

Bruno LECOUTRE  
Jacques POITEVINEAU

The Significance Test  
Controversy Revisited:  
The fiducial Bayesian Alternative

July 26, 2014

Springer



# Contents

<b>1</b>	<b>Introduction</b>	3
1.1	The fiducial Bayesian Inference	4
1.2	The Stranglehold of Significance Tests	5
1.3	Beyond the Significance Test Controversy	6
1.4	The Feasibility of Fiducial Bayesian Methods	6
1.5	Plan of the Book	7
<b>2</b>	<b>Preamble - Frequentist and Bayesian Inference</b>	9
2.1	Two Different Approaches to Statistical Inference	9
2.2	The Frequentist Approach: From Unknown to Known	10
2.2.1	Sampling Probabilities	10
2.2.2	Null Hypothesis Significance Testing in Practice	11
2.2.3	Confidence Interval	12
2.3	The Bayesian Approach: From Known to Unknown	12
2.3.1	The Likelihood Function and the Bayesian Probabilities	12
2.3.2	An Opinion-Based Analysis	14
2.3.3	A “No Information Initially” Analysis	16
<b>3</b>	<b>The Fisher, Neyman-Pearson and Jeffreys views of Statistical Tests</b>	21
3.1	The Fisher Test of Significance	21
3.2	The Neyman-Pearson Hypothesis Test	23
3.3	The Jeffreys Bayesian Approach to Testing	25
3.4	Different Views of Statistical Inference	28
3.4.1	Different Scopes of Applications: The Aim of Statistical Inference	28
3.4.2	The Role of Bayesian Probabilities	30
3.4.3	Statistical Tests: Judgment, Action or Decision?	32
3.5	Is It possible to Unify the Fisher and Neyman-Pearson Approaches?	34
3.6	Concluding Remarks	35

<b>4</b>	<b>GHOST: An officially Recommended Practice</b>	37
4.1	Null Hypothesis Significance Testing	37
4.1.1	An Amalgam	37
4.1.2	Misuses and Abuses	38
4.2	What About the Researcher's Point of View?	40
4.3	An Official Good Statistical Practice	40
4.3.1	Guidelined Hypotheses Official Significance Testing	41
4.3.2	A Hybrid Practice	43
<b>5</b>	<b>The Significance Test Controversy Revisited</b>	45
5.1	Significance Tests vs Pure Estimation	45
5.2	The Null Hypothesis: A Straw Man	46
5.3	Usual Two-sided Tests Do Not Tell the Direction	47
5.4	Determining Sample Size	48
5.5	Critique of $p$ -values: A Need to Rethink	49
5.6	Decision and Estimation	53
5.7	The Role of Previous Information and the Sample Size	54
5.8	The Limited Role of Significance Problems	55
5.9	Other Issues	56
5.9.1	Non-inferiority and Equivalence Questions	56
5.9.2	Stopping Rules and the Likelihood Principle	56
<b>6</b>	<b>Reporting Effect Sizes: The New Star System</b>	59
6.1	What Is an Effect Size?	59
6.2	Abuses and Misuses Continue	60
6.3	When Things Get Worse	64
6.3.1	A Lot of Choices for a Standardized Difference	64
6.3.2	A Plethora of ES Indicators	66
6.3.3	Don't Confuse a Statistic with a Parameter	67
6.4	Two Lessons	69
<b>7</b>	<b>Reporting Confidence Intervals: A Paradoxical Situation</b>	71
7.1	Three views of Interval Estimates	71
7.1.1	The Bayesian Approach (Laplace, Jeffreys)	71
7.1.2	Fisher's Fiducial Inference	73
7.1.3	Neyman's Frequentist Confidence Interval	74
7.2	What Is a Good Interval Estimate?	76
7.2.1	Conventional Frequentist Properties	76
7.2.2	The Fatal Disadvantage of "Shortest Intervals"	76
7.2.3	One-sided Probabilities are Needed	77
7.2.4	The Jeffreys Credible Interval is a Great Frequentist Procedure	77
7.3	Neyman-Pearson's Criterion Questioned	77
7.3.1	The Inconsistencies of Noncentral $F$ Based Confidence Intervals for ANOVA Effect Sizes	78

7.3.2	The Official Procedure for Demonstrating Equivalence . . . . .	80
7.4	Isn't Everyone a Bayesian? . . . . .	81
<b>8</b>	<b>Basic Fiducial Bayesian Procedures for Inference About Means . . . . .</b>	<b>83</b>
8.1	Fiducial Bayesian Methods for an Unstandardized Contrast . . . . .	84
8.1.1	The Student Pharmaceutical Example . . . . .	84
8.1.2	Specific Inference . . . . .	84
8.2	Fiducial Bayesian Methods for a Standardized Contrast . . . . .	87
8.2.1	A Conceptually Straightforward Generalization . . . . .	87
8.2.2	Inference About the Proportion of Population Differences . . . . .	88
8.3	Inference About Pearson's Correlation Coefficient . . . . .	89
8.4	A Coherent Bayesian Alternative to GHOST . . . . .	90
8.4.1	NHST: The Fiducial Bayesian Interpretation of the $p$ -Value . . . . .	90
8.4.2	Interval Estimates: The Fiducial Bayesian Interpretation of the Usual CI . . . . .	90
8.4.3	Effect Sizes: Straight Bayesian Answers . . . . .	90
8.4.4	Making Predictions . . . . .	92
8.4.5	Power and Sample Size: Bayesian Data Planning and Monitoring . . . . .	93
8.5	Our Guidelines . . . . .	94
<b>9</b>	<b>Generalizations and Methodological Considerations for ANOVA . . . . .</b>	<b>95</b>
9.1	From $F$ tests to Fiducial Bayesian Methods for ANOVA Effect Sizes . . . . .	95
9.1.1	The Traditional Approach . . . . .	96
9.1.2	Fiducial Bayesian Procedures . . . . .	97
9.1.3	Some Conceptual and Methodological Considerations . . . . .	100
9.2	Alternatives to the Inference About ANOVA ES . . . . .	102
9.2.1	The Scheffé Simultaneous Interval Estimate and Its Bayesian Justification . . . . .	102
9.2.2	Contrast Analysis . . . . .	105
9.3	An Illustrative Example: Evaluation of the "0.05 Cliff Effect" . . . . .	105
9.3.1	Numerical Results . . . . .	106
9.3.2	A Cliff Effect Indicator . . . . .	107
9.3.3	An Overall Analysis Is not Sufficient . . . . .	109
9.3.4	A simultaneous inference about all contrasts . . . . .	110
9.3.5	An Adequate Analysis . . . . .	110
9.3.6	What about standardized effects? . . . . .	111
9.4	Our Guidelines for ANOVA . . . . .	112
<b>10</b>	<b>Conclusion . . . . .</b>	<b>113</b>
	References . . . . .	115
	<b>Index . . . . .</b>	<b>123</b>



## Acronyms

APA	American Psychological Association
fB	fiducial Bayesian
CI	Confidence Interval
ES	Effect Size
GHOST	Guidelined Hypotheses Official Significance Testing
HPD	Highest Posterior Density
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
NCF	Noncentral $F$ based
NCF-CI	Noncentral $F$ based Confidence Interval
NHST	Null Hypothesis Significance Testing
TOST	Two-One Sided Tests procedure
U-CI	Usual Confidence Interval





”i



# Chapter 1

## Introduction

**We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions (Fisher, 1955, p. 77).**

A critical aspect of experimental data analysis is that results must be accepted by the scientific community. This can be the reason why Bayesian methods of analyzing experimental data are, at best constantly ignored, at worst explicitly discarded. Most potential users feel that they are too complicated to use and too subjective to be scientifically acceptable. It must be stressed that these *a priori* reasons are completely unjustified:

A common misconception is that Bayesian analysis is a subjective theory; this is neither true historically nor in practice. The first Bayesians, Bayes (see Bayes (1763)) and Laplace (see Laplace (1812[1840])) performed Bayesian analysis using a constant prior distribution for unknown parameters, although the motivations of each in doing so were considerably more sophisticated than simply stating that each possible value of the parameter should receive equal prior weight. Indeed, this approach to statistics, then called “inverse probability” (see Dale (1991)) was central to statistics for most of the nineteenth century, and was highly influential in the early part of the twentieth century (Berger, 2004, p. 3).

Following the lead of Bayes and Laplace, Jeffreys (1931) aimed at proposing a general methodology for “learning from data and experience”. The key feature of his approach is to assign prior probabilities when *we have no information initially* about the value of the parameter. In practice, these “noninformative” probabilities are vague prior distributions, which do not favor any particular value: they let the data “speak for themselves”.

In this form the Bayesian paradigm provides *reference* methods appropriate to report experimental results. However, the potential contribution of Bayesian inference to experimental data analysis and scientific reporting is obscured by the fact that many today’s Bayesian proponents focus on individual decision making. Indeed, it should be acknowledged with Rozeboom that

the primary aim of a scientific experiment is not to precipitate decisions (Rozeboom, 1960, p. 420).

So Jeffreys' approach has been embedded into a Bayesian decision-theoretic framework, without concern for the roles he assigned to significance tests and estimation in experimental data analysis. Moreover, within this context, many Bayesians place emphasis on a subjective perspective.

## 1.1 The fiducial Bayesian Inference

Our motivation may be found in Efron's assertion:

A widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance. [...] A successful objective Bayes theory would have to provide good frequentist properties in familiar situations, for instance, reasonable coverage probabilities for whatever replaces confidence intervals (Efron, 1998, pp. 106 and 112).

We suggest that such a theory is by no means a speculative viewpoint but on the contrary a desirable and perfectly feasible project. For many years we have worked with colleagues in France in order to develop routine Bayesian methods for the most familiar situations encountered in experimental data analysis. These methods can be taught and used easily and offer promising new ways in statistical methodology. In order to promote them, it is important to give these methods an explicit name. Berger (2004) proposed the name *objective Bayesian analysis*. With the same incentive, we argued for the name *fiducial Bayesian methods* (Lecoutre, in Rouanet et al., 2000; Lecoutre, Lecoutre & Poitevineau, 2001). This deliberately provocative, politically incorrect, name pays tribute to Fisher's work on scientific inference for research workers (Fisher, 1990a), which was highly influential on Jeffreys' works.

It must be stressed that, if Fisher and Jeffreys have first contested the validity of each other's approach (see Jeffreys, 1933, p. 535), Jeffreys's position considerably evolved and, in the first edition of *Theory of Probability*, he came to emphasize his practical agreement with Fisher (see Aldrich, 2005):

I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher has already grasped the essentials by some brilliant piece of common sense, and that his results would be either identical with mine or would differ only in cases where we should both be very doubtful (Jeffreys, 1939, p. 324).

In actual fact, "fiducial Bayesian" indicates the aim to let the statistical analysis express what the data have to say independently of any outside information.

In short, **fiducial Bayesian inference uses Bayesian approach with a fiducial motivation**. Nowadays, thanks to the computer age, fiducial Bayesian routine methods for the familiar situations of experimental data analysis are easy to implement and use. They fulfill the requirements of experimental data reporting and they fit in better with scientists' spontaneous interpretations of data than frequentist significance tests and confidence intervals.

## 1.2 The Stranglehold of Significance Tests

In spite of some recent changes, “Null hypothesis significance testing” is again conventionally used in experimental literature. In practice, each experimental result is dichotomized: *significant* (the null hypothesis is rejected) vs *nonsignificant* (the null hypothesis is not rejected). This appears as a hybrid theory, an amalgam of two different views, the Fisher test of significance and the Neyman-Pearson hypothesis test, the latter being the “official” theory of testing.

This hybrid is essentially Fisherian in its logic, but it pays lip service to the Neyman-Pearson theory of testing (Spielman, 1974, p. 211).

This ignores the fact that sharp controversies have constantly opposed Fisher and Neyman (and Pearson to a lesser extent) on the very foundations of statistical inference. A detailed account is given in Lehmann (2011).

Several empirical studies emphasized the widespread existence of common misuses of significance tests among students and scientists (for a review, see Lecoutre, Lecoutre & Poitevineau, 2001). Many methodology instructors who teach statistics, including professors who work in the area of statistics, appear to share their students’ misconceptions. Moreover, even professional applied statisticians are not immune to misinterpretations, especially if the test is nonsignificant. It is hard to interpret these findings as an individual’s lack of mastery: they reveal that scientists cannot find in null hypothesis significance testing appropriate answers to their precise questions. In order to interpret their data in a reasonable way, they must resort to a more or less naive mixture of significance tests outcomes and other information. But this is not an easy task!

It is not surprising that, from the outset (e.g. Boring, 1919), significance tests have been subject to intense criticism. Their use has been explicitly denounced by the most eminent and most experienced scientists. In the 1960s, more and more publications have stressed their shortcomings, especially in the behavioral and social sciences:

*The significance test controversy* (Morrison & Henkel, 1970).

Nowadays, almost all papers that examine the current practice in experimental publications, or discuss alternative solutions, begin with a more or less detailed section on the significance test shortcomings. Moreover, many papers are replete with ill-informed, secondary and even tertiary sources, or ill-considered claims, and first and foremost concerning Fisherian and Bayesian inferences.

Criticisms about significance tests are endlessly repeated and extended to virtually all fields, not to mention the controversies on the foundations of statistical inference that continue to divide frequentists and Bayesians. This gives a discouraging feeling of *déjà-vu* and is without doubt detrimental to the impact of new proposals, if not to the image of statistical inference.

### 1.3 Beyond the Significance Test Controversy

Changes in reporting experimental results are more and more enforced within guidelines and editorial policies.

Most of these changes are explicitly intended to deal with the essential question of effect sizes. The term “effect size” has become increasingly popular in recent years. For instance, this term did not appear in the subject index of the book *The significance test controversy* (Morrison & Henkel, 1970). By contrast, 18 lines are devoted to it in the index of the book *What if there were no significance tests* (Harlow, Mulaik & Steiger, 1997).

Reporting an effect size estimate is one of the first necessary steps in overcoming the abuses of significance tests. It can effectively prevent users from unjustified conclusions in the conflicting cases where a nonsignificant result is associated with a large observed effect size. However, small observed effect sizes are often illusorily perceived by researchers as being *favorable* to a conclusion of no effect, when they can’t in themselves be considered as sufficient proof.

Consequently, the majority trend is to advocate the use of confidence intervals, in addition to or instead of significance tests. In practice, two probabilities can be routinely associated with a specific interval estimate computed from a particular sample. The first, *frequentist*, probability is “the proportion of repeated intervals that contain the parameter”. It is usually termed the coverage probability. The second, *Bayesian*, probability is the “posterior probability that this interval contains the parameter”. In the frequentist conception it is *forbidden* to use the second probability while in the Bayesian conception, the two probabilities are valid.

In actual practice, reporting effect sizes and confidence intervals appear to have very little impact on the way the authors interpret their data. Most of them continue to focus on the statistical significance of the results. They only wonder whether the interval includes the null hypothesis value, rather than on the full implications of confidence intervals: **the steamroller of significance tests cannot be escaped.**

### 1.4 The Feasibility of Fiducial Bayesian Methods

Fiducial Bayesian methods are concrete proposals in order to bypass the inadequacy of NHST. For more than thirty years now, with other colleagues in France we have worked in order to develop routine procedures for the most familiar situations encountered in experimental data analysis (see e.g., Rouanet & Lecoutre, 1983; Lecoutre, Derzko & Grouin, 1995; Lecoutre, 1996; Lecoutre & Charron, 2000; Lecoutre & Poitevineau, 2000; Lecoutre & Derzko, 2001). These procedures can be learned and used as easily, if not more, as the  $t$ ,  $F$  or  $\chi^2$  tests. We argued

that they offer promising new ways in statistical methodology (Rouanet et al., 2000; Lecoutre, 2006, 2008).

We especially developed Bayesian methods in the analysis of variance framework, which is an issue of particular importance for experimental data analysis. Experimental investigations frequently involve complex designs, especially repeated-measures designs. Bayesian procedures have been developed on the subject, but they are generally thought difficult to implement and not included in the commonly available computer packages. As a consequence the possibility of using them is still largely questionable for many investigators.

We have developed the statistical software LePAC (Lecoutre & Poitevineau, 1992; Lecoutre, 1996). It incorporates both traditional frequentist practices (significance tests, confidence intervals) and routine Bayesian methods (including the use of conjugate priors) for univariate and multivariate analysis of variance. LePAC also includes Bayesian methods for inference about proportions. Extensive applications to real data have been done. From the outset, they have been accepted well in experimental publications (e.g. Ciancia et al., 1988).

## 1.5 Plan of the Book

We are conscious that warnings about common misconceptions and unsound statistical practices have been given many times before, apparently without much effect. Our ambition is not only to revisit the “significance test controversy”, but also to provide a conceptually sounder alternative. Thus the presentation will be methodologically oriented. The book is organized as follows.

Chapter 2 serves as an overall introduction to statistical inference concepts. The basic notions about the frequentist and Bayesian approaches to inference are presented and the corresponding terminology is introduced. Chapter 3 presents *normative* aspects: the three main views of statistical tests – Fisherian, Neyman-Pearsonian and Jeffreys’ Bayesian – are discussed.

Chapters 4 and 5 are devoted to *descriptive* aspects: what is the current practice in experimental research? The misuses of null hypothesis significance tests are reconsidered in the light of Jeffreys’ Bayesian conceptions about the role of statistical inference in experimental investigations.

Chapters 6 and 7 examine *prescriptive* aspects: what are the recommended “good statistical procedures?” The effect size and confidence interval reporting practices are discussed. The risks of misuses and misinterpretations of the usual ANOVA ES indicators (Cohen’s  $d$ , eta-squared, etc.) are stressed. Frequentist confidence intervals commonly proposed for these indicators are also seriously questioned.

Chapter 8 introduces basic routine procedures for inference about means and demonstrates that the fiducial Bayesian paradigm is appropriate to report experimental results: **don’t worry, be Bayesian**. Of course, this does not mean that by adopting the Bayesian approach one could actually “stop thinking about data”. This is not our message: the opposite is actually true!

Chapter 9 generalizes the basic procedures to the usual unstandardized and standardized ANOVA effect sizes indicators. Then methodological aspects are discussed and appropriate alternatives to these indicators are developed.

This book should not be read from the perspective of providing an introduction to Bayesian statistics as such. It aims at discussing the uses of statistical procedures, conceptually appropriate to report experimental results, especially in the familiar ANOVA framework. The objective is to equip the reader with appropriate procedures in order to bypass the common misuses of significance tests.



## Chapter 2

# Preamble - Frequentist and Bayesian Inference

### 2.1 Two Different Approaches to Statistical Inference

Statistical inference is typically concerned with both known quantities - the observed data - and unknown quantities - the parameters. How to assign a probability to them? Two main broad approaches are available (Jaynes, 2003).

1. In the **frequentist** conception probability is the *long-run frequency* of occurrence of an event, either in a sequence of repeated trials or in an ensemble of “identically” prepared systems.
2. In the **Bayesian** conception probability is a measure of the *degree of confidence* (or belief) in the occurrence of an event or in a proposition.

The common statistical inference procedures in scientific publications – null hypothesis significance tests and confidence intervals – are based on the frequentist conception. Owing to this domination that goes back to the first part of the 20th century, the frequentist inference has been inappropriately called “classical:”

A great deal of the basis of classical inference was forged in this period [1920-1935] (Barnett, 1999, p. 124).

To debate which of the two approaches, frequentist or Bayesian, is the more classical would be futile. From the outset the concept of probability was considered as essentially dualistic, for being related either to degrees of confidence or to systems leading to produce frequencies in the long run.

#### A simple illustrative example

It is well known that most statistical users confuse frequentist and Bayesian probabilities when interpreting the results of statistical inference procedures:

Inevitably, students (and everyone else except for statisticians) give an inverse or Bayesian twist to frequentist measures such as confidence intervals and  $p$ -values (Berry, 1997, page 242).

All the attempts made by frequentists to rectify these misinterpretations have been a losing battle. Nevertheless, imagine the following situation. Two colleagues of us, Reynald and Jerry, statistical instructors in psychology, claimed to have developed a prospective individual teaching method that yields promising results. We are very skeptical and we suggest to them to apply their method in a classroom of  $N = 16$  students. We agree with them that a success rate greater than 50% – at least  $M = 9$  successes out of 16 – would be very encouraging. Even a rate greater than 25% – at least  $M = 5$  successes out of 16 – could still be a reasonable initial step for trying to improve the method.

Consider the following simple conceptual context: assume that  $n = 4$  individuals in the classroom have received the new teaching method. Three successes and one failure have been observed, hence the observed success rate  $f = 3/4$ . This is an encouraging result: can it be generalized to the entire classroom? This is the familiar problem of trying to predict the actual number of white balls in an urn containing  $N = 16$  balls in total, each either black or white, based on an observed sample of size  $n = 4$ .

For this purpose the data are considered as a random sample from the entire classroom, that is a finite population of  $N = 16$  individuals, where each individual falls into one of the two types: 1 (success) or 0 (failure). Let  $\varphi$  be the success rate in this population.

The statistical reasoning is fundamentally a generalization from a known quantity – here the data  $f = 3/4$  – to an unknown quantity – here the parameter  $\varphi$ .

## 2.2 The Frequentist Approach: From Unknown to Known

In the frequentist approach, we have no probabilities and consequently no possible inference... unless we *fix* a parameter value and *imagine* repetitions of the observations. This requires reversing the reasoning, from the unknown parameter  $\varphi$  to the known data. But it is very different to learn something about data when the parameter is assumed to be known, and to learn something about the unknown parameter when all that is known is a data sample.

### 2.2.1 Sampling Probabilities

So, if we assume for instance  $\varphi = .25$  (4 successes in the population), we get sampling probabilities:  $\Pr(f | \varphi = .25)$ . These sampling probabilities can be empirically generated by repeated random sampling without replacement from a dichotomous population containing  $M = 4$  successes out of  $N = 16$  individuals. Alternatively, we can consider all 1 820 possible samples of size  $n = 4$ : 495 of them contain zero success, 880 contain one success, etc. Hence we get the *sampling distribution*:

successes	samples	sampling probabilities
0	495	$\Pr(f = 0/4   \varphi = .25) = .2720$
1	880	$\Pr(f = 1/4   \varphi = .25) = .4835$
2	396	$\Pr(f = 2/4   \varphi = .25) = .2176$
3	48	$\Pr(f = 3/4   \varphi = .25) = .0264$
4	1	$\Pr(f = 4/4   \varphi = .25) = .0005$
1 820		1

Formally, the probability of observing  $a$  successes is given by a Hypergeometric distribution  $HG(N, n, M)$ , with  $N = 16$ ,  $n = 4$  and  $M = 4$ , so that

$$\Pr(a|M) = \frac{M!(N-M)!n!(N-n)!}{a!(M-a)!(n-a)!(N-M-n+a)!N!} \quad [0 \leq a \leq n].$$

In the frequentist inference all probabilities are conditional on parameters that are assumed known. This leads in particular to Null Hypothesis Significance Tests, where the value of the parameter of interest is fixed by hypothesis, and confidence intervals.

### 2.2.2 Null Hypothesis Significance Testing in Practice

The sampling probabilities can serve to define a significance test of the null hypothesis  $H_0 : \varphi = .25$ . Assuming that this hypothesis is true, the expected value of  $f$  is .25 (1/4). The more distant from 1 is the observed number of successes, the less *plausible* is the null hypothesis. Plausible must be understood as “occurring by random sampling – i.e. by chance – from the population if the null hypothesis is true.” If  $\varphi = .25$ , the sampling probability of getting a value  $f \geq 3/4$  as *least as extreme* as the observed success rate is  $.0264 + .0005 = .0269$ . Consequently, the test is said to be *significant*:  $p = .0269$  (“ $p$ -value”). In other words the null hypothesis  $H_0 : \varphi = .25$  is *rejected*.

Note that we do not enter here in the one-sided (more extreme in a direction) vs two sided (more extreme in the two directions) test distinction, which is irrelevant for the moment.

Consider another example of null hypothesis,  $H_0 : \varphi = .50$ . The corresponding sampling distribution is:

$\Pr(f = 0/4   \varphi = .50) = .0385$
$\Pr(f = 1/4   \varphi = .50) = .2462$
$\Pr(f = 2/4   \varphi = .50) = .4308$
$\Pr(f = 3/4   \varphi = .50) = .2462$
$\Pr(f = 4/4   \varphi = .50) = .0385$

In this case, if the null hypothesis is true, the sampling probability of getting a value  $f \geq 3/4$  is  $.2462 + .0385 = .2847$ . The test is said to be *nonsignificant*:  $p = .2847$ . In other words the null hypothesis  $H_0 : \varphi = .50$  is *not rejected* (is “accepted”).

The dichotomy between significant and nonsignificant is the most often based on the conventional level  $\alpha = .05$ . Our two colleagues agree on this convention, but

are divided about the words to use: significant vs nonsignificant for Reynald and rejected vs accepted for Jerry. Moreover, Reynald claims that reporting the  $p$ -value gives more information. Jerry considers this practice to be superfluous and prefers to take into account the *power* of the test. He assumes the *alternative* hypothesis  $H_a : \phi = .75$  and computes the probability of rejecting the null hypothesis  $H_0 : \phi = .50$  if  $H_a$  is true. This probability is .272, and he states that the nonsignificant result is due to the “lack of power” of the test.

### 2.2.3 Confidence Interval

The null hypothesis  $H_0 : \phi = .50$  is not rejected. Has it been proved that  $\phi = .50$ ? **Certainly not:** many other null hypotheses are not rejected! So, the set of all possible parameter values that are not rejected at (one-sided) level  $\alpha = .05$  is  $\{\frac{5}{16}, \frac{6}{16} \dots \frac{15}{16}\}$ . This set constitutes a  $100(1 - \alpha)\% = 95\%$  *confidence interval* for  $\phi$ . How to interpret the confidence level 95%? The frequentist interpretation is based on the universal statement:

Given a fixed value of the parameter, whatever this value is, 95% (at least) of the intervals computed for all possible samples include this value.

In the frequentist interpretation, the confidence level 95% is based on all possible samples, and *does not depend on the data in hand*.

## 2.3 The Bayesian Approach: From Known to Unknown

Assigning a frequentist probability to a single case event requires imagining a reference set of events or a series of repeated experiments. This can easily be done for obtaining the sampling probabilities of  $f$ . However, during this repeatable process the underlying parameter  $\phi$  remains fixed. In consequence, the assignment of probabilities to a parameter is simply rejected by frequentists. By contrast, it is not conceptually problematic to assign a Bayesian probability to a parameter.

### 2.3.1 The Likelihood Function and the Bayesian Probabilities

So, let us return to the natural order of statistical reasoning, from the known data to the unknown parameter  $\phi$ . Adopting a Bayesian viewpoint, we first reconsider the sampling probabilities. Instead of the probabilities of imaginary samples given a fixed parameter value, Bayesian inference involves the probabilities of *the observed*

*data* ( $f = 3/4$ ) for *each possible value* of the parameter  $\varphi$ :  $\Pr(f = 3/4 | \varphi)$ . This is the *likelihood function* that is denoted by  $\ell(\varphi | \text{data})$ :

$\Pr(f = 3/4   \varphi = 0/16) = 0$	$\Pr(f = 3/4   \varphi = 9/16) = .3231$
$\Pr(f = 3/4   \varphi = 1/16) = 0$	$\Pr(f = 3/4   \varphi = 10/16) = .3956$
$\Pr(f = 3/4   \varphi = 2/16) = 0$	$\Pr(f = 3/4   \varphi = 11/16) = .4533$
$\Pr(f = 3/4   \varphi = 3/16) = .0071$	$\Pr(f = 3/4   \varphi = 12/16) = .4835$
$\Pr(f = 3/4   \varphi = 4/16) = .0264$	$\Pr(f = 3/4   \varphi = 13/16) = .4714$
$\Pr(f = 3/4   \varphi = 5/16) = .0604$	$\Pr(f = 3/4   \varphi = 14/16) = .4000$
$\Pr(f = 3/4   \varphi = 6/16) = .1099$	$\Pr(f = 3/4   \varphi = 15/16) = .2500$
$\Pr(f = 3/4   \varphi = 7/16) = .1731$	$\Pr(f = 3/4   \varphi = 16/16) = 0$
$\Pr(f = 3/4   \varphi = 8/16) = .2462$	

In the Bayesian inference parameters can also be probabilized. This results in distributions of probabilities that express our uncertainty:

1. about the parameter before observation (they do not depend on data): **prior** probabilities;
2. about the parameter after observation (conditional on data): **posterior** (or *revised*) probabilities;
3. about future data: **predictive** probabilities.

The choice of the prior distribution is fundamental. The flexibility of the Bayesian paradigm allows for different approaches. In the *personalistic* view (de Finetti, 1937; Savage, 1954), the prior is based mainly on personal opinion. For experimental data analysis, such prior can be derived by elicitation from “experts”, but this is obviously controversial. U.S. Food and Drug Administration guidelines for the use of Bayesian statistics in medical device clinical trials recently recommended to use “good prior information” (Food and Drug Administration, 2010). These guidelines mentioned the following possible sources of prior information:

- clinical trials conducted overseas,
- patient registries,
- clinical data on very similar products,
- pilot studies.

However, they admitted that

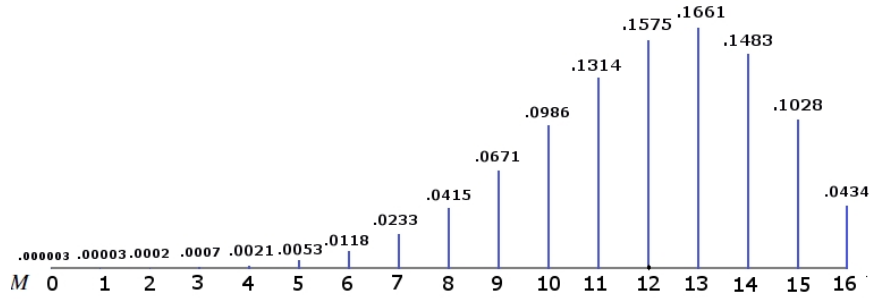
the evaluation of “goodness” of the prior information is subjective (Food and Drug Administration, 2010, p. 22).

So, it is tempting to consider a completely different approach and to use the prior to express the fact that we have *no information initially* (Jeffreys, 1967).

### 2.3.2 An Opinion-Based Analysis

#### From prior to posterior probabilities

For illustration purposes, let us assume that our colleagues' *a priori* opinion about the number of successes  $M$  in the population, or equivalently the unknown rate  $\varphi = M/16$  can be expressed by the probabilities given in Figure 2.1.



**Fig. 2.1** Opinion-based analysis: Prior probabilities  $Pr(M = 0, 1 \dots 16)$  [ $\varphi = M/16$ ].

They have a prior probability .915 that  $\varphi$  exceeds .50 (at least  $M = 9$  successes out of 16 in the population). Then, by a simple product, we get the joint probabilities of the parameter values and the data:

$Pr(\varphi \text{ and } f = 3/4) = Pr(f = 3/4   \varphi) \times Pr(\varphi) = \ell(\varphi   \text{data}) \times Pr(\varphi)$	
$Pr(\varphi = 0/16 \text{ and } f = 3/4) = 0$	$Pr(\varphi = 9/16 \text{ and } f = 3/4) = .0217$
$Pr(\varphi = 1/16 \text{ and } f = 3/4) = 0$	$Pr(\varphi = 10/16 \text{ and } f = 3/4) = .0390$
$Pr(\varphi = 2/16 \text{ and } f = 3/4) = 0$	$Pr(\varphi = 11/16 \text{ and } f = 3/4) = .0596$
$Pr(\varphi = 3/16 \text{ and } f = 3/4) = .000005$	$Pr(\varphi = 12/16 \text{ and } f = 3/4) = .0761$
$Pr(\varphi = 4/16 \text{ and } f = 3/4) = .00005$	$Pr(\varphi = 13/16 \text{ and } f = 3/4) = .0783$
$Pr(\varphi = 5/16 \text{ and } f = 3/4) = .0003$	$Pr(\varphi = 14/16 \text{ and } f = 3/4) = .0593$
$Pr(\varphi = 6/16 \text{ and } f = 3/4) = .0013$	$Pr(\varphi = 15/16 \text{ and } f = 3/4) = .0257$
$Pr(\varphi = 7/16 \text{ and } f = 3/4) = .0040$	$Pr(\varphi = 16/16 \text{ and } f = 3/4) = 0$
$Pr(\varphi = 8/16 \text{ and } f = 3/4) = .0102$	

The sum of the joint probabilities gives the marginal predictive probability of the data, before observation:

$$Pr(f = 3/4) = \sum_{\varphi} Pr(\varphi \text{ and } f = 3/4) = .3756.$$

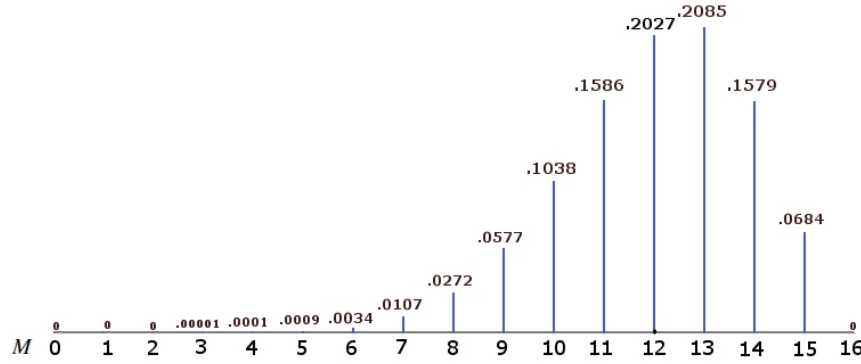
This predictive probability is very intuitive: it is a weighted average of the likelihood function, the weights being the prior probabilities.

