A critical analysis of the role of statistical significance testing in education research: With special attention to mathematics education

Yui-kin Ng

A critical analysis of the role of statistical significance testing in education research: With special attention to mathematics education

By

Yui-kin Ng

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Education

in the

School of Education

University of Durham

2005

A critical analysis of the role of statistical significance testing in education research: With special attention to mathematics education

Yui-kin Ng

Submitted for the degree of Doctor of Education

2005

Abstract

This study analyzes the role of statistical significance testing (SST) in education. Although the basic logic underlying SST – a hypothesis is rejected because the observed data would be very unlikely if the hypothesis is true – appears so obvious that many people are tempted to accept it, it is in fact fallacious. In the light of its historical background and conceptual development, discussed in Chapter 2, the Fisher's significance testing, Neyman-Pearson hypothesis testing and their hybrids are clearly distinguished. We argue that the probability of obtaining the observed or more extreme outcomes (p value) can hardly act as a measure of the strength of evidence against the null hypothesis. After discussing the five major interpretations of probability, we conclude that if we do not accept the subjective theory of probability, talking about the probability of a hypothesis that is not the outcome of a chance process is unintelligible. But the subjective theory itself has many intractable difficulties that can hardly be resolved. If we insist on assigning a probability value to a hypothesis in the same way as we assign one to a chance event, we have to accept that it is the hypothesis with low probability, rather than high probability, that we should aim at when conducting scientific

research. More important, the inferences behind SST are shown to be fallacious from three different perspectives. The attempt to invoke the likelihood ratio with the observed or more extreme data instead of the probability of a hypothesis in defending the use of p value as a measure of the strength of evidence against the null hypothesis is also shown to be misleading because it can be demonstrated that the use of tail region to represent a result that is actually on the border would overstate the evidence against the null hypothesis.

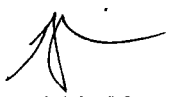
Although Neyman-Pearson hypothesis testing does not involve the concept of the probability of a hypothesis, it does have some other serious problems that can hardly be resolved. We show that it cannot address researchers' genuine concerns. By explaining why the level of significance must be specified or fixed prior to the analysis of data and why a blurring of the distinction between the p value and the significance level would lead to undesirable consequences, we conclude that the Neyman-Pearson hypothesis testing cannot provide an effective means for rejecting false hypotheses.

After a thorough discussion of common misconceptions associated with SST and the major arguments for and against SST, we conclude that SST has insurmountable problems that could misguide the research paradigm although some other criticisms on SST are not really as justified. We also analyze various proposed alternatives to SST and conclude that confidence intervals (CIs) are no better than SST for the purpose of testing hypotheses and it is unreasonable to expect the existence of a statistical test that could provide researchers with algorithms or rigid rules by conforming to which all problems about testing hypotheses could be solved. Finally, we argue that falsificationism could

eschew the disadvantages of SST and other similar statistical inductive inferences and we discuss how it could bring education research into a more fruitful situation in which teachers and other educational professionals would find the research output really matters to their practices. Although we pay special attention to mathematics education, the core of the discussion in the thesis might apply equally to other educational contexts.

Declaration

I declare that the work presented in this thesis is, to my best of my knowledge, original and that the material has not been submitted, in whole or in part, for the award of any other degree at this or any other university.



Yui-kin Ng

December 2005

Contents

List of tables

List of figures

ACKNOWLEDGMENTS

I wish to thank my supervisors, Professor Peter Tymms and Dr Robert Coe, for their generous advice and critical suggestions on this thesis. They are also responsible for the timely completion of the project. I don't think it is possible for this thesis to be free of obscurities and mistakes, but I am sure that there are fewer of them because of their valuable comments.

Next, I would like to express my intellectual debts to Steven N.S. Cheung, who demonstrates how social sciences could become genuine scientific, and John R. Searle, whose work exemplifies how the most difficult concepts in philosophy could be expressed lucidly, without whose writings the present work, whatever its worth, would never have been.

Finally, to my wife Sheena, I owe more than I can express.

Yui-kin Ng

December, 2005

Chapter 1 Introduction

Nowadays there are abundant educational studies that are claimed to be scientific. At one level, scientific enquiry could allow us to confirm the existence of phenomena in the educational arena and, as a result, it should give teachers information whether their teaching strategies would really enhance the learning and teaching effectiveness or provide policy makers with data whether a policy will lead to better education for the students. At another level, when researchers propose a theory or hypothesis, the theory or hypothesis has to be tested to see if it true or not. Hypothesis testing is thus an inevitable process in the scientific inquiry. At both levels, we usually have to approach the analysis of experimentation with the tools of statistics – descriptive and inferential statistics. There is little controversy surrounding the use of descriptive statistics in describing various states of nature. But almost all inferential statistical methods are subject to serious discussions. Statistical significance testing (SST) is one of the controversial statistical methods.

