusive events is the sum of the individual probabilities.

But the mathematical definition of a probability is of no help when you want to assign probabilities to everyday events. Here the frequentist approach is useful, but its real usefulness is in the way it builds traditional (frequentist) statistics. Consider a coin, which may be biased. Denote the probability that it falls heads by π. How do I determine π when I cannot toss the coin infinitely many times? The frequentist statistician turns this question around and tells you to toss it a finite number of times, say n. That does not allow you to give a definite answer as to what π is, but you can make statements about what π almost certainly is not. Such statements are made with a certain degree of confidence which represents your knowledge about π. You may, for example, make a statement about whether you believe the coin is unbiased (which means that $\pi = 0.5$), or not. Your confidence in this statement can then be expressed in the p-value, which is a probability in the tradition of the frequentist definition above, namely the number of experiments with an unbiased coin that will produce the outcome you observed or 'worse'. This p-value is used to express your confidence in the statement that the coin is unbiased. Similarly, you can express confidence in the actual parameter value π using confidence intervals, to be described in more detail in later chapters.

But confidence is not a probability. It lies between 0 and 1, and it is intimately related to probabilities, but the relation is that my confidence in a particular statement about a parameter is at least 30% if more than 30% of all experiments are expected to provide a point estimate that fulfills that statement. The main problem is that any particular experiment we do is unique, and we cannot determine true probabilities with any accuracy from a single experiment. What do we do if an experiment cannot be repeated? The frequentist approach has problems with assigning probabilities to non-repeatable events, and may therefore say there is no such thing.

Bayesian statisticians take a different view. They differ from the frequentists already in their view of what a probability is and they acknowledge a wider use of the probability concept

which includes what is usually called subjective probabilities. This leads to a very different view on what one should be doing in statistics and science. Whereas the frequentist is someone who is 100% focused on science as the process of falsifying hypotheses, the Bayesian views science more as an inductive process.

To see the difference, take a simple example. Brazil and Argentine are about to play a World Cup qualifying soccer game. What is the probability that Brazil will win? To estimate this, the frequentist will need to find a sample of matches already played which is such that the new game can be considered a random sample from this set. One way would be to look at all previous games between the nations. But is such an estimate relevant? The new match is not played with the same players as the previous matches, and the motivation for the two teams on this particular occasion may differ for the two teams depending on table position, whether it is a home game, etc. To account for such factors, the frequentist needs to make a more elaborate model of the previous games. When he has done that, he will arrive at an estimate of the probability that Brazil wins a game of this sort, accounting for factors such as which is the home team, table positions, etc. But it will refer to some historical average and some of the factors that matter cannot be taken care of, including the particular coaches and players participating on this occasion.

Actually you may, as a frequentist, deny there is a relevant probability concept at all for this situation because there is no relevant space to take a sample from. Alternatively you may say that since I can construct a thought experiment in which this game is played simultaneously in parallel universes, there should be a sensible probability, though its estimation still poses unsurmountable problems.

The Bayesian takes a different view. He says that there is a probability that Brazil will win, but it should not be calculated as a frequency. Instead each individual has a subjective estimate of this probability. To get an overall probability we may take the average of these subjective estimates, which is more or less what the betting companies do when they set the odds. So the Bayesian can actually talk about the probability of sunny weather tomorrow, even though that is not a repeatable experiment. Or the probability of a new Ice Age within the next millennium.

The different views on the nature of probability between a frequentist and a Bayesian really boil down to whether it is a real, physical tendency of the event to occur, or just a measure of how strongly you believe it will occur. They do agree that whatever it is, it should follow Kolmogorov's axiom system, which is the starting point of all the mathematical calculations in probability theory.

The mathematics of Bayesian statistics is the inductive (recursive) computation of new probabilities from old ones, accounting for new data. It therefore needs a starting point, the specification of an *a priori* probability, or prior for short. This is typically obtained from a consideration of available data on the matter. The issue is that for a given problem, there are in general multiple ways to assess such data, and choosing one is a matter of judgement; different people may assign different prior probabilities.