Finally we compute the posterior probabilities after observation, by application of the definition of conditional probabilities. They are simply the normalized prod-

uct of the prior and the likelihood, which is a form of the “principle of inverse probability” (Jeffreys, 1967, p. 29), or equivalently of the Bayes’ formula:

$$\Pr(\varphi | f = 3/4) \propto \ell(\varphi | \text{data}) \times \Pr(\varphi) = \frac{\Pr(\varphi \text{ and } f = 3/4)}{\Pr(f = 3/4)}.$$

These posterior probabilities are given in Figure 2.2. The posterior probability that  $\varphi$  exceeds .50 (at least  $M = 9$  successes) is updated to .958.



**Fig. 2.2** Opinion-based analysis: Posterior probabilities  $\Pr(M = 0, 1 \dots 16 | f = 3/4) [\varphi = M/16]$ .

### A few technical considerations

It is convenient here to choose for the number of successes  $M$  a *Beta-Binomial* prior distribution. A *Beta-Binomial* [*BBin*] distribution is a discrete probability distribution. Formally,  $X$  has the *BBin*( $u, v, K$ ) distribution if

$$\Pr(X = x) = \frac{1}{B(u, v)} \frac{\Gamma(K+1)\Gamma(x+u)\Gamma(K+v-x)}{\Gamma(x+1)\Gamma(K-x+1)\Gamma(K+u+v)} \quad [0 \leq x \leq K]$$

where  $\Gamma(z)$  is the gamma function and  $B(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$  is the beta function.

Its mean is

$$\frac{u}{u+v} K.$$

The advantage is that it is a *conjugate* distribution of the Hypergeometric distribution: after having observed  $a$  successes out of  $n$  (which implies  $a \leq M \leq N - n + a$ ), the number of successes  $M' = M - a$  [ $a \leq M' \leq N - M$ ] in the unknown part of the population is also a *Beta-Binomial* distribution. So, assuming the prior

$$M \sim \text{BBin}(a_0, b_0, N),$$

the posterior distribution is given by

$$M' | a \sim BBin(a + a_0, n - a + b_0, N - n) \quad [M = a + M'].$$

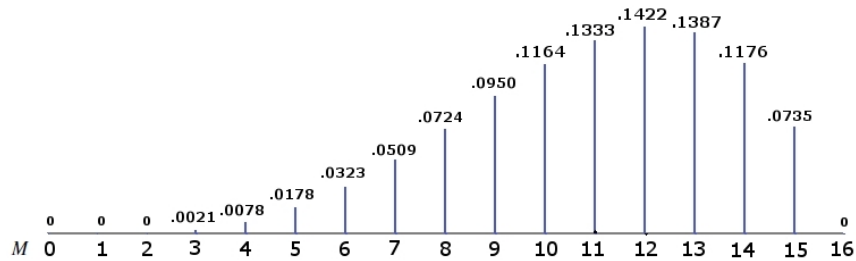
The prior weights  $a_0$  and  $b_0$  are simply added to the observed counts  $a$  and  $n - a$ . The above example of an opinion-based prior (see Figure 2.1) corresponds to the case  $a_0 = 12$  and  $b_0 = 4$ . This  $BBin(12, 4, 16)$  distribution has mean 12, hence an expected prior rate  $\bar{\varphi} = 12/16 = .75$ . The posterior distribution for  $\varphi$  in Figure 2.2 follows from the posterior for  $M' = M - 3$  ( $0 \leq M' \leq 12$ ), which is the  $BBin(15, 5, 12)$  distribution, with mean 9. Consequently, the expected posterior rate is  $\bar{\varphi} = (3+9)/16 = .75$  (unchanged).

### 2.3.3 A “No Information Initially” Analysis

Even if it is in accordance with their opinion, our colleagues doubt that the above analysis can convince the scientific community. We suggest them to act as if they have *no information initially* (in Jeffreys’s terms), and to consider a vague prior distribution.

#### The uniform prior

Typically, within the Beta-Binomial family, such a distribution is defined by small weights  $a_0$  and  $b_0$ , included between 0 and 1. In particular, the uniform prior for  $\varphi$ , which assigns probabilities  $1/17$  on all possible values  $0/16, 1/16, 2/16, \dots, 16/16$ , is the  $BBin(1, 1, 16)$  distribution for  $M$ . For this prior, the posterior distribution for  $M'$  is  $BBin(4, 2, 12)$ , with mean 8. The corresponding posterior probabilities for  $M$ , or equivalently for  $\varphi = M/16$ , are given in Figure 2.3. They follow from the posterior distribution  $BBin(4, 2, 12)$  for  $M' = M - 3$ . The posterior mean is  $\bar{\varphi} = (3+8)/16 = .6875$ .



**Fig. 2.3** Uniform prior analysis: Posterior probabilities  $Pr(M = 0, 1 \dots 16 | f = 3/4) [\varphi = M/16]$ .



Clearly, we are not interested in the probabilities of the particular values .25, .50, .75 that have been used for defining the statistical hypotheses about  $\varphi$ . What is needed is to evaluate the plausibility of specified regions of interest for  $\varphi$  bracketed by these values. For instance, we have the posterior probabilities:

$$\Pr(\varphi \leq 4/16 \mid f = 3/4) = .100 \quad \Pr(\varphi > 8/16 \mid f = 3/4) = .817 \quad \Pr(\varphi > 12/16 \mid f = 3/4) = .330$$

### Bayesian procedures are no more arbitrary than frequentist ones

Frequentist methods are full of more or less ad hoc conventions. Thus, in our example, the  $p$ -value has been computed as the sampling probability of getting a value *as least as extreme* as the observed success rate (under the null hypothesis). The convention to include the observed success rate results in a *conservative* test: if the null hypothesis is true, this test is significant (rejects the null hypothesis) for less than 5% of the samples. On the contrary, if the observed rate would be excluded, the test would be significant for more than 5% of the samples, else *anti-conservative*.

This choice has an exact counterpart in the Bayesian approach. For the prior weights  $a_0 = 0$  and  $b_0 = 1$ , we have the posterior distribution for  $M' = M - 3$ ,  $BB(3, 2, 12)$ , and the posterior probabilities:

$$\Pr(\varphi \leq 4/16 \mid f = 3/4) = .027 \quad \Pr(\varphi > 8/16 \mid f = 3/4) = .715 \quad \Pr(\varphi > 12/16 \mid f = 3/4) = .245$$

so that the posterior probability  $\Pr(\varphi \leq .25 \mid f = 3/4) = .0269$  is exactly equal to the  $p$ -value of the significance test of the null hypothesis  $H_0 : \varphi = .25$ . The numerical results coincide, but the Bayesian interpretation clearly shows that a non-significant outcome cannot be interpreted as “proof of no effect.” Our two colleagues are very intrigued by this result.

Many potential users of Bayesian methods continue to think that they are too subjective to be scientifically acceptable. The Bayesian interpretation of the  $p$ -value, and consequently of the frequentist confidence level, in terms of data dependent probabilities, clearly show that it is not the case: the “no information initially” analysis is no less objective than frequentist inference.

### Some remarks about exchangeability and hierarchical models

Many Bayesians place emphasis on the notion of exchangeability, introduced by de Finetti (1937), which can be viewed as a counterpart to the frequentist notion of repeated trials. According to de Finetti,

if we assume the [random elements]  $X_h$  to be exchangeable, this means that we attribute the same probability to an assertion about any given number of them, no matter how their indices are chosen or in what order (de Finetti, 1972, page 213).

The practical implications of exchangeability for designing experiments and analyzing data were examined in the above mentioned Food and Drug Administration guidelines.

In a clinical trial, patients within the trial are usually assumed to be exchangeable. [...] If patients in the trial are exchangeable with patients in the population from which they were sampled (e.g. the intended use population), then inferences can be made about the population on the basis of data observed on the trial patients. Thus, the concept of a representative sample can be expressed in terms of exchangeability (Food and Drug Administration, 2010, p. 17).

So, in our illustrative example, the probability of getting any sequence of successes and failures, given by the hypergeometric distribution, depends only on the number of successes and failures. It does not depend on the order in which the outcomes were observed. Future students must be assumed to be exchangeable with the students who have already been observed in order to make predictive probabilities reasonable. In the same way, similar experiments must be assumed to be exchangeable for a coherent integration of the information.

The assumption of trial exchangeability enables the current trial to “borrow strength” from the previous trials, while acknowledging that the trials are not identical in all respects. (Food and Drug Administration, 2010, p. 17).

Exchangeability is a key concept in the Bayesian framework. Using multilevel prior specifications, it allows a flexible modeling of related experimental devices by means of *hierarchical models*.

When incorporating prior information from a previous study, the patients in the previous study are rarely considered exchangeable with the patients in the current study. Instead, a hierarchical model is often used to “borrow strength” from the previous studies. At the first level of the hierarchy, these models assume that patients are exchangeable within a study but not across studies. At a second level of the hierarchy, the previous studies are assumed to be exchangeable with the current study, which acknowledges variation between studies (Food and Drug Administration, 2010, p. 23).

Hierarchical models are useful for analyzing the data from a multi-center experiment. They are also particularly suitable for *meta-analysis* in which we have data from a number of relevant studies that may be exchangeable on some levels but not on others.

## Epilogue

We have an interpretation for the fact that the teaching method has been effective for three of the four individuals. Indeed, these students had been exposed to an introduction to Bayesian inference before they received the new teaching method. We completely agree with Berry (1997), who ironically concludes that students exposed to a Bayesian approach come to understand  $p$ -values and confidence intervals better than do students exposed only to a frequentist approach.

Since the frequentist definition seems to make probability an objective property, existing in the nature independently of us, frequentists self proclaim to be *objective*. Most of them firmly reject the Bayesian inference as being *necessarily subjective*. However, the Bayesian definition can also serve to describe “objective knowledge”, in particular based on symmetry arguments or on frequency data.

Use the quite natural Bayesian interpretations of significance tests and confidence intervals: you will more clearly understand the common misuses and abuses of NHST and you will be able to overcome the usual difficulties encountered with the frequentist approach.



## Chapter 3

# The Fisher, Neyman-Pearson and Jeffreys views of Statistical Tests

This chapter briefly reviews the rationale of the three main views of statistical tests. Current practice is based on the Fisher “test of significance” and the Neyman-Pearson “hypothesis test.” Jeffreys’ approach is a Bayesian alternative based on the use of “objective” prior probabilities of hypotheses. The main similarities and dissimilarities of these three approaches will be considered from a methodological point of view: what is the aim of statistical inference, what is the relevance of significance tests in experimental research? The dangers inherent in uncritical application of the Neyman-Pearson approach will also be stressed.

### 3.1 The Fisher Test of Significance

Sir Ronald Fisher’s primary field was genetics, but he also made decisive contributions to statistics. His three books *Statistical Methods for Research Workers* (Fisher, 1990a), *The Design of Experiments* (Fisher, 1990b) and *Statistical Methods, Experimental Design, and Scientific Inference* (Fisher, 1990c) were first published in 1925, 1935 and 1956 respectively. They were primarily intended for scientific workers and they received considerable success and positive feedbacks.

#### An objective method for reporting experimental results

Fisher expanded the practices already in use: the celebrated Karl Pearson’s chi-square and Student’s  $t$  papers were respectively published in 1900 and 1908. He structured them into a new paradigm, the test of significance, presented as an objective method for reporting experimental results:

Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an *objective basis* in that the probability statement on which it is based is a fact *communicable to, and verifiable by, other rational minds* (Fisher, 1990c, p. 46, italics added).

### The null hypothesis

A single hypothesis, called the “null hypothesis,” is challenged:

the hypothesis that the phenomenon to be demonstrated is in fact absent (Fisher, 1990b, p. 13),

It is a hypothesis to be disproved, to be *nullified*, and not necessarily the hypothesis that the parameter has a null value, even if this is the most usual case.

### The outcome of the test of significance

The experimental result is judged to be

1. either *significant*, the null hypothesis is disproved;
2. or *nonsignificant*, the null hypothesis is not disproved.

### The test statistic and the level of significance $p$

An appropriate test statistic, whose sampling distribution when the null hypothesis is true is exactly known, is considered. Once the data have been collected, its observed value is calculated. The sampling distribution gives the probability that this observed value “is exceeded by chance” (Fisher, 1990b, p. 38), *if the null hypothesis is true*. It is the *level of significance*, nowadays called the  $p$ -value. The experimental result is judged to be *significant* when  $p$  is considered to be small enough. This is a consequence of the logical disjunction:

*Either* the hypothesis is untrue, *or* the value of  $\chi^2$  has attained by chance an exceptionally high value (Fisher, 1990a, p. 80).

### How to evaluate the smallness of $p$ ?

Fisher often used 5% as a reasonable, convenient, threshold for evaluating  $p$ ,

... it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not (Fisher, 1990a, p. 44),

However, he came to firmly reject the conception of an absolute, fixed level and he even stated the possibility of using different levels for the same data:

The value for  $Q$  is therefore significant on the higher standard (1 per cent) and that for  $N_2$  at the lower standard (5 per cent) (Fisher, 1990b, pp. 152-153).

For Fisher, the level of significance (the  $p$ -value) is a fundamental characteristic and its actual value for *the particular data* under consideration

indicates the *strength of the evidence* against the [null] hypothesis (Fisher, 1990a, p. 80, italics added).

## 3.2 The Neyman-Pearson Hypothesis Test

Since it does not provide precise guidelines to decide about research hypotheses, Fisher's test of significance can seem frustrating.

### Rational decision rules

Jerzy Neyman, a Polish mathematician, and Egon Pearson (Karl Pearson's son), a British statistician, collaborated with the aim to give rules of rational behavior for *taking statistical decisions* about hypotheses. Their basic articles were published in 1928 and 1933. This collaboration on hypothesis tests led Neyman later to formulate his method of confidence intervals within the same perspective.

### The hypothesis to be tested and alternative hypotheses

Neyman and Pearson rejected Fisher's conception of a single hypothesis and emphasized the necessity of alternative hypotheses:

It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from [the population] would not in itself justify the rejection of the hypothesis that it has been so drawn, *if there were no other more probable hypotheses conceivable* (Neyman & Pearson, 1928, p. 178, italics added).

They considered mutually exclusive hypotheses and introduced for them explicit notations,  $H_0, H_1, \dots, H_m$ ,  $H_0$  being called *the hypothesis to be tested* (Neyman & Pearson, 1933b, p. 495).

### The outcome of the hypothesis test

Given the sample space  $W$  and the sample point (the observed event)  $\Sigma$ , a hypothesis test is a decision rule, based on the division of the sample space into two regions (Neyman and Pearson, 1933b, p. 493):

- (a) Reject  $H_0$  if  $\Sigma$  falls into the *critical region*  $w$ ;
- (b) Accept  $H_0$  if  $\Sigma$  falls into the *region of acceptance*  $w' = W - w$

The no-decision case – *remain in doubt* – was envisaged as a further subdivision of the region of acceptance  $w'$ , but not really treated.

### A long-run control

The problem to be solved by a hypothesis test is to control the errors in the following sense.

If he [the practical statistician] makes *repeated use of the same statistical tools* when faced with a similar set of admissible hypotheses, in what sense can he be sure of certain long run results? A certain proportion of correct decisions, A certain proportion of errors, and if he so formulates the tests a certain proportion of cases left in doubt? (Neyman & Pearson, 1933b, p. 494, italics added).

So, when repeated under identical circumstances, the test is viewed as *a rule of rational behavior*. It is intended to minimize the *long-run* proportion of erroneous decisions regarding the hypotheses considered.

### Two types of errors and their long-run frequencies

There are two types of errors (Neyman and Pearson, 1933b, p. 493):

Type I we reject  $H_0$  when it is true;

Type II we accept  $H_0$  when some alternative  $H_i$ , is true

In most applications, there exists an essential difference in nature between the two types of errors, the Type I being the most important to avoid. Neyman and Pearson suggested that all errors of Type I may be regarded as equivalent, because if  $H_0$  is wrongly rejected the consequences are generally the same whatever the sample size. On the contrary, if  $H_0$  is wrongly accepted the consequences depend on the true alternative  $H_i$ :

Generally it will not be of serious consequence if we accept  $H_0$  falsely when the true hypothesis  $H_i$  only differs only very slightly, but the danger will steadily increase as this difference increases (Neyman & Pearson, 1933b, p. 497).

### Power of the test and best critical region

The *power* of a critical region  $w$  of size  $\alpha$  with regard to an alternative simple hypothesis  $H_i$  is the probability  $p(w|H_i)$  of rejecting the hypothesis tested  $H_0$  when the true hypothesis is  $H_i$ . The “best critical region”  $w_0$  maximizes the power  $P(w|H_i)$  under the condition that  $P(w_0|H_0)$  is fixed. If  $w_0$  possesses this property for a certain class of alternatives, it is called “a best critical region for the whole class”. The now-famous Neyman-Pearson lemma (1933a, 1936b) provides, at least under certain conditions (testing point hypotheses), a way of finding a best critical region, which defines a “uniformly most powerful” test.

Neyman and Pearson (1933a and 1933b) called  $P(w|H_0)$  the “size of the critical region” and denoted it by  $\varepsilon$ . The notation  $\alpha$ , now used for the Type I error rate, was introduced later in Neyman and Pearson (1936a). Neyman (1938, p. 79) introduced



the symbol  $\beta(\theta|w_n)$  to designate the “power function of the critical region  $w_n$ ,” a function of the parameter value  $\theta$ .  $\beta$  is now used for the Type II error rate (and  $1 - \beta$  for the power).

In practice, a conventional rule is that  $\alpha$  is generally set at .05. The problem of computing the power of a test is much more complicated, since there is not a single  $\beta$  (see Neyman, 1977, p. 107). One reason is that the alternative hypothesis is usually composite. Furthermore, the power function depends on the choice of  $\alpha$ .

In the Neyman-Pearson approach, the hypotheses, the  $\alpha$ -level and the critical region are fixed *before observations*. The data serve only to determine whether or not they fall in the critical region. To be coherent with this conception, the  $p$ -value should even not be considered:

... a  $p$ -value from a Fisherian significance test has no place in the Neyman-Pearson hypothesis-testing framework (Hubbard, 2004, p. 320).

### 3.3 The Jeffreys Bayesian Approach to Testing

Sir Harold Jeffreys was a world authority in mathematical physics and theoretical geophysics. His two statistical books, *Scientific Inference* (Jeffreys, 1973) and *Theory of Probability* (Jeffreys, 1967), were first published in 1931 and 1939 respectively. The later can be viewed as the first attempt to develop a fundamental theory of statistical inference based on the Bayesian approach. Two extended editions appeared in 1948 and 1961 (reprinted in 1967 with some corrections and in 1998).

Following the lead of Bayes (1763) and Laplace (Laplace, 1840), Jeffreys worked at developing objective Bayesian methods, applicable when nothing is known about the value of the parameter (“no information initially”):

we are aiming chiefly at a theory that can be used in the early stage of a subject (Jeffreys, 1967, p. 252).

Bayesian prior probabilities are used for this purpose:

The answer is really clear enough when it is recognized that a probability is merely a number associated with a degree of reasonable confidence and has no purpose except to give it a formal expression. If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none (Jeffreys, 1967, p. 118).

### The Jeffreys rule

...how can we assign the prior probability when we know nothing about the value of the parameter, except the very vague knowledge just indicated? (Jeffreys, 1967, p. 118).

The so-called *Jeffreys' rule*, based on the Fisher information, is used to obtain a prior that is appropriate to answer this question. This prior has the essential property to be invariant under one-to-one reparameterization. For instance, for a Normal sampling distribution  $N(\mu, \sigma^2)$ , the Jeffreys prior is uniform for  $(\mu, \log(\sigma^2))$ . It is noteworthy to mention the work of Ernest Lhoste, a captain in the French army, who developed a similar approach concerning Bayesian inference and the choice of the prior. Several years before, he derived results identical to those of Jeffreys for the Normal distribution (Lhoste, 1923; see Broemeling & Broemeling, 2003).

The Jeffreys prior, usually called noninformative, objective or default, is a reasonable choice in most usual situations of experimental data analysis. In more complex situations, its use is more controversial and alternative approaches have been developed (for a recent review, see Ghosh, 2011).

### The function of significance tests

For Jeffreys the function of significance tests was

to compare a suggested value of a new parameter, often 0, with the aggregate of other possible values (Jeffreys, 1967, p. 245).

Consequently, he considered two complementary hypotheses, he denoted by  $q$  and  $q'$ :

$q$ , that the parameter has the suggested value, and  $q'$ , that it has some other value to be determined from the observations (Jeffreys, 1967, p. 246).

As Fisher he used the term null hypothesis:

We shall call  $q$  the *null hypothesis*, following Fisher, and  $q'$  the *alternative hypothesis* (Jeffreys, 1967, p. 246).

It is worthwhile to note that this sentence has replaced the original one: “ $q$  would always be what Fisher calls the null hypothesis”, which appeared in the first edition (1939, p. 194).

### A specific prior for testing precise hypothesis

A uniform (and more generally a continuous) prior distribution is inappropriate for testing of a *precise* (“point null”) hypothesis, since it gives it a zero probability:

The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. If a new parameter is being considered, the uniform distribution of prior probability for it would practically always lead to the result that the most probable value is different from zero (Jeffreys, 1967, p. 117).

In order “to say that we have no information initially,” it seemed an evidence to Jeffreys that the two hypotheses are initially equally probable:

The essential feature is that we express ignorance of whether the new parameter is needed by taking half the prior probability for it as concentrated in the value indicated by the null hypothesis and distributing the other half over the range possible (Jeffreys, 1967, p. 246).

Consequently, if  $H$  is “the set of propositions accepted throughout an investigation,”

we must take  $P(q|H) = P(q'|H) = \frac{1}{2}$  (Jeffreys, 1967, p. 246).

A prior that does not favor any particular parameter value is used on the complementary alternative hypothesis. For usual sample sizes, it follows that when the null hypothesis is rejected by a frequentist test (small  $p$ -value), the Bayesian posterior probability of the null hypothesis is generally dramatically higher than the  $p$ -value. For Berger (2003, p. 3), this demonstrates that “the too-common misinterpretation of  $p$ -values as error probabilities very often results in considerable overstatement of the evidence against  $H_0$ .”

### A measure of evidence against the null hypothesis

Jeffreys suggested to measure evidence against the null hypothesis with the ratio of posterior to prior odds, in his notations:

$$K = \frac{P(q|\theta H)}{P(q'|\theta H)} / \frac{P(q|H)}{P(q'|H)}$$

where  $\theta$  is the “observational evidence.”  $K$  is nowadays called the *Bayes factor* and reduces to the *likelihood ratio* if  $P(q|H) = P(q'|H) = \frac{1}{2}$ .

For practical purposes, Jeffreys (1967, p. 432) proposed to “grade the decisiveness of the evidence” as follows:

Grade 0	$K > 1$	Null hypothesis supported
Grade 1	$1 > K > 0.3162$ ( $10^{-1/2}$ )	Evidence against $q$ , but not worth more than a bare mention
Grade 2	$0.3162 > K > 0.1$	Evidence against $q$ substantial
Grade 3	$0.1 > K > 0.0316$ ( $10^{-3/2}$ )	Evidence against $q$ strong
Grade 4	$0.0316 > K > 0.01$	Evidence against $q$ very strong
Grade 5	$0.01 > K$	Evidence against $q$ decisive

### An averaged risk of error

Jeffreys criticized the Neyman-Pearson approach:

I do not think that they have stated the question correctly (Jeffreys, 1967, p. 395).

He advocated the use of an averaged risk of errors, where the averaging is performed over the possible values of the parameter, according to their Bayesian probability:

But if the actual value is unknown the value of the power function is also unknown; the total risk of errors of the second kind must be compounded of the power functions over the possible values with regard to their risk of occurrence (Jeffreys, 1967, p. 396).

Neyman and Pearson also considered this notion, under the name of *resultant power*, but they discarded it because it is dependent of the probabilities *a priori* and cannot often be known:

It is seen that while the power of a test with regard to a given alternative  $H_i$  is independent of the probabilities *a priori*, and is therefore known precisely as soon as  $H_i$  and  $w$  [the critical region of the test] are specified, this is not the case with the resultant power (Neyman and Pearson, 1933b, p. 499).

However, this objection is essentially theoretical, and Jeffreys' criticism appears to be relevant, since the specified  $H_i$  is hypothetical.

Nowadays, many authors dogmatically oppose the Jeffreys and Fisher approaches to testing, claiming that they can lead to quite different conclusions in actual practice:

The discrepancy between the numbers reported by Fisher [the *p*-value] and Jeffreys [the Bayes factor] are dramatic (Berger, 2003, p. 1).

However, in so far as experimental data analysis is concerned, this was not the Jeffreys viewpoint!

In spite of the difference in principle between my tests and those based on the P integrals [Fisher's tests]. . . it appears that there is not much difference in the practical recommendations (Jeffreys, 1967, p. 435).

## 3.4 Different Views of Statistical Inference

### 3.4.1 Different Scopes of Applications: The Aim of Statistical Inference

#### To avoid wrong decisions

Although they discussed the case of "scientific investigation," the typical example of application that Neyman and Pearson (1933b) considered in detail for illustration concerned the process of *quality control*:

$H_0$  is the hypothesis that the consignment which is sampled is of quality above a certain standard. From the *producer's* point of view it is important that the sample should not be

rejected when  $H_0$  is true; he wishes  $P_I$  [the Type I error rate] to be reduced to a minimum. To the *consumer* on the other hand it is important that the sample should not pass the test when  $H_0$  is false, the quality of the consignment being below standard; his object will be to reduce  $P_{II}$  [the Type II error rate] (Neyman & Pearson, 1933b, p. 498).

Consequently, when answering questions raised by hypothesis testing, their main concern was to avoid errors in decisions:

Any attempts to answer will be associated with a wish to avoid being wrong (Neyman, in Fisher, 1935, pp. 75-76).

### Learning from data and experience

Fisher did not dispute this and acknowledged the usefulness of “acceptance” tests for decision making in some fields:

In various ways what are known as acceptance procedures are of great importance in the modern world. When a large concern such as the Royal Navy receives material from its makers, it is, *I suppose*, subjected to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments (Fisher, 1990c, p. 80, italics added).

However, “I suppose” clearly reveals that Fisher felt himself to be not concerned with such kinds of applications. For him, the attempt to reinterpret the test of significance as a means of making decisions was not suitable for experimental research:

It is not therefore at all in disdain of an artifice of proved value, in commerce and technology, that I shall emphasize some of the various ways in which this operation differs from that by which improved theoretical knowledge is sought in experimental research (Fisher, 1990c, pp. 79-80).

Even if Fisher gave examples of observational data, his main concern was the “experimental sciences,” and especially “the natural sciences:”

It is noteworthy, too, that the men who felt the need for these tests [of significance] who first conceived them, or later made them mathematically precise, were all actively concerned with researches in the natural sciences (Fisher, 1990c, pp. 79-80).

Consequently, he considered that the aim of tests of significance was not to take decisions, but *to learn from experimental data*:

The conclusions drawn from such tests constitute the steps by which the research worker gains a better understanding of his experimental material, and of the problems which it presents (Fisher, 1990c, p. 79).

Jeffreys went further and aimed at proposing a general methodology for learning from data and experience, applicable to research in all fields of science. Bayesian probabilities are successively updated when new data become available:

Starting with any distribution of prior probability and taking account of successive batches of data by the principle of inverse probability, we shall in any case be able to develop an account of the corresponding probability at any assigned state of knowledge (Jeffreys, 1967, p. 118).

Jeffreys distinguished *estimation problems*,

concerned with the estimation of the parameters in a law, the form of the law itself being given (Jeffreys, 1967, p. 245), [in which] we want the probability distribution of these parameters, given the observations (Jeffreys, 1967, p. 117),

from *significance tests*, which involve

a specially suggested value of a new parameter (Jeffreys, 1967, p. 246).

So, he had a conception of significance tests, related to what is commonly referred to as “model selection:”

The function of significance tests is to provide a way of arriving, in suitable cases, at a decision that at least one new parameter is needed to give an adequate representation of the data and valid inferences to future ones (Jeffreys, 1967, p. 245).

However, for Jeffreys, the question asked in a significance test,

Is the new parameter supported by the observations, or is any variation expressible by it better interpreted as random? (Jeffreys, 1967, p. 245).

was not relevant in “agricultural experiments”, which he regarded *to be very largely problems of pure estimation*. We will consider his views further in Chapter 5.

It should be recognized that, according to his approach to statistical inference for experimental data, Fisher

seems to have assigned the tests a rule-of-thumb status as largely preparatory to estimation (Rosenkrantz, 1973, p. 304),

a status which was unfortunately ignored in practice. Indeed, the hypothesis test method has attracted the interest of experimental scientists, because they have been unduly

encouraged to expect final and definite answers from their experiments in situations in which only slow and careful accumulation of information could be hoped for. And some of them, indeed, came to regard the achievement of a significant result as an end in itself (Yates, 1964, p. 320).

### 3.4.2 The Role of Bayesian Probabilities

As a Bayesian, Jeffreys considered probability as *the degree of confidence that we may reasonably have in a proposition*. Moreover, because our degrees of confidence in a proposition change when new observations or new evidence become available, he stated that probability is always conditional and “must be of [a proposition]  $p$  on data  $q$ :”

It is no more valid to speak of the probability of a proposition without stating the data than it would be to speak of the value of  $x + y$  for a given  $x$ , irrespective of the value of  $y$  (Jeffreys, 1967, p. 15).

He used the formal notation  $P(p|q)$  to mark the fundamental role of this conditional probability.

### Fisher and Bayes

Fisher has always acknowledged that the Bayesian argument should be used “when knowledge *a priori* in the form of mathematically exact probability statements is available” (Fisher, 1990b/1935 p. 198). What he contested was the relevance of this case in scientific research:

A more important question, however, is whether in scientific research, and especially in the interpretation of experiments, there is cogent reason for inserting a corresponding expression representing probabilities *a priori* (Fisher 1990c, p. 17).

His aim was to avoid the use of prior probabilities about hypotheses:

Familiarity with the actual use made of statistical inference in the experimental sciences shows that in the vast majority of cases the work is completed without any statement of mathematical probability being made about the hypothesis under consideration (Fisher 1990c, p. 40).

Nevertheless, Fisher considered the probability level (the  $p$ -value) as characterizing a “unique sample”. Actually, he defined the  $p$ -value of the  $t$  test (Fisher, 1990a, p. 118) as a *predictive probability*, and not as a frequentist probability: see Lecoutre Lecoutre, Lecoutre & Poitevineau, 2010, pp. 161–162. Furthermore, Fisher (1959) came later to write very explicitly that he used probability as a measure of degree of uncertainty:

The subject of a probability statement if we know what we are talking about, is singular and unique; we have some degree of uncertainty about its value, and it so happens that we can specify the exact nature and extent of our uncertainty by means of the concept of Mathematical Probability as developed by the great mathematicians of the 17<sup>th</sup> century Fermat, Pascal, Leibnitz, Bernoulli and their immediate followers (Fisher, 1959, p. 22).

### Neyman and Pearson and Bayes

Neyman and Pearson’s general principles underlying their “most efficient tests of statistical hypotheses” pertain to the same preoccupation to avoid the use of prior probabilities. Their aim was to find

what statements of value to the statistician in reaching his final judgment can be made from an analysis of observed data, which would not be modified by any change in the probabilities *a priori* (Neyman and Pearson, 1933b, p. 492).

However, it must be emphasized that, in their early writings, Neyman and Pearson acknowledged the dualistic view on probability:

In the long run of statistical experience the frequency of the first source of error (*or in a single instance its probability*) can be controlled... (Neyman and Pearson, 1928, p. 177, italics added).

Even if Neyman explicitly aired his opposition to Fisher, and clearly advocated a frequentist conception of probability,

For Fisher, probability appears as a measure of uncertainty applicable in certain cases but, regrettably, not in all cases. For me, it is solely the answer to the question ‘How frequently this or that happens’ (Neyman, 1952, p. 187),

he also emphasized later that it was not a systematic opposition to the use of Bayesian inference:

Perhaps because of lack of clarity in some of my papers, certain authors appear to be under the impression that, for some reason, I condemn the use of Bayes’ formula and that I am opposed to any consideration of probabilities a priori. This is a misunderstanding. What I am opposed to is the dogmatism which is occasionally apparent in the application of Bayes’ formula when the probabilities a priori are not implied by the problem treated and an author attempts to impose on the consumer of statistical methods the particular a priori probabilities invented by himself for this particular purpose (Neyman, 1957, p. 19).

It should at the least be agreed that Fisher’s conception of probability

was in fact much closer to the ‘objective Bayesian’ position than that of the frequentist Neyman (Zabell, 1992, p. 381).

### 3.4.3 Statistical Tests: Judgment, Action or Decision?

#### Fisher: An aid to judgment

For Fisher, statistical inference involved both deductive and inductive reasoning:

The statistical examination of a body of data is thus *logically similar to the general alternation of inductive and deductive method* throughout the sciences. A hypothesis is conceived and defined with all necessary exactitude; its logical consequences are ascertained by a deductive argument; these consequences are compared with the available observations; if these are completely in accord with the deductions, the hypothesis is justified at least until fresh and more stringent observations are available (Fisher, 1990a, p. 8, italics added).

Within this perspective, the tests of significance constitute an “aid to judgment”:

for the tests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests, or ‘decision functions’ (Fisher, 1990a, p. 128).



**Neyman-Pearson: Automatic decisions viewed as inductive behavior**

Neyman and Pearson emphasized the fact that a hypothesis test is not intended to make a judgment about the truth or falsity of a hypothesis:

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis (Neyman and Pearson, 1933a, pp. 290-291).

Contrary to Fisher's view of statistical reasoning, their approach is only deductive. When he developed later the notion of confidence interval, Neyman came to reject the phrase "inductive reasoning:"

the term inductive reasoning does not seem appropriate to describe the new method of estimation because all the reasoning behind this method is clearly deductive (Neyman, 1951, p. 85).

He introduced (Neyman, 1938) the term "comportement inductif" (*inductive behavior*), by opposition to Fisher's inductive reasoning:

... the term 'inductive reasoning' is out of place and, if one wants to keep the adjective 'inductive', it seems most appropriate to attach to it the noun 'behavior' (Neyman, 1951, p. 85).