The history of the use of SST can date back to a paper published almost 300 years ago, 'An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes.' (Arbuthnott, 1710). SST was not prevalent until the first half of the twentieth century and has dominated education research for the past 70 years (Kirk, 1996). In the meantime, various concepts underlying SST were developed. To understand how the cluster of concepts has been brought together in the current practice of SST, we should have a proper understanding of its historical development. In this

connection, the historical background and conceptual paradigms underlying the development will be discussed in Chapter 2.

The use and interpretation of SST have, however, increasingly come under attack in recent decades. For example, many recent studies have argued that the long-standing practice of reliance on SST is logically indefensible and corrupts our research enterprise in social sciences (Cohen, 1990, 1994; Falk & Greenbaum, 1995; Schmidt, 1996; Schmidt and Hunter, 1997) and the theoretical bases of SST are widely misunderstood by researchers (Macdonald, 1997). In the arena of education, Carver (1978) has argued that 'even properly used in scientific method, education research would still be better off without statistical significance testing.' (p.398) The controversy has recently come to a head. Many clarion calls, as Thompson (1999a) described, have been published urging researchers to abandon, or to at least supplement, the use of SST. The growing uneasiness can also be reflected in the change of editorial policies of some top journals. For instance, the previous edition of American Psychological Association (APA) (1994) style manual has prompted closer scrutiny of contemporary analytic practices:

> Neither of the two types of probability values[1] [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size. You are encouraged to provide effect-size information (APA, 1994, p.18).

[1] The two types of probability here are referring to the "alpha level" (or "significance level") and the *p* value (APA, 1994).

A group of researchers had requested the American Psychological Association (APA) to consider banning the SST altogether in its journals (Shrout, 1997). The APA Board of Scientific Affairs thus launched a Task Force on Statistical Inference in March of 1996 to work on related recommendations for improving research practices. (Azar, 1997; Shea, 1996).[2] In its latest edition, APA (2001) made the following remark:

> The field of psychology is not of a single mind on a number of issues surrounding the conduct and reporting of what is commonly known as *null hypothesis significance testing.* These issues include, but are not limited to, the reporting and interpretation of results of hypothesis tests, the selection of effect size indicators, the role of hypothesis-generating versus hypothesis studies, and the relative merits of multiple degree-of-freedom tests (APA, 2001, p.21).

This manual did not target at resolving these issues and it merely directed the discussion of these and other issues to Wilkinson and the Task Force on Statistical Inference (1999)[3]. It is noteworthy that before and after the publication of this manual, different journal editors began adopting requirements that authors have to report

[2] A ban of SST has taken place in *American Journal of Public Health.* (Shrout, 1997) For changes in editorial policies or author guidelines of other journals, see Levin and Robinson, 1999; Thompson, 1996, 1999a, 1999b, 1999c, 1999d; Thompson and Snyder, 1997.

[3] Here the reference was quoted wrongly as "Wilkerson (1999)" in APA (2001, p.21) (it should be "Wilkinson and the Task Force on Statistical Inference (1999)"). Corrections could be found in APA's official website: http://www.apastyle.org/pubman-reprint.pdf. Similar prompt for the report of effect size and strength of relationship could also be found in its latest edition: "Neither of the two types of probability values directly reflects the magnitude of an effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section." (APA, 2001, p.25).

information in additional to SST.[4] Some researchers have argued that the ban is justified because SST has violated the scientific principle that a scientist should espouse whereas others believe that the problem stems from misuses of SST, instead of from SST itself. SST, properly used, can be beneficial to the development of education research. The controversy, to which a special lead section of the January 1997 issue of *Psychological Science* is devoted, continues (Shrout, 1997; Hunter, 1997; Harris, 1997; Abelson, 1997a; Estes, 1997).

The controversy is not new. A number of criticisms of the use and interpretation of statistical significance tests have been made since the 1960s (Bakan, 1966; Morrison & Henkel, 1970).[5] Perhaps very few methodological issues have generated as much controversy as the use and interpretation of SST (Pedhazur and Schmelkin, 1991), and this issue is certainly one of the hottest developments in methodology over the past 40 years. If the criticisms against the use of SST are correct, how on earth could the use of SST still prevail now? Is it simply, as Jacob Cohen says (Shea, 1996), due to academic inertia? Or, are the criticisms not completely correct? Do we have good reasons to make our judgment on this controversy? What are the implications of our judgment for future research studies in education? If SST really has serious limitation or is being misused, what practices should replace or supplement it? These are the major questions that we are going to address in this thesis. But before addressing these questions, we have to consider the different interpretations of probability first. As Grayson (1998) has argued,

---

[4] For example, authors have to report effect sizes as indices of practical significance in *Journal of Experimental Education*, *Journal of Educational and Psychological Consultation*, *Research in Schools*, etc. See Kennedy, 2002 for a list of 17 journals now require the reporting of effect sizes.