This difference in point of view has many implications. When comparing two hypotheses and using some data, frequency methods would typically result in the rejection or non-rejection of the null hypothesis at a particular significance level, and frequentists would all agree (in the best of all possible worlds) whether this hypothesis should be rejected or not at that level of significance. Bayesian methods would suggest that one hypothesis was more probable than another, but individual Bayesians might differ about which was the more probable and by how much, if they use different *a priori* probabilities. Bayesians would argue that this

is right and proper – if the contemporary knowledge is such that reasonable people can put forward different, but plausible, priors and the data do not swamp the prior, then the issue is not resolved unambiguously on available knowledge. They would argue that any approach that aims at producing a single, definitive answer in these circumstances is flawed.

In the mainstream application of biostatistics the Bayesian view is seldom listened to. Most experimenters want a clear-cut answer, in black and white, and the uncertainty imposed by the use of *p*-values is at the limit of what they can take. To actually acknowledge that there are different interpretations of the results, depending on your prior view of the matter, is usually considered an unwelcome complication, best given a wide berth.

1.11 Comments and further reading

There are a number of great scientists mentioned in this book, who have contributed to the science of clinical trials. I have in general avoided discussing these scientists, their lives, deeds and the context they worked in. For such a treatise, see the book *Dicing with Death* (Senn, 2003), which also discusses much of what we address in our first three chapters.

Most books on philosophy probably have something to say about the philosophy of science. In his autobiography (Popper, 1976), Karl Popper includes a chapter on the philosophy around induction and falsification. The quote by Richard Feynman is taken from the last chapter of what may be called his autobiography (Feynman and Leighton, 1992). Wootton (2007) gives a historian's view of the impact of physicians on people's welfare in history. For more on the suspicion of fraud indicated in Box 1.6, see Fisher (1936). (R. A. Fisher was not only the inventor of modern statistics, but also a first rank geneticist.)

We have chosen to describe the statistical output in this chapter in terms of the *p*-value. This is not necessarily a choice made because it is a very good summary, but because of its role in the medical literature. There are a lot of ways in which *p*-values are misinterpreted (Gigerenzer, 2004), some (but not all) of which we have discussed in this chapter, and in the statistical community there is often a very negative attitude (Royall, 1997) toward its use. Hopefully we can explain what it means without entering into a debate on what the statistical summaries in the medical literature should or should not be. The citations by Fisher in Box 1.2 were from Fisher (1979, p. 80) and (Fisher, 1929), respectively.

The exact nature of the FDA rule discussed in the text is unclear, and our discussion may not be fully valid. In fact, it is not clear that anyone knows what the rule really is; it is probably somewhat flexible (Senn, 2007, Section 12.2.8). We know that the rule stems from the FDA interpretation of the 1962 amendment to the Federal Food, Drug, and Cosmetics Act, which required 'adequate and well controlled investigations'. However, a further amendment in 1997 permits the FDA to require only one such study, as long as there is other substantial evidence for the benefit of the drug. This seems to mean that, by law at least, approval is not only about a low significance level. To what extent this has had any impact on the FDA approval process is unclear (at least to me).

The comment about the non-inferiority study type may not go down well with every statistician, because finding the non-inferiority margin has provided food for numerous statistical publications, including regulatory guidelines from health authorities. The non-inferiority study type was designed to solve one problem: that absence of statistical significance was taken as proof of equality. However, the solution is almost as bad. As far as I know the concept was introduced in connection with a wider attempt to

harmonize the regulatory requirements all over the world, in a document labeled ICH E10 (International Conference on Harmonisation, 2000).