Fisher has always expressed his opposition to an approach that leads to automatic decisions in scientific research:

The idea that this responsibility [the detailed interpretation of verifiable observations] can be delegated to a giant computer programmed with Decision Functions belongs to a phantasy of circles rather remote from scientific research (Fisher, 1990c, p. 105).

**Jeffreys: A decision based on a measure of evidence**

As Fisher, Jeffreys was convinced that deductive logic is insufficient for the scientific method:

I reject the attempt to reduce induction to deduction (Jeffreys, 1967, p. B).

Although the role of the test is to decide if a new parameter is needed, this is not an automatic decision making procedure. Rather the decision is based on a measure of evidence against the null hypothesis that leads to a graduate judgment.

In regard to the effective use of the statistical test method in experimental research, we can agree with Rozeboom that

its most basic error lies in mistaking the aim of a scientific investigation to be a decision, rather than a cognitive evaluation of propositions (Rozeboom, 1960, p. 428).

However, it should be recognized that this use is not in agreement with the Fisher and Jeffreys views. Actually, significance tests, as commonly used, are uninformative because

In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences (Cochran & Cox, 1957, p. 5).

### 3.5 Is It possible to Unify the Fisher and Neyman-Pearson Approaches?

Lehmann argued that a unified approach is possible:

despite basic philosophical differences, in their main practical aspects the two theories are complementary rather than contradictory and that a unified approach is possible that combines the best features of both (Lehmann, 1993, p. 1242).

This was seriously questioned by Perlman and Wu (1999), who showed that, in several composite null hypothesis testing problems, optimal tests in the Neyman-Pearson sense are flawed.

#### Demonstrating equivalence

This issue has been specifically investigated in the framework of clinical equivalence trials. In order to evaluate equivalence between two experimental treatments, a small, positive value  $\Delta$  is used to define an “equivalence region”  $[-\Delta, \Delta]$  for the difference  $\mu_1 - \mu_2$  between the two treatment means. An appropriate hypothesis test procedure for demonstrating equivalence is to consider the composite null hypothesis  $H_0 : |\mu_1 - \mu_2| \geq \Delta$  (which is to be rejected), and the alternative  $H_a : |\mu_1 - \mu_2| < \Delta$  (“equivalence”).

The seemingly natural solution consists in using the absolute value of the usual  $t$ -test statistic (or, equivalently its square, the  $F$ -ratio). Then the test is to reject  $H_0$  (to demonstrate equivalence) if  $|t|$  is small enough, formally if  $|t|$  is smaller than the  $(1 - \alpha)\%$  lower point of its sampling distribution given  $|\mu_1 - \mu_2| = \Delta$ , that is the absolute value of a noncentral  $t$ -distribution (or equivalently, if  $F$  is used, a non-central  $F$ -distribution). When the error variance is known, this test is the uniformly most powerful test for testing  $H_0$  against the alternative  $H_a$ .

### Neyman-Pearson's criterion leads to incoherent and inadmissible procedures

As a matter of fact this test, and many other closely related tests, have always been considered unacceptable and rejected by applied statisticians (e.g. Selwyn, Hall & Dempster, 1985; Schuirman, 1987; Food and Drug Administration, 2001; Lecoutre & Derzko, 2001, 2014).

- When the observed difference is null, the observed significance level is always null, whatever  $\Delta$ , the sample size and  $\sigma$ , leading to the automatic conclusion of equivalence (rejection of  $H_0$ ).
- For a given observed difference, the critical rejection region varies in a nonmonotonic way as a function of the sampling error variance. Moreover, it may include values of the observed difference that lie *outside* the equivalence region.
- Schervish (1995, problem 42, p. 291) demonstrated that they are incoherent and inadmissible in the sense of the decision theory.

### Theoretical debates: counterintuition or good sense?

The defenders of the Neyman-Pearson “optimal tests” dismissed the warnings made by applied statisticians who have rejected their use in clinical equivalence trials. They argued that their undesirable properties are only “counterintuitions”:

we believe that notions of size, power, and unbiasedness are more fundamental than ‘intuition’ (Berger & Hsu, 1996, p. 192).

This was seriously challenged by Perlman and Wu (1999), who demonstrated that such tests are *scientifically inappropriate*.

Perlman and Wu advocated the pragmatism and good sense of Fisher, Cox and many others (among which we include Jeffreys):

we hope that we have alerted statisticians to the dangers inherent in uncritical application of the NP [Neyman & Pearson] criterion, and, more generally, convinced them to join Fisher, Cox and many others in carefully weighing the scientific relevance and logical consistency of any mathematical criterion proposed for statistical theory (Perlman & Wu, 1999, p. 381).

## 3.6 Concluding Remarks

Fisher's genius is recognized:

R.A. Fisher was certainly the hero of 20th century statistics (Efron, 1998, p. 95).

However and unsurprisingly, it was the so-called “sound mathematical foundation” of the Neyman-Pearson theory that attracted the interest of most frequentist statisticians.

Criteria such as most (or more) powerful test, unbiasedness or alpha-admissibility have hardened into dogma, often without concern for the needs of scientists. Is it more important to know that the two-sided  $t$ -test is uniformly most powerful among all unbiased tests or is it more important to ask if it is relevant?

In a similar way, Jeffreys’ approach has been embedded into a Bayesian decision-theoretic framework, which perpetuates the “reject/accept” dichotomy of significance tests, without concern for the role Jeffreys assigned to estimation in experimental data analysis.

## Chapter 4

# GHOST: An officially Recommended Practice

This chapter gives a brief account of the misuses and abuses of Null Hypothesis Significance Testing [NHST]. It also examines the most often recommended “good statistical practice”, which we call *Guidelined Hypotheses Official Significance Testing* [GHOST]. GHOST is a hybrid practice that appears as an amalgam of Fisherian and Neyman-Pearsonian views. It does not ban the use of significance testing, but the choice of the sample size should be justified and estimates of the size of effects and confidence intervals should also be reported.

### 4.1 Null Hypothesis Significance Testing

#### 4.1.1 An Amalgam

Most statisticians and scientific workers do not clearly distinguish the Fisher and Neyman-Pearson views. Furthermore, in statistics textbooks significance tests are often anonymously presented as a mixture of the two approaches, and controversies are ignored.

#### NHST: A hybrid logic

Following Cohen, this mixture is designated by the acronym NHST, for *Null Hypothesis Significance Testing*:

NHST; I resisted the temptation to call it statistical hypothesis inference testing (Cohen, 1994, p. 997).

It has no real, theoretical or methodological, justification and results in many problems. It has been denounced by Gigerenzer as the “hybrid logic” of statistical inference that Fisher, Neyman, and Pearson would all have rejected:

It is an incoherent mishmash of some of Fisher's ideas on one hand, and some of the ideas of Neyman and E.S. Pearson on the other (Gigerenzer, 1993, p. 314).

### A mixture of terms and notations

The terminology and notations reveal this amalgam. For instance, to speak of the “the null hypothesis  $H_0$ ” is so common that it is surprising to learn that Fisher never used the notation  $H_0$  and Neyman and Pearson never used “null hypothesis.” This reveals that these authors had very different views on its role.

The  $\alpha$ -level level is supposed to have been selected *a priori* (Neyman and Pearson), but several levels are implicitly used to qualify the outcome of the test, according to different reference values (Fisher). So it is a common practice to report the results as significant ( $p \leq .05$ ), highly significant ( $p \leq .01$ ), extremely significant ( $p \leq .001$ ), and even sometimes quasi-significant, marginally significant or near significance ( $.05 < p \leq .10$ ). In many publications, tables are labeled with stars to indicate degrees of significance.

Null Hypothesis Significance Testing is a unjustified amalgam of the Fisher and Neyman-Pearson views of statistical tests, which is nowadays the most frequently used statistical procedure in many, if not in most, scientific journals.

### 4.1.2 Misuses and Abuses

#### The dictatorship of significance

Experimental research can be compared to a game or a fight (Freeman, 1993, used the adjective “gladiatorial”): only the significant results win. Nonsignificant ones are theoretically only statements of ignorance, and thus perceived as failures, as illustrated by the common expression “we fail to reject the null hypothesis”. It must be recognized that Fisher has paved the way by writing:

Personally, the writer prefers to . . . *ignore entirely* all results which fail to reach that [significance] level (Fisher, 1926, p. 504, italics added)

#### A typical game

Here is an example of a typical game between an author and a referee.

1. **The author** (first version of the article): A  $t$ -test comparing the scores between the control condition and the three other ones showed each time a significant difference.
2. **The referee**: Has a correction (e.g. Bonferroni) been made for multiple comparisons?
3. **The author** (added in the final version): Dunnett's correction was applied to account for the use of multiple comparisons.

The referee expressed doubts: have the game rules been correctly applied? If the author had used the Bonferroni correction, one of the three  $t$ -tests had turned to be nonsignificant at .05 level (he had lost the game!). Fortunately, he was an experienced gamer and knew that Dunnett's correction was appropriate in this case (multiple comparisons to a same control group), leading to the magic *significant at .05 level* for the three tests.

The difference between “significant” and “not significant” is not itself statistically significant (Gelman & Stern, 2006).

### Interpreting significance as proof of no effect

Inappropriate null conclusions (“there is no effect”) based on nonsignificant tests are not uncommon, even in prestigious journals (e.g. Harcum, 1990). It is extremely surprising that some experimental publications, such as the *International Journal of Psychology*, explicitly instruct authors to adopt this improper practice:

Results of statistical tests should be given in the following form: “...results showed an effect of group,  $F(2,21) = 13.74$ ,  $MSE = 451.98$ ,  $p < .001$ , but *there was no effect of repeated trials*,  $F(5,105) = 1.44$ ,  $MSE = 17.70$ , and *no interaction*,  $F(10,105) = 1.34$ ,  $MSE = 17.70$ ” (International Journal of Psychology, 2014, italics added).

### The sizeless scientists and the star system

Many publications seem to have been written by *sizeless scientists*:

Sizeless scientists act as if they believe the *size* of an effect does not matter. In their hearts they do care about size, magnitude, oomph. But strangely they don't measure it (Ziliak & McCloskey, 2008, p. x).

The  $p$ -value, magnified by the ritual symbols \*, \*\*, and \*\*\* (Meehl, 1978), is used as an implicit substitute for judgment about the meaningfulness of research results:

Furthermore, even when an author makes no claim as to an effect size underlying a significant statistic, the reader can hardly avoid making an implicit judgment as to that effect size (Oakes, 1986, p. 86).

We face a paradoxical situation. On the one hand, NHST is often regarded as an objective criterion of scientificity. On the other hand, it leads to innumerable misuses – the misinterpretations of results being the most visible – and entails publication bias (e.g. Sterling, 1959) and considerable distortions in the designing and monitoring of experiments.

## 4.2 What About the Researcher's Point of View?

### A cognitive filing cabinet

NHST reporting practices (and their misuses) are undoubtedly reinforced by a natural cognitive tendency – and also a necessity – to take a position when being published. It follows that experimental results are in some way arranged in a *cognitive filing cabinet*, where significance goes under “there is an effect” and nonsignificance is improperly filed under “there is no effect” (see the significance hypothesis of Oakes, 1986). It is not really a *rule of behavior* in the sense of Neyman and Pearson or a *decision* in the sense of the Bayesian decision-theoretic approach.

### It is the norm

However, the users' attitudes are far from being as homogeneous as might be inferred from the reporting practices. This was revealed by our empirical studies about the way accustomed users – psychological researchers and professional applied statisticians – interpret NHST outcomes (Lecoutre, Lecoutre & Poitevineau, 2001; Poitevineau & Lecoutre, 2001; Lecoutre, Poitevineau & Lecoutre, 2003). Most users appear to have a real consciousness of the stranglehold of NHST: they use them because “**it is the norm**”.

When faced to experimental results, only a minority of accustomed users has a systematically clear-cut attitude. Actually, most NHST users try to qualify their interpretation in relation to other information.

## 4.3 An Official Good Statistical Practice

Changes in reporting experimental results are more and more enforced within guidelines and editorial policies. So the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use [ICH]



which brings together the regulatory authorities and pharmaceutical industry of Europe, Japan and the US, has developed guidelines for clinical trials (ICH E9 Expert Working Group, 2006).

### A clinical trial example

The following example of application will serve as a typical illustration of the recommended practice. It concerns the inference about a proportion in a prospective clinical trial (see Lecoutre, Derzko & Grouin, 1995) and is a direct extension of the example treated in Chapter 1. The patients under study were post-myocardial infarction subjects treated with a new drug, a low molecular weight heparin. The trial aimed at assessing the potential efficacy of this drug as a prophylaxis of an intracardiac left ventricular thrombosis.

The drug was expected to reduce thrombosis rate. It was considered that .70 was the success rate (no thrombosis) below which the drug would be of no interest and further development would be abandoned. Consequently, investigators planned a one-sided binomial test of  $H_0 : \phi = .70$  versus  $H_1 : \phi > .70$  at the prespecified significance level  $\alpha = .05$  (Type I error probability).  $H_0$  is the tested hypothesis and  $H_1$  is the alternative hypothesis, the set of all admissible hypotheses for the one-sided test.

#### 4.3.1 Guidelined Hypotheses Official Significance Testing

##### Sample size determination

The ICH E9 guidelines prescribe to determine an appropriate sample size  $n$ :

The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed (ICH E9 Expert Working Group, 2006, p. 19).

In our example, the power-based Neyman-Pearson approach is recommended. This needs to specify a particular value  $\tilde{\phi}$ , reflecting a “working” (alternative) hypothesis  $H_a$ . Due to its focus on two particular hypotheses, a *tested hypothesis* and a *working one*, we suggest to call this procedure “Guidelined Hypotheses Official Significance Testing”.

The investigators retained  $\tilde{\phi} = .85$ : the success rate above which the drug would be really attractive. This was in accordance with the often recommended practice to base the choice on “a judgment concerning the minimal effect which has clinical relevance” (ICH E9 Expert Working Group, 2006, p. 19). A target value .80 of the power of the test when  $\phi = .85$  was specified, that is, a maximum Type II error probability  $\beta = .20$ .

Here, there is no additional (“nuisance”) parameters and the power of the test is only a function  $\pi(\phi, n, \alpha)$ . The required sample size  $n = 59$  is the smallest integer

$n$  such that  $\pi(.85, n, .05) \geq .80$ . It can be obtained by successive iterations. The probability of observing  $a$  successes is given by the Binomial distribution:

$$\Pr(a | \varphi) = \binom{n}{a} \varphi^a (1 - \varphi)^{n-a}.$$

For  $n = 59$ , the binomial test rejects  $H_0$  at level .05 if the observed number of success  $a$  is greater or equal to 48:

$$\Pr(a \geq 48 | H_0 : \varphi = .70) = .035 < .05 (= \alpha)$$

while the probability of rejecting  $H_0$  if  $H_a : \varphi = .85$  is true is

$$\Pr(a \geq 48 | H_a : \varphi = .85) = .834 > .80 (= 1 - \beta)$$

This determines the critical region of the test: if at least 48 successes are observed,  $H_0$  is rejected, otherwise it is accepted. Note that, due to the discreteness of the distribution, the actual Type I error rate is only .035. Similarly, the actual Type II error rate is smaller than .20 (the actual power is larger than .80).

### Reporting and interpreting $p$ -values

The Neyman-Pearson based justification of GHOST should make the use of  $p$ -values irrelevant. So, it should not matter that  $a = 48$  or  $a = 51$  successes would be observed: in each case,  $H_0$  is rejected at .05 level and the alternative is accepted. Nevertheless, the Fisherian practice to report  $p$ -values is strongly recommended: for instance, if  $a = 48$ ,  $p = .035$  and if  $a = 51$ ,  $p = .003$ .

When reporting the results of significance tests, precise  $p$ -values (e.g. ' $p = 0.034$ ') should be reported rather than making exclusive reference to critical values (ICH E9 Expert Working Group, 2006, p. 32).

Moreover, it is suggested that they can be used to make judgments about differences.

The calculation of  $p$ -values is sometimes useful either as an aid to evaluating a specific difference of interest, or as a 'flagging' device applied to a large number of safety and tolerability variables to highlight differences worth further attention (ICH E9 Expert Working Group, 2006, p. 31).

This may be much more problematic, even if it is not stupid to consider that, for a fixed  $n$ , a smaller  $p$ -value is more in favor of further development of the drug.

### Reporting effect size estimates and confidence intervals

In addition to the  $p$ -values, estimates of the size of effects and confidence intervals should also be reported:

it is important to bear in mind the need to provide statistical estimates of the size of treatment effects together with confidence intervals (in addition to significance tests) (ICH E9 Expert Working Group, 2006, p. 28).

Since it does not depend of  $n$ , the observed proportion –  $f = .814$  if  $a = 48$  and  $f = .864$  if  $a = 51$  – gives another piece of information than the  $p$ -value. To be coherent with the use of the binomial test, we consider the 95% Clopper-Pearson interval:  $[.691, .903]$  if  $f = .814$  and  $[.750, .940]$  if  $f = .864$ . Reporting this interval should remove the temptation to conclude that the drug is really attractive ( $\phi \geq .85$ ) when  $f = .864$  and  $p < .003$ .

### 4.3.2 A Hybrid Practice

GHOST is obviously a hybrid practice. Actually, it appears to justify an amalgam of Fisherian and Neyman-Pearsonian views, in spite of the criticisms against this hybrid logic. A major objection is that it involves using point null  $H_0$  and working  $H_a$  hypotheses: in our example, the investigators are clearly not interested in the precise values  $\phi = .70$  ( $H_0$ ) and  $\phi = .85$  ( $H_a$ ). They are concerned with the pre-specified regions:  $\phi \leq .70$ ,  $.70 < \phi < .85$ ,  $\phi \geq .85$ . Nevertheless, GHOST is considered as “good statistical practice” in many fields.

#### The American Psychological Association Task Force

So the following extracts of the recommendations made by the *American Psychological Association [APA] Task Force on Statistical Inference* (Wilkinson & APA Task Force on Statistical Inference, 1999) are in accordance with the ICH guidelines.

**Hypothesis tests.** It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual  $p$  value or, better still, a confidence interval.

**Power and sample size.** Provide information on sample size and the process that led to sample size decisions.

**Effect sizes.** Always provide some effect-size estimate when reporting a  $p$  value.

**Interval estimates.** Interval estimates should be given for any effect sizes involving principal outcomes.

Similar recommendations have been reiterated in the 6th edition of the publication manual of the APA (American Psychological Association, 2010, pp. 33-35).

#### A notable exception

The above ICH E9 recommendations concern *superiority trials*, designed to demonstrate that one treatment is more effective than another. Were also considered *equivalence* and *non-inferiority* trials, designed to demonstrate, respectively, that two or

more treatments differ by an amount which is clinically unimportant, and that a treatment is not clinically inferior to a comparative treatment. In these (less frequent) cases, it was recommended to “normally” base the statistical analysis on the use of confidence intervals.

The sample size of an equivalence trial or a non-inferiority trial [...] should normally be based on the objective of obtaining a confidence interval for the treatment difference that shows that the treatments differ at most by a clinically acceptable difference (ICH E9 Expert Working Group, 2006, p. 20).

This notable exception to the GHOST procedure is valuable, but adds again its hybridism .

Guidelined Hypotheses Official Significance Testing, recommended by the ICH E9 guidelines and the APA Task Force report, is both partially technically redundant and conceptually incoherent. It completes the ritual of Null Hypothesis Significance Testing by **another set of rituals**, without supplying a real statistical thinking. The consequence is that NHST continues to resist all warnings.

## Chapter 5

# The Significance Test Controversy Revisited

This chapter revisits the significance test controversy in the light of Jeffreys' views about the role of statistical inference in experimental investigations. These views have been clearly expressed in the third edition of his *Theory of Probability*. We will quote and comment the relevant passage. We will consider only the elementary inference about the difference between two means, but our conclusions will be applicable to most of the usual situations encountered in experimental data analysis.

### 5.1 Significance Tests vs Pure Estimation

Jeffreys (1967, chapter VII, p. 389)

But what are called significance tests in agricultural experiments seem to me to be very largely *problems of pure estimation*. When a set of varieties of a plant are tested for productiveness or when various treatments are tested, it does not appear to me that the question of presence or absence of difference comes into consideration at all (*italics added*).

The relation between estimation and significance tests is at the heart of Jeffreys' methodology. Implications for experimental data analysis can be stated. So, if we are interested in comparing two treatment means, a significance test – in Jeffreys' sense – should not be used “if there is no question whether the difference is zero” (more generally whether a parameter has a specific value).

#### The Meehl Paradox

Meehl contrasted the uses of NHST in social sciences and physics and found an apparent paradox, which he summarized as follows.

In physics, one typically compares the observed numerical value with the theoretically predicted one, so a significant difference refutes the theory. In social science, the theory being too weak to predict a numerical value, the difference examined is that between the observed

value and a null (“chance”) value, so statistical significance speaks for the theory (Meehl, 1990, p. 108).

Due to the logic of NHST, the null hypothesis may virtually always be rejected with a sufficiently large sample. Consequently, Meehl argued that increasing the experimental precision leads to a weaker corroboration of a theory in social science and to a stronger corroboration in physics.

Jeffreys objected in advance that if we are not interested in a particular numerical value of the parameter – or in other terms “if there is no doubt initially about the relevance of the parameter” – it is a problem of pure estimation. Consequently, the following question is asked, of course in Bayesian terms:

If there is nothing to require consideration of some special values of the parameter, what is the probability distribution of that parameter given the observations? (Jeffreys, 1967, p. 388).

Moreover, even when the theory predicts a precise value, as in the laws of physics, Jeffreys emphasized the need for clearly stated alternative hypotheses, a further restriction to adopt a rejection rule.

Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts (Jeffreys, 1967, p. 390).

The Meehl paradox results only from the use of NHST as a decision rule to reject the null hypothesis in a situation that is in fact a problem of pure estimation.

## 5.2 The Null Hypothesis: A Straw Man

Jeffreys (1967, chapter VII, p. 389)

It is already known that varieties habitually differ and that *treatments have different effects*... (italics added)

It is almost universally recognized that, in experimental research, the usual point null hypothesis of no effect is known to be false before the data are collected.

- In many experiments, it seems obvious that the different treatments must produce some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: The real problem is to obtain estimates of the sizes of the differences (Cochran & Cox, 1957, p. 5).
- ... in typical applications, one of the hypotheses – the null hypothesis – is known by all concerned to be false from the outset (Edwards, Lindman & Savage, 1963, p. 214).

- In many experiments [...] it is known that the null hypothesis customarily tested, i.e. that the treatments produce no effects, is certainly untrue (Yates, 1964, p. 320).
- All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking ‘Are the effects different?’ is foolish (Tukey, 1991, p. 100).

It follows that the null hypothesis is unrealistic: it is a *straw man* that NHST tries to knock down (Carver, 1978, p. 380).

### 5.3 Usual Two-sided Tests Do Not Tell the Direction

Jeffreys (1967, chapter VII, p. 389)

... and the problem is to decide which [treatment] is the best; that is to put the various members, as far as possible, *in their correct order* (italics added).

Most experiments are designed to demonstrate that one treatment is more effective than another. Let's call this issue a *superiority question*, by analogy with the terms *superiority trials* used in clinical research.

#### Two-sided vs one-sided tests and their shortcomings

When comparing two treatment means  $\mu_A$  and  $\mu_B$ , the only allowable conclusions of the conventional two-sided test are either to reject the null hypothesis  $H_0 : \mu_A = \mu_B$  or to fail to reject it. Two-sided tests do not tell the direction and consequently cannot lead to the desired conclusions.

Few experimenters, of whom we are aware, want to conclude “there is a difference” Rather, they are looking to conclude “the new treatment is better.” Thus, for the most part, there is a direction of interest in almost any experiment, and saddling an experimenter with a two-sided test will not lead to the desired conclusions (Casella & Berger, 1987, p. 106).

On the other hand, a one-sided test does not allow to conclude that the result is statistically significant if the sign of the observed effect is opposite to that expected. Furthermore, it is often suspected that a one-sided test has been used in order to get significant results more easily.

#### Jones and Tukey's three-alternative conclusion procedure

While recognizing that it should be better to formulate experimental problems in terms of estimation, with the establishment of confidence intervals (Jones, 1955, p. 407), Jones and Tukey considered another view to NHST as a three-alternative conclusion procedure:

- (a) act as if  $\mu_A - \mu_B > 0$ ;
- (b) act as if  $\mu_A - \mu_B < 0$ ;
- (c) act as if the sign of  $\mu_A - \mu_B$  is indefinite, i.e., is not (yet) determined  
(Jones & Tukey, 2000, p. 412).

Consequently, they proposed to report the  $p$ -value as “the area of the  $t$ -distribution more positive or more negative (but not both)” than the observed value of the  $t$ -test statistic. This procedure avoids the unrealistic formulation of a point null hypothesis. Consequently there is no Type I error: the only possibility of error is to conclude in a direction when the truth is the other direction. To conclude that the direction is not determined is not an error.

This procedure has not received much attention. A simple reason is that it is usual to conclude about the direction, even *when a two-sided test is performed*, typically:

group A is superior to group B,  $F(1,14) = 5.87$ ,  $p < .03$ .

For a superiority question – to demonstrate that one treatment is more effective than another – the first requirement of an appropriate statistical inference procedure should be to allow a conclusion about the direction of the effect.

## 5.4 Determining Sample Size

Jeffreys (1967, chapter VII, pp. 389–390)

The design of the experiment is such that the order of magnitude of the uncertainty of the result can be predicted from similar experiments in the past, and especially from uniformity trials, and has been chosen so that any differences large enough to be interesting would be expected to be revealed on analysis. The experimenter has already a very good idea of how large a difference needs to be before it can be considered to be of practical importance; the design is made so that the uncertainty will not mask such differences.

In the Bayesian framework, questions about sample size can be stated in a natural way: how big should be the experiment to have a reasonable chance of demonstrating a given conclusion? This question may be viewed, either as unconditional in that it requires consideration of all possible values of parameters, and predictive probabilities give a direct answer, or conditional to some particular values of interest, and power calculations for sample size determination can be reconsidered from a Bayesian point of view.

Jeffreys was optimistic about his “very good idea” that the experimenter has about the practical importance of a difference. However, specifying an effect size of scientific interest is an essential requirement:

The treatment difference to be detected may be based on a judgment concerning the minimal effect which has clinical relevance in the management of patients or on a judgment



concerning the anticipated effect of the new treatment, where this is larger (ICH E9 Expert Working Group, 2006, p. 19).

Other approaches, which failed to recognize this requirement, such as the methods that control the length of the interval estimates of the difference, are not recommendable (see Grouin et al., 2007).

In his description of the design of an experiment, Jeffreys gave some legitimacy to the official GHOST practice for determining sample size described in 4.3.

## 5.5 Critique of $p$ -values: A Need to Rethink

Jeffreys (1967, chapter VII, p. 390)

the  $P$  integral found from the difference between the mean yields of two varieties gives correctly the probability on the data that the estimates are in the wrong order, which is what is required.

There are repeated warnings about the misinterpretations of  $p$ -values that can result from the relation between significance and sample size. Typical examples are the following.

- A given degree of significance – say  $p = .01$  – is not the same evidence whether the sample size is small or large.
- In some situations, the sample size is so large that even a trivial difference can turn to be statistically significant.
- A large  $p$ -value is not an evidence in support of the absence of difference, since it may result from inadequate sample size.

### Jeffreys' answer to the problem of pure estimation

For the Jeffreys prior, the posterior – fiducial Bayesian – distribution (see Section 8.1.2) of the difference  $\delta = \mu_A - \mu_B$ , given the data, is a generalized, or scaled,  $t$ -distribution. It is centered on the mean observed difference and has a scale factor equal to the denominator of the usual  $t$ -test statistic.

This demonstrates the technical link with the NHST procedure and with the usual confidence interval, and this is Jeffreys' answer to the problem of pure estimation:

the  $t$  rule gives the complete posterior probability distribution of a quantity to be estimated from the data, provided again that there is no doubt initially about its relevance; and the integral gives the probability that it is more or less than some assigned value (Jeffreys, 1967, p. 387).

### The Bayesian interpretation of the $p$ -value

As a particular case, we get the posterior probability that  $\delta$  is more or less than zero. The probability that  $\delta$  has the opposite sign of the observed difference is exactly the halved  $p$ -value of the usual two-sided  $t$ -test. This is in close agreement with the above Jones and Tukey procedure. The advantages are that the Jeffreys solution does not resort to statistical hypotheses and can be expressed in the natural language of Bayesian probability: if, say, the observed difference is positive, there is a  $p/2$  posterior probability of a negative difference and a  $(1 - p/2)$  complementary probability of a positive difference.

### Student's conception

'Student' is the pseudonym used by William S. Gosset, a chemist at Guinness brewery. It must be emphasized that, in his original article on what was called "the Student's  $t$ -test" (the notation  $t$  was introduced by Fisher), he had the same conception as Jeffreys. He considered a pharmaceutical example designed to compare the "additional hour's sleep" gained by the use of two soporifics [1 and 2]. Clearly, the procedure aimed at obtaining a judgment about the sign of the effect (the word hypothesis did not appear in the paper), and this judgment was expressed in terms of Bayesian probabilities.

First let us see what is the probability that 1 will on the average give increase of sleep; i.e. *what is the chance that the mean of the population of which these experiments are a sample is positive*. ... we find ... .8873 [in our notations  $1 - p/2$ ] or the odds are .887 to .113 [ $p/2$ ] that the mean is positive (italics added). ... the probability is .9985 or the odds are about 666 to 1 that 2 is the better soporific (Student, 1908, pp. 20–21).

At least, it must be acknowledged that "a somewhat loosely defined conception of inverse probability seems to underlie the argument" (Pearson, 1939, p. 223). Student, as Jeffreys, was primarily interested in an inference conditional on the data (see Zabell, 2008, p. 2). This is also revealed by the words "unique sample" in the title of his later paper:

Tables for estimating the probability that the mean of a *unique sample* of observations lies between  $-\infty$  and any given distance of the mean of the population from which the sample is drawn (Student, 1917, italics added).

### Jaynes' Bayesian test

Jaynes, another physicist, argued on behalf of using Bayesian inference in a perspective close to that adopted by Jeffreys. For him the Bayesian test for comparing the means  $b$  and  $a$  (in his notations) of two normal distributions was based on the Jeffreys prior and consisted in computing the posterior probability that  $b > a$ :

If the question at issue is whether  $b > a$  [ $b$  and  $a$  being the two means], the way to answer it is to calculate the *probability* that  $b > a$ , conditional on the available data (Jaynes, 1976, p. 182).

Moreover, he firmly argued against the use of a preassigned significance level and he considered that the difference between this Bayesian test and the Fisher test of significance using the  $p$ -value was in this case

only a verbal disagreement as to whether we should use the word ‘probability’ or ‘significance’ (Jaynes, 1976, p. 185).

Jaynes also advocated that “the best confidence interval for any location or scale parameter” was the Bayesian posterior probability interval.

### The methodological shortcomings of NHST clearly pointed out

It must be stressed that the Bayesian interpretation *does not depend on sample size*. It becomes apparent that, *in itself*, the  $p$ -value says nothing about the magnitude of  $\delta$ . A given  $p$ -value gives the same evidence *in favor of a positive difference* (nothing else), whatever the sample size is.

- A small  $p$ -value (even “highly significant”) only establishes that  $\delta$  has the same sign as the observed difference.
- A “nonsignificant” outcome is hardly worth anything, as exemplified by the Bayesian interpretation  $\Pr(\delta < 0) = \Pr(\delta > 0) = 1/2$  of the *perfectly nonsignificant* test obtained in the case of a null observed difference.

### The Bayesian interpretation of the two-sided $p$ -value

The “counternull value” (Rosenthal & Rubin, 1994) is the alternative effect size that results in the observed  $p$ -value when it is taken as the null hypothesis. It follows that the posterior probability that the difference  $\delta$  exceeds this counternull value is also equal to  $p/2$ . Consequently, the posterior probability that  $\delta$  lies outside the interval bounded by 0 (the null hypothesis value) and twice the observed difference (the counternull value) is exactly equal to the two-sided  $p$ -value.

This alternative interpretation is more informative, since it gives also an upper bound for  $\delta$ . With a very high experimental precision (large sample size and/or small variance), a significant outcome can lead to the conclusion of a difference of small magnitude in the direction of the observed difference.

### Killeen’s $p_{\text{rep}}$

Killeen recommended to report the *probability of replication*  $p_{\text{rep}}$  of an experimental result, which he defined as the probability of finding in a replication (same sample size) of an experiment

an effect of the same sign as that found in the original experiment (Killeen, 2005, p. 346).

The probability  $p_{\text{rep}}$  is conditional on the data in hand (and not on unknown quantities) and goes to the unknown future observations (the replication). Its justification is exactly the same as the Jeffreys justification of the one-sided  $p$ -value, but instead of the posterior distribution about the parameter  $\delta$ , the posterior predictive distribution about the statistic is considered (Lecoutre, Lecoutre & Poitevineau, 2010). Consequently,  $p_{\text{rep}}$  points out the methodological shortcomings of NHST in exactly the same way as the Jeffreys Bayesian interpretation of the one-sided  $p$ -value. The analogue of the two-sided  $p$ -value is the probability of finding in a replication a difference lying inside the interval bounded by 0 and the counternull value.

Following Killeen's paper, the Association for Psychological Science recommended that articles submitted to Psychological Science and their other journals report  $p_{\text{rep}}$ . It was also included in the list of statistical abbreviations and symbols of the 6th edition of the publication manual of the American Psychological Association (2010, p. 120). It follows that for the first time a Bayesian probability was routinely reported in some psychological journals. Unfortunately,  $p_{\text{rep}}$  was not really taken into consideration: it was simply used in place of or in addition to the  $p$ -value, with very little impact on the way the authors interpreted their data.