[5] The earliest serious criticism on the logic and usefulness of SST, according to Kirk (1996), appeared in a 1938 article by Joseph Berkson in the *Journal of the American Statistical Association*.

a major issue underlying the controversy of SST revolves around the meaning of the probability of a hypothesis. Indeed, when using SST we often invoke the concept of the probability of a hypothesis. But what do we mean by 'the probability of a hypothesis'? We will address this question in Chapter 3. There are many different interpretations of probability of which five major interpretations, namely the classical theory, the logical theory, the frequency theory, the subjective theory and the propensity theory, will be discussed in Chapter 3. We will see whether some could be resistant to criticisms and how the clarification of the concept of probability is of crucial importance to the understanding of SST.

Apart from the interpretations of probability, the null hypotheses, p value, statistical significance and Type I and Type II error are all important concepts underlying SST. Misunderstanding of these concepts could have disastrous consequences for the use of SST. These concepts will be clarified in Chapter 4, together with an in-depth discussion of the distinctions between Fisher's significance testing and Neyman-Pearson hypothesis testing. Only a thorough discussion of these issues could enable us to give a critical examination of the arguments for and against the use of SST. We will show in Chapter 5 that not all arguments against SST are tenable and advance some arguments to show that SST has many insurmountable problems that could misguide the research paradigm.

Many critics of SST have proposed different alternatives to SST. We will see in Chapter 6 whether the proposals are successful. There is one deeper reason for why SST fails. We will show in Chapter 6 that most critics of SST ignore this deeper reason for

the failure of SST when they are trying to find alternatives to SST. In order to explicate this reason, we made a contrast between SST and falsificationism. According to the traditional view, such as those advocated by many falsificationists, SST is only a methodological rule for falsifying statistical hypotheses. We will argue, however, that there are indeed subtle differences between SST and falsificationism.

As Menon (1993) has argued, the use of mathematics in SST might be the reason that many researchers have assumed that SST is supported by rigorous mathematical logic. The non-mathematically inclined education researchers seem to have good excuses for their failure of examining the logic and the sophisticated statistical techniques underlying SST. How about the researchers in mathematics education? Do they take a lead in throwing off the shackles of SST, as Menon (1993) has urged almost 12 years ago? In order to answer this question, although the analysis in this dissertation could be applied to education research in general, the focus will be placed on mathematics education research.

A lot of money is spent each year on education research (Pring, 2000). It is, however, said that this money is not well spent (Hillage, 1998; Tooley and Darby, 1998) and education research is continually being criticized for its poor quality (Levin and O'Donnell, 1999). Criticisms of education research have been emerging for a number of years. The reputation of education research has been regarded as awful (Kaestle, 1993). There are many explanations for the existence of these criticisms. We hope that the discussions of the SST issue could throw some light on the methodological foundations of education research so that we could positively respond to the criticisms of education

research and suggest some ways how education research could develop in a scientific way. And this constitutes our last task in Chapter 6.

Chapter 2 A historical review of statistical significance testing

Statistical significance testing (SST) or null hypothesis significance testing[6] has a long history. Underlying it various concepts were developed at different times and in different conceptual paradigms. SST is thus not a single and simple concept that a careful examination of its current practice could be adequate for a thorough understanding of the logic underlying it. In this connection, we will first review the historical background and conceptual paradigms behind its development.

2.1 The first published test of a statistical hypothesis

The history of the use of SST can date back to John Arbuthnott's 1710 paper, 'An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes.' (Arbuthnott, 1710[7]). In this short paper, which is now commonly regarded as the first paper on the formal test of significance or inferential statistics (Baird, 1981, p.48; Campbell, 2001, p.607; Eisenhart & Birnbaum, 1967, p.24; Gigerenzer et al., 1989, p.79; Hald, 1998, p.65; Tankard, 1984, p.11)[8], Arbuthnott provided arguments for

[6] To be more precise, 'significance testing' (or 'significance test') is used to designate the approach developed by Ronald A. Fisher and 'hypothesis testing' (or 'hypothesis test') to that by Jerzy Neyman and Egon Pearson in the twentieth century (Berger, 2003; Hubbard & Bayarri, 2003a; 2003b; Hubbard, 2004; Lehmann, 1993). The distinction between these two approaches will be discussed later. When there is no need to distinguish between these two approaches, we will use the term 'statistical significance testing' (or 'SST')in this thesis.

[7] The electronic version of the original paper is available on the Web at JSTOR and its edited version is at http://panoramix.univ-paris1.fr/CHPE/Textes/Arbuthnot/arbuth.html# (the data set is not free of typos, e.g. in 1662, the number of females should be 4803 rather than 4823). It has also been reprinted elsewhere (e.g., David & Edwards, 2001, pp.13-17; Kendall & Plackett (Eds.), 1977, pp.30-34). This paper might not be Arbuthnott's first contribution to probability or significance tests, see Bellhouse (1989) for details of his earlier manuscript on probability and significance tests.