There is much more to be said about the nature of probabilities and its implications for a proper treatment of p-values. Often the need for a probabilistic discussion stems from lack of information, as in the game show. If we only had complete information, we would often not need probabilities, like if we only knew all the initial conditions when we toss a coin, we can predict the outcome with certainty. In fact, you could argue that there are few cases when there are pure random events. The notable exception are some deep aspects of contemporary quantum physics. Deterministic processes may appear probabilistic to us, simply because we cannot obtain sufficient knowledge to explore the deterministic nature of the problem, a subject mathematicians discuss in chaos theory. Accepting that we need to compute probabilities, it becomes important to understand the conditions under which the computed probability is valid. A very rare event occurs, seen from a prospective vantage point, with a very small probability. Retrospectively the probability is one. To compute that probability when we know that it has occurred is basically meaningless. However, we should not confuse that with what we do when we compute p-values. These are probabilities for the outcome, computed under an assumption, and we use the p-value as indirect evidence for or against that assumption. Note that the p-value computes the probability of the outcome given that the null hypothesis is true, not the transposed conditional, the probability that the null hypothesis is true given the outcome we have observed.

When it comes to error control with multiple testing, the original suggestion by Bonferroni on how to allocate parts of the available α (significance level) to the different tests was improved upon considerably by Sture Holm in 1979. He showed that the testing could be done in a stepwise manner in the order of increasing individual p-values, where these p-values were compared with successively larger fractions of α. After that it took another 28 years for the next major step, made independently by Guilbaud and Strassburger-Bretz, which was the development of confidence intervals corresponding to Holm's and related testing procedures. A modern review of this subject, by Dmitrienko et al. (2010), covers not only basic/traditional approaches but also novel ones. In connection with the multiplicity problem we also touched upon one of the present hypes in medical statistics, the adaptive designs. We will say no more on this subject, and refer the reader who wants to learn more to the vast literature on the subject (there are also plenty of conferences he or she can go to). Both Whitehead (1997) and Chang (2008) offer useful starting points. The illustration in Example 1.7 is an adaptation of the main result in Calverley et al. (2007).

References

Calverley, P.M., Anderson, J.A., Celli, B., Ferguson, G.T., Jenkins, C., Jones, P.W., Yates, J.C. and Vestbo, J. (2007) Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *New England Journal of Medicine*, **356**(8), 775–789.

Chang, M. (2008) *Adaptive Design Theory and Implementation Using SAS and R,* CRC Biostatistics Series. Boca Raton, FL: Chapman & Hall/CRC.

Dmitrienko, A., Tamhane, A.C. and Bretz, F. (2010) *Multiple Testing Problems in Pharmaceutical Statistics,* CRC Biostatistics Series. Boca Raton, FL: Chapman & Hall/CRC.

Feynman, R.P. and Leighton, R. (1992) *Surely You're Joking Mr Feynman! Adventures of a Curious Character*. London: Vintage.

Fisher, R. (1929) The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, **39**, 189–192.

Fisher, R. (1936) Has Mendel's work been rediscovered?. *Annals of Science*, **1**, 115–137.

Fisher, R. (1979) *Statistical Methods for Research Workers* 10th edn. Edinburgh: Oliver & Boyd.

Gigerenzer, G. (2004) Mindless statistics. *Journal of Socio-Economics*, **33**, 587–606.

International Conference on Harmonisation (2000) *ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials*. Geneva: International Conference on Harmonisation.

Popper, K. (1976) *Unended Quest: An Intellectual Autobiography*. London: Fontana/Collins.

Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm,* vol. 71 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Senn, S. (2003) *Dicing with Death: chance, risk and health*. Cambridge: Cambridge University Press.

Senn, S. (2007) *Statistical Issues in Drug Development*. Chichester: John Wiley & Sons, Ltd.

Whitehead, J. (1997) *The Design and Analysis of Sequential Trials*. Chichester: John Wiley & Sons, Ltd.

Wootton, D. (2007) *Bad Medicine: Doctors Doing Harm since Hippocrates*. Oxford: Oxford University Press.