Killeen's  $p_{\text{rep}}$  was the object of criticism and the Association for Psychological Science abandoned its recommendation. However, most critics misunderstood its meanings (and its limitations) and the attempts to reinterpret  $p_{\text{rep}}$  as a frequentist probability revealed misconceptions about predictive probabilities. So, Iverson, Lee, and Wagenmakers (2009) confused the *conditional* Bayesian predictive probability of replication of an observed direction of effect with the frequentist *joint* probability that two future experiments will return the same sign (see Lecoutre & Killeen, 2010).

A further discussion about predictive inference and its advantages over parametric inference can be found in Geisser (1983).

Most of those, frequentists as well as Bayesians, who discuss the misuses and misinterpretations of  $p$ -values seem to ignore Jeffreys' lesson that

several of the  $P$  integrals have a definite place in the present theory, *in problems of pure estimation* (Jeffreys, 1967, p. 387, italics added: not in significance tests!).

The consequence is the existence of technical and conceptual links between fiducial Bayesian and frequentist procedures: the Bayesian interpretation of  $p$ -values and confidence levels.

## 5.6 Decision and Estimation

Jeffreys (1967, chapter VII, p. 390)

If the probability that they are misplaced is under 0.05 we may fairly trust the decision.

The significant/nonsignificant dichotomy inevitably suggests using NHST as the binary decision rule: “there is an effect/there is no effect”.

### The decision making viewpoint: A very controversial issue

This common use is undoubtedly encouraged by the decision making viewpoint often advocated in statistical literature. This is explicit within the Neyman-Pearson approach, but one could also consider that

the methods associated with [Fisher’s] test of significance constitute [...] a decision- or risk-evaluation calculus (Bakan, 1966, pp. 435-436).

There was a change of emphasis towards decision making in the middle of the 20th century. Today, many Bayesians concentrate on the decision-theoretic principles. So, in his book *The Bayesian Choice*, Robert argued that

the overall purpose of most inferential studies is to provide the statistician (or a client) with a decision (Robert, 2007, p. 51).

Without dismissing the merits of this approach in some problems, it must be recognized that this is a very controversial issue.

- I have been concerned for a number of years with the tendency of decision theory to attempt the conquest of all statistics. This concern has been founded, in large part, upon my belief that science does not live by decisions alone – that its main support is a different sort of inference. [...] I believe that conclusions are even more important to science than decisions (Tukey, 1960, p. 423).
- [NHST] most basic error lies in mistaking the aim of a scientific investigation to be a decision, rather than a cognitive evaluation of propositions (Rozeboom, 1960, p. 428).
- Scientific investigation uses statistical methods in an iteration in which controlled data gathering and data analysis alternate. [...] In problems of scientific inference we would usually, were it possible, like the data “to speak for themselves” (Box & Tiao, 1973, p. 2).
- [in many epidemiological studies and randomized controlled trials] the issue tends more to be whether the direction of an effect has been reasonably firmly established and whether the magnitude of any effect is such as to make it of public health or clinical importance (Cox, 2001, p. 1469).
- In many cases published medical research requires no firm decision: it contributes incrementally to an existing body of knowledge (Sterne & Smith, 2001, p. 229).

### Jeffreys' Bayesian methodology

Scientists cannot find in the binary decision rule – “there is an effect/there is no effect” – all the answers to the questions of primary interest in experimental investigations:

this decision making process is antithetical to the information accumulation process of scientific inference (Morrison & Henkel, 1970, p. 309).

Jeffreys' Bayesian methodology recognizes the primacy of estimation problems in experimental data analysis and lets the data *to speak for themselves*. This does not preclude to express clear-cut conclusions in a publication, given that they are never definitely accepted and that they can always be challenged in the light of new results.

It is a reasonable strategy to “decide” first about the direction of the difference and then to estimate the magnitude of this difference. This is again in accordance with the recommended GHOST practice of reporting a *p*-value, an effect size estimate and a confidence interval. The basic difference is that the frequentist inference involves three distinct procedures, while in the Bayesian approach there is just one coherent procedure – computing the posterior distribution – which answers the different questions.

## 5.7 The Role of Previous Information and the Sample Size

Jeffreys (1967, chapter VII, p. 390)

It is hardly correct in such a case to say that previous information is not used; on the contrary, previous information relevant to the orders of magnitude to be compared has determined the whole design of the experiment. What is not used is previous information about the differences between the actual effects sought, usually for the very adequate reason that there is none; ...

Jeffreys argued that the information used for designing the experiment should not be used twice. One reason is that previous information pertains only to the orders of magnitude and not to the actual difference. This seems to be a reasonable position.

In experimental investigations, previous information is generally used more or less explicitly for selecting the sample size with the aim to find an acceptable compromise between the chance of finding “a significant difference” and the cost resulting from a large sample size. With regard to this practice, Meehl's affirmation concerning superiority questions,

In most psychological research, improved power of a statistical design leads to a prior probability approaching  $1/2$  of finding a significant difference in the theoretically predicted direction (Meehl, 1967, p. 103),

is not paradoxical, but merely contingent.

The Bayesian interpretation of  $p$ -values makes clear that a sample may be “too big” for economic or ethical reasons but *cannot be too big for statistical analysis*.

## 5.8 The Limited Role of Significance Problems

It is noteworthy that Jeffreys’ views about the role of statistical tests in experimental research has never been seriously considered. For instance, in the very detailed review of Jeffreys’ book, published by Robert, Chopin and Rousseau (2009), the concerned section is acknowledged as “the most famous part of the chapter”. However, the authors only quoted Jeffreys’ criticism about the frequentist interpretation of the  $p$ -value. They insisted on his emphasis on the need for alternative hypotheses, in relation to their own conviction that testing hypotheses is *the central issue*, but they omitted to mention Jeffreys’ conception that this issue is only relevant when the theory predicts a precise value. There was no more mention of Jeffreys’ views about experimental research in the comments made by eminent statisticians that followed this review.

Jeffreys (1967, chapter VII, p. 390)

If they are genuine questions of significance in agricultural experiments it seems to me that they must concern only the higher interactions.

It could be of interest to use a Bayesian test (see Section 3.3) for higher interactions. In particular, when the null hypothesis of no interaction would be retained, this could be used to estimate the error term. This practice is sometimes recommended. However, to be really justifiable, it should imply that, when the null hypothesis is rejected, the alternative hypothesis of interaction would be considered as meaningful. But this is rarely the case. Higher interactions are usually very difficult to interpret and actually are rarely interpreted in experimental publications, even when they are significant.

Actually, Jeffreys’ views applied to any situation where the objective is to learn from experimental or observational data without precise predictions associated with a sharp model.

## 5.9 Other Issues

### 5.9.1 *Non-inferiority and Equivalence Questions*

We have focused on superiority questions, but Jeffreys' Bayesian methodology is appropriate for answering other typical questions raised by experimental data analysis.

- *Non-inferiority* questions: to demonstrate that one treatment is not substantially worse than another;
- *Equivalence* questions: to demonstrate that the difference between two treatments is not large in either positive or negative direction.

These questions are also problems of pure estimation: we are not interested in a particular numerical value of the parameter. It should also be acknowledged that demonstrating a good fit for a theoretical model should generally be treated as an equivalence problem:

With regard to a goodness-of-fit test to answer whether certain ratios have given exact values, 'we know a priori this is not true; no model can completely capture all possible genetic mechanisms' (Matloff, 1991, p. 1247).

### 5.9.2 *Stopping Rules and the Likelihood Principle*

A recurrent criticism made by Bayesian against  $p$ -values, and more generally frequentist procedures, is that they do not conform to the likelihood principle. So, suppose that in the clinical trial example of inference about a proportion (Section 4.3) 2 successes and 8 failures have been observed. This can correspond to different sampling models, for instance:

- the sample size, fixed in advance, was  $n = 10$ ;
- the investigators had planned to stop the trial after 8 failures were observed;
- the sample size was  $n = 59$ , as in the real trial, but the investigators had planned an interim analysis to stop the trial if more than 7 failures were observed after the inclusion of 10 patients (and to continue elsewhere).

In the three cases, the likelihood is proportional to

$$\varphi^2(1 - \varphi)^8$$

The strong likelihood principle implies that the inference should be based only on the information "2 successes and 8 failures have been observed" and should be identical for the three models (e.g. Robert, 2007, p. 16). An extremist Bayesian position is that stopping rules are irrelevant. For instance, Kruschke (2011) rejected the use of  $p$ -values because different values are obtained in each of the above cases:



It is wrong to speak of “the”  $p$ -value for a set of data, because any set of data has many different  $p$ -values depending on the intent of the experimenter (Kruschke, 2011, p. 305).

However, this can be seriously questioned:

Information without knowledge concerning its production does not support probabilities (Fraser, 1980, p. 58).

Actually, many investigators feel that the design possibility of early stopping cannot be ignored, since it may induce a bias on the inference that must be explicitly corrected. A reasonable point of view is that the experimental design, incorporating the stopping rule, is prior to the sampling information and that the information on the design is one part of the evidence.

A comprehensive discussion can be found in Box and Tiao (1973, pp. 45–46). de Cristofaro (2004, 2006) persuasively argued that the Bayes’ formula must integrate the design information, in particular the sampling rule, as well as the initial evidence prior to designing (see also Bunouf & Lecoutre; 2006, 2010). This is in accordance with the Jeffreys conception of the prior, which is explicitly conditional to “the set of propositions accepted throughout an investigation” (see Section 3.3).

Of course, it would be illusory to claim that Jeffreys’ methodology for learning from experience and data is a completely objective and coherent methodology, a not attainable goal (Berger, 2004). Any widely accepted inferential method cannot avoid more or less arbitrary conventions; in this sense, Jeffreys’ Bayesian approach provides, if not objective methods, at least *reference* methods appropriate for situations involving scientific reporting.



## Chapter 6

# Reporting Effect Sizes: The New Star System

This chapter demonstrates the shortcomings of the widespread practice that consists of simply reporting effect size [ES] indicators in addition to NSHT (without interval estimates). It also questions the consequences of restricting the use of ES to standardized measures, as commonly done in psychology and related fields.

### 6.1 What Is an Effect Size?

Consider the following basic situation. A study is designed to evaluate the effect of a treatment by comparing the mean of a treated group of individuals to the mean of a control group. A natural measure of the treatment effect is the *simple* difference  $\mu_t - \mu_c$  between the two population means. However, it is frequently claimed that the “natural effect size” parameter is the *standardized* difference  $(\mu_t - \mu_c)/\sigma$ , where the parameter  $\sigma$  is the within group standard deviation. It is common to call it *Cohen’s d*, following Cohen’s book on *Statistical Power Analysis for the Behavioral Sciences*. In this book ES is used to mean

the *degree* to which the phenomenon is present in the population,  
or  
the degree to which the null hypothesis is false (Cohen, 1977, pp. 9–10).

In the current context, the first definition is undoubtedly preferable to the second one, which focuses on NHST.

#### A definition restricted to standardized measures

Cohen restricted the use of ES to *metric-free*, and hence standardized, indicators, in part for the (bad) reason that they facilitate power computations:

a necessity demanded by the practical requirements of table making (Cohen, 1977, p. 20).

Nowadays, many methodologists considered “effect size” as a statistical term whose definition and usage is restricted to standardized measures, either of population or sample effects, for instance:

- An effect-size measure is a standardized index (Olejnik & Algina, 2003, p. 434).
- Most effect-sizes are standardized values. That is, similar to a z-score or a standard score, standardized effect-sizes are scale free (Robey, 2004, p. 311).
- An effect size is simply an objective and (usually) standardized measure of the magnitude of observed effect (Field & Miles, 2010, p. 56).

## 6.2 Abuses and Misuses Continue

Simply reporting a standardized effect size indicator, in addition to significance tests, is often considered good statistical practice, even when no interval estimate is reported.

### A psychological example

The following extract from an article published in a major psychology journal will serve us to illustrate the fact that this practice does not actually overcome the abuses of null hypothesis significance tests.

Subjects in the two conditions performed significantly better than expected by chance: respectively  $t(9) = 2.56$ ,  $p = .031$ ,  $d = 0.81$  and  $t(9) = 2.66$ ,  $p = .026$ ,  $d = 0.84$ . Furthermore, there was no significant difference between the two conditions:  $t(9) = -0.237$ ,  $p = .82$ ,  $d = 0.075$  [to preserve anonymity, the phrasing and the numerical results have been slightly modified].

The design is comparable to the Student pharmaceutical example considered in Section 5.5.  $n = 10$  subjects were submitted to each of the two conditions. Each subject's performance was measured by the number of correct responses out of 50. The (exact) observed mean percentages of correct responses in the two conditions were respectively 59.8% (29.9) and 60.6% (30.3). The standard deviations were not reported. For inferential purposes were reported the  $t$ -test statistic, with its two-tailed  $p$ -value, and an ES indicator, ‘Cohen's  $d$ ’, but no interval estimate. The sample size was not justified. For subsequent computations, the  $t$  values will be used as exact numbers.

Within each condition, the mean percentage is compared to 50% (“chance”). In this case, ‘Cohen's  $d$ ’ is the ratio of the observed differences from chance,  $59.8 - 50 = +9.8\%$  and  $60.6 - 50 = +10.6\%$ , to the standard deviation of the ten individual respective percentages of correct responses. For the comparison of the two conditions, ‘Cohen's  $d$ ’ is the ratio of the mean difference,  $59.8 - 60.6 = -0.8\%$ , to

the standard deviation of the ten individual differences between percentages. In the three cases, ‘Cohen’s  $d$ ’ is related to the  $t$ -test statistic by the formula

$$\text{‘Cohen’s } d\text{’} = t \sqrt{\frac{1}{n}}, \text{ for instance } -0.075 = -0.237 \sqrt{\frac{1}{10}}.$$

### An ES indicator that does not tell the direction

The absolute value of the standardized difference was reported. This is the most usual practice, in accordance with Cohen’s recommendation to interpret the difference “without sign. . . for the nondirectional (two-tailed) test” (Cohen, 1977, p. 67). However, this does not convey the information given by the negative  $t$  value.

### Disregarding the robust beauty of simple effect sizes

The use of a standardized ES indicator appears to disregard what Baguley (2009, p. 610) called “the robust beauty of simple effect sizes”. This is revealed by the fact that most authors who advocate this use frequently refer to the APA task force (see Section 4.3.2), “always present effect sizes for primary outcomes,” but fail to mention the sentence that followed this recommendation:

If the units of measurement are meaningful on a practical level (e.g. number of cigarettes smoked per day), then *we usually prefer an unstandardized measure* (regression coefficient or mean difference) to a standardized measure ( $r$  or  $d$ ) (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599, italics added).

So, the *simple* differences are here the meaningful, easily interpretable, effect size. Dividing a difference by a standard deviation cannot magically reveal its real-world implications (Jaccard & Guillamo-Ramos, 2002), and

being so disinterested in our variables that we do not care about their units can hardly be desirable (Tukey, 1969, p. 89)

Standardization would be a miraculous panacea to compare and combine (meta-analysis) the results of multiple studies.

The fact that the measure is standardized just means that we can compare effect sizes across different studies that have measured different variables, or have used different scales of measurement (Field & Miles, 2010, p. 56).

However, this is highly questionable and it can be argued that

we will generally be better of using simple, unstandardised effect size metrics (Baguley, 2010, p. 122)

### Heuristic benchmarks: A new star system

Cohen (1977) suggested a “common conventional frame of reference” for judging the magnitude of a standardized difference: the difference is *small*, *medium*, or *large* if it is 0.20, 0.50, or 0.80 respectively. He cautiously warned that it was

an operation fraught with many dangers . . . recommended for use only when no better basis for estimating the ES index is available (Cohen, 1977, pp. 12 and 25).

Nevertheless these heuristic benchmarks are more and more often used without consideration of the context. The star system is improved: so, our illustrative article included a table labeled with symbols, added to the significance stars, to indicate the range of values:

$$^1d > 0.20 \quad ^2d > 0.50 \quad ^3d > 0.80.$$

However, the reference to small, medium and large differences was only implicit.

### Observed ES indicators can be misleading

The significant “large” observed values  $d = 0.81$  and  $d = 0.84$  suggest that the study demonstrated a large departure from chance in each condition. This is contradicted by the 95% interval estimates for the corresponding population standardized signed differences (see Section 8.2):  $[+0.07, +1.51]$  and  $[+0.10, +1.55]$ .

On the other hand, the “small” observed value  $d = 0.075$ , added to the ritual rhetoric of NHST – “there was no significant difference” – strongly suggests that the results demonstrated a small difference, if not no difference, between conditions. Moreover, the conclusion retained in the final discussion section was: “performance was identical for the two conditions”. This conclusion is not justified, as clearly shown by the 95% interval estimates: respectively  $[-0.69, +0.55]$  and  $[-8.4\%, +6.8\%]$  for the population standardized and unstandardized signed differences.

### A good adaptive practice is not a good statistical practice

Reporting ES indicators could prevent researchers from unjustified conclusions in the *conflicting* cases where a significant result is associated with a small observed value or a nonsignificant result is associated with a large observed value. It is revealing that the present experiment was designed to avoid such conflict. Indeed,  $n = 10$  appears to be about the smallest integer such that an observed standardized difference of 0.80 is significant at .05 level. It can be verified that  $d = 0.80$  is significant for  $n \geq 9$ : for  $n = 9$ ,  $t(8) = 2.400$ ,  $p = .043$  (this does not depend on the standard deviation).

Hence, for  $n = 10$ , it is known in advance that all  $d$  larger than 0.80 will be significant, and moreover that all  $d$  smaller than 0.50 will be nonsignificant. Choosing

a too small, *ad hoc* sample size, is a typical illustration of a *good adaptive practice* that protects the authors from the risk of conflicting cases, while taking into account conventionally accepted target ES. It is certainly not a good statistical practice.

### The need for a more appropriate sample size

This could explain the small, inadequate, sample sizes used in most studies published by psychology journals, constantly denounced, following Cohen (1962). Consider here the power-based Neyman-Pearson approach (see Section 4.3.1) with  $\alpha = .05$  and  $\beta = .20$  (power = .80). If it had been applied with the respective target ES, 0.20, 0.50 and 0.80, the following sample sizes had been used:  $n = 199$ ,  $n = 34$  and  $n = 15$ . Sample size determination with “canned” (Lenth, 2001) effect sizes has evident shortcomings:

Thus, asking for a small, medium, or large standardized effect size is just a fancy way of asking for a large, medium, or small sample size, respectively. If only a standardized effect is sought without regard for how this relates to an absolute effect, the sample size calculation is just a pretense (Lenth, 2001, p. 191).

However, it must be recognized that a sample size of about 200 subjects would have been a more appropriate choice for demonstrating a small difference between the two conditions. So, with  $n = 200$ , for the same observed means and standard deviations, the 95% interval estimates would be respectively  $[-0.21, +0.06]$  and  $[-2.3\%, +0.7\%]$  for the population standardized and unstandardized differences.

The use of canned – small, medium, large – effect sizes can be seriously misleading. It causes important distortions in the designing of experiments and in the interpretations of statistical findings.

### The shortcomings of the phi coefficient

As another example, consider an epidemiological study designed to determine whether women who had been examined using x-ray fluoroscopy during treatment for tuberculosis had a higher rate of breast cancer than those who had not been examined using x-ray fluoroscopy (Rothman & Greenland, 1998). There were respectively 28,010 and 19,017 person-years at risk in the treatment and placebo groups. The corresponding observed cases of breast cancer were 41 and 15, hence the two rates  $f_1 = 41/28,010 = .00146$  and  $f_2 = 15/19,017 = .00079$ . The observed difference  $f_1 - f_2 = .00067$  can only be interpreted by reference to the placebo rate .00079, hence the relative difference (or relative risk increase)  $(f_1 - f_2)/f_2 = (f_1/f_2) - 1 = 0.856$ : the observed rate of breast cancer is 85.6% higher in the treatment group. Equivalently, the ratio (or relative risk) is  $f_1/f_2 = 1.818$ .

It is often recommended to use a standardized ES such as phi for assessing the relationship between two dichotomous variables. The  $\phi$  coefficient is related to the  $\chi^2$  test statistic by the formula (note the analogy with the relation between ‘Cohen’s  $d$ ’ and  $t$ ):

$$\phi = \sqrt{\frac{\chi^2}{n}}, \text{ where } n \text{ is the total sample size.}$$

We have here  $\phi = .0096$ , or a proportion of variance explained  $r^2 = \phi^2 = .00009$ , which does not reflect the real effect of the treatment.

Moreover, as for the psychological example, it is not sufficient to only report an observed ES indicator, and an interval estimate is needed. So, assuming a Poisson model, the 95% interval estimate for the population relative risk  $\tau$  (Lecoutre & Derzko, 2009), based on the Jeffreys Bayesian approach (see Section 7.2) is [1.05, 3.42]. It can be concluded that women examined using x-ray fluoroscopy in have a higher rate of breast cancer ( $\tau > 1$ ), which is not surprising, due to the large sample size. Nevertheless, for rare events, the interval estimate shows that this sample size is not sufficient for assessing the magnitude of the effect.

The shortcomings of the phi coefficient reinforces the contention that standardized ES should be used with the utmost caution for interpreting the magnitude of effects.

## 6.3 When Things Get Worse

### 6.3.1 A Lot of Choices for a Standardized Difference

#### What denominator for ‘Cohen’s $d$ ’?

In our illustrative example, the denominator of the reported ‘Cohen’s  $d$ ’ was the usual sample standard deviation, corrected for degrees of freedom. This definition seems to be the more frequently used one (e.g. Smithson, 2003; Kirk, 2007; Howell, 2010). It is in accordance with Cohen’s definition (Cohen, 1977, pp. 66–67) of the standardized sample mean difference for two independent groups. Expressed as a function of the  $t$  statistic, it is

$$t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ and for the one sample case } t \sqrt{\frac{1}{n}}.$$

However, some authors considered an uncorrected standard deviation (for instance, Rosnow & Rosenthal, 2009, p. 8), which gives the alternative formulae



$$t \frac{n_1 + n_2}{\sqrt{n_1 n_2 (n_1 + n_2 - 2)}} \text{ and } t \sqrt{\frac{1}{n-1}}.$$

### Descriptive statistic or point estimate?

Another practice could be to report a “good estimate” of the population parameter. It is often argued that this should be an *unbiased* estimate. While the unstandardized difference meets this requirement, it is not the case of the standardized difference. Furthermore, at least two formulae are available, the exact one, given by the mean of the noncentral  $t$ -distribution and involving the Gamma function, and a frequently used very accurate approximation. For instance, in the one sample case, they are respectively

$$\sqrt{\frac{2}{df}} \frac{\Gamma(df/2)}{\Gamma((df-1)/2)} t \sqrt{\frac{1}{n}} \text{ and } \left(1 - \frac{3}{4df-1}\right) t \sqrt{\frac{1}{n}}.$$

In our example, the reported value for the first condition would be 0.74 – a *two-star* value – instead of the *three-star* value 0.81.

### What standardizer for a difference between means?

Consider for instance the observed difference between two means in a between-subject design (independent groups). Several alternatives have been proposed for the *standardizer*, in particular :

- the pooled standard deviation of the two compared groups: ‘Cohen’s  $d$ ’;
- the standard deviation of one of the compared groups, especially if it is a control group: ‘Glass’s  $\Delta$ ’ (Glass, 1976);
- the pooled standard deviation of all the groups in the design: ‘Hedges’s  $g$ ’ (Hedges, 1981).

They correspond to two different assumptions about the equality of the population standard deviations. Note that several generalizations of Cohen’s  $d$  have been developed in the case of unequal standard deviations (heteroscedasticity).

Moreover, in our example of a within-subject design (two repeated measures), two kinds of standardizers have been proposed:

- the standard deviation of the individual differences (e.g. Gibbons, Hedeker & Davis, 1993);
- a standard deviation obtained by treating the design as a between-subject one (e.g. Bird, 2004).

### 6.3.2 A Plethora of ES Indicators

#### Expressing the standardized difference as a correlation or as a proportion of variance

In the case of two independent groups, one can compute the Pearson *product moment coefficient correlation*  $r$  between the dependent variable and the predictor variable with, for instance, the value 0 for one group and 1 for the other. The observed  $r$  can be expressed as a function of the  $t$  statistic:

$$r = \frac{t}{\sqrt{t^2 + df}}.$$

A similar approach is to consider ES indicators intended to estimate the proportion of variance in the dependent variable that can be explained by the independent variable. They appear in the context of analysis of variance (ANOVA). The *partial eta-squared*, calculated by common statistical software such as SPSS, is a popular indicator, frequently reported, for instance:

There was a statistically significant main effect for time,  $F(1, 83) = 14.83$ ,  $p < .001$ , partial eta squared = .15.

The observed value .15 is the ratio of the *between* sum of squares to the *total* sum of squares. Consequently, it is related to the  $F$ -test (with  $df1$  and  $df2$  degrees of freedom) by the formula

$$\text{observed partial eta-squared} = \frac{F}{F + \frac{df2}{df1}}.$$

In this case ( $df1 = 1$ ), it is equal to  $r^2$ , since  $F = t^2$ .

It is often argued that these correlational and proportion of variance accounted for approaches are easy to interpret. Again, it can be questioned whether a further, somewhat artificial, transformation of the mean difference would reveal its real-world implications.

#### Plenty of ES indicators and further difficulties

While our review is not exhaustive, plenty of standardized ES indicators can be associated with a simple difference between means by combining the different alternatives. For more complex effects in general ANOVA designs (without speaking of multivariate analyses), the number of indicators explodes exponentially.

Further issues arise. So, it is well known that standardized ES estimates are affected by the research design and by methodological artifacts, such as the difference of reliability and the range of sampled values, which requires appropriate corrections (e.g. Hunter & Schmidt, 2004):

Anything that influences a sample SD but not the population SD has the potential to distort standardized effect size as measure of the population effect (Baguley, 2010, p. 123).

It would be unrealistic to expect that statistical users could understand the subtleties of all ES indicators. The plethora of definitions has entailed endless debates about “what is the best ES indicator to use” that obscure the real problems. This is a considerable source of misuses and misinterpretations.

### 6.3.3 Don't Confuse a Statistic with a Parameter

#### The need for appropriate definitions and notations

Many recent methodological papers fail to explicitly distinguish between statistics (sample ES) and parameters (population ES), by lack of appropriate definitions and notations. For instance, Rosnow and Rosenthal (2009, p. 8) wrote “the null hypothesis is  $M_1 - M_2 = 0$ ”, after having defined  $M_1$  and  $M_2$  as statistics. Furthermore, in the same sentence, “Cohen's  $d$ ” referred both to the parameter – “a 95% CI for Cohen's  $d$ ” – and to the statistic – “the variance of Cohen's  $d$ ” (Rosnow & Rosenthal, 2009, p. 10). This is at least misleading, if not incomprehensible.

Note that Cohen (1977) explicitly distinguished between the population standardized mean difference denoted by  $\mathbf{d}$  (p. 20), and the “standardized mean difference for the sample” denoted by  $\mathbf{d}_s$  (p. 66).

#### One or two parameters?

Many of the recent discussions about *the proportion of variance accounted for* in ANOVA are also very confusing. It is usual to distinguish between two main “ES estimates”, eta-squared and omega-squared, commonly denoted by  $\eta^2$  and  $\omega^2$ . So, for a one-way ANOVA, Howell (2010) defined them as two statistics (pp. 345–347):

$$\eta^2 = \frac{SS_{\text{treatment}}}{SS_{\text{total}}} = .447 \text{ and } \omega^2 = \frac{SS_{\text{treatment}} - (k-1)MS_{\text{error}}}{SS_{\text{total}} - MS_{\text{error}}} = .393.$$

He interpreted them as the estimates of two (not defined) parameters, also denoted by the letters  $\eta^2$  and  $\omega^2$ :

The estimate of  $\omega^2$  in this case (.393) is noticeably less than the estimate of  $\eta^2 = .447$ .

So, he suggested the existence of two distinct parameters. However, the sequel of the sentence – “reflecting the fact that the latter is more biased” – may alert the reader that the two statistics are in reality two different estimates of the *same* parameter.

The “old” papers about the magnitude of effects were much more explicit. For instance, Fleiss (1969) clearly distinguished between:

- the parameter  $\omega^2 = \theta^2 / (\sigma^2 + \theta^2)$ , the proportion of the total variance  $\sigma^2 + \theta^2$  attributable to experimental effect;
- its two estimates  $\hat{\omega}^2$  and  $\hat{\eta}^2$ .

He justified his preference for  $\hat{\omega}^2$  (recommended by Hays, 1963) by the fact that it is obtained as the ratio of unbiased estimates of the numerator and of the denominator (however, this is not an unbiased estimate, and even not a good estimate since it takes negative values when  $F < 1$ ). This does not erase the confusion, since Fleiss used the notation  $\omega^2$  for the parameter when the more usual notation, following Pearson, is  $\eta^2$ . Moreover, while Steiger and Fouladi (1997) used  $\eta^2$ , Steiger (2004) used  $\omega^2$ .

### A crazy parameter

Fidler and Thompson (2001, pp. 592–593) added to the confusion by computing, not only a confidence interval for the parameter  $\eta^2$ , but also a distinct confidence interval for  $\omega^2$ . These two 95% confidence intervals are respectively  $[0, .5346]$  and  $[-.14, .4543]$  (note the curious lower limit  $-.14$  for a positive parameter). Their parameter  $\omega^2$  is not defined, but it can be deduced that it is linked to  $\eta^2$  by the same relation as the one that links the two estimates  $\hat{\omega}^2$  and  $\hat{\eta}^2$ . So this “new” parameter has no rational basis and cannot be interpreted.

### When the proponents disagree

Smithson (2001, p. 619 and 2003, p. 44) and Steiger (2004, p. 171) analyzed the same illustrative data from a two-way  $2 \times 7$  ANOVA, with 4 observations per cell. They considered the partial eta-squared, but they gave it

- two different notations,  $\eta^2$  and  $\omega^2$  respectively,
- two different names, “a squared partial correlation” and “the proportion of the variance remaining that is explained by the effect,”
- two different 90% confidence intervals  $[.0093, .3455]$  and  $[0.0082, .3160]$  for the interaction effect.

There is no doubt that they considered the same parameter. The different confidence intervals resulted from the fact that Smithson used a formula that is inappropriate in the case of a fixed effects ANOVA. We have found mention of this unfortunate disagreement only in the SAS/STAT user’s guide (SAS Institute Inc., 2010, p. 3059).

Many users tend to confuse the sample ES indicator with the population ES. This is obviously another source of misuses, probably encouraged by the common expression “effect size estimate” and by the lack of appropriate definitions and notations.

## 6.4 Two Lessons

### The new star system

The current focus on the magnitude of effects is without doubt welcome. However, recent reviews of ES reporting practices in educational and psychological journals have emphasized the lack of substantive discussions and interpretations of effect size (e.g. McMillan & Foley, 2011). This can be viewed as a consequence of the undue emphasis on standardized ES and heuristic benchmarks. A *new star system* has been created that jeopardizes the results and conclusions of experimental research:

if people interpreted effect sizes with the same rigidity that .05 has been used in statistical testing, we would merely be being stupid in another metric (Thompson, 2001, pp. 82–83).

### Should standardized effect sizes ever be used?

In a lucid paper, Baguley (2009, p. 612) bluntly asked this question. It is beyond the scope of this book to discuss the pro and con arguments, but we fully agree that

careless and routine application of standardization in psychology (without any awareness of the potential pitfalls) is dangerous (Baguley, 2010, p. 123)

It is not sufficient to only report an observed (standardized or unstandardized) ES indicator, ignoring the variability of this indicator. This does not answer questions about the magnitude of the population effect, and consequently does not avoid erroneous inferences. In particular, in the case of a nonsignificant result, this practice seems to support the conclusion of no effect, while there is even no evidence of a small effect. It is indispensable to include a real estimation of the magnitude of the population effect, taking explicitly into account the sampling variability, in particular, but not only an interval estimate.



## Chapter 7

# Reporting Confidence Intervals: A Paradoxical Situation

This chapter reviews the different views and interpretations of interval estimates. It discusses their methodological implications – what is the right use of interval estimates? The usual confidence intervals are compared with the so-called “exact” or “correct” confidence intervals for ANOVA effect sizes. While the former can receive both frequentist and Bayesian justifications and interpretations, the latter have logical and methodological inconsistencies that demonstrate the shortcomings of the uncritical use of the Neyman-Pearson approach. In conclusion, we have to ask: *Why isn’t everyone a Bayesian?*

### 7.1 Three views of Interval Estimates

The frequentist theory of statistical estimation was essentially developed by Neyman. So, usual confidence intervals pertain to the Neyman and Pearson conception, and their interpretation is at odds with the alternative Jeffreys Bayesian and Fisher fiducial approaches.

#### 7.1.1 The Bayesian Approach (Laplace, Jeffreys)

Historically, one of the first interval estimate was proposed by Laplace in 1812. He estimated the mass of Saturn, compared to the mass of the sun taken as unity, given (imperfect) astronomical data. In modern terms, he derived the Bayesian posterior distribution for the mass, using a uniform prior. He presented the results as follows:

il y a onze mille à parier contre un, que l’erreur de ce résultat [la masse de Saturne est égale à la 3512<sup>e</sup> partie de celle du soleil] n’est pas à un centième de sa valeur (it is a bet of 11 000 against 1 that the error of this result [the mass of Saturn is equal to 1/3512 of the mass of the sun] is not 1/100 of its value) (Laplace, 1840, p. 99).

In other terms, there is a posterior probability .99991 (1-1/11 000) that the unknown mass of Saturn, estimated to be 1/3512 of the mass of the sun, is within 1% of this point estimate.

For Jeffreys the estimate of a parameter is the “complete posterior probability distribution” of this parameter, given the data (see Chapter 5). It may be inferred that he was opposed to the use of a point estimate. However, he did not really develop the use of interval estimates.