[8] Hogben (1957) attributes, however, the earliest use of probable error as a form of significance test in the biological arena to Jules Gavarrett (1840) (p. 324), and consider Venn (1889) as one of the earliest users of the terms 'test' and 'significant' in their now current meaning (p.325). Isolated application of statistical hypothesis testing could also be found centuries earlier in the Royal Mint's Trial of Pyx (Stigler, 1999, pp. 383-402).

the presence of divine providence in the determination of the sex ratio.[9] First, he considered the toss of an even number (say, n) of fair two-sided dice (or coins) whose sides were marked M and F. By calculating the binomial coefficients in the expansion of $(M + F)^n$, he showed that the probability[10] of getting exactly as many M as F would become very small when n got large and thus concluded that 'in the vast number of mortals, there would be but a small part of all the possible time, that an equal number of males and females should be born' (p.187). In his words, 'it is very improbable (if mere chance govern'd) that they [the outcomes] would never reach as far as the extremities' (p.188). But in fact the Bills of Mortality indicated that no such vast preponderance of M over F, or F over M ever occurred. He thus concluded that 'this equality of males and females is not the effect of chance but divine providence, working for a good end' (p.186).

Second, Arbuthnott noticed that the sex ratio was not exactly equal to one. He argued further that if only chance governed, then in a given year the probability that the number of male births exceeded that of female births would be smaller than or equal to ½. Assuming that the probability is equal to 1/2 , he reckoned, with the use of the table of logarithms, the probability of more males than females every year over an 82 year period would be $1/2^{82} = \frac{1}{4\,836\,000\,000\,000\,000\,000\,000\,000}$ (p.188). This is certainly a very small number, but the fact he observed was: over 82 consecutive years from 1629

[9] According to Hacking (1975), the 'argument' is consisting of three inferences of which two are statistical and one is about the very nature of statistical stability (p.167). And, according to Bellhouse (1989), Arbuthnott has provided two arguments for the presence of divine providence in this paper. For detailed commentaries on Arbuthnott's arguments, see, for instance, Bellhouse, 1989; Hacking, 1965, 1975; Hald, 1990, 1998; Shoesmith, 1985, 1987.

[10] In the 17th century, the terms 'probability' and 'chance' were used in way that may be different from our present usage. 'Probability', resembling subjective probability, was related to beliefs, opinions, proposition, whereas 'chance' was used to mean objective probability (Hald, 1990, p.246). For our present purpose, there is no need for us to distinguish its different meanings. We will defer the discussion of its different notions to Chapter 3.

to1710[11], more males than females were born (or, strictly speaking, christened) in London. The existence of this phenomenon with such a minute probability was interpreted as decisive evidence against the hypothesis that merely chance governs the distribution of sexes and for the alternative of a regular excess of male births. This rejection of a hypothesis because the observed data would be unlikely if the hypothesis were true is cited as 'the first published test of significance of a statistical hypothesis'. Arbuthnott also tried to justify this constant regularity (i.e. more male births than female births but not the converse): it served to offset the higher death rate of males due to more external accidents which males were subject to, and preserving the equal proportions of adult males and females required by the institution of monogamy.

Arbuthnott's arguments did not single him out from his contemporaries in the advancement of religious argument for divine providence, nor were his observations of the sex ratio innovative (Shoesmith, 1987). But these arguments still provoked controversy during the succeeding few years amongst many reputed mathematicians, such as Abraham de Moivre, Willem 'sGravesande, Pierre R de Montmort, Nicholas (or Niklaus) Bernouilli and Bernard Nieuwentijt (Hacking, 1965; Eisenhart & Birnbaum, 1967; Shoesmith, 1985, 1987). And, several years after the publication of this paper, Daniel Bernoulli[12] used Arbuthnott's method to test a hypothesis about the inclination of

[11] Hacking (1975) thus argued that the paper must have been printed in 1711 since it included data on births going to the end of 1710. And, according to Hald's (1990) study, the paper was published in 1712.

[12] Bernoullis are probably the most famous family in the history of statistics. James (or Jacob, Jakob, Jacques) Bernoulli (1654-1705), who was regarded by Keynes (1921) as the real founder of mathematical probability (p.41), was the one who first proves the first limit theorem in probability theory - the weak law of large numbers (Hald, 1990, p.225; Hacking, 1978, p.154) and we will return to this theorem later. 'Bernoulli trials' was named after him (Feller, 1968, p.251). John (Johann, Jean) Bernoulli (1667-1748) was a brother of James Bernoulli. John worked for many years on the same problems as James. Daniel Bernoulli (1700-1782) was John's son, and his elder cousin was Nicholas (Niklaus) Bernoulli (1687-1759)