### Evaluating the probability of specified regions

Rather, Jeffreys proposed to use the posterior distribution to obtain the probability that the parameter is more or less than some assigned value. Consider, for instance our clinical trial example (Section 4.3) involving the inference about a proportion. Using the Jeffreys prior for the Binomial sampling model (see Lecoutre, 2008), if 51 out 59 successes have been observed, it could be stated that there are respective posterior probabilities:

- .002 that the drug would be of no interest ( $\phi < .70$ )
- .606 that the drug would be really attractive ( $\phi > .85$ )
- .392 that  $\phi$  would be in the intermediate region ( $.70 < \phi < .85$ ).

If we want a term to qualify these probabilities and distinct them from frequentist probabilities, we can use words such as *chance* or *guarantee*.

These results can be interpreted as probabilities of (composite) hypotheses, given data, which can satisfy the researcher’s *Ego* (Gigerenzer, 1993). However, following Jeffreys, it is unnecessary to regard the statistical analysis of these data as a problem of hypotheses *testing*.

### Bayesian credible intervals

It may also be of interest to summarize the posterior distribution by reporting an interval estimate, associated with a given probability, denoted by  $\gamma$  (or by  $1 - \alpha$  as is customary for frequentist confidence intervals). In the Bayesian framework, such an interval is usually termed a *credible* (or *credibility*) interval. For instance, here the Jeffreys 95% credible intervals for  $\phi$  is [.760, .934]. It is an *equal-tailed* interval: the posterior probabilities that  $\phi < .760$  and  $\phi > .934$  are both equal to .025. Of course, a one-tailed interval could be preferred.

Its flexibility makes the Bayesian approach particularly suitable for estimation purpose. We can get an estimate (credible) interval associated with a given probability. We can as well compute the probabilities of specified regions of interest.



### 7.1.2 Fisher's Fiducial Inference

Fisher's fiducial inference is an attempt to conciliate his reluctance to use prior probabilities with his motivation "for making correct statements of probability about the real world" in "the absence of knowledge *a priori*."

#### An attempt to make the Bayesian omelet without breaking the Bayesian eggs

The fiducial argument gives a posterior distribution about the parameter, without having to specify a prior:

the fiducial argument uses the observations only to change the logical status of the parameter from one in which nothing is known of it, and no probability statement about it can be made, to the status of a random variable having a well-defined distribution (Fisher, 1990c, p. 54).

The interpretation is explicitly in terms of Bayesian probabilities:

The concept of probability involved is entirely identical with the classical probability of the early writers, such as Bayes (Fisher, 1990c, p. 54).

#### Fiducial interval and null hypotheses

For Fisher, the 95% (for instance) fiducial interval was linked with the test of significance. Indeed, he alternatively viewed it as a simultaneous statement about all null hypotheses concerning the parameter. So, in the case of the inference about a mean  $\mu$ ,

variation of the unknown parameter,  $\mu$ , generates a continuum of hypotheses each of which might be regarded as a null hypothesis (Fisher, 1990b, p. 192).

The continuum of hypotheses is divided into two portions by the data. The values of the parameter that "are not contradicted by the data", at the 5% (two-sided) level of significance, constitutes the 95% fiducial interval for  $\mu$ . Inverting a statistical test to construct a frequentist confidence interval is a very common technique. However, Fisher did not give the resulting interval a frequentist interpretation, but a Bayesian one:

the probability of  $\mu$  actually lying in the outer zone is only 5 percent; any other probability could equally have been chosen (Fisher, 1990b, p. 192).

#### Fisher's biggest blunder or a big hit?

Fiducial inference is admittedly considered by most modern statisticians to be "Fisher's one great failure" (Zabell, 1992).

The expressions “fiducial probability” and “fiducial argument” are Fisher’s. Nobody knows just what they mean, because Fisher repudiated his most explicit, but definitely faulty, definition and ultimately replaced it with only a few examples. (Savage, 1976, p. 466)

However the story is not ended, as exemplified by the attempts to generalize the fiducial argument (e.g. Hannig, 2009).

Maybe Fisher’s biggest blunder will become a big hit in the 21st century (Efron, 1998, p. 107).

### 7.1.3 Neyman’s *Frequentist Confidence Interval*

The term *confidence* was introduced by Neyman, who developed “a theory of statistical estimation based on the classical [frequentist] theory of probability” (Neyman, 1937). Reviewing the previous attempts to solve the problem of estimation, he presented the Bayesian approach as a “theoretically perfect solution,” but that “may be applied in practice only in quite exceptional cases.” His main argument was that prior probabilities are usually unknown.

Even if the parameters to be estimated, [...] could be considered as random variables, the elementary probability law a priori [...] is usually unknown, and hence the [Bayes] formula cannot be used because of the lack of the necessary data (Neyman 1937, p. 344).

Neyman (1937) also explicitly rejected the Jeffreys approach as being “not justifiable on the ground of the theory of probability adopted in this paper.”

#### **The meaning of the confidence interval needs to be clarified**

Later, Neyman (1977, pp. 116-119) took great pains to clarify the meaning of the confidence intervals. He defined the lower and upper confidence limits (or bounds) for an unknown parameter  $\vartheta$  as two functions of the observables, denoted by  $Y_1(X)$  and  $Y_2(X)$ , hence the confidence interval [CI]:  $I(X) = [Y_1(X), Y_2(X)]$ . All these quantities are random variables and are considered as *tools of inductive behavior*.

Being functions of the random variable  $X$ , the two confidence bounds and the confidence interval  $I(X)$  will be random variables also (Neyman 1977, p. 116).

As for the Neyman-Pearson hypothesis test, the frequentist justification of a CI involves long run frequency properties. The assertions about the unknown number  $\vartheta$  must be

FREQUENTLY correct, and this irrespective of the value that  $\vartheta$  may possess (Neyman 1977, p. 117).

Given the “confidence coefficient”  $\alpha$ , “acceptably close to unity”, this requirement can be formalized as

$$P\{Y_1(X) \leq \vartheta \leq Y_2(X) | \vartheta\} \equiv \alpha,$$

where the conditioning on  $\vartheta$ , unfortunately dropped in most of the recent presentations, is made explicit. Note that the confidence coefficient was denoted by  $\alpha$ , and not  $1 - \alpha$ .

### The difficulties of the frequentist interpretation

Neyman was clearly aware of the difficulties of the frequentist interpretation and of the need to “anticipate certain misunderstandings.” So, he carefully explained this interpretation. He made explicit that, in the above formula, “the probability of the two confidence bounds ‘bracketing’ the true value of  $\vartheta$ ”, which is today named *the coverage probability*,

- is written not in terms of the *observed*  $x$  but in terms of the *observable*  $X$  (italics added),
- is true whatever may be the value of the unknown  $\vartheta$  (Neyman, 1977, pp. 117–118).

Moreover, he stressed the fact that there was no frequentist probability assigned to a single CI computed from a particular sample.

However, if one substitutes [...] the observed  $x$  in the place of the observable  $X$ , the result would be *absurd*. In fact, the numerical results of the substitution may well be

$$(4) \quad P\{Y_1(x) \leq \vartheta \leq Y_2(x) | \vartheta\} = P\{1 \leq 5 \leq 3 | 5\} = 0.95$$

or alternatively,

$$(5) \quad P\{1 \leq 2 \leq 3 | 2\} = 0.95$$

It is essential to be clear that both (4) and (5) are *wrong*. The probability in the left hand side of (4) has the value zero (and thus not 0.95), and that in the left hand side of (5) is unity, *neither of any interest* (Neyman, 1977, pp. 118–119, italics added).

It would be optimistic to think that Neyman’s efforts to explain their correct interpretation could reduce the misunderstandings about frequentist confidence intervals. The reason is that most users think they understand them, albeit they interpret them in Bayesian terms.

## 7.2 What Is a Good Interval Estimate?

### 7.2.1 *Conventional Frequentist Properties*

The most common approach to the evaluation of an interval estimate for a parameter  $\vartheta$  is to see whether it yields confidence (or credible) limits that have good frequentist coverage properties. However, the basic identity

$$P\{Y_1(X) \leq \vartheta \leq Y_2(X) \mid \vartheta\} \equiv \alpha$$

cannot be always satisfied, i.e. “without introducing certain artificialities” (Neyman, 1977, pp. 118). This occurs in particular when the observable variables  $X$  are so-called *discrete*.

In cases of this kind, rather than require the exact equality to  $\alpha$  ... one can require ‘at least equal’ or ‘approximately equal’ (Neyman, 1977, pp. 118).

For discrete data, it results that there are a plethora of solutions. Some of them, ambiguously called “exact”, require a coverage probability ‘at least equal’ to the nominal level, hence too large (*conservative*). The others are approximate and are generally preferred for experimental data analysis reporting (e.g. Agresti & Coull, 1998). Of course, the coverage probability should be close to the nominal level, even for small sample size or for extreme parameter values.

### 7.2.2 *The Fatal Disadvantage of “Shortest Intervals”*

The length of the interval must be “in a sense, just as small as possible” (Neyman, 1977, pp. 117). However, this requirement can result in intervals that are not invariant under transformation. So many Bayesians recommend to consider the highest posterior density [HPD] credible interval. For such an interval, which can be in fact an union of disjoint intervals (if the distribution is not unimodal), every point included has higher posterior probability density than every point excluded. The aim is to get the shortest possible interval. However, except for a symmetric distribution, each of the two one-sided probabilities of a  $100(1 - \alpha)\%$  HPD interval is different from  $\alpha/2$ , a property generally not desirable in experimental data analysis. Moreover, such an interval is not invariant under transformation (except for a linear transformation), which can be considered with Agresti and Min (2005, p. 3) as “a fatal disadvantage.”

### 7.2.3 *One-sided Probabilities are Needed*

Actually, Neyman acknowledged the fact that, in practical cases, questions of interest are frequently one-sided.

The application of the regions of acceptance having the above properties is found useful in problems which may be called those of one-sided estimation. In frequent practical cases we are interested only in one limit which the value of the estimated parameter cannot exceed in one or in the other direction (Neyman, 1937 pp. 374).

It follows that, even if a two-tailed interval is retained, it is essential to consider, not only the coverage probability, but also the frequentist probabilities that both the lower and upper limit exceed the parameter value. In the Bayesian approach, one-tailed or equal two-tailed credible intervals should be privileged.

### 7.2.4 *The Jeffreys Credible Interval is a Great Frequentist Procedure*

The Jeffreys credible intervals for the Binomial proportion  $\varphi$  (Section 7.1.1) has remarkable frequentist properties. Its coverage probability is very close to the nominal level, even for small-size samples. Moreover, it can be favorably compared to most frequentist confidence intervals (Brown, Cai & DasGupta, 2001). Similar results have been obtained for other discrete sampling models (e.g. Lecoutre & Charron, 2000; Berger, 2004; Agresti & Min, 2005; Cai, 2005; Lecoutre & ElQasr, 2008; Lecoutre & Derzko, 2009; Lecoutre, Derzko & ElQasr, 2010).

This demonstrates that the Jeffreys credible interval

is actually a great frequentist confidence procedure (Berger, 2004, p. 6).

## 7.3 Neyman-Pearson's Criterion Questioned

Some criterion of optimality is required to get the best interval estimate. Neyman (1977) acknowledged the existence of “delicate conceptual points” in the definition of optimality. Many frequentist CIs are constructed by inverting a statistical test. We have questioned in Section 3.5 the Neyman-Pearson “optimal” tests of composite null hypothesis. The CIs based on these tests are also scientifically inappropriate. This is the case of CIs for ANOVA effect sizes that have been extensively developed in the recent years.

### ***7.3.1 The Inconsistencies of Noncentral $F$ Based Confidence Intervals for ANOVA Effect Sizes***

It will be sufficient, with no loss of generality, to consider the basic case of the inference about a difference between two means. All results also apply to a contrast between means. A common procedure for constructing a CI for an ANOVA effect size parameter consists in defining this interval as the set of values for the parameter of interest that cannot be rejected by the data, using the ANOVA  $F$ -test (Venables, 1975). The precise technical developments are not necessary. It will be sufficient to know that in the standardized case – generally the only one considered by the proponents of this procedure – the derivation involves the noncentrality parameter of a noncentral  $F$ -distribution. Hence this test and its associated CI will be called hereafter “noncentral  $F$  based”: in short NCF-test and NCF-CI.

#### **A scientifically inappropriate procedure**

As a matter of fact the NCF-test, as well as many other closely related tests (including the case of unstandardized ES), has always been considered as a scientifically inappropriate procedure and rejected by applied statisticians: see Section 3.5. Nevertheless, in spite of repeated warnings, it has been recurrently “rediscovered” in various contexts. So-called “exact confidence” (e.g. Steiger, 2004) or “correct confidence” (e.g. Smithson, 2001) intervals, derived from the NCF-test, can be theoretically easily computed for a variety of ANOVA ES parameters, mathematically equivalent. This includes, in particular, the partial eta-squared (see Section 6.3.2), but also the Cohen  $f$  (or its square, the “signal-to-noise ratio”,  $f^2$ ) and its variants such as the “root-mean-square standardized effect” (see Steiger & Fouladi, 1997).

These NCF-CIs are offered by their proponents as “good statistical practice.” Moreover, they are generally presented as a seemingly natural generalization of the usual confidence interval [U-CI] for a standardized difference between two means (involving the noncentral  $t$ -distribution). This is not the case, and NCF-CIs have logical and methodological inconsistencies that support the contention that their use should be discouraged.

#### **An enlightening comparison of the U-CI and NCF-CI**

For a comparison between two means (and more generally for a contrast between means), the above-mentioned ANOVA ES parameters are all equivalent to the absolute value of the standardized difference (contrast). Consider again the basic situation of our psychological example (Section 6.2), which corresponds also to the Student pharmaceutical example (Section 5.5). In such case, the U-CI, either for the unstandardized (preferably) or standardized signed difference (see Sections 8.1.2 and 8.2 respectively), seems to be the solution of choice. Nevertheless, NCF-CIs

have been proposed as a suitable alternative routine procedure, even in the case of one degree of freedom effects (e.g. Fidler & Thompson, 2001; Smithson, 2003; Steiger, 2004). Consider the 95% U-CI for the signed standardized difference and the 95% NCF-CI for its absolute value, associated with different  $t$  values and their corresponding observed 'Cohen's  $d$ '.

Data		95% confidence interval	
$t$	'Cohen's $d$ '	U-CI	NCF-CI
-0.010	-0.0032	$[-0.623, +0.617]$	$[0, 0]$
+0.033	+0.0104	$[-0.610, +0.630]$	$[0, 0.069]$
-0.237	-0.0750	$[-0.694, +0.548]$	$[0, 0.637]$
-2.500	-0.791	$[-1.491, -0.058]$	$[0, 1.491]$
+4.062	+1.285	$[+0.415, +2.118]$	$[0.414, 2.118]$

### Troublesome properties

Fidler and Thomson (2001) gave the following characterization of interval construction

CI's are typically computed by adding and subtracting from a given parameter estimate the standard error (SE) of that estimate times some  $\alpha$  or  $\alpha/2$  centile of a relevant test distribution (Fidler & Thomson, 2001, p. 579).

So, the U-CI for Cohen's  $d$  is approximately centered around the observed difference and its width, which reflects the precision of estimate, is approximately constant. This looks reasonable. On the contrary, the width of the NCF-CI dramatically decreases for small observed value, as if the precision of estimate was superior in this case. This does not look justified.

### The shortcomings of NCF-CI lower limits

Clearly, a *lower* limit for an unsigned ANOVA ES is not suitable for demonstrating "largeness". When  $t = +4.062$ , as in the Student example, significant at two-sided level .05, the 95% NCF-CI rightly excludes zero, but gives in itself very poor information: it means that the population difference can be larger than 0.414 in *either a positive or negative* direction. Moreover, when  $t = -2.50$ , also significant at two-sided level .05, the 95% NCF-CI surprisingly includes zero. Of course, the 95% U-CI  $[-1.491, -0.058]$  rightly takes into account the significant outcome.

### NCF-CIs lead to unacceptable inferences

Demonstrating the "smallness" of an ANOVA effect is an important methodological issue, as illustrated by our psychological example. Typical relevant situations are to demonstrate the equivalence of drugs, to show that an interaction effect is negligible, or again to demonstrate a "good-enough" fit for a theoretical model. An *upper* limit

for an unsigned ANOVA ES is suitable for these purposes. Unfortunately, it is well known that it can lead to unacceptable inferences.

So, in our example, when  $t = +0.033$  the 95% NCF-CI  $[0, +0.069]$  corresponds to the interval  $[-0.069, +0.069]$  for the signed difference, which is considerably shorter than the U-CI  $[-0.610, +0.630]$ . The NCF-CI is even empty for smaller observed standardized differences. In particular, it is empty for a null difference whatever the sample size is.

Another undesirable property is that the NCF-CI upper limit may vary in a non-monotonous way when the sample size increases. So, when the observed standardized difference is 0.10, the upper limit of the 95% NCF-CI is for instance 0 ( $n \leq 10$ ), 0.0759 ( $n = 11$ ), 0.1915 ( $n = 30$ ), 0.1502 ( $n = 105$ ). So, for the same observed means, it can be concluded, for example, that the population difference is smaller than 0.15 in absolute value with  $n \leq 13$  or  $n \geq 106$ , but not with  $14 \leq n \leq 105$ .

The defenders of NCF-CIs argue that, with a suitable minimum sample size, the practical risk of unacceptable inferences can be reduced. However, it is unfortunate that a “bad planned” experiment could result in a seemingly well supported conclusion and that the procedure may always be under suspicion. Most experimental investigations involve a complex design in which a NCF-CI could be used for instance to demonstrate the negligibility of an interaction effect. Unfortunately, the design is generally planned for other purposes, e.g. demonstrating large main effects, and consequently has not the required sample size for demonstrating a small effect.

### 7.3.2 The Official Procedure for Demonstrating Equivalence

In the context of equivalence clinical trials, the need to specify a *smallness margin* of scientific relevance, not a conventional benchmark, is stressed. This margin must be defined according to the relative magnitude of the differences.

An equivalence margin should be specified in the protocol; this margin is the largest difference that can be judged as being clinically acceptable and should be smaller than differences observed in superiority trials of the active comparator... The choice of equivalence margins should be justified clinically (ICH E9 Expert Working Group, 2006, p. 18).

The officially recommended procedure is to use the U-CI, or equivalently two simultaneous one-sided tests – the so-called *Two One-Sided Tests* [TOST] procedure (e.g. Schuirmann, 1987):

For equivalence trials, two-sided confidence intervals should be used. Equivalence is inferred when the entire confidence interval falls within the equivalence margins. Operationally, this is equivalent to the method of using two simultaneous one-sided tests to test the composite null hypothesis that the treatment difference is outside the equivalence margins versus the alternative hypothesis that the treatment difference is within the margins (ICH E9 Expert Working Group, 2006, p. 18).



### How to get $100(1 - \alpha)\%$ confidence from a $100(1 - 2\alpha)\%$ U-CI

Following Westlake (1981), many authors (e.g. Deheuvels, 1984; Schuirmann, 1987; Rogers, Howard & Vessey, 1993; Steiger, 2004) have argued that the appropriate CI for demonstrating equivalence is the  $100(1 - 2\alpha)\%$ , not  $100(1 - \alpha)\%$ , U-CI. So with the traditional .05 criterion, the recommended procedure is to compute the 90% U-CI. The rationale is as follows: if the  $100(1 - 2\alpha)\%$  interval is symmetrized (hence enlarged) around zero by considering only the largest in absolute value of the two limits, the resulting interval is a (conservative)  $100(1 - \alpha)\%$  CI.

Of course, the frequentist interpretation of the confidence level requires that the procedure be decided independently of the data. If this has not been explicitly done before experiment, it will be suspected that the above procedure has been used in order to get a shorter interval. This looks like the endless one-sided *vs* two-sided tests debates.

The “optimal” noncentral  $F$  (NCF) based test and confidence interval procedures have always been discarded by biostatisticians. By definition, the  $p$ -value of the recommended procedure (the TOST) is the *larger* of the two  $p$ -values associated with each of the two one-sided tests, while the  $p$ -value of the NCF-test is the absolute value of *their difference*: a strange definition that explains its undesirable properties.

## 7.4 Isn't Everyone a Bayesian?

### The ambivalence of statistical instructors

Treating the data as random even after observation is so strange that the “correct” frequentist interpretation does not make sense for most users, who spontaneously use the Bayesian interpretation of CIs. This *heretic* interpretation is encouraged by the ambivalence of most frequentist statistical instructors. So, in his book *Statistics with Confidence*, Smithson (2005, pp. 160–161) characterized a 95% interval for a population mean as follows:

... so our interval is ...  $Pr(101.4 < \mu < 104.6) = 0.95$ .

This is obviously wrong, unless we abandon the frequentist requirement that  $\mu$  is a fixed quantity.

In another popular textbook that claims the goal of “understanding statistics”, we find the ambiguous definition:

A confidence interval is a range of values that probably contains the population value (Pagano, 2007, p. 309).

It is hard to imagine that the reader can understand that “a range of values” is a random variable and does not refer to the particular limits computed from the data in hand.

### A frequentist should avoid colloquialisms

Moreover, many authors claim that a CI can be characterized by a statement such as:

- she or he is 95% confident that the true percentage vote for a political candidate lies somewhere between 38% and 48% (Smithson, 2003, p. 1);
- the confidence interval has determined with 90% confidence that the main effect accounts for between 26.1% and 56.5% of the variance in the dependent variable (Steiger, 2004, pp. 169–170);
- the researchers [...] can be 90% confident that the true population mean is in an interval from 1.17 to 3.23 (Gravetter & Wallnau, 2010, p. 341).

The frequentist interpretation advocated by these authors assumes that these colloquialisms are true whatever the value of the parameter may be. Consequently, the conditioning on the parameter cannot be dropped, which leads to the *absurd* (in Neyman’s words) statements (see Section 7.1.3):

- if  $\mu = 2$ , we can be 90% confident that  $\mu$  is in an interval from 1.17 to 3.23;
- if  $\mu = 4$ , we can be 90% confident that  $\mu$  is in an interval from 1.17 to 3.23;

Actually, these colloquialisms give to understand that the confidence level may be a measure of uncertainty *after the data have been seen*, which it may not be.

### A typical confusion between frequentist and Bayesian probabilities

In a methodological paper, Rosnow and Rosenthal (1996) considered the example of an observed difference between two means +0.266, associated with a  $p$ -value .23. They defined the counternull value as twice the observed difference (see Section 5.5) and they interpreted the specific null-counternull interval  $[0, +0.532]$  as “a 77% confidence interval”, that is as a  $100(1 - p)\%$  CI. This cannot be a frequentist procedure, because the confidence level 77% has been determined by the data in hand. Clearly, .77 is here a data dependent probability, which needs a Bayesian approach to be correctly interpreted.

Virtually all users interpret frequentist confidence intervals in a Bayesian fashion. What a paradoxical situation: **Isn’t Everyone a Bayesian?**

## Chapter 8

# Basic Fiducial Bayesian Procedures for Inference About Means

We have extensively developed routine Bayesian procedures for inference about means. They are included in the LePAC package and are applicable to general experimental designs (in particular, repeated measures), with equal or unequal cell sizes, with univariate (ANOVA) or multivariate (MANOVA) data, and covariables. Some relevant references are: Lecoutre (1981, 1984, 2006); Rouanet and Lecoutre (1983). Rouanet (1996) and Le Roux and Rouanet (2004) also contain useful information, but in our opinion with too much emphasis on standardized ES.

In this chapter we will consider the basic fiducial Bayesian procedures for a contrast between means, which is an issue of particular importance for experimental data analysis. The presentation will be essentially non-technical. Within this perspective, we will give only intuitive justifications and we will focus on the computational and methodological aspects. Formal justifications can be found in Lecoutre (1984, 1996). A similar presentation for inferences about proportions is available elsewhere (Lecoutre, 2008).

We will only consider here the simplest and fastest ways to use the LePAC package. Actually, all fB procedures can be performed with very little effort. In most cases it is sufficient to know the values of an appropriate descriptive statistic (a contrast or an ANOVA ES) and of a valid test statistic (Student's  $t$  or  $F$  ratio). These values can be computed from usual statistical packages, or again obtained from a publication, which allows to reanalyze data with fB procedures. Of course, in order to get accurate results, it is important to enter "exact", or at least sufficiently accurate, numerical values.

## 8.1 Fiducial Bayesian Methods for an Unstandardized Contrast

### 8.1.1 *The Student Pharmaceutical Example*

Consider again the Student pharmaceutical example mentioned in 5.5. Given, for each of the  $n=10$  patients the two “additional hour’s sleep” gained by the use of two soporifics [1 and 2], Student illustrated his method for comparing the two treatment means.

#### A Bayesian answer

The terms in which the analysis was reported demonstrate the similarity between the Student and Jeffreys Bayesian conceptions, pointed out in Section 5.5:

But I take it the real point of the authors that 2 is better than 1. This we must test by making a new series, subtracting 1 from 2. The mean value of this series is  $+1.58$  while the S.D. is  $1.17$  [the uncorrected standard deviation], the mean value being  $+1.35$  times the S.D. From the table *the probability is .9985 or the odds are about 666 to 1 that 2 is the better soporific*. The low value of the S.D. is probably due to the different drugs reacting similarly on the same patient, so that there is correlation between the results (Student, 1908, p. 21, italics added).

The artifice of the null hypothesis was completely avoided in this presentation, which involved only the hypothesis of interest, “the real point of the authors that 2 is better than 1.” Student’s table provided a direct answer – the Bayesian probability .9985 – to the right question: “*What is the probability that 2 is better than 1?*” Note that the value .9985, obtained by interpolation, was remarkably accurate, the exact value being .99858...

Student introduced the standardized mean,  $+1.58/1.17 = +1.35$  (he used the uncorrected S.D. 1.17) as an intermediate value for his table. However, it must be emphasized that he did not comment about it; rather he interpreted the standard deviation as reflecting a high correlation between the two measures. Actually, the Pearson correlation coefficient was  $r = +.795$ .

### 8.1.2 *Specific Inference*

Student’s analysis is a typical example of specific inference about a contrast between means (see Rouanet & Lecoutre, 1983; Lecoutre, 2006). The basic data are for each of the  $n = 10$  patients the difference between the two “additional hour’s sleep gained by the use of hyoscyamine hydrobromide [an hypnotic],” the hour’s sleep being measured without drug and after treatment with either [1] “dextro hyoscyamine hydrobromide” or [2] “laevo hyoscyamine hydrobromide” (note that they already are derived data).

### The relevant data

The derived relevant data are obtained “by making a new series, subtracting 1 from 2.” They consist of the following ten individual differences of differences (in hours).

Patient	1	2	3	4	5	6	7	8	9	10	Mean	S.D.
1	+0.7	-1.6	-0.2	-1.2	-.1	+3.4	+3.7	+0.8	0	+2.0	+0.75	1.70
2	+1.9	+0.8	+1.1	+0.1	-.1	+4.4	+5.5	+1.6	+4.6	+3.4	+2.33	1.90
Individual difference (2-1)	+1.2	+2.4	+1.3	+1.3	0	+1.0	+1.8	+0.8	+4.6	+1.4	+1.58	1.17

We can apply to the relevant data the elementary Bayesian inference about a Normal mean, with only two parameters, the population mean difference  $\delta$  and the standard deviation  $\sigma$ . These data are summarized by the observed (unstandardized) difference  $d_{obs} = +1.58$  (do not confuse  $d_{obs}$  with Cohen’s  $d$ ) and the (corrected) standard deviation  $s_{obs} = 1.23$ . Of course, the difference  $+1.58$  must be reported to the gains of each soporific, respectively  $+0.75$  and  $+2.33$ , and these gains can only be interpreted by reference to the baseline sleep duration without soporific. The observed value of the usual  $t$  test statistic for the inference about a normal mean is  $t_{obs} = +1.58/(1.23/\sqrt{10}) = +4.06$  (9 df). For further generalization, this can be written


$$t_{obs} = \frac{d_{obs}}{b s_{obs}} \quad \text{with here } b = 1/\sqrt{10}.$$

### The fiducial Bayesian distribution

Assuming the Jeffreys prior (see Section 3.3), we get the posterior – or *fiducial Bayesian* [fB] distribution of  $\delta$ . This is a *generalized* (or scaled)  $t$ -distribution (which must not be confused with the noncentral  $t$ -distribution, familiar to power analysts). It is centered on  $d_{obs} = +1.58$  and has scale factor  $e = s_{obs}/\sqrt{n} = 0.39$ . The distribution has the same degrees of freedom  $q = 9$  as the  $t$ -test. This is written

$$\delta | \text{data} \sim d_{obs} + e t_q, \text{ or again } \delta | \text{data} \sim t_q(d_{obs}, e^2) \text{ by analogy with the Normal distribution.}$$

This distribution can be easily obtained in LePAC: see Figure 8.1.

Run LePAC, click on the icon  and enter the appropriate values for  $d_{obs}$ ,  $n$ ,  $s_{obs}$  and  $q$ , as in Figure 8.1. Then click on the distribution to get a new windows in which probability statements about this distribution can be interactively computed. Either the limits associated with a fixed probability (or *guarantee*) or the probability associated with one or two fixed limits can be obtained.

#### Remarks

- The fB distributions of the standard deviation  $\sigma$  and of the standardized difference  $\delta/\sigma$  are also obtained.
- The notation  $d$  for the raw difference can be changed in the Option menu.

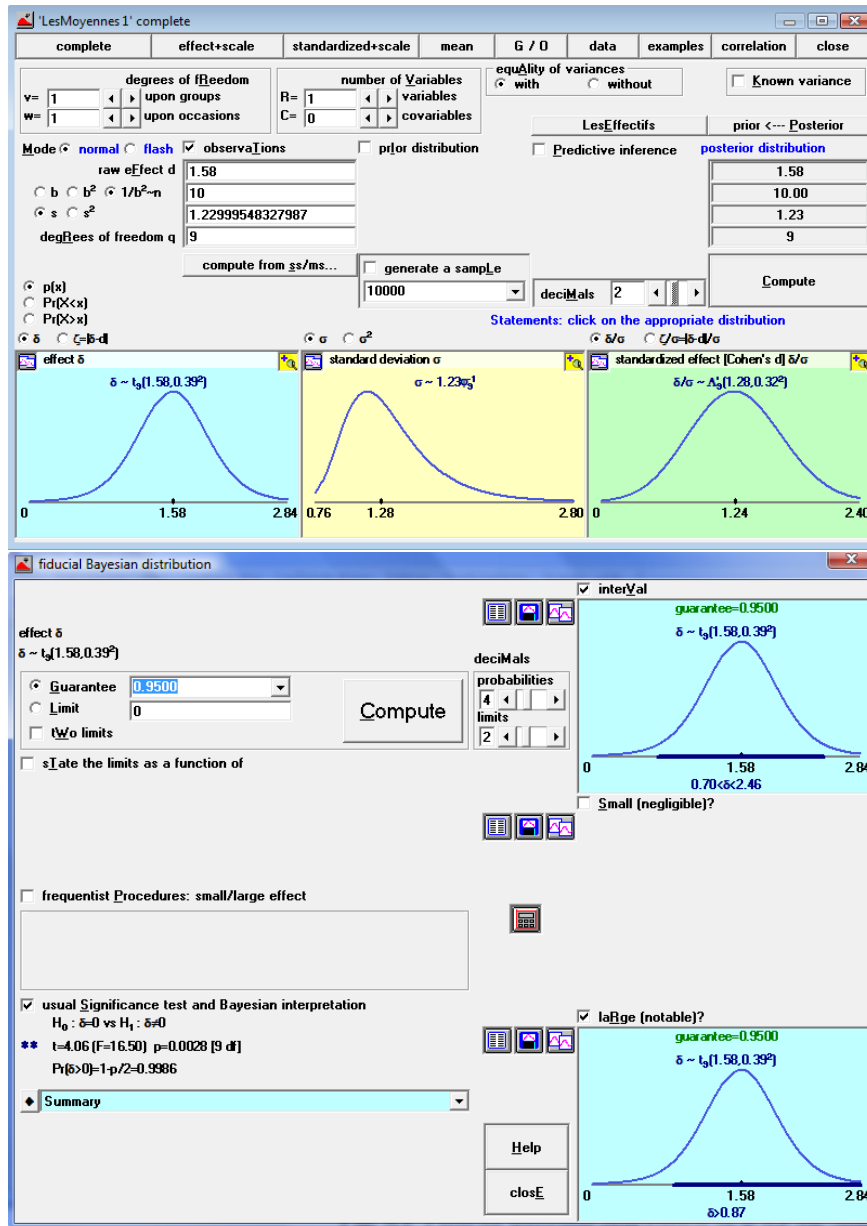


Fig. 8.1 Fiducial Bayesian distribution of  $\delta$  for the Student example computed by LePAC.

The scale factor  $e$  is the denominator of  $t$ -test statistic, that is

$$e = \frac{d_{obs}}{t_{obs}} \text{ (assuming } d_{obs} \neq 0 \text{)}.$$

The fiducial Bayesian distribution for a difference, and more generally for a contrast, between means can be directly derived from  $t_{obs}$ :

$$\delta \mid \text{data} \sim t_q(d_{obs}, e^2), \text{ where } e = bs_{obs} = \frac{d_{obs}}{t_{obs}} \quad (d_{obs} \neq 0).$$

This result brings to the fore the fundamental property of the  $t$ -test statistic of being an estimate of the experimental accuracy, *conditionally on the observed value*  $d_{obs}$ . More precisely,  $(d_{obs}/t_{obs})^2$  estimates the sampling error variance of the difference.

## 8.2 Fiducial Bayesian Methods for a Standardized Contrast

In Chapter 6 we argued against routine application of standardization (Section 6.4). However an inference about a standardized contrast may be of interest in some cases.

### 8.2.1 A Conceptually Straightforward Generalization

For deriving and computing an interval estimate for a standardized difference (or contrast)  $\delta/\sigma$  ('Cohen's  $d$ '), the traditional frequentist procedure involves the non-central  $t$ -distribution familiar to power analysts. One of its pre-eminent conceptual difficulties is the lack of explicit formula. Although the considerable advances in computing techniques are supposed to render the task easy, they do not solve the conceptual difficulties.

This is all the more deceptive in that, when the number of degrees of freedom is large enough, the confidence limits are given by the percent points of a Normal distribution, as for the simple difference. The fB distribution is (approximately) a Normal distribution, centered on  $d_{obs}/s_{obs}$ , with scale factor  $b = (d_{obs}/s_{obs})/t_{obs}$ . The exact solution is again a conceptually straightforward, only technically more complex, generalization. The distribution, which was considered (with no name) by Fisher (1990c, pp. 126–127) in the fiducial framework, was called *Lambda-prime* in Lecoutre (1999). It is an asymmetric distribution, the asymmetry being more pronounced when  $t_{obs}$  increases. The distribution has the same degrees of freedom  $df=9$  as the  $t$ -test.

The fiducial Bayesian distribution for a standardized difference, and more generally for a standardized contrast, between means can be directly derived from  $t_{obs}$ . This is written by analogy with the Normal distribution

$$\frac{\delta}{\sigma} | \text{data} \sim \Lambda_q^*(\frac{d_{obs}}{s_{obs}}, b^2), \text{ or in the standard form } \frac{\delta}{\sigma} | \text{data} \sim b\Lambda_q^*(t_{obs}),$$

where  $b = \frac{d_{obs}}{s_{obs}}$  ( $d_{obs} \neq 0$ ).

Here again, the Jeffreys' Bayesian credible, fiducial, and usual frequentist confidence intervals all coincide, and the distribution has no probability interpretation in the frequentist conception. The link between the frequentist and fB approaches is demonstrated in Lecoutre (2007), and a very accurate approximation is given.

### Numerical illustration

For the Student example ( $d_{obs}/s_{obs} = +1.285$ ), the fB distribution of  $\delta/\sigma$  (see Figure 8.1) is slightly asymmetric, with mean +1.249 and median +1.243. We have for instance the 90% interval estimate  $[+0.545, +1.975]$ . The two limits are respectively the 5 and 95 percent points of the  $\Lambda_9^*(+1.28, 0.32^2)$  distribution. This explicit and conceptually simple result can be contrasted to the process involved in the frequentist approach, which is described by Thompson (2002, p. 27) as “extremely technical. . . because a formula cannot be used for this process”.

## 8.2.2 Inference About the Proportion of Population Differences

It can be observed that 9 of the 10 individual differences are positive and are at least 0.8 hours, which explains the large value of the standardized difference. Actually, assuming a Normal population distribution of differences  $N(\delta, \sigma^2)$ , there is a one to one transformation between  $\delta/\sigma$  and  $\pi_{[0]}$ , the proportion of positive differences in the population. This determines the fB distribution of  $\pi_{[0]}$ . An interesting property is that the mean of this distribution is the predictive probability .874 (see Section 8.4.4) to find a positive difference in an additional experimental unit (Gertsbakh & Winterbottom, 1991). The interval bounds are easily deduced from those of  $\delta/\sigma$ . So, for a Normal distribution with a positive mean equal to 0.545 times its standard deviation, 70.7% of the values are positive, and consequently:

$$\Pr(\pi_{[0]} > .707 | \text{data}) = 0.95$$

Note that .707 is again an exact frequentist confidence limit.


In the same way, we can get the fB distribution of  $\pi_{[x]}$ , the proportion of population differences larger than  $x$  (and more generally included in a given range). It is deduced from the distribution of  $(\delta - x)/\sigma$ , which is obtained by simply replacing  $d_{obs}$  with  $d_{obs} - x$ . For instance  $\pi_{[+0.5]}$ , the proportion of differences larger than half an hour has a distribution with mean .788, the predictive probability to find a difference larger than 0.5 in an additional experimental unit. Instead of computing a confidence bound, we can select a minimum value of interest for  $\pi_{[+0.5]}$ , say 2/3,

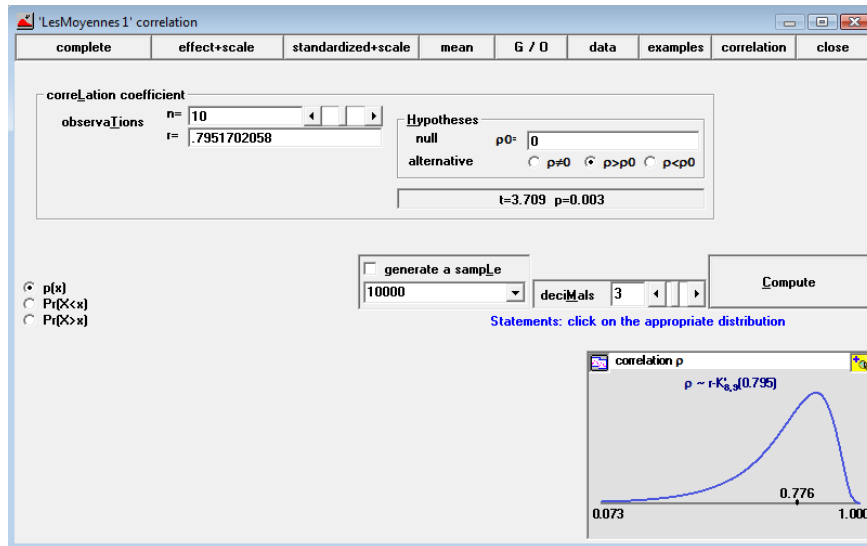


and compute the posterior probability that  $\pi_{[+0.5]}$  exceeds this value. We find here a .87% probability that the proportion of population differences larger than half an hour exceeds  $2/3$ .

### 8.3 Inference About Pearson's Correlation Coefficient

The fB inference about the Pearson correlation coefficient is also a conceptually straightforward generalization, (Poitevineau & Lecoutre, 2010). For instance, for the Student data, consider the correlation coefficient  $\rho$  between the two series. The fB distribution of  $\rho$  can be obtained in LePAC: see Figure 8.2.

Run LePAC, click on the icon , click on the button 'correlation', and enter the appropriate values for  $n$  and  $r$ , as in Figure 8.2. Probability statements about this distribution can be interactively computed as for  $\delta$ .



**Fig. 8.2** fB distribution of  $\rho$  for the Student example:  $\rho | data \sim rK_{8,9}(+.795)$ , median = .776,  $Pr(\rho > 0) = .997$ ,  $Pr(\rho > +.417) = .95$ .

It can be stated that

there is a 99.7% probability of a positive correlation and a 95% probability of a correlation larger than +.417.

The probability that  $\rho$  has the opposite sign of the observed coefficient (.003) is exactly the one-sided  $p$ -value of the usual test of a null correlation (given by a  $t$ -distribution). And the 95% equal two-tailed credible interval  $[+.312, +.940]$  is the (exact) frequentist 95% usual CI.

## 8.4 A Coherent Bayesian Alternative to GHOST

### 8.4.1 NHST: The Fiducial Bayesian Interpretation of the $p$ -Value

The fB probability that the population difference  $\delta$  has the opposite sign of the observed difference is exactly the one-sided  $p$ -value of the  $t$ -test. The Bayesian interpretation clearly points out the methodological shortcomings of NHST. It becomes apparent that the  $p$ -value *in itself* says nothing about the magnitude of  $\delta$ . On the one hand, even a very *small*  $p$  (“highly significant”) only establishes that  $\delta$  has the same sign as the observed difference  $d_{obs}$ . On the other hand, a *large*  $p$  (nonsignificant) is hardly worth anything, as exemplified by the fB interpretation  $\Pr(\delta < 0) = \Pr(\delta > 0) = 1/2$  of a “perfectly nonsignificant” test (i.e.  $d_{obs} = 0$ ).

### 8.4.2 Interval Estimates: The Fiducial Bayesian Interpretation of the Usual CI

When the number of degrees of freedom is large enough, so that the Normal approximation holds, the  $100(1 - \alpha)\%$  usual confidence interval is given by the formula:

$$d_{obs} \pm z_{\alpha/2} \frac{d_{obs}}{t_{obs}}, \text{ where } z_{\alpha/2} \text{ is the } 100\alpha/2\% \text{ upper point of the Normal distribution.}$$

Otherwise,  $z_{\alpha/2}$  is replaced with the upper point of the  $t$ -distribution with  $df$  degrees of freedom. It results that, for a contrast between means, the Jeffreys credible, Fisher fiducial, and Neyman confidence intervals *all coincide*, but in the fB framework, *it is correct* to say that

there is a 95% probability (or *guarantee*) of  $\delta$  being included between the fixed bounds of the interval (given data), i.e. for the Student example between +0.70 and +2.46 hours.

### 8.4.3 Effect Sizes: Straight Bayesian Answers

#### Asserting largeness

We can compute the probability that  $\delta$  exceeds a fixed, easier to interpret, extra time of sleep. For instance, there is a .915 probability of  $\delta$  exceeding one hour. Since the

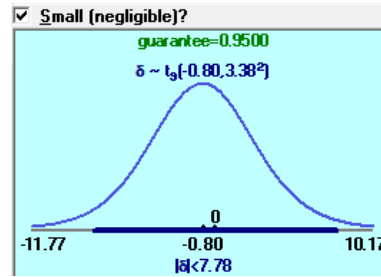
units of measurement are meaningful, the practical significance of the magnitude of  $\delta$  can be assessed. To summarize the results, it can be reported that

there is a .915 posterior probability of a large positive difference ( $\delta > +1$ ), a .084 probability of a positive but limited difference ( $0 < \delta < +1$ ), and a .001 probability of a negative difference.

Such a statement has no frequentist counterpart and should have a real impact on the way the authors and their readers interpret experimental results. This should escape the shortcomings of frequentist CIs, which are only used by most users “to do a significance test,” only wondering whether the CI includes zero.

### Asserting smallness

Given smallness (or equivalence) margins, we can compute the probability that  $\delta$  lies within these margins. So, in the psychological example, (Section 6.2). assume that a difference  $\delta$  of 5% between the two conditions, either in the positive or negative direction, could be considered as relatively small. From  $d_{obs} = -0.80\%$  and  $t_{obs} = -0.237$  ( $p = .818$ , two-sided), we get, in the same way as for the Student example, the fB distribution in Figure 8.3.



**Fig. 8.3** fB distribution of  $\delta$  for the psychological example:  $\delta|data \sim t_9(-0.80, 3.38^2)$ ,  $\Pr(\delta < -5\%) = .122$ ,  $\Pr(\delta > +5\%) = .060$ ,  $\Pr(-7.78\% < \delta < +7.78\%) = .95$ .

There is a .409 ( $p/2$ ) probability of a positive difference and a .591 probability of a negative difference.

There is a .122 probability of a negative difference smaller than -5%, a .060 probability of a positive difference larger than 5%, and a .818 probability of a difference lying within the smallness range  $[-5\%, +5\%]$ .

Alternatively, we can compute a 95% credible interval centered on zero (and not on the observed difference):  $[-7.78\%, +7.78\%]$ , which includes differences larger than 5% in absolute value. Given the very small sample size, this can be viewed as an encouragement to perform a more decisive experiment, with a higher sample size and, likely, a more stringent smallness criterion. Of course, the decision should take into consideration the scientific interest of the study.

We get again a clear understanding of the frequentist procedures. The probabilities .122 and .060 are the significance levels of the two one-sided tests involved in the “official” TOST procedure (see Section 7.3.2). This procedure gives for  $\delta$  a 95% CI centered on zero  $[-6.99\%, +6.99\%]$ , deduced from the 90% U-CI  $[-6.99\%, +5.39\%]$ . The 95% fB credible interval  $[-7.78\%, +7.78\%]$  appears as an acceptable compromise with the 95% U-CI,  $[-8.44\%, +6.84\%]$ , avoiding the debates about the choice of the confidence level (see Section 7.3.2).

#### 8.4.4 Making Predictions

An important aspect of statistical inference is making predictions. For instance, what can be said about the difference  $d'_{obs}$  that would be observed for additional data? The fB posterior predictive distribution for this difference in a future sample of size  $n'$  is again a generalized  $t$ -distribution, naturally centered on  $d_{obs}$ ,

$$d' \mid \text{data} \sim t_q(d_{obs}, e^2 + e'^2), \text{ where } e' = s_{obs}/\sqrt{n'}.$$

Of course, this predictive distribution is more scattered than the fB distribution of the population difference  $\delta$ . The uncertainty about  $\delta$  given the available data, reflected by  $e^2$ , is added to the uncertainty about the additional data, reflected by  $e'^2$ . This is all the more true since the size  $n'$  of the future sample is smaller. For instance, in Student's example, we get the predictive distributions.

- For an additional experimental unit ( $n' = 1$ ),  $d' \sim t_9(+1.58, 1.29^2)$ . There is a .874 predictive probability of a positive difference and a .788 probability of a difference exceeding half one hour.
- For ten additional experimental units ( $n' = 10$ ,  $e' = e$ ),  $d' \sim t_9(+1.58, 0.55^2)$ . There is a .991 predictive probability of a positive difference (Killeen's *probability of replication*: see Section 5.5) and a .959 probability of a difference exceeding half one hour.

The fB posterior predictive distribution is obtained in the same way as the distribution of  $\delta$ . Activate the checkbox ‘Predictive inference’ and enter the future  $n$  or, if appropriate, activate the checkbox ‘Replication’: see Figure 8.4.

##### Remark

- The predictive distributions of the standard deviation, and of the  $t$  test statistic or of the confidence (or fB credible) limits are also obtained.

The predictive fiducial Bayesian distribution for a difference, and more generally for a contrast, between means in an exact replication of an experiment (same sample size) can be directly derived from  $t_{obs}$ :

$$d' \mid \text{data} \sim t_q\left(d_{obs}, 2\left(\frac{d_{obs}}{t_{obs}}\right)^2\right) \quad (d_{obs} \neq 0).$$

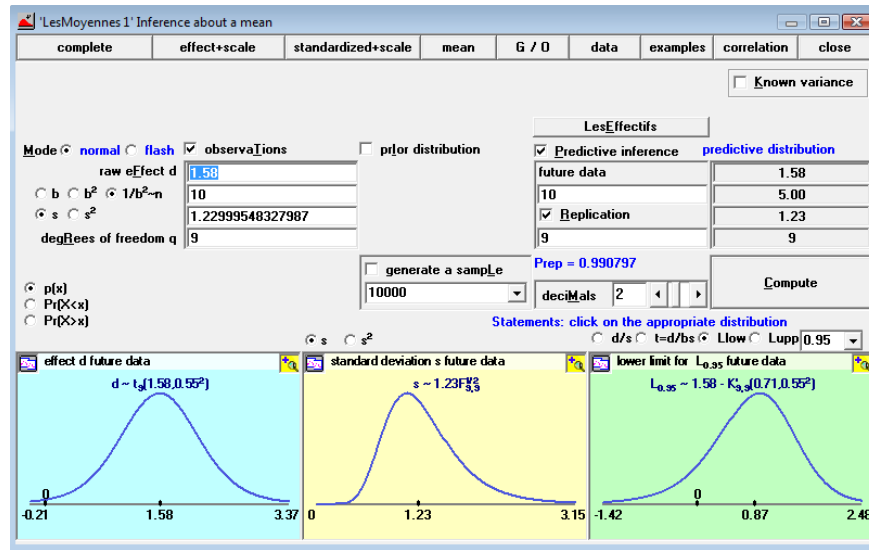


Fig. 8.4 fB predictive distributions for a replication of Student experiment ( $n' = 10$ ).

### 8.4.5 Power and Sample Size: Bayesian Data Planning and Monitoring

Bayesian predictive procedures give users a very appealing method to answer essential questions such as:

- How big should be the experiment to have a reasonable chance of demonstrating a given conclusion?
- Given the current data at an interim analysis, what is the chance that the final result will be in some sense conclusive, or on the contrary inconclusive?

Predictive probabilities give them direct answers. In particular, from a pilot study, the predictive probabilities on credibility limits are a useful summary to help in the choice of the sample size of an experiment. Predictive procedures can also be used to aid the decision to abandon an experiment if the predictive probability appears poor (see Lecoutre, Derzko & Grouin, 1995; Lecoutre, 2001; Lecoutre, Mabika & Derzko, 2002).

## 8.5 Our Guidelines

- Consider experimental data analysis as a problem of pure estimation in Jeffreys' sense (null hypothesis is not needed).
- Don't use noncentral  $F$  based tests and confidence intervals.
- **Don't Worry, Be Bayesian:** Think about  $p$ -values and usual confidence intervals in Bayesian terms and use their fiducial Bayesian interpretation.

## Chapter 9

# Generalizations and Methodological Considerations for ANOVA

The elementary fB procedures for a difference between means only involve the observed effect and the  $t$ -test statistic. Consequently, they are directly applicable to a contrast in ANOVA design for which a valid  $t$  or  $F$ -test is available – recall that in this case  $F = t^2$ .

For a contrast in ANOVA design, replace  $\frac{d_{obs}}{t_{obs}}$  with  $\frac{|d_{obs}|}{\sqrt{F_{obs}}}$ .

In this chapter, we will present straightforward generalizations of the fB procedures for a contrast between means easily applicable to the usual unstandardized and standardized ANOVA ES indicators. First, we will put aside methodological considerations about the risks of misuses and misinterpretations of these indicators (see Chapter 6). Then we will discuss methodological aspects and consider appropriate alternatives.

### 9.1 From $F$ tests to Fiducial Bayesian Methods for ANOVA Effect Sizes

Frequentist procedures for constructing confidence intervals for the conventional ANOVA ES indicators have received considerable attention in the last years. However, in spite of several recent presentations claiming that these intervals “can be easily formed” (e.g. Kelley, 2007), many potential users continue to think that deriving and computing them is a very complex task. The traditional frequentist intervals involve the noncentral  $F$  distributions, familiar to power analysts. Again, the lack of explicit formula (without speaking here of the inconsistencies of noncentral  $F$  based CIs discussed in Section 7.3.1) renders the task conceptually difficult.

### 9.1.1 The Traditional Approach

#### The Reaction Time Example

Consider the following example, derived from Holender and Bertelson (1975). In a psychological experiment, the subject must react to a signal. The experimental design involves two crossed repeated factors: Factor *A* (signal frequency) with two levels (*a1*: frequent and *a2*: rare), and Factor *B* (foreperiod duration), with two levels (*b1*: short and *b2*: long). The main research hypothesis is a null (or about null) interaction effect between factors *A* and *B* (*additive model*). There is also a between-subject Factor *G*, classifying the 12 subjects into three groups of 4 subjects each.

#### The ANOVA table

Here the basic data consists of three “groups” and four “occasions” of measure. The dependent variable is the reaction times in ms (averaged over trials). These data have been previously analysed in detail with Bayesian methods in Rouanet and Lecoutre (1983), Rouanet (1996), Lecoutre and Derzko (2001) and Lecoutre (2006). Consider here the main effect of factor *G* (with 2 df).

Frequentist CIs for ANOVA ES are derived from the sampling distribution of the *F* ratio, which can be deduced from the traditional ANOVA table for the between subjects source of variation:

Source	<i>SS</i>	df	<i>MS</i>	<i>E(MS)</i>	<i>F<sub>obs</sub></i>	<i>p</i> -value
<i>G</i>	7 960.17	2	3 980.08	$16\sigma_G^2 + 4\sigma_{\text{error}}^2$	0.5643	.588
Error	63 480.56	9	7 053.40	$4\sigma_{\text{error}}^2$		

#### ANOVA ES indicators

In this table, the effect variance  $\sigma_G^2$  is an unstandardized ES indicator. It is the variance (corrected for df) of the population group means  $\mu_g$ . An indicator expressed in the unit of measurement is usually preferred. Let *k* be the number of groups (here *k* = 3). Three possible alternatives, which are of the general form *c*  $\sigma_G$ , are

- the corrected standard deviation

$$\sqrt{\frac{(\mu_1 - \bar{\mu})^2 + (\mu_2 - \bar{\mu})^2 + \dots + (\mu_k - \bar{\mu})^2}{k-1}} = \sigma_G,$$

- the uncorrected standard deviation

$$\sqrt{\frac{(\mu_1 - \bar{\mu})^2 + (\mu_2 - \bar{\mu})^2 + \dots + (\mu_k - \bar{\mu})^2}{k}} = \sqrt{\frac{k-1}{k}} \sigma_G,$$



- the quadratic mean of all the partial differences between the population means

$$\sqrt{\frac{(\mu_1 - \mu_2)^2 + (\mu_2 - \mu_3)^2 + \dots + (\mu_k - \mu_1)^2}{\frac{k(k-1)}{2}}} = \sqrt{2} \sigma_G,$$

which is a direct generalization of the case of two means: when  $k = 2$  it reduces to the absolute value of the difference  $|\mu_1 - \mu_2|$ . For this reason it is our favorite indicator, implemented in the LePAC software.

In the ANOVA table,  $\sigma_{\text{error}}^2$  is the error variance. The usual standardized ES indicators are functions of  $\sigma_G^2 / \sigma_{\text{error}}^2$ . Corresponding to the three above alternatives, we have respectively

- the root-mean-square standardized effect (Steiger, 2004, p. 169)  $\frac{\sigma_G}{\sigma_{\text{error}}}$ ,
- the Cohen's  $f$  (Cohen, 1977, p. 275)  $\sqrt{\frac{k-1}{k}} \frac{\sigma_G}{\sigma_{\text{error}}}$ ,
- the standardized quadratic mean  $\sqrt{2} \frac{\sigma_G}{\sigma_{\text{error}}}$ .

There are simple relationships between these indicators and the partial eta-squared, in particular

$$\eta_p^2 = \frac{f^2}{f^2 + 1}$$

### Sampling distributions

In the ANOVA table, the coefficients 16 and 4 are respectively the number of observations for every level of  $G$  and the number of occasions. The sampling distribution of the effect mean-square  $MS_G$ , with  $m = 2$  df, is a noncentral chi-square distribution with noncentrality parameter  $8\sigma_G^2 / \sigma_{\text{error}}^2$  and scale factor  $\sigma_{\text{error}}^2 / 2$ . The sampling distribution of the error mean-square  $MS_{\text{error}}$ , with  $q = 9$  df, is a central chi-square distribution with scale factor  $\sigma_{\text{error}}^2 / 9$ .

We get the sampling distribution of the  $F$  ratio, a noncentral  $F$  distribution

$$F = \frac{MS_G}{MS_{\text{error}}} | \sigma_G^2, \sigma_{\text{error}}^2 \sim F_{2,9} \left( 8 \frac{\sigma_G^2}{\sigma_{\text{error}}^2} \right).$$

from which frequentist (“noncentral  $F$  based”) CIs for standardized ES indicators are derived. Due to the nuisance parameter  $\sigma_{\text{error}}^2$ , no exact solution is available for the unstandardized case. This could partly explain why this case is rarely considered.

### 9.1.2 Fiducial Bayesian Procedures

#### The relevant statistics

Here the relevant data consist of the following twelve individual means, averaged on the four occasions (in ms)

Group	$g_1$				$g_2$				$g_3$			
Subject	1	2	3	4	5	6	7	8	9	10	11	12
Individual data	427.75	342.00	360.75	379.75	434.00	382.25	360.75	443.75	412.25	381.50	356.25	468.25
Mean	377.5625				405.1875				404.5625			
SD	36.8371				40.0751				48.2396			

Let us denote by the general notation  $\lambda$  any population parameter proportional to  $\sigma_G$ :

$$\lambda = c \sigma_G.$$

For a given unstandardized ES indicator  $\lambda$ , let  $\ell$  be the corresponding observed indicator, obtained by replacing the population means  $\mu_g$  by the observed means. For instance, the observed value of the quadratic mean of all the partial differences ( $c = \sqrt{2}$ ) is

$$\ell_{obs} = \sqrt{\frac{(377.5625 - 405.1875)^2 + (405.1875 - 404.5625)^2 + (404.5625 - 377.5625)^2}{3}} = \sqrt{497.5104} = 22.3049 \text{ ms.}$$

For simplicity, let us write again  $\sigma^2$  in place of  $\sigma_{\text{error}}^2$ . Let  $s^2$  denote its usual estimate, i.e. the within-group variance of the individual data. The observed value is

$$s_{obs}^2 = \frac{36.8371^2 + 40.0751^2 + 48.2396^2}{2} = 1763.3490 \quad (s_{obs} = 41.9922 \text{ ms}).$$

The statistics  $\ell^2$  and  $s^2$  are proportional to the mean-squares. With general notations, this can be written

$$\ell^2 = a^2 b^2 MS_{\text{effect}} \quad \text{and} \quad s^2 = a^2 MS_{\text{error}},$$

where  $a^2$  and  $b^2$  are appropriate constants. In the above example, it can easily be verified that

$$a^2 = \frac{1}{4} \text{ and } b^2 = \frac{1}{2}.$$

Their means (or expectations) are

$$E(\ell^2) = \lambda^2 + b^2 \sigma^2 \text{ and } E(s^2) = \sigma^2.$$

and the sampling distribution of the ratio  $\ell^2/s^2$  is the noncentral  $F$  distribution

$$\frac{\ell^2}{s^2} = b^2 F \mid \lambda^2, \sigma^2 \sim b^2 F_{m,q}^* \left( m \frac{\lambda^2}{b^2 \sigma^2} \right).$$

It follows that the observed value of the  $F$  ratio can be written

$$F_{obs} = \left( \frac{\ell_{obs}}{b s_{obs}} \right)^2, \text{ here } F_{obs} = \left( \frac{22.3049}{(1/\sqrt{2}) 41.9922} \right)^2 = 0.5643$$

### Remarks

The notations ensure a direct generalization of the inference about a contrast between means.

- The constant  $a^2$  plays no role in the inference. It simply ensures the link with the ANOVA mean squares.
- When  $m = 1$ ,  $\ell_{obs}$  is proportional to the absolute value of a contrast, so that  $b$  is precisely the constant considered in this case. For instance, in the above example, if we compare groups  $g_1$  and  $g_2$ , using the above estimate of  $\sigma$ , we have  $d_{obs} = 377.5625 - 405.1875 = -27.6250$  and  $\ell_{obs} = |d_{obs}| = 27.6250$ . The  $F$  statistic is the square of the usual  $t$  test statistic. So here

$$t_{obs} = \frac{377.5625 - 405.1875}{41.9922\sqrt{1/4 + 1/4}} = -0.9304 \text{ and } F_{obs} = \left( \frac{27.6250}{(1/\sqrt{2})41.9922} \right)^2 = 0.8656 = t_{obs}^2$$

- The above relationship demonstrates that, in the balanced case with equal group sizes  $\bar{n}$ , the constant  $b^2$  for the quadratic mean of the differences is

$$b^2 = \frac{2}{\bar{n}}$$

### fb distributions of $\lambda^2$ and $(\lambda/\sigma)^2$

The fiducial Bayesian solution is a straightforward generalization, which introduces no additional conceptual difficulty, but only involves two new distributions.

For any population parameter  $\lambda^2$  proportional to the ANOVA effect variance  $\sigma_G^2$ , The fb distributions of  $\lambda^2$  and  $(\lambda/\sigma)^2$  can be directly derived from  $F_{obs}$ . They are respectively a *Psi-square* and a *Lambda-square* distribution.

- $\lambda^2 | \text{data} \sim e^2 \Psi_{m,q}^2(mF_{obs})$ , where  $e = bs_{obs} = \frac{\ell_{obs}}{\sqrt{F_{obs}}}$  ( $F_{obs} \neq 0$ ).
- $(\frac{\lambda}{\sigma})^2 | \text{data} \sim b^2 \Lambda_{m,q}^2(mF_{obs})$ , where  $b = \frac{\ell_{obs}}{\sqrt{F_{obs}}}$  ( $F_{obs} \neq 0$ ).

In particular, for the standardized ES we have a simple and intuitive result: the  $F$  ratio and the noncentrality parameter are permuted and the noncentral  $F$  distribution is replaced with the Lambda-square distribution

The Lambda-square distribution has been used by Geisser (1965), with no name. It has been considered by Schervish (1992, 1995) under the name of alternate chi-square distribution, with a different scale factor (see also Lecoutre & Rouanet, 1981). The Psi-square distribution has been introduced in Lecoutre (1981). It has been considered by Schervish (1992, 1995) under the name of alternate  $F$  distribution. Note that in central case, the Lambda-square and Psi-square distributions are respectively the chi-square (up to a constant of proportionality) and  $F$  distributions, which justify Schervish's names.

These two distributions have been characterized as particular cases of the  $K$ -square distribution in Lecoutre (1999) and algorithms (see Lecoutre et al., 1992; Poitevineau & Lecoutre, 2010) have been implemented in the LePAC package.

The fB distribution of the partial eta-squared is derived from the lambda-square distribution by a one-to-one transformation of the form  $\eta_p^2 = \lambda^2 / (\lambda^2 + 1)$ . Note that this relation is analogous to the relation between Beta and  $F$  distributions, so that the fB distribution of  $\eta_p^2$  is a kind of alternate Beta distribution.

### Numerical Application

The above results apply to any ES of the form  $\lambda = c \sigma_G$  and  $\lambda / \sigma$ . The appropriate constant  $b$  is determined from the observed corresponding values. For instance, for the quadratic mean of all the partial differences ( $c = \sqrt{2}$ ), we have


$$\ell_{obs} = 22.3049 \quad \frac{\ell_{obs}}{s_{obs}} = 0.5312 \quad F_{obs} = 0.5643.$$

hence the fB distributions

$$\lambda = \sqrt{2} \sigma_G | data \sim 29.69 \psi_{2,9}(1.129) \quad (e = \frac{22.3049}{\sqrt{0.5643}} = 29.69),$$

$$\frac{\lambda}{\sigma} | data \sim 0.707 \Lambda_{2,9}(1.129) \quad (b = \frac{0.5312}{\sqrt{0.5643}} = 0.707).$$

These distributions reflect a great uncertainty about the ES. For instance, there is a 90% fB probability that  $\lambda$  is smaller than 60.6 ms. Clearly, in spite of the nonsignificant test, it cannot be concluded to a small effect

The simplest and fastest way to get these distributions with LePAC is to simply enter the observed ES and the  $F$  ratio. For obtaining the fB distribution of  $\lambda$ , Run LePAC, click on the icon , click on the button ‘effect+scale’, and enter the appropriate values for  $\ell_{obs}$ ,  $F_{obs}$  and  $q$ , as in Figure 9.1. Probability statements about this distribution can be interactively computed as in the above cases. For obtaining the fB distribution of  $\lambda / \sigma$ , proceed in the same way, but use the button ‘standardized+scale’.

#### Remark

- The fB distributions of the parameter  $\zeta$  (“deviation from observed effect”) is also obtained: see Section 9.2.1.

### 9.1.3 Some Conceptual and Methodological Considerations

Interval estimates for standardized ANOVA ES are often offered as a natural generalization of the usual confidence interval for a single contrast (e.g., Steiger and Fouladi 1997, p. 244). However this claim is not justified. Certainly, such interval estimates, as well as those for unstandardized ES, are simultaneous estimates for all contrasts, but they are a generalization of the confidence interval for the *absolute*

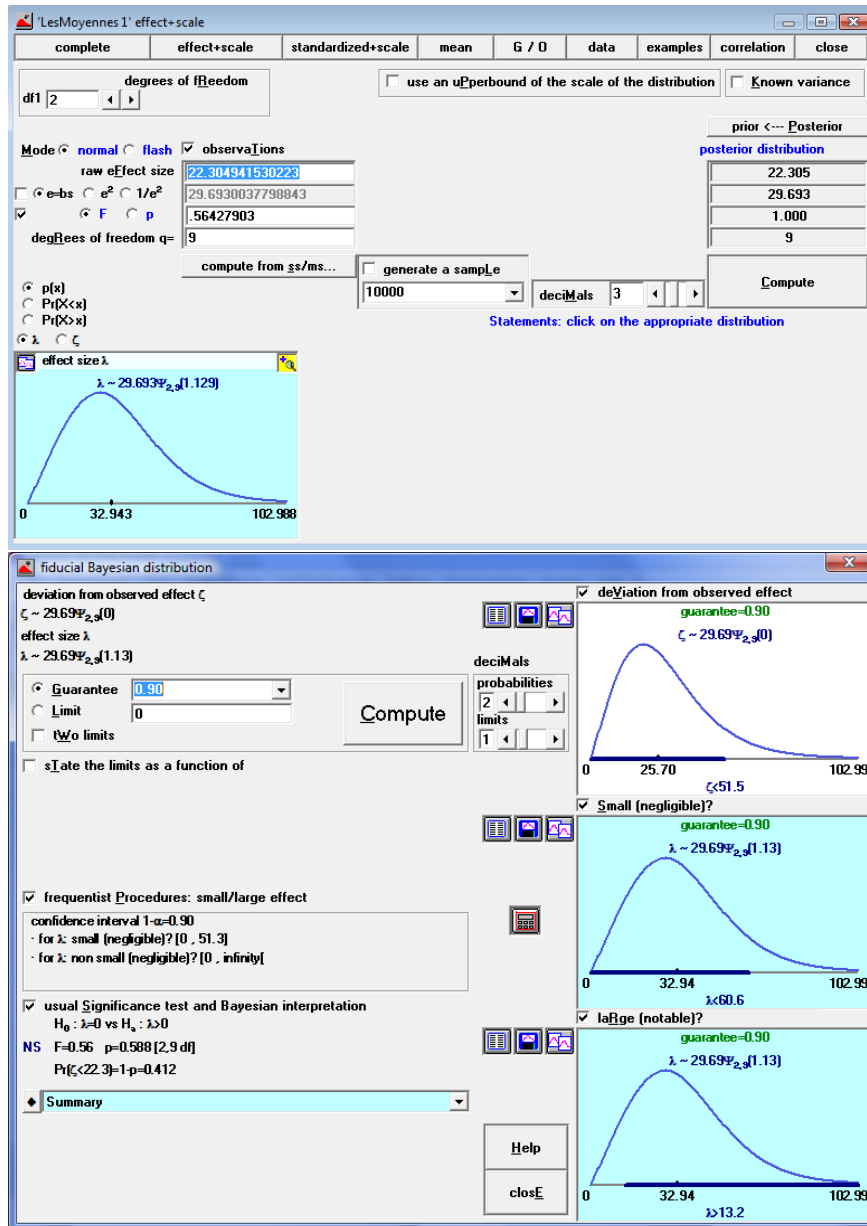


Fig. 9.1 Fiducial Bayesian distribution of  $\lambda$  for the reaction time example.

value of a contrast, that is of an interval centered on zero and not on the observed signed contrast. So they give no indication about the direction of each contrast. This

is the reason why they are appropriate for asserting smallness, but not really for asserting largeness.