the planetary orbits (Freudenthal, 1970). Arbuthnott's attempt in providing a statistical proof of his assertions, based upon a quantitative notion of chance, has secured him a place in the history of statistics because the structure of his arguments is regarded as that of a SST, which can still be identified in modern statistics texts[13] (Shoesmith, 1987). However, are Arbuthnott's arguments really sound? In the first argument, although Arbuthnott calculated only the probabilities for n = 2, 4, 6, 8, 10 in his short paper and asserts, without giving any proof, that the probability of getting exactly as many M as F, for large n, would be vanishingly small, this conclusion is indeed correct.[14] But as shown in Appendix 1, the relative frequency of getting extreme outcomes, for large n, will also be very small. It is thus incorrect to conclude that "it is very improbable that the outcomes would never reach as far as the extremities.' That is to say, the fact that no vast preponderance of M over F, or F over M ever occurs could simply be an effect of chance, or a result of the Bernoulli's limit theorem.

In his second argument, Arbuthnott studied the hypothesis:

H : It is an even chance, whether a child be born male or female.

He argued that if H was true, the distribution of births would be similar to the outcomes obtained from the toss of fair two-sided dice (or fair coins). Believing that there was a constant regularity in the births of both sexes, he also admitted that there could be slight preponderance of males over females. Instead of testing H directly, a new hypothesis H_0 was put to the test:

---

who had refined his uncle's limit theorem. They all spent most of their lives as professors at Basle. For details, see Hald, 1990, pp.220-223; 1998, p.83; Pearson, 1978, pp.221-237; Stigler, 1986, p63.

[13] See, for example, Lehmann, 1986, pp.106-107; Miller & Miller, 1999, pp.529-531; Smithson, 2000, p.182.

[14] The details are shown in Appendix 1.

H_0 : The distribution of chances of male years is binomial with $P(M)=1/2$, where 'male year' denoted the year in which more boys were born than girl.

He noticed that there were indeed 82 male years in London. He then calculated the chance of getting this outcome under the assumption that H_0 is true, and the chance was found to be extremely small($1/2^{82}$.) He then rejected H_0, and subsequently H. The whole argument can be put in this way:

(P_1) If H_0 is true, then the chance of getting 82 male years happening is extremely small, i.e. $1/2^{82}$.

(P_2) 82 male years happened in London.[15]

$\therefore$ (C_1) H_0 is rejected.

$\therefore$ (C_2) H is rejected.

(P_3) If H is rejected, then it must be art, not chance, that governed in the matter of birth.

Therefore,

(C_3) In the matter of birth it must be art, not chance, that governed.

Before assessing the validity of the arguments, we have to assess the truth of the premises. Let us consider (P_2) first. Arbuthnott's assertion of (P_2) was based on christening instead of birth statistics.[16] There are at least three problems arising from this set of data. First, since the data was based on christenings instead of birth statistics, only persons subscribing to the Church of England were counted. Second, there might be time

[15] According to Arbuthnott, the actual probability were indeed much smaller than this estimation for the inequality in the sex ratio had been observed in several other places and at other times (p.188).
[16] See Appendix 2 for the data.

delay between birth and christening. Third, there was no data on population size and it will certainly impose constraints on further statistical analysis. Hence, Arbuthnott's assertion was not so well-founded. Nevertheless, an excess of male births over female births is now a widely held belief that has been demonstrated in many other studies (Campbell, 2001, p.608), we could grant here the truth of (P_2) for the sake of argument. And we will, in turn, analyze Arbuthnott's calculation of the chance in (P_1)

Although it seemed as if Arbuthnott had shown that the probability of the observed data (82 male years) was very small, assuming the truth of H_0, the precise manner in which this small probability was to be interpreted so as to yield the conclusion was not well-articulated. For example, suppose the first year was male year, the next was female, the next to next was again male, and so on, in some definite order, giving a total of 60 male years and 22 female years. What is the probability of this outcome provided H_0 was true? It seems that the probability was not different from that of the original data:

$$\frac{1}{2}\times\frac{1}{2}\times\cdots\times\frac{1}{2}=\frac{1}{2^{82}}.$$

If it were the case, then H_0 would be rejected regardless of the outcomes. Obviously, the probability of getting 60 male years and 22 female years, irrespective of the order, should be greater than that of getting of 82 male years. Hence, Arbuthnott should have given a more explicit statement of the calculation and the test. In other words, there are 2^{82} different possible outcomes. If order mattered, Arbuthnott should have stated clearly which subclass of these results was being considered so that H_0 would be rejected if and only if the actual outcome lay in this subclass. In the writing of Arbuthnott's

contemporary mathematician Willem (or Guillaume) 'sGravesande, probably under the influence of Bernard Nieuwentijt, a clear rejection class for the test had been proposed. Based on this concept, 'sGravesande argued that Arbuthnott's argument could have put the case for divine providence even more strongly than he had done if he could engage in a more detailed calculation using the binomial distribution (Shoesmith, 1987).[17] We could see here that the rejection subclass is the seed of the later developed concept 'rejection region' in SST.