### Asserting smallness

For asserting smallness, it must be demonstrated that all contrasts are in a sense simultaneously close to zero. Consequently, an *upper limit* for an ANOVA ES is clearly appropriate.

### Asserting largeness

In the one df case ( $m = 1$ ), a *lower limit* for an ANOVA ES does not provide any indication about the direction. It is unquestionable that an inference about the signed contrast should be preferred. In the several df case, it cannot be expected that the generalization of an inappropriate procedure could be a “good statistical practice.” At the best it could provide a rough indication, but it should always be followed by a more detailed investigation.

### The right use of simultaneous interval estimates

Actually, there is a well-known straightforward generalization of the U-CI, available both for the unstandardized and standardized cases, namely the Scheffé simultaneous interval estimate (Scheffé 1953). If we are really interested in a simultaneous inference about all contrasts, except for the purpose of asserting overall smallness, this is obviously the appropriate procedure. Moreover, it can receive both a frequentist and a fB interpretation.

## 9.2 Alternatives to the Inference About ANOVA ES

### 9.2.1 The Scheffé Simultaneous Interval Estimate and Its Bayesian Justification

It is worthwhile to note that the Scheffé marginal interval estimate for a particular contrast can be viewed (and computed) as a  $100(1 - \tilde{\alpha})\%$  U-CI. Of course,  $1 - \tilde{\alpha}$  is sharply larger than  $1 - \alpha$  in order to ensure for the set of all contrasts:

- the frequentist interpretation of a simultaneous confidence level  $100(1 - \alpha)\%$ ,
- the fB interpretation of a joint posterior probability  $1 - \alpha$ , the marginal probability for each contrast being  $1 - \tilde{\alpha}$ .

An interval estimate for an ANOVA ES is a simultaneous interval estimate for the absolute value of all contrasts. So it is only appropriate for asserting overall smallness. The generalization of the usual  $100(1 - \alpha)\%$  confidence interval for a signed contrast is the Scheffé interval. In practice, it can be computed as the  $1 - \tilde{\alpha}$  U-CI, where the appropriate value  $\tilde{\alpha}$  is given by the relationship

$$F_{1,q;\tilde{\alpha}} = mF_{m,q;\alpha}.$$

For instance, when  $1 - \alpha = .90$ ,  $1 - \tilde{\alpha} = .9634$  ( $q = 9$ ),  $1 - \tilde{\alpha} = 0.9660$  ( $q = 20$ ),  $1 - \tilde{\alpha} = 0.9677$  ( $q = 100$ ) and  $1 - \tilde{\alpha} = 0.9681$  ( $q = \infty$ ).

### Numerical application

For illustration, assume that the above reaction time statistics – means and standard deviations – have been obtained from an experiment with 100 subjects in each group. In this case, we have here  $F_{obs} = 14.1070$  ( $p = .000001$ ), hence  $e = 22.3049/\sqrt{14.1070} = 5.94$ . For  $1 - \alpha = .90$  we get in particular the marginal interval estimates for the three raw differences between groups, which are the .9680 ( $q = 297$ ) U-CI:

$$\begin{array}{ll} g_2, g_1 & [+14.8, +40.4] \\ g_2, g_3 & [-12.2, +13.4] \\ g_3, g_1 & [+14.2, +39.8] \end{array}$$

These estimates are not very precise, due to the large error variance. Nevertheless, it can be concluded that the average reaction time is lower in group  $g_1$  and that the two other groups are relatively equivalent. Note that the original experiment was designed to study the within-group factors for which the variability is weak. This justify the small sample size.

### The fB Justification

Here again the fB justification is straightforward. Consider the (infinite) set of contrasts  $c_1\mu_1 + c_2\mu_2 + c_3\mu_3$  such that  $c_1^2 + c_2^2 + c_3^2 = 2$ , which includes the partial differences between two means. The Scheffé simultaneous interval estimate implies that all these contrasts lie within a circle (an hypersphere for higher  $m$ ) centered on the observed contrasts. Consequently, the procedure is equivalent to an inference about the radius (or a proportional quantity) of this circle.

While  $\lambda$  is an average deviation of the  $m$ -dimensional population effect from zero, it is relevant for the generalization of the U-CI to consider the average deviation  $\zeta$  from the observed effect. In particular, let us define  $\zeta$  as the quadratic mean of the deviations between the partial population and observed differences, hence here

$$\zeta = \sqrt{\frac{\left((\mu_1 - \mu_2) - (377.5625 - 405.1875)\right)^2 + \left((\mu_2 - \mu_3) - (405.1875 - 395.7708)\right)^2 + \left((\mu_3 - \mu_1) - (395.7708 - 377.5625)\right)^2}{3}}.$$

The relevant parameter for a simultaneous inference about all signed contrasts is the quadratic mean  $\zeta$  of the deviations between the partial population and observed differences (or a proportional parameter). The fB distributions of  $\zeta^2$  and  $(\zeta/\sigma)^2$  are respectively a central Psi-square, hence a usual  $F$  distribution, and a central Lambda-square, hence a usual chi-square distribution.

- $\zeta^2 | \text{data} \sim e^2 \psi_{m,q}^2(0)$  [or  $e^2 F_{m,q}$ ].
- $(\frac{\zeta}{\sigma})^2 | \text{data} \sim b^2 \lambda_{m,q}^2(0)$  [or  $b^2 \frac{\chi_m^2}{m}$ ].

### Numerical application

For the modified reaction time example, the fB distribution  $\zeta$ , shown in Figure 9.2 is obtained as for the original data (Figure 9.1), with  $F = 14.1070$  and  $q = 297$ .

$$\zeta | \text{data} \sim 5.94 \psi_{2,297}(0),$$

hence  $\Pr(\zeta < 9.05) = .90$ .

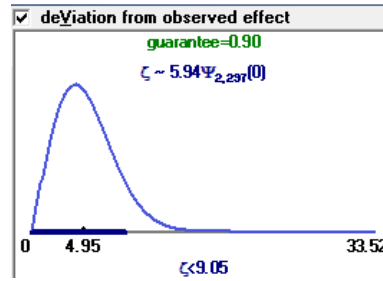


Fig. 9.2 fB distribution of  $\zeta$  for the modified reaction time example.

We can conclude with a .90 fB guarantee that the average deviation between the partial population and observed differences is less than 9.05 ms. This limit 9.05 is equal to the common half-width of the marginal Scheffé interval estimates, divided by  $\sqrt{m}$ , which takes into account the dimensionality of the effect, for instance:

$$9.05 = \frac{\frac{1}{2}(40.4 - 14.8)}{\sqrt{2}} = \frac{12.8}{\sqrt{2}}.$$

### The fiducial Bayesian interpretation of the $p$ -value

The fB probability that  $\zeta$  exceeds  $\ell_{obs}$  is exactly the  $p$ -value of the  $F$ -test, so here

$$\Pr(\zeta > 22.3049) = .000001.$$



This generalizes the Bayesian interpretation of the two-sided  $p$ -value for a difference between means, given in Section 5.5, which is the posterior probability that  $\delta$  lies outside the interval bounded by 0 (the null hypothesis value) and twice the observed difference (the counternull value). For instance, if  $d_{obs} > 0$

$$\Pr(\delta < 0) + \Pr(\delta > d_{obs}) = \Pr(|\delta - d_{obs}| > |d_{obs}|) = p.$$

### 9.2.2 Contrast Analysis

We have progressively realized that the conceptual difficulties raised by the interpretation of multidimensional effects are considerable and generally underestimated. This is in accordance with the APA guidelines:

Multiple degree-of-freedom effect-size indicators are often less useful than effect-size indicators that decompose multiple degree-of-freedom tests into meaningful one degree-of-freedom effects – particularly when the latter are the results that inform the discussion (American Psychological Association, 2010, p. 34).

However, in the frequentist paradigm a contrast analysis involves subtle methodological considerations (planned or unplanned comparisons, orthogonal or non-orthogonal contrasts, etc.), which engenders endless debates. On the contrary the Bayesian approach is particularly straightforward. The marginal posterior distribution of a particular contrast always give valid probability statements. So, for the modified reaction time example, instead of the Scheffé intervals, we can compute for instance the two marginal probabilities

$$\begin{aligned} g_{2,g1} \quad \Pr(\delta_{g_{2,g1}} > 15 \text{ ms}) &= .983 \\ g_{3,g1} \quad \Pr(\delta_{g_{3,g1}} > 15 \text{ ms}) &= .978 \end{aligned}$$

If we are interested in a simultaneous inference for these two differences, we compute the joint probability. Of course, this may involve more sophisticated computational procedures, but for practical purpose, it is often sufficient to use the Bonferroni inequality. So, we know that the joint probability that each of the two differences exceeds 15 ms is at least  $1 - (1 - .983) - (1 - .978) = .961$ .

## 9.3 An Illustrative Example: Evaluation of the “0.05 Cliff Effect”

In one of the first experiments on the use of significance tests (Rosenthal & Gaito, 1963, 1964), researchers in psychology were asked to state their degree of belief in the hypothesis of an effect as a function of the associated  $p$ -values and sample sizes. The degree of belief decreased when the  $p$ -value increased, and was on average approximately an exponential function. However the authors emphasized a *cliff effect* for the .05 level, i.e. “an abrupt drop” in confidence just beyond this level. This cliff effect was invoked by Oakes (1986, p. 83) in support of his *significance hypothesis* according to which the outcome of the significance test is interpreted in

terms of a dichotomy: An effect either “exists” when it is significant, or “does not exist” when it is nonsignificant. Similar results were obtained in subsequent studies (Beauchamp & May, 1964; Minturn et al., 1972, quoted by Nelson et al., 1986; Nelson et al., 1986).

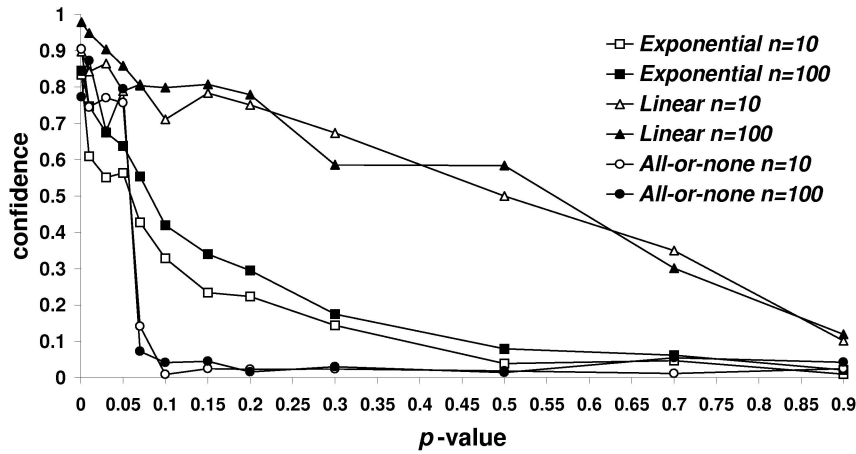
Poitevineau and Lecoutre (2001) replicated this experiment, with the aim of identifying distinct categories of subjects, possibly corresponding to different views of statistical inference, referring in particular to Neyman-Pearson, Fisher and Bayes. 12  $p$ -values (.001, .01, .03, .05, .07, .10, .15, .20, .30, .50, .70, .90) combined with two sample sizes ( $n = 10$  and  $n = 100$  as in the original experiment) were presented at random, on separate pages of a notebook. It was specified that the test was a Student's  $t$  for paired groups. The subjects were asked to state their degree of belief in the hypothesis that “experimental treatment really had an effect”. They were asked to tick off a point on a non-graduated segment line of 10 centimeters, from null confidence (left extremity) to full confidence (right extremity of the scale). The participants' responses were measured in the  $[0, 1]$  interval. 18 psychology researchers carried out this experiment.

### 9.3.1 Numerical Results

Although the experiment was conducted about 35 years after the original one and in another country, the average curves appeared to be similar. As in Rosenthal and Gaito's (1963) study, a .05 cliff effect was apparent for the two sample sizes. However the average curves were fairly well fitted by an exponential function, and furthermore the study of individual curves revealed that participants could actually be classified into three clearly distinct categories, the classification being identical for the two curves ( $n = 10$  and  $n = 100$ ) of each individual (see Figure 9.3).

1. 10 out of 18 participants presented a decreasing exponential curve, as if these subjects considered the  $p$ -values as a physical measure of weight of evidence.
2. 4 participants presented a negative linear curve, which is compatible with the common misinterpretation of a  $p$ -value as the complement of the probability that the alternate hypothesis is true.
3. 4 participants presented an all-or-none curve with a very high degree of belief when  $p \leq 0.05$  and with nearly a null degree of belief otherwise. Only these stepwise curves clearly referred to a decision making attitude.

The larger sample size gave more confidence to the participants in the first category, whereas all the other participants had almost the same degree of belief for a given  $p$ , whatever the sample size.



**Fig. 9.3** Confidence in the hypothesis that experimental treatment really had an effect, as a function of the  $p$ -value and the sample size  $n$ .

### 9.3.2 A Cliff Effect Indicator

The cliff effect was measured in the same line as Nelson, Rosenthal and Rosnow (1986). According to these authors, a test for the .05 cliff effect “controlling for the ordinary decline in confidence as  $p$  increases” can be based on the cubic contrast assigned to the four consecutive  $p$ -values .03, .05, .07, and .10. The reason for this procedure is that it is equivalent to test whether the patterns of the population means associated with these  $p$ -values and the predicted means based on a second-degree polynomial equation are the same, in which case there is a null .05 cliff effect.

Consequently, a natural measure of effect size is given by an index of departure (“goodness of fit”). Given the second-degree polynomial equation that best fits the data for the  $p$ -values .03, .05, .07, and .10, the unstandardized cliff effect can be estimated by the quadratic mean of the residuals between the observed and predicted means.

For instance, for the all-or-none group, we have the respective observed means .724, .776, .107, .025, and the corresponding predicted means .808, .543, .301, −.019. The quadratic mean of the residuals

$$\sqrt{\frac{(-0.083)^2 + (+0.233)^2 + (-0.194)^2 + (+0.044)^2}{4}} = 0.159$$

is equal to the (absolute) numerical value of the usual cubic contrast, up to a constant of proportionality. Taking into account the unequal spacing of the four  $p$ -values and introducing the appropriate constant of proportionality, a cliff effect indicator is the cubic contrast with coefficients +.1310, −.3668, +.3057, and −.0699, that is

$$d_{obs} = +.1310 \times .724 - .3668 \times 0.776 + .3057 \times .107 - .0699 \times .025 = -.159.$$

The sign of this contrast is obviously relevant: since the cliff effect is a drop in confidence, it is natural to represent it by a negative value. Note that the maximum cliff effect is  $+.1310 - .3668 = -.2358$ , corresponding to the pattern of responses 1, 1, 0, 0.

The hypothesis of an *exact* model is of little practical interest. For instance, a single observation different from zero or one is sufficient to falsify the *strict* all-or-none model. The issue is not to accept or reject a given *exact* model, but rather to evaluate the departure from the model. This is a problem of pure estimation, in Jeffreys' terms (see Section 5.1).

### The relevant data

The derived relevant data consist of the *individual* cliff effects, reported in Table 9.3.2. For simplicity, we have ignored here the "sample size" factor and the data are averaged on the two modalities ( $n = 10$  and  $n = 100$ ).

#### Exponential group (10 participants)

	1	2	3	4	5	6	7	8	9	10
.03	.8625	.9450	.4850	.6475	.1900	.4875	.4100	.6500	.7300	.7200
.05	.7900	.9250	.4500	.6400	.1850	.5000	.4850	.5925	.7150	.7275
.07	.7200	.9325	.1775	.4925	.1400	.2775	.4900	.4200	.6675	.5900
.10	.5850	.8725	.0400	.4400	.1500	.2225	.1400	.3300	.5300	.4325
Cliff effect	+.0024	+.0086	-.0501	-.0301	-.0107	-.0503	+.0158	-.0269	+.0004	-.0224

#### Linear group (4 participants)

	1	2	3	4
.03	.9025	.8500	.8500	.9350
.05	.8475	.7950	.7800	.8725
.07	.9125	.6250	.7800	.9050
.10	.7925	.7150	.6825	.8300
Cliff effect	+.0309	-.0392	+.0160	+.0210

#### All or none group (4 participants)

	1	2	3	4
.03	.8275	.2075	.8625	1
.05	.8800	.4050	.8200	1
.07	.2475	.1800	0	0
.10	.0800	.0200	0	0
Cliff effect	-.1443	-.0677	-.1878	-.2358

**Table 9.1** Relevant data: Individual cliff effects.

### Descriptive summary

The adequate descriptive statistics are the means and standard deviations of the relevant data.

	Mean	Standard deviation
Exponential group	-.0163	.0234
Linear group	+.0072	.0315
All or none group	-.1589	.0713
Weighted mean	-.0428	
Within-group SD		.0393

If we examine the individual patterns of responses in Table 9.3.2, it seems reasonable to consider that a cliff effect less than .04 in absolute value is small and that a cliff effect more than .06 in the negative direction is large. Consequently, the observed cliff effect is small for the exponential and linear groups, showing a large departure from the all-or-none model, and very large for the all-or-none group.

### 9.3.3 An Overall Analysis Is not Sufficient

#### The average cliff effect

For the inference about the average cliff effect, we can apply the same procedures as for the Student example. Simply, we use the weighted mean and the within-group standard deviation, hence

$$d_{obs} = -.0428 \quad s_{obs} = .0393 \text{ (15 df)} \quad e = s_{obs}/\sqrt{18} = 0.0093 \quad t_{obs} = d_{obs}/e = -4.615.$$

We obtain the 90% interval estimate [-.059, -.027] for the overall (unstandardized) cliff effect that appears to be rather moderate.

#### An ES indicator for comparing the three groups

Then we could be tempted to compare the cliff effects of the three groups. This in an interaction effect with 2 df, for which we can apply the results of Section 9.1.2. For instance, an indicator of the observed effect size is the weighted quadratic mean of all the partial differences. The weights are the product of the group sizes, which generalizes the balanced case, so that

$$\ell_{obs} = \sqrt{\frac{10 \times 4(-.0163 - .0072)^2 + 4 \times 4(.0072 + .1589)^2 + 4 \times 10(-.1589 + .0163)^2}{10 \times 4 + 4 \times 4 + 4 \times 10}} = \sqrt{.0133} = .1153.$$

This is undoubtedly a large observed effect. On the one hand, this is not surprising since the participants have been classified in view of the data in order to maximize the differences between the groups. On the other hand, an inference about the population ES is not suitable for generalizing the descriptive conclusions about the respective magnitudes of the group cliff effects.

#### Remarks

- If a specific ANOVA is performed on the relevant derived data, we find the mean-squares  $MS_G = .0355$  ( $\ell_{obs}^2 = .3750MS_G = .1153$ ) and  $MS_{error} = .0015$  ( $s_{obs}^2 = MS_G$ ). They are proportional to the mean-squares that would be obtained from the analysis of the complete data, using a mixed-model, so that the  $F$  ratio is the same:  $F_{obs} = 22.94$  ( $p = .00003$ , 2 and 15 df). The two analyses are equivalent, but the former is more easily understandable. The design structure

of the relevant data is much simpler than the original design structure, and the number of *nuisance* parameters is drastically reduced. Consequently, necessary and minimal assumptions specific to each particular inference are made explicit. Here, they are simply the usual assumptions of a one-way ANOVA design. When these assumptions are under suspicion, alternative procedures can be envisaged (see Section 9.3.5).

- For the specific analysis we have  $a^2 = 1$ . The constant  $b^2 = .0133/.0355 = .3750$  is given by

$$b^2 = \frac{2}{\bar{n} - \frac{\check{n}}{n}}.$$

where  $\bar{n}$  is the mean of the group sizes – here  $\bar{n} = 6$  – and  $\check{n}$  is their variance, corrected for df – here  $\check{n} = 12$ .

### 9.3.4 A simultaneous inference about all contrasts

Clearly, an inference about  $\zeta$  (see Section 9.2), the (weighted) quadratic mean of the deviations between the partial population and observed differences is more suitable. The fB distribution is

$$\zeta | \text{data} \sim .0241 \psi_{2,15}(0) \quad [e = \frac{.1153}{\sqrt{22.94}} = .0241].$$

We can conclude with a .90 fB guarantee that the average deviation between the partial population and observed differences is less than .040:  $\Pr(\zeta < .040) = .90$ . So the descriptive conclusions of a very large difference between the all-or-none group and each of the two other groups and a relatively moderate difference between the exponential and linear groups can be generalized. Marginal probability statements about the partial differences could be obtained in the same way as for the reaction time data. Note the interpretation of the  $p$ -value,  $\Pr(\zeta < .1153) = 1 - p = .99997$ : not really informative.

### 9.3.5 An Adequate Analysis

A simultaneous inference about all contrasts is undoubtedly preferable to the inference about an ANOVA ES. However, it is not the best alternative, since it does not directly answer the relevant question: to evaluate the departure from the model. For this purpose, we have to consider the inference about the cliff effect (the cubic contrast) separately for each group.

Since it is likely that the all-or-none model results in larger individual variability, the assumption of equal group variances made in the previous analyses is not realistic. It can be easily relaxed here by considering separate variance estimates for each group. Then the relevant statistics are

$$\begin{array}{llll}
\text{Exponential group} & d_{obs} = -.0163 & s_{obs} = .0234 & e = \frac{.0234}{\sqrt{10}} = .0074 \quad t_{obs} = \frac{-.0163}{.0074} = -2.202 \\
\text{Linear group} & d_{obs} = +.0072 & s_{obs} = .0315 & e = \frac{.0315}{\sqrt{4}} = .0158 \quad t_{obs} = \frac{+.0072}{.0158} = +0.457 \\
\text{All or none group} & d_{obs} = -.1589 & s_{obs} = .0713 & e = \frac{.0713}{\sqrt{4}} = .0357 \quad t_{obs} = \frac{-.1589}{.0357} = -4.455
\end{array}$$

### Relevant inferences

The fB distribution for the respective population contrasts and the corresponding statements for asserting the relative magnitudes of effects are

$$\begin{array}{lll}
\text{Exponential group} & t_9(-0.0163, 0.0074^2) & \Pr(|\delta| < .027) = .90 \\
\text{Linear group} & t_3(+0.0072, 0.0158^2) & \Pr(|\delta| < .039) = .90 \\
\text{All-or-none group} & t_3(-0.1589, 0.0357^2) & \Pr(\delta < -.100) = .90
\end{array}$$

These statements clearly demonstrate the major finding of this experiment: the cliff effect has been overstated, it is only large for a minority of “all-or-none” respondents. On the contrary, it is of limited magnitude for the other participants who expressed *graduated* confidence judgments about *p*-values.

Note that, since they are based on different variance estimates, the three marginal distributions are independent. So their joint probability is simply  $.90^3 = .729$ .

#### 9.3.6 What about standardized effects?

Nelson, Rosenthal and Rosnow (1986) only reported an averaged standardized effect, the product moment coefficient correlation  $r = .34$ . In Poitevineau & Lecoutre (2001), we were asked by the editor to also report this indicator. It is only function of the *t* statistic and the number of df ( $r = t / \sqrt{t^2 + \text{df}}$ ) and its square is the observed partial eta-squared. We have here for the whole set of participants  $r = -.77$ , far larger than that obtained in the previous study. From the above specific analyses for each group, we get  $r = .59$  (exponential),  $r = +.26$  (linear), and  $r = -.93$  (all-or-none). According to Cohen’s conventions, the observed cliff effect should be considered as large (more than .50), not only for the all-or-none group, but also for the whole set of participants and for the exponential group. It could not be considered to be small (less than .10) for the linear group. Note that the respective observed Cohen’s *d* are -2.93 (whole set), -.696 (exponential), +.228 (linear), and -2.23 (all-or-none). The conclusion would be slightly different: a medium effect (less than .80) for the exponential group.

This again reinforces our contention against the use of standardized ES and heuristics benchmarks. The usual claims that standardization is useful (and even needed) for comparing effect sizes across different conditions or different studies are highly questionable.

## 9.4 Our Guidelines for ANOVA

The guidelines of the previous chapter can be completed, following Baguley's (2009) guidelines.

- *Consider experimental data analysis as a problem of pure estimation in Jeffreys' sense (null hypothesis is not needed).*
- Prefer simple [signed] effect size to standardized effect size.
- Avoid reporting effect sizes for multiple effects [*except for asserting smallness, otherwise prefer contrast analysis*].
- *Don't use noncentral  $F$  based tests and confidence intervals.*
- ***Don't Worry, Be Bayesian:*** *Think about  $p$ -values and usual confidence intervals in Bayesian terms and use their fiducial Bayesian interpretation.*
- Always include adequate descriptive statistics (e.g. sufficient statistics).
- Comment on the relative rather than the absolute magnitude of effects.
- Avoid using 'canned' effect sizes [*heuristic benchmarks*] to interpret an effect.

(adapted from Baguley, 2009, p. 615, italicized terms are ours)



## Chapter 10

### Conclusion

Despite all criticisms, Null Hypothesis Significance Testing [NHST] continues to be required in most experimental publications as an unavoidable norm. This is an amalgam of the Fisher and Neyman-Pearson views of statistical tests. It is used to strengthen data and convince the community of the value of the results. NHST can be seen with Salsburg (1985) as the “religion of statistics” with rites such as the use of the “profoundly mysterious symbols of the religion NS, \*, \*\*, and *mirabile dictu* \*\*\*” (the *star system*). The degree of statistical significance – the *p*-value – is for most users a substitute for judgment about the meaningfulness of experimental results: they behave like “star worshippers” (Guttman, 1983) and “sizeless scientists” (Ziliak & McCloskey, 2008).

NHST is such an integral part of scientists’ behavior that its misuses and abuses should not be discontinued by flinging it out of the window. Actually, the official guidelines for experimental data analysis do not ban its use. Rather, they appear to reinforce its legitimacy by placing it at the center of a hybrid reporting strategy. This strategy includes other practices, especially effect sizes and their confidence intervals. We name it *Guidelined Hypotheses Official Significance Testing* [GHOST], because it focuses on the Neyman-Pearson (power based) justification of sample size, involving two precise (point null) hypotheses.

GHOST is only a set of recipes and rituals and does not supply a real statistical thinking. As a consequence, it has created a *new star system*, based on the use of pre-established heuristic benchmarks for standardized effect sizes. Experimental literature reveals that most scientists continue to behave like star worshipers and sizeless scientists.

Due to his great influence on experimental research, Fisher’s responsibility in the today’s practices cannot be discarded. One of the most virulent attacks came from Ziliak and McCloskey (see also Meehl, 1978, p. 817):

After Fisher, then, the sizeless sciences neither test nor estimate (Ziliak & McCloskey, 2008, p. 17).

However, Fisher’s conception of probability and his works on the fiducial theory are a fundamental counterpart to his emphasis on significance tests, and he should

not be treated as guilty (Lecoutre, Poitevineau & Lecoutre, 2004). This was clearly acknowledged by Jeffreys:

But it seems to me that the cases that chiefly concern Fisher are problems of estimation, and for these the fiducial and inverse probability approaches are completely equivalent (Jeffreys, 1940, p. 51).

The gentlemen's agreement between him and Fisher was made explicit:

The general agreement between Professor R.A. Fisher and myself has been indicated in many places. The apparent differences have been much exaggerated. . . (Jeffreys, 1967, p. 393).

Following Jeffreys, experimental data analysis must be regarded as a problem of “pure estimation”, and significance tests of precise hypotheses should have a very limited role. Within this perspective, there is no sense to search for an interpretation of the  $p$ -value as the probability of the null hypothesis. Rather, for the usual test of no difference between means (for instance), the halved  $p$ -value of the usual two-sided  $t$ -test is the posterior probability that the population difference has the opposite sign of the observed difference. This was also Student's conception.

Nowadays Bayesian routine methods for the familiar situations of experimental data analysis are easy to implement and use. Based on more useful working definitions than frequentist procedures, the Bayesian approach makes all choices explicit and offers considerable flexibility. This gives statistics users a real possibility of thinking sensibly about statistical inference problems, so that they behave in a more reasonable manner.

In particular, Fiducial Bayesian methods emphasize the need to think hard about the information provided by the data in hand (“what the data have to say?”), instead of applying ritual, ready-made procedures. This does not preclude using other Bayesian techniques *when appropriate*. In some situations, it may be essential to use objective prior information external to the data. An opinion-based analysis can serve for individual decision making, such as to publish a result or to replicate an experiment.

In all cases, fiducial Bayesian methods have a privileged status in order to gain “public use” statements, acceptable by the scientific community. Our consulting and teaching practices, especially in psychology, have shown us that they are much closer to scientists' spontaneous interpretations of data than frequentist procedures. Using the fiducial Bayesian interpretation of the  $p$ -value in the natural language of probability about unknown effects comes quite naturally, and the common misuses and abuses of NHST can be clearly understood. The need for estimation becomes evident, and users' attention can be focused to more appropriate strategies, such as consideration of the practical significance of results and replication of experiments. Fiducial Bayesian users are also well equipped for a critical reading of experimental publications.