Assuming the truth of (P1) and (P2), we still could not establish the truth of (C_1) unless the argument from (P1) and (P2) to (C) is valid. Although many commentators on Arbuthnott's second argument had assumed the validity of this argument (see, for example, Hacking, 1975, p.168; Shoesmith, 1987, p.138), its validity has to be critically assessed as this sort of argument still plays a very important role in SST. We will leave the discussion to Chapter 5.[18]

2.2 Development before Fisher, Neyman and Pearson

As argued in previous section, Arbuthnott did not give a rigorous treatment to the SST. We will discuss briefly in this section how the concept of SST was developed before our modern forms of SST emerged.

[17] The details are shown in Appendix 3.
[18] Besides, Nicholas Bernoulli has proposed an argument against Arbuthnott's second argument. But this is not the point in question, we will leave the details to Appendix 4.

James Bernoulli's posthumous book *Ars Conjectandi* (*The Art of Conjecturing*, 1713)[19] consists of four parts and it is the fourth part that revolutionizes probability theory. The revolution, according to Hacking (1975), is two-fold. First, a 'subjective' conception of probability was explicitly elaborated. Second, James Bernoulli gave a formal treatment to the vague conception that the greater accumulation of observations about the proportion of cases, the closer we are to certain knowledge about the proportion. This is also the first limit theorem proved in probability theory.[20] One of the great achievements by Bernoulli is the commencement of the journey toward a mathematical quantification of uncertainty. But whether a mathematical quantification of uncertainty could be used to justify induction will be critically reviewed in later Chapters.

Thomas Bayes' paper 'An essay towards solving a problem in the doctrine of chances'[21] was presented to the Royal Society in an edited form by Richard Price in 1763, half a century after the publication of *Ars conjectandi*. In the presentation, Price briefly mentioned that although James Bernoulli and de Moivre had attained important results they had not demonstrated how to solve the problem of induction.[22] According to Price, Bayes was the first person who provides 'a sure foundation for all our reasonings concerning the past facts' (Bayes, 1763/1958, p.296). It is now well-known that Thomas

---

[19] *Ars Conjectandi* was only nearly completed when James Bernoulli died in 1705. It was not until eight years later in 1713 that the manuscript was published with a short preface by Nicholas Bernoulli, a nephew of James. The original version could be found in Speiser (Ed.), 1975, pp.107-286. No complete English translation of this manuscript is available (Shafer & Vovk, 2001, p.376). English translations of the extracts could however be found in Adams (1974, pp.10-14), David (1962/1998, pp.130-139), Shafer (1978, 326-339), Uspensky (1937, pp.105-107).

[20] See Appendix 5 for the details of the theorem and its implication.

[21] The essay was original published in *Philosophical Transactions of the Royal Society of London*, *53*, 370-418 and was reprinted in *Biometrika* , *45*, 293-315 with commentary by G. A. Barnard and elsewhere (e.g., Pearson and Kendall (Eds.), 1970, 131-153; Swinburne (Ed.), 2002, 122-149).

[22] For detailed discussion of Bayes's essay and Price's introduction, see Dale (1999).

Bayes and Peirre Simon Laplace[23] had independently proved different versions of the inverse of Bernoulli's theorem. In modern notation, the theorem is usually stated as:

$$P(H \mid E) = \frac{P(H \wedge E)}{P(E)} \quad \text{(provided } P(E) \neq 0\text{)}$$

or

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \quad \text{(provided } P(E) \neq 0\text{)}^{24}.$$

If 'H' and 'E' denotes a hypothesis and evidence respectively, then Bayes's theorem says that the posterior probability of H (its probability, given E, i.e. $P(H \mid E)$) equals its likelihood ($P(E \mid H)$, sometimes called the 'predictive power of H'[25]) multiplied by its prior probability ($P(H)$) divided by the prior probability of E ($P(E)$). It is controversial that all propositions should have probabilities. For instance, for a definite hypothesis H, either it is true or false, does it make sense for us to talk about the prior probability of a hypothesis? That depends on what we mean by 'probability', a point to which we will return in the next chapter.[26]

Pierre Simon Laplace also made a contribution to the development of SST. His treatment of testing a simple hypothesis against a simple alternative hypothesis was very

[23] Laplace's 1774 memoir on the probability of the causes of events is regarded as one of the revolutionary papers in the history of statistical inference and his first proof of Bayes's formula could be found in his 1781 memoir (Hald, 1998, p.164).

[24] For those who think that all probabilities should be conditional probabilities, Bayes's theorem will be expressed in the form: $P(H|E\wedge K) = P(H\wedge E|K)/P(E\wedge K)$ or $P(H|E\wedge K) = P(E|H\wedge K)\ P(H|K)/P(E|K)$, where '$K$' denotes our background knowledge.

[25] 'Likelihood' is a terminology derived from R. A. Fisher and is widely used in current literature. Swinburne (2002) has argued that this terminology is misleading and suggested to use 'predictive power' to denote this probability (p.10).

[26] For a preliminary discussion of the relation between Bayes's theorem and SST, see Appendix 6.

simple[27]. Assuming equal prior probabilities of the two mutually exclusive and exhaustive hypotheses (say, H_1 and H_2), he obtained

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)}, \text{ where '}E\text{' denotes the evidence.}$$

Later statisticians (such as Good, 1983, p.158) defined the Bayes factor in favour of the hypothesis H_1 provided by the evidence E as the ratio of the final odds to the initial odds[28]:

$$\frac{O(H_1 \mid E)}{O(H_1)} = \frac{\dfrac{P(H_1 \mid E)}{1-P(H_1 \mid E)}}{\dfrac{P(H_1)}{1-P(H_1)}} = \frac{\dfrac{P(H_1 \mid E)}{P(H_2 \mid E)}}{\dfrac{P(H_1)}{P(H_2)}} = \frac{P(H_1 \mid E)}{P(H_2 \mid E)} \frac{P(H_2)}{P(H_1)} = \frac{P(H_1 \mid E)}{P(H_2 \mid E)}.$$

The Bayes factor in favour of H_1 against H_2 can thus, as Laplace has shown, be expressed as:

$$\frac{P(E \mid H_1)}{P(E \mid H_2)},$$

which in frequentist terminology is called the 'likelihood ratio'. For example, in the two-urns example[29], the evidence E is that there are 4010 drawings in which a white pebble is taken out when there is a total 6450 drawings have been made. The Bayes factor is thus:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)} = \frac{1.751243 \times 10^{-5}}{3.164495 \times 10^{-1184}} = 5.53404 \times 10^{1178},$$

a very large number which indicates that E is strongly in favour of H_1 or against H_2.

[27] See Hald (1998, 167-183) for a presentation of Laplace's testing of hypothesis in modern terminology.
[28] The odds corresponding to a probability p are defined as $p / (1-p)$.
[29] One of the two urns contains 30 white and 20 black pebbles and the other contains 10 white and 40 black pebbles. We first randomly select one urn and take out one pebble after another (with replacement). See Appendix 6 for details.

As Hald (1998) pointed out, the Bayes factor should have been called 'Laplace factor' since it did not occur in Bayes's work. Laplace's test of hypotheses was fundamentally different from the SST that was derived from direct probability arguments.[30] He laid down the foundation of the modern theory of SST. His theory of estimation and testing was a large-sample theory based on the normal distribution. Although the posterior probabilities could be calculated for small samples, he did not give any examples of SST for small samples. He has indeed derived large-sample tests for one-sided hypotheses for both one and two samples. For example, in testing the one-sided hypothesis for one sample $\theta \le \theta_0$ against $\theta > \theta_0$, where θ denotes the unknown probability of success in a series of independent binomial trials, he compared the posterior probability:

$$P(\theta \le \theta_0) = \Phi(\frac{\theta_0 - m}{c})$$

with its complement, where m and c^2 are respectively the mean and variance of the posterior distribution of θ (which is assumed to be normal for large samples) and Φ is the standard normal cumulative distribution function (i.e. $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$)..

And in testing the one-sided hypothesis for two samples $\theta_2 - \theta_1 \le 0$ against $\theta_2 - \theta_1 > 0$, he used:

$$P(\theta_2 - \theta_1 \le 0) = \Phi(u), \text{ where } u = \frac{m_1 - m_2}{\sqrt{{c_1}^2 + {c_2}^2}},^{31}$$

where m_1, m_2 and ${c_1}^2$, ${c_2}^2$ are respectively the means and variances of the posterior distributions of θ calculated from the two samples. These results have ramifications for

[30] Laplace also extended his analysis to more general cases. For details, see Hald (1998, pp.232-247).
[31] In his proof, Laplace only assumed the asymptotic normality of the two variables. Details could be found in Hald (1998, pp.229-247).

both Bayesians and non-Bayesians since the effect of the prior distribution diminishes sharply for large samples and the resulting distribution is normal[32] so that the parameter and its estimate are symmetrically involved. In other words, non-Bayesians are content with the interpretation that the resulting distribution is a distribution of m about θ, and Bayesians will interpret it as a distribution of θ about m (Hald, 1998, p.246).

Before the nineteenth century, measurement did not play much of a role in sciences (even in physics and chemistry) (Kuhn, 1977, p.220). Kuhn has argued that only around 1840 was the practice of measurement in bloom. The emphasis on measurement gave rise to the avalanche of numbers (Hacking, 1990, p.5). Although Newtonian laws might still be regarded as exact, measurements would no longer provide people with an exact quantitative picture of the world. When measurements become more and more exact, more and more errors appear. Errors and deviations from the mean became the 'norm'. The positivist philosophy of science began spreading to the realm of social sciences. Social scientists represented social facts with the use of statistics in the similar way that natural scientists did. As a consequence, statistical propositions were interpreted in a more realistic way while the Laplacian subjective interpretation began to fade. And gradually, science began to have a new paradigm called the 'the statistical model of reality' (Salsburg, 2001, p.VIII).