## References

1. Agresti, A., Coull, B.: Approximate is better than exact for interval estimation of binomial proportions. *Amer. Statist.* **52**, 119–126 (1998)
2. Agresti, A., Min, Y.: Frequentist performance of Bayesian confidence intervals for comparing proportions in  $2 \times 2$  contingency tables. *Biometrics* **61**, 515–523 (2005)
3. Aldrich, J.: The statistical education of Harold Jeffreys. *Int. Stat. Rev.* **73**, 289–307 (2005)
4. American Psychological Association: Publication Manual of the American Psychological Association (6th edition). American Psychological Association, Washington, DC (2010)
5. Baguley, T.: Standardized or simple effect size: What should be reported? *Brit. J. Psychol.* **100**, 603–617 (2009)
6. Baguley, T.: When correlations go bad. *The Psychologist* **23**, 122–123 (2010)
7. Bakan, D.: The test of significance in psychological research. *Psychol. Bull.* **66**, 423–437 (1966)
8. Barnett, V.: *Comparative Statistical Inference* (3rd edition). John Wiley & Sons, New York. (1999)
9. Bayes, T.: An essay towards solving a problem in the doctrine of chances. *Phil. Trans.* **53**, 370–418 (1763)
10. Beauchamp, K.L., May, R.B.: Replication report: Interpretation of levels of significance by psychological researchers. *Psychol. Rep.*, **14**, 272 (1964)
11. Berger, J.O.: Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* **18**, 1–32 (2003)
12. Berger J.: The case for objective Bayesian analysis. *Bayesian Analysis* **11**, 1–17 (2004)
13. Berger, R.L., Hsu, J.C.: Bioequivalence trials, intersection-union tests and equivalence confidence sets (with comments). *Stat. Sci.* **11**, 283–319 (1996)
14. Berry, D.A.: Teaching elementary Bayesian statistics with real applications in science. *Amer. Statist.* **51**, 241–246 (1997)
15. Bird, K.: *Analysis of Variance via Confidence Intervals*. Sage, London (2004)
16. Boring, E.G.: Mathematical versus scientific significance. *Psychol. Bull.* **16**, 335–338 (1919)
17. Box, G.E.P., Tiao, G.C.: *Bayesian Inference in Statistical Analysis*. Addison Wesley, Reading, MA (1973)
18. Broemeling, L., Broemeling, A.: Studies in the history of probability and statistics XLVIII: The Bayesian contributions of Ernest Lhoste. *Biometrika* **90**, 728–731 (2003)
19. Brown, L.D., Cai, T., DasGupta, A.: Interval estimation for a binomial proportion (with discussion). *Stat. Sci.* **16**, 101–133 (2001)
20. Bunouf, P., Lecoutre, B.: Bayesian priors in sequential binomial design. *Cr. Acad. Sci. I-Math.* **343**, 339–344 (2006)
21. Bunouf, P., Lecoutre, B.: An objective Bayesian approach to multistage hypothesis testing. *Sequential Anal.* **29**, 88–101 (2010)
22. Cai, T.: One-sided confidence intervals in discrete distributions. *J. Stat. Plan. Infer.* **131**, 63–88 (2005)
23. Carver, R.P.: The case against statistical significance testing. *Harvard Educ. Rev.* **48**, 378–399 (1978)
24. Casella, G., Berger, L.: Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Am. Stat. Assoc.* **82**, 106–111 (1987)
25. Ciancia F., Maitte M., Honoré J., Lecoutre B., Coquery J.-M.: Orientation of attention and sensory gating: An evoked potential and RT study in cat. *Exp. Neurol.* **100**, 274–287 (1988)
26. Cochran, W.G., Cox, G.M.: *Experimental Designs* (2nd edition). John Wiley & Sons, New York (1957)
27. Cohen, J.: The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Soc. Psych.* **65**, 145–153 (1962)
28. Cohen, J.: *Statistical Power Analysis for the Behavioral Sciences* (revised edition). Academic Press, New York (1977)
29. Cohen, J.: The earth is round ( $p < .05$ ). *Am. Psychol.* **49**, 997–1003 (1994)

30. Cox, D.R.: Another comment on the role of statistical methods. *Brit. Med. J.* **322**, 231 (2001)
31. Dale, A.: *A History of Inverse Probability*. Springer-Verlag, New York (1991)
32. de Cristofaro, R.: On the foundations of likelihood principle. *J. Stat. Plan. Infer.* **126** 401–411 (2004)
33. de Cristofaro, R.: Foundations of the ‘Objective Bayesian Inference’. In: *First Symposium on Philosophy, History and Methodology of ERROR*. Virginia Tech., Blacksburg VA (2006). Available <http://www.error06.econ.vt.edu/Christofaro.pdf>. Cited 13 March 2014
34. de Finetti, B.: La prévision: Ses lois logiques, ses sources subjectives. *Annales de l’Institut Henri Poincaré* **7**, 1–68 (1937). (English translation: Foresight: Its logical laws, its subjective sources. In: Kyburg, H.E. and Smokler H.E. (eds.) *Studies in Subjective Probability*, pp. 53–118. John Wiley & New York, 1964)
35. de Finetti, B.: *Probability, Induction and Statistics*. Wiley, London (1972)
36. Deheuvels, P.: How to analyze bio-equivalence studies? The right use of confidence intervals. *J. Organ. Behav. Stat.* **1**, 1–15 (1984)
37. Edwards, W., Lindman, H., Savage, L.J.: Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242 (1963)
38. Efron, B.: R.A. Fisher in the 21st century (with discussion). *Stat. Sci.* **13**, 95–122 (1998)
39. Fidler, F., Thompson, B.: Computing correct confidence intervals for ANOVA fixed and random-effects effect sizes. *Educ. Psychol. Meas.* **61**, 575–604 (2001)
40. Field, A., Miles, J.: *Discovering Statistics Using SAS*. Sage, London (2010)
41. Fisher, R.A.: The arrangement of field experiments. *J. Ministry Agriculture Great Britain* **33**, 503–513 (1926)
42. Fisher, R.A.: The logic of inductive inference (with discussion). *J. R. Stat. Soc. A* **98**, 39–82 (1935)
43. Fisher, R.A.: Statistical methods and scientific induction. *J. R. Stat. Soc. B* **17**, 69–78 (1955)
44. Fisher, R.A.: Mathematical probability in the natural sciences. *Technometrics* **1**, 21–29 (1959)
45. Fisher, R.A.: *Statistical Methods for Research Workers* (reprinted 14th edition, 1970). In: Bennett, J.H. (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford (1990a)
46. Fisher, R.A.: *The Design of Experiments* (reprinted 8th edition, 1966). In: Bennett, J.H. (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford (1990b)
47. Fisher, R.A.: *Statistical Methods and Scientific Inference* (reprinted 3rd edition, 1973). In: Bennett, J.H. (ed.) *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford (1990c)
48. Fleiss, J.L.: Estimating the magnitude of experimental effects. *Psychol. Bull.* **72**, 273–276 (1969)
49. Food and Drug Administration: *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (2001). Available <http://www.fda.gov/downloads/Drugs/.../Guidances/ucm070244.pdf>. Cited 13 March 2014
50. Food and Drug Administration: *Guidance for Industry and FDA Staff: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health (2010). Available <http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM071121.pdf>. Cited 13 March 2014
51. Fraser, D.A.S.: Discussion of Hill’s paper, on some statistical paradoxes and non-conglomerability. *Trab. Estad. Investig. Oper.* **31**, 56–58 ((1980)
52. Freeman, P.R.: The role of *P*-values in analysing trial results. *Stat. Med.* **12**, 1443–1452 (1993)
53. Geisser, S.: Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159 (1965).
54. Geisser, S.: *Predictive Inference: An Introduction*. Chapman & Hall, New York (1993)

55. Gelman, A., Stern, H.: The difference between “significant” and “not significant” is not itself statistically significant. *Amer. Statist.* **60**, 328–331 (2006)
56. Gertsbakh, I., Winterbottom, A., 1991: Point and interval estimation of normal tail probabilities. *Commun. Stat. A-Theor* **20**, 1497–1514 (1991)
57. Ghosh, M.: Objective priors: An introduction for frequentists (with discussion). *Stat. Sci.* **26**, 187–202.
58. Gibbons, R.D., Hedeker, D.R., Davis, J.M.: Estimation of effect size from a series of experiments involving paired comparisons. *J. Educ. Stat.* **18**, 271–279 (1993)
59. Gigerenzer, G.: The superego, the ego, and the id in statistical reasoning. In: Keren, G., Lewis, C. (eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, pp. 311–339. Erlbaum, Hillsdale, NJ (1993)
60. Glass, G.V.: Primary, secondary, and meta-analysis of research. *Educ. Researcher* **5**, 3–8 (1976)
61. Goodman, S.N., Berlin, J.A.: The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* **121**, 200–206 (1994)
62. Gravetter, F.J., Wallnau, L.B.: *Statistics for the Behavioral Sciences* (8th edition). Wadsworth, Belmont (2009)
63. Grouin, J.-M., Coste, M., Bunouf, P., Lecoutre, B.: Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Stat. Med.* **26**, 26, 4914–4924 (2007)
64. Guttman, L.: What is not what in statistics? *The Statistician* **26**, 81–107 (1983)
65. Hannig J.: On generalized fiducial inference. *Statist. Sinica* **19**, 491–544 (2009)
66. Harcum, E.R.: Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *Am. Psychol.* **45**, 404–405 (1990)
67. Harlow, L.L., Mulaik, S.A., Steiger, J.H. (eds.): *What if there were no significance tests?* Lawrence Erlbaum, Mahwah, NJ (1997)
68. Hays W.L.: *Statistics for Psychologists*. Holt, Rinehart & Winston, New York (1963)
69. Hedges, L.V.: Distribution theory for Glass’s estimator of effect size and related estimators. *J. Educ. Stat.* **7**, 107–128 (1981)
70. Holender, D., Bertelson, P.: Selective preparation and time uncertainty. *Acta Psychol.* **39**, 193–203 (1975)
71. Howell D.C.: *Fundamental Statistics for the Behavioral Sciences* (7th edition). Wadsworth, Belmont, CA (2010)
72. Hubbard, R.: Alphabet soup: Blurring the distinctions between  $p$ ’s and  $\alpha$ ’s in psychological research. *Theor. Psychol.* **14**, 295–327 (2004)
73. Hunter, J.E., Schmidt, F.L.: *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (2nd edn.). Sage, Thousand Oaks, CA (2004)
74. ICH E9 Expert Working Group: Statistical principles for clinical trials: ICH harmonised tripartite guideline, current step 4 version (1998). Available [http://www.ich.org/fileadmin/Public\\_Web\\_Site/ICH\\_Products/Guidelines/Efficacy/E9/Step4/E9\\_Guideline.pdf](http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf). Cited 13 March 2014
75. International Journal of Psychology: Instructions for authors (2014) Available via Taylor & Francis <http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291464-066X/homepage/ForAuthors.html>. Cited 13 March 2014
76. Iverson, G.J., Lee, M.D., Wagenmakers, E.-J.:  $p_{rep}$  misestimates the probability of replication. *Psychon. B. Rev.* **16**, 424–429 (2009)
77. Jaccard, J., Guilamo-Ramos, V.: Analysis of variance frameworks in clinical child and adolescent psychology: Advances issues and recommendations. *J. Clin. Child Adolesc.* **31**, 278–294 (2002)
78. Jaynes, E.T.: Confidence intervals vs Bayesian intervals (with discussion). In: W. L. Harper and C.A. Hooker (eds.) *Statistical Inference and Statistical Theories of Science Volume 2*, pp. 175–257. D. Reidel, Dordrecht, The Netherlands (1976)
79. Jaynes, E.T.: *Probability Theory: The Logic of Science* (Edited by G.L. Bretthorst). Cambridge University Press, Cambridge, England (2003)

80. Jeffreys, H.: Probability, statistics, and the theory of errors. *Proc. R. Soc. Lond. A* **140**, 523–535 (1933)
81. Jeffreys, H.: Note on the Behrens-Fisher formula. *Ann. Eugen.* **10**, 48–51 (1940)
82. Jeffreys, H.: *Theory of Probability* (3rd edition, 1st edition 1939). Clarendon, Oxford (1967)
83. Jeffreys, H.: *Scientific Inference* (3rd edition, 1st edition 1931). Cambridge University Press, Cambridge, England (1973)
84. Jones, L.V.: Statistics and research design. *Annu. Rev. Psychol.* **6**, 405–430 (1995)
85. Jones, L.V., Tukey, J.W.: A sensible formulation of the significance test. *Psychol. Methods* **5**, 411–414 (2000)
86. Kelley, K.: Confidence intervals for standardized effect sizes: Theory, application, and implementation. *J. Stat. Softw.* **20**, 2–24 (2007)
87. Killeen, P.R.: An alternative to null-hypothesis significance tests. *Psychol. Sci.* **16**, 345–353 (2005)
88. Kirk, R.E.: Effect magnitude: A different focus. *J. Stat. Plan. Infer.* **137**, 1634–1646 (2007)
89. Kruschke, J.K.: Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* **6**, 299–312 (2011)
90. Laplace, P.S.: *Essai Philosophique sur les Probabilités* (6th edition) (English translation: *A Philosophical Essay on Probability*, Dover, New York, 1952). Bachelier, Paris (1840)
91. Le Cam, L.: Comparison of experiments – A short review. In: Ferguson, T., Shapley, L. (eds.) *Statistics, Probability and Game Theory* (IMS Lecture Notes Monogr. Ser. 30), pp. 127–138. Institute of Mathematical Statistics, Hayward, CA (1996)
92. Lecoutre, B.: Extensions de l'analyse de la variance: L'analyse bayésienne des comparaisons. *Math. Sci. Hum.* **75**, 49–69 (1981)
93. Lecoutre, B.: *L'Analyse Bayésienne des Comparaisons*. Presses Universitaires de Lille, Lille, FR (1984)
94. Lecoutre, B.: Reconsideration of the F test of the analysis of variance: The semi-Bayesian significance tests. *Commun. Stat. A-Theor* **14**, 2437–2446 (1985)
95. Lecoutre, B.: *Traitement statistique des données expérimentales: des pratiques traditionnelles aux pratiques bayésiennes*. SPAD, Suresnes, FR (1996). Bayesian Windows programs by B. Lecoutre and J. Poitevineau, freely available <http://www.univ-roen.fr/LMRS/Persopage/Lecoutre/Eris>. Cited 13 March 2014
96. Lecoutre, B.: Two useful distributions for Bayesian predictive procedures under normal models. *J. Stat. Plan. Infer.* **79**, 93–105 (1999)
97. Lecoutre, B.: Bayesian predictive procedure for designing and monitoring experiments. In: *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pp. 301–310. Office for Official Publications of the European Communities, Luxembourg (2001)
98. Lecoutre, B.: Training students and researchers in Bayesian methods. *J. Data Sci.* **4**, 207–232 (2006)
99. Lecoutre, B.: Another look at confidence intervals for the noncentral t distribution. *J. Mod. Appl. Stat. Methods* **6**, 107–116 (2007)
100. Lecoutre, B.: Bayesian methods for experimental data analysis. In: Rao, C.R., Miller, J., Rao, D.C. (Eds.) *Handbook of statistics: Epidemiology and Medical Statistics* (Vol 27), pp. 775–812. Elsevier, Amsterdam (2008)
101. Lecoutre, B., Charron, C.: Bayesian procedures for prediction analysis of implication hypotheses in  $2 \times 2$  contingency tables. *J. Educ. Behav. Stat.* **25**, 185–201 (2000)
102. Lecoutre, B., Derzko, G.: Asserting the smallness of effects in ANOVA. *Methods Psychol. Res.* **6**, 1–32 (2001)
103. Lecoutre, B., Derzko, G.: Intervalles de confiance et de crédibilité pour le rapport de taux d'événements rares. 4èmes Journées de Statistique, SFdS, Bordeaux (2009). Available <http://hal.inria.fr/docs/00/38/65/95/PDF/p40.pdf>. Cited 13 March 2014
104. Lecoutre, B., Derzko, G.: Tester les nouveaux médicaments: Les statisticiens et la réglementation. *Stat. Société* **2**, 61–67 (2014)
105. Lecoutre, B., Derzko, G., ElQasr, K.: Frequentist performance of Bayesian inference with response-adaptive designs. *Stat. Med.* **29**, 3219–3231 (2010)

106. Lecoutre, B., Derzko, G., Grouin, J.-M.: Bayesian predictive approach for inference about proportions. *Stat. Med.* **14**, 1057–1063 (1995)
107. Lecoutre, B., ElQasyr, K.: Adaptive designs for multi-arm clinical trials: The play-the-winner rule revisited. *Commun. Stat. B-Simul.* **37**, 590–601 (2008)
108. Lecoutre, B., Guigues, J.-L., Poitevineau, J.: Distribution of quadratic forms of multivariate Student variables. *Appl. Stat.-J. Roy. St. C* **41**, 617–627 (1992)
109. Lecoutre, B., Killeen, P.: Replication is not coincidence: Reply to Iverson, Lee, and Wagenmakers (2009). *Psychon. B. Rev.* **17**, 263–269 (2010)
110. Lecoutre, B., Lecoutre, M.-P., Poitevineau, J.: Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *Int. Stat. Rev.* **69**, 399–418 (2001)
111. Lecoutre, B., Lecoutre, M.-P., Poitevineau, J.: Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. *Psychol. Methods* **15**, 158–171 (2010)
112. Lecoutre, B., Mabika, B., Derzko, G.: Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups: A Bayesian approach with Weibull modeling. *Stat. Med.* **21**, 663–674 (2002)
113. Lecoutre, B., Poitevineau, J.: PAC (Programme d'Analyse des Comparaisons): Guide d'Utilisation et Manuel de Référence. CISIA-CERESTA, Montreuil, FR (1992). Freely available <http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris>. Cited 12 Sep 2013
114. Lecoutre, B., Poitevineau, J.: Aller au delà des tests de signification traditionnels: Vers de nouvelles normes de publication. *Ann. Psychol.* **100**, 683–713 (2000)
115. Lecoutre, B., Poitevineau, J., Lecoutre, M.-P.: Fisher: Responsible, not guilty. Discussion of D. Denis, The modern hypothesis testing hybrid: R. A. Fisher's fading Influence. *J. SFdS*, **145**, 55–62 (2004)
116. Lecoutre, M.-P., Poitevineau, J., Lecoutre, B.: Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *Int. J. Psychol.* **38**, 37–45 (2003)
117. Lecoutre, B., Rouanet, H.: Deux structures statistiques fondamentales en analyse de la variance univariée et multivariée. *Math. Sci. Hum.* **75**, 71–82 (1981)
118. Lehmann, E.L.: The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Am. Stat. Assoc.* **88**, 1242–1249 (1993)
119. Lehmann, E.L.: Fisher, Neyman, and the Creation of Classical Statistics. Springer, New York (2011)
120. Lenth, R.V.: Some practical guidelines for effective sample size determination. *Amer. Statist.* **55**, 187–193 (2001)
121. Le Roux, B., Rouanet, H.: Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis. Kluwer Academic Publisher, New York (2004)
122. Lhoste, E.: Le calcul des probabilités appliqué à l'artillerie. *Revue d'Artillerie*, **91**, 405–423 (mai), 516–532 (juin), 58–82 (juillet), 153–179 (août) (1923)
123. Matloff, N.S.: Statistical hypothesis testing: problems and alternatives. *Environ. Entomol.* **20**, 1246–1250 (1991)
124. McMillan, J.H., Foley, J.: Reporting and discussing effect size: Still the road less traveled? *Pract. Ass., Res. & Eval.* **16**(14) (2011) Available <http://pareonline.net/pdf/v16n14.pdf>. Cited 13 March 2014
125. Meehl, P.E.: Theory testing in psychology and physics: A methodological paradox. *Philos. Sci.* **34**, 103–115 (1967)
126. Meehl, P.E.: Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Counseling and Clinical Psychology* **46**, 806–834 (1978)
127. Meehl, P.E.: Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* **1**, 108–141 (1990)
128. Meng, X.-L.: Posterior predictive p-values. *Ann. Stat.* **22**, 1142–1160 (1994)
129. Minturn, E.B., Lansky, L.M., Dember, W.N.: The interpretation of levels of significance by psychologists: A replication and extension. Paper presented at the meeting of the Eastern Psychological Association, Boston (1972)

130. Morrison, D.E., Henkel, R.E. (eds.): The Significance Test Controversy – A Reader. Butterworths, London (1970)
131. Nelson, N., Rosenthal, R., Rosnow, R.L.: Interpretation of significance levels and effect sizes by psychological researchers. *Am. Psychol.* **41**, 1299–1301 (1986)
132. Neyman, J.: Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Tr. R. Soc. A* **236**, 333–380 (1937)
133. Neyman, J.: L'estimation statistique traitée comme un problème classique de probabilité. *Actualities Sci. Indust.* **739**, 25–57 (1938)
134. Neyman, J.: First Course in Probability and Statistics Henry Holt and Co., New York (1950)
135. Neyman, J.: Foundations of the general theory of estimation. *Actualités Sci. Indust.* **1146**, 83–95 (1951)
136. Neyman, J.: Lectures and Conferences on Mathematical Statistics and Probability (2nd edition) Graduate School U.S. Department of Agriculture, Washington (1952)
137. Neyman, J.: “Inductive behavior” as a basic concept of philosophy of science. *Rev. Inst. Stat.*, **25**, 7–22 (1957)
138. Neyman, J.: Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131 (1977)
139. Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. *Biometrika* **20A**, 175–240 (1928)
140. Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. *Philos. Tr. R. Soc. A* **231**, 289–337 (1933a)
141. Neyman, J., Pearson, E.S.: The testing of statistical hypotheses in relation to probabilities *a priori*. *P. Camb. Philos. Soc.* **29**, 492–510 (1933b)
142. Neyman, J., Pearson, E.S.: Contributions to the theory of testing statistical hypotheses. *Stat. Res. Mem.* **1**, 1–37 (1936a)
143. Neyman, J., Pearson, E.S.: Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Stat. Res. Mem.* **1**, 113–137 (1936b)
144. Oakes, M.: Statistical Inference: A Commentary for the Social and Behavioural Sciences. Wiley, New York (1986)
145. Olejnik, S., Algina, J.: Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychol. Methods* **8**, 434–447 (2003)
146. Pagano, R.R.: Understanding Statistics in the Behavioral Sciences (8th edn.). Wadsworth Publishing Co Inc, (1997)
147. Pearson, E.S.: “Student” as statistician. *Biometrika*, **30**, 210–250 (1939)
148. Pearson, K.: On the criterion that a given system of deviations form the probable in the case of correlated systems of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philos. Mag.* 5th ser. **50**, 157–175 (1900)
149. Perlman, M.D., Wu, L.: The emperor’s new tests. *Stat. Sci.* **14**, 355–369 (1999)
150. Poitevineau, J., Lecoutre, B.: Interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated. *Psychon. B. Rev.* **8**, 847–850 (2001)
151. Poitevineau, J., Lecoutre, B.: Implementing Bayesian predictive procedures: The K-prime and K-square distributions. *Comput. Stat. Data An.* **54**, 723–730 (2010)
152. Robert, C.P.: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer, New York (2007)
153. Robert, C.P., Chopin, N., Rousseau, J.: Harold Jeffreys’s theory of probability revisited (with comments). *Stat. Sci.* **24**, 141–194 (2009)
154. Robey, R.R.: Reporting point and interval estimates of effect-size for planned contrasts: fixed within effect analyses of variance (tutorial). *J. Fluency Disord.* **29**, 307–341 (2004)
155. Rogers, J.L., Howard, K.I., Vessey, J.: Using significance tests to evaluate equivalence between two experimental groups. *Psychol. Bull.* **113**, 553–565 (1993)
156. Rosenkrantz, R.D.: The significance test controversy. *Synthese* **26**, 304–321 (1973)
157. Rosenthal, R., Gaito, J.: The interpretation of levels of significance by psychological researchers. *J. Psychol.* **55**, 33–38 (1963)
158. Rosenthal, R., Gaito, J.: Further evidence for the cliff effect in the interpretation of levels of significance. *Psychol. Rep.*, **15**, 570 (1964)



159. Rosenthal, R., Rubin, D.B.: The counternull value of an effect size: A new statistic. *Psychol. Sci.* **5**, 329–334 (1994)
160. Rosnow, R.L., Rosenthal, R.: Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychol. Methods* **1**, 331–340 (1996)
161. Rosnow, R.L., Rosenthal, R.: Effect sizes: Why, when, and how to use them *J. Psychol.* **217**, 6–14 (2009)
162. Rothman, K.J., Greenland S.: *Modern Epidemiology*, 2nd edn. Lippincott-Raven, Philadelphia (1998)
163. Rouanet, H.: Bayesian procedures for assessing importance of effects. *Psychol. Bull.* **119**, 149–158 (1996)
164. Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B.: *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2nd edition). Peter Lang, Bern, SW (2000)
165. Rouanet, H. & Lecoutre, B.: Specific inference in ANOVA: From significance tests to Bayesian procedures. *Brit. J. Math. Stat. Psy.* **36**, 252–268 (1983)
166. Rozeboom, W.W.: The fallacy of the null hypothesis significance test. *Psychol. Bull.* **57**, 416–428 (1960)
167. Salsburg, D.S. The religion of statistics as practiced in medical journals. *Amer. Statist.* **39**, 220–223 (1985)
168. SAS Institute Inc.: *SAS/SAT 9.22 User's Guide*. SAS Institute Inc., Cary, NC (2010)
169. Savage, L.J.: *The Foundations of Statistical Inference*. John & Sons, New York (1954)
170. Savage, L.J.: On Rereading R.A. Fisher. *Ann. Stat.* **4**, 441–500 (1976)
171. Scheffé, H.: A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104 (1953)
172. Schervish, M.J.: Bayesian analysis of linear models. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.) *Bayesian Statistics IV*, pp. 419–434. Oxford University Press, Oxford (1992)
173. Schervish, M.J.: *Theory of statistics*. Springer Verlag, New York (1995)
174. Schuirman, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biop.* **15**, 657–680 (1987)
175. Selwyn, W.J., Hall N.R. Dempster, A.P.: Letter to the Editor. *Biometrics* **41**, 561 (1985)
176. Smithson, M.: Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educ. Psychol. Meas.* **61**, 605–632 (2001)
177. Smithson, M.: *Confidence intervals*. Sage, Thousand Oaks, CA (2003)
178. Smithson, M.: *Statistics with Confidence* (reprint). Sage, London (2005)
179. Spielman, S.: The logic of tests of significance. *Philos. Sci.* **41**, 211–226 (1974)
180. Steiger, J.H.: Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychol. Methods* **9**, 164–182 (2004)
181. Steiger, J.H., Fouladi, R.T.: Noncentrality interval estimation and the evaluation of statistical models. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H. (eds.) *What If There Were No Significance Tests?*, pp. 221–257. Erlbaum, Hillsdale, NJ (1997)
182. Sterling, T.D.: Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J. Am. Stat. Assoc.* **54**, 30–34 (1959)
183. Sterne, J.A.C., Smith, G.D.: Sifting the evidence-what's wrong with significance tests? *Brit. Med. J.* **322**, 226–231 (2001)
184. Student: The probable error of a mean. *Biometrika* **6**, 1–25 (1908)
185. Student: Tables for estimating the probability that the mean of a unique sample of observations lies between  $-\infty$  and any given distance of the mean of the population from which the sample is drawn. *Biometrika* **11**, 414–417 (1917)
186. Thompson, B.: Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *J. Exp. Educ.* **70**, 80–93 (2001)

187. Thompson, B.: What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educ. Researcher* **31**, 24–31 (2002)
188. Tukey, J.W.: Conclusions vs decisions. *Technometrics* **2**, 1–11 (1960)
189. Tukey, J.W.: Analyzing data: Sanctification or detective work? *Am. Psychol.* **24**, 83–91 (1969)
190. Tukey, J.W.: The philosophy of multiple comparisons. *Stat. Sci.* **6**, 100–116 (1991)
191. Venables, W. (1975): Calculation of confidence intervals for non-centrality parameters. *J. R. Stat. Soc. B* **37**, 406–412 (1975)
192. Westlake, W.J.: Response to bioequivalence testing: A need to rethink (reader reaction response). *Biometrics* **37**, 591–593 (1981).
193. Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs: Statistical Methods in Psychology Journals: Guidelines and Explanations. *Am. Psychol.* **54**, 594–604 (1999)
194. Yates, F.: Sir Ronald Fisher and the design of experiments. *Biometrics* **20**, 307–321 (1964)
195. Zabell S.L.: R. A. Fisher and the fiducial argument *Stat. Sci.* **7**, 369–387 (1992)
196. Zabell S.L.: On Student's 1908 paper "The probable error of a mean." *J. Am. Stat. Assoc.* **103**, 1–7 (2008)
197. Ziliak S.T., McCloskey D.: *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* University of Michigan Press, Ann Arbor (2008)

# Index

- Alternative, *see* Hypothesis
- Amalgam, *see* Statistical tests
- ANOVA, *see* Confidence intervals, Effect size
- APA, *see* Guidelines
  - APA Task Force, 43, 61
- Bayes, *see* Credible intervals
  - Bayes factor, 27
  - Bayes' formula, 15, 57
  - Degree of confidence, 9, 25, 30
  - Fiducial Bayesian, *see* Fiducial Bayesian
  - Objective Bayesian analysis, 4
- Behavior
  - Inductive behavior, 33, 74
  - Rules of behavior, 23, 24
- Clinical trials, 41–43
  - Equivalence trials, 34, 44, 56, 80–81
  - Non inferiority trials, 44, 56
  - Superiority trials, 43, 47
- Cohen's *d*, 59–65, 79, 87
  - And *t*-test statistic, 61, 66
  - Confidence interval, 87
  - Denominator, 64
  - Heuristic Benchmarks, 62
  - Population and sample, 67
  - Standardizer, 65
  - Unsigned, 61
- Confidence interval [CI], 6, 12, 43, 44, 74
  - Bayesian misinterpretation, 10, 81–82
  - CI for demonstrating equivalence, 81
  - Clopper-Pearson CI for a proportion, 43
  - Conservative/anti-conservative CI, 76, 81
  - Exact CI for discrete data, 76
  - Frequentist definition, 74–75
  - Frequentist interpretation, 12, 81
  - NCF-CI for ANOVA effect sizes, 68, 78–81
  - Usual CI, 78, 90
- Conservative, *see* Confidence interval, statistical tests
- Correlation coefficient, 84, 89–90
- Counternull value, 51–52, 82
- Coverage
  - Coverage probability, 4, 6, 75–77
  - Coverage properties, 76
- Credible interval, *see* Interval estimate
  - Definition, 72
  - Highest posterior density, 76
- Decision, 23–24, 32
  - Decision making, 29, 33, 53
  - Reject/accept rule, 23–24, 43–46
- Distribution, *see* Posterior, Predictive, prior
  - Beta-Binomial, 15
  - Binomial, 42
  - Generalized *t*, 49, 85, 92
  - Hypergeometric, 11
  - Lambda-prime, 87–88
  - Noncentral *F*, 34, 78, 95
  - Noncentral *t*, 34, 65, 78, 85, 87
  - Normal, 26, 87, 88, 90
  - r*-K-prime, 89
  - Uniform, 26–27
- Effect size [ES], 6, 39, 43
  - ANOVA ES indicators, 66–68
    - Eta-squared, 66–68, 78
    - Omega-squared, 67
    - Proportion of variance explained, 64, 66
  - Canned ES, 63, 112
  - Cohen's *d*, *see* Cohen's *d*
  - Definition, 59
  - ES estimate, 6, 43, 68
  - Glass's  $\Delta$ , 65

- Hedge's  $g$ , 65
- Heuristic benchmarks, 62, 69, 112
- Phi coefficient, 64
- Relative risk, 63
- Sample/population ES indicator, 67, 68
- Simple ES, 61, 112
- Standardized ES, 59–61, 64, 112
- Unstandardized ES, 61, 112
- Equivalence, *see* Clinical trials
- Errors, 24
  - Risk of errors, 28
  - Type I and Type II, 24–25, 41–42, 48
- Estimate, *see* Interval estimate
  - Point estimate
    - Unbiased estimate, 68
  - Unbiased point estimate, 65
- Estimation, 30
  - Estimation/significance tests problems, 30
  - Pure estimation, 30, 45–46, 49, 52, 56, 94, 112
- Examples
  - A clinical trial example, 41, 56, 72
  - A psychological example, 60–61, 78–91
  - A simple illustrative example, 10
  - An epidemiological study, 63
  - Student's example, 50, 79, 84–90, 92
- Fiducial Bayesian
  - Fiducial Bayesian inference, 4
  - Fiducial Bayesian methods, 4, 49, 85
- Fiducial inference, 4, 73–74
  - Fiducial argument, 73
  - Fiducial interval, 73
- Frequentist, *see* Confidence interval, Probability
  - Good frequentist properties, 4, 76, 77
- Guideline Hypotheses Official Significance Testing, 37, 41–44
  - Bayesian alternative, 90
- Guidelines, 6, 44
  - APA publication manual, 43, 105
  - ICH guidelines for clinical trials, 41–43
- Heuristic benchmarks, *see* Effect size
- Hybrid
  - Hybrid logic of statistical inference, 37
  - Hybrid practice, 43–44
  - Hybrid theory of testing, 5
- Hypothesis, *see* Null hypothesis
  - Alternative hypothesis, 12, 23, 26
  - Neyman-Pearson's tested hypothesis, 23
  - Working hypothesis, 41
- Interval estimate, 6, 43, 69, 71–72, 76, 77
  - Confidence, *see* Confidence intervals
  - Equal-tailed interval, 72
  - For a contrast between means, 90
    - Centered on zero, 91
  - For a correlation coefficient, 90
  - For a proportion of population differences, 88
  - For a relative risk, 64
  - For a standardized contrast, 87–88
  - One-tailed interval, 72, 77
  - Shortest intervals, 76
- Inverse, *see* probability
- Jeffreys, *see* Prior, Statistical tests
  - Jeffreys' rule, 26
- Judgment, 32–33
- Killeen's probability of replication, 51
- Learning from data, 29, 55
- Likelihood
  - Likelihood function, 13
  - Likelihood principle, 56–57
- Meehl's paradox, 45, 54
- Misinterpretations, *see* Confidence intervals, p-values
- Neyman-Pearson lemma, 24
- Noncentral  $F$  based [NCF], *see* Confidence intervals
- Noninformative, *see* Prior
- Null hypothesis, 11
  - A straw man, 47
  - Composite, 34
  - Fisher's null hypothesis, 22
  - Jeffrey's null hypothesis, 26
  - Notation  $H_0$ , 38
- Null hypothesis Significance Testing
  - Misinterpretations, 5
- Null hypothesis Significance Testing [NHST], *see* Statistical tests
- Objective, *see* Prior
  - Objective Bayesian analysis, 4
  - Objective Bayesian position, 32
  - Objective methods, 19, 21, 25, 57
- Odds
  - Prior and posterior odds, 27
- One-tailed, *see* Interval estimate
- p-value, 11, 22
  - Bayesian interpretation, 17, 50, 51, 90

- Jones and Tukey's procedure, 48
  - Misinterpretation, 10, 49–52
  - Reporting, 42
- Population, *see* Effect size
- Finite population, 10
- Posterior, *see* Fiducial BayesianMethods
  - Posterior distribution
    - for  $\delta$ , 85
  - Posterior probability, 6, 13, 14
  - Posterior probability of specified regions, 17, 72
  - Predictive posterior, *see* Predictive
- Power, *see* Sample size
- Power function, 25
- Power of a test, 12, 24, 41, 54, 63
- Resultant power, 28
- Predictive, *see* Distribution
  - Posterior predictive distribution, 52
  - Posterior predictive distribution for  $d'_{obs}$ , 92
  - Predictive probability, 13
- Prior
  - Default prior, 26
  - Jeffreys' prior, 26, 49, 72, 85
  - Noninformative prior, 26
    - Noninformative prior probabilities, 3
  - Objective prior, 26
  - Opinion-based prior, 14
  - Prior probability, 13
  - Uniform prior, 16, 71
  - Vague prior distribution, 3, 16
- Probability, *see* Posterior, Predictive, prior
  - Bayesian conception, 9
  - Frequentist conception, 9, 32
  - Inverse probability, 3, 50
    - Principle of inverse probability, 15, 29
  - Probability of hypotheses, 27, 31, 72
  - Probability of replication, *see* Killeen
- Quality control, 28
- Reasoning
  - Deductive/inductive reasoning, 32–33
  - From data to parameter, 12
  - From parameter to data, 10
- Risk, *see* Errors
- Sample size, 43, 54, 80
  - Ad hoc sample size, 63
  - For equivalence/non-inferiority trials, 44
  - Sample size and significance, 49, 51
  - Sample size determination, 41, 48, 63, 93
- Sampling
  - Sampling distribution, 10
  - Sampling probabilities, 10–11
- Significance, *see* Sample size, Statistical tests
  - Level of significance, 22
    - Reference levels of significance, 22, 38
  - Nonsignificance as proof of no effect, 39, 40
  - significant/nonsignificant, 5, 11–12, 22, 39, 53
  - The dictatorship of significance, 38
- Smallness, *see* Clinical equivalence trials
  - Demonstrating smallness, 79–81, 91–92
- Statistical tests
  - Amalgam, 5, 37–38, 43
  - Conservative/anti-conservative, 17
  - Frequentist conception, 11
  - Goodness-of-fit test, 56, 79
  - Jaynes' Bayesian test, 50
  - NHST, 5, 11, 37
  - One-sided tests, 11, 41, 47
  - Student's [W.S. Gosset] conception, 50
  - Test of composite hypotheses, 34, 77, 80
  - The Fisher test of significance, 5, 21
  - The Jeffreys Bayesian approach, 25
  - The Neyman-Pearson hypothesis test, 5, 23
  - Two One-Sided Tests procedure, 80, 92
  - Two-sided tests, 11, 47
- Subjective
  - Bayesian subjective perspective, 3–4, 17, 19
- The significance test controversy, 5, 45–57
- The sizeless scientists, 39
- The star system, 38, 39
  - The new star system, 62, 69
- Uniformly most powerful test, 24, 34, 36