Measurements have been expressed as multiples of a variety of basic units that reveal the dispersion of the range of possible scores since the time de Moivre (Cowles,

[32] For the discussion of the nature of normal distribution, see Appendix 7.

2001, p.127). Standard deviation is the one that is commonly used nowadays.[33] In the early nineteenth century, probable error, which is defined as one half of the quantity that encompasses the middle 50% of a normal distribution of measurements, was first used by Friedrich Wilhelm Bessel[34] to reflect the dispersion[35]. Gauss has subsequently developed several methods of computing it (Cowles and Davis, 1982, p.555) and we now know that one probable error unit equals approximately 0.6744898 of a standard deviation[36]. By comparing probable errors, Bessel tried to determine if a difference was genuine or due to observational error. It is regarded as the earliest SST that involved the probable error measurement. Hermann Ebbinghaus, an experimental psychologist in the nineteenth century, used probable error to interpret the data he had gathered in his important study of memory around 1885[37]. Ebbinghaus regarded that a difference of six times the probable error would give a solid proof and a difference of twice the probable error is noteworthy. According to Ebbinghaus, we should be certain that a difference exists if the observed difference is six times the probable error but the observed difference is probably not the exact size of the true difference. This is perhaps an earlier conception of effect size that many researchers currently use.

Near the end of the nineteenth century, Francis Edgeworth developed a test of significance in which he compared the difference of the means with the modulus *M*, which is $\sqrt{2}$ times the standard deviation. Edgeworth has tried to determine how far the

[33] Standard deviation was first used in connection with normal distribution in 1896 by Karl Pearson.
[34] See Hald, 1998, pp.360-361.
[35] Other notions that have been used include: modulus, precision, fluctuation, mean error, error of the mean square, etc. (Baird, 1981, p.123).
[36] The figure is obtained with the use of Maple 8:

```
fsolve(0.25=Int(NormalPDF(0,1,x),x=0..a),a);
```

[37] For details, see Stigler, 1986, pp.254-261. Weigle (1994, p.7) also gave a brief introduction to Ebbinghaus's work.

difference between proposed means is accidental or indicative of a law. There are at least two cases that we could now be able to distinguish clearly[38]. First, for instance, if we wish to test the psychics' claim that they can systematically determine the colour of a playing card which is out of their view (Edgeworth, 1919). Suppose the card is drawing from a pack of normal playing card and we can assume that the a priori probability that the psychics could guess correctly the colour is 1/2. We can then perform a series of trials to determine if the psychics' performance could outweigh this a priori probability. This case is indeed a sort of goodness of fit. Having a hypothesis about how some observables are distributed, we deduce expectations from this hypothesis about a series of observations. We use goodness of fit tests to decide whether a difference observed between these expectations and the actual observations made could arise merely from sampling fluctuations.

For the second case, there are, in some situations, no such a priori probabilities on hand. We merely want to compare two means taken from what is supposed or hypothesized to be the same population. Using Edgeworth's terminology, we have to determine two a posteriori probabilities and compare them to see whether chance alone could give rise to the differences. Edgeworth (1885) gave this example: based on the observations of 65 deaths in the year, we knew that the rate of mortality amongst young farmers between the ages of 15 and 25 exceeded that in all other professions by 0.3 percent. He asked 'How far is such an extent of deviation based upon such a number of observations significant of a real difference in respect of healthiness between the conditions of young farmers and the rest of the industrial community?' (p.182). As Baird

[38] We have, however, no evidence that Edgeworth was able to distinguish these two types of cases.

(1981) noted, this question contained the earliest use of the word 'significant' with essentially its current sense. According to Edgeworth, a difference of twice the modulus (i.e. $2\sqrt{2}$ times the standard deviation) was regarded as 'significant' and differences of 1.5 times the modulus was noteworthy. In a nutshell, what we concern in this case is to see if the two samples could have been drawn from the same population instead of the exact shape of the population from which samples are taken. Student's t test is the first modern test for this problem.

Karl Pearson was clearly influenced by the work of Francis Galton, the first person who provided a scientific basis to the use of fingerprinting.[39] But it was his competitive relationship with Edgeworth that highly motivated Pearson to work.[40] As a result, even though Pearson had emphasized Edgeworth's significance testing methods in a series of lectures held in 1893, he deliberately measured differences in terms of a new measure of variation which he called the 'standard deviation' instead of the modulus which was proposed by Edgeworth. Person's one great contribution to SST was the creation of the first 'goodness of fit test' – the chi-square goodness of fit test[41]. This test allows us to determine the probability of occurrences of discrepancies between observed and expected frequencies in a distribution. The expected frequencies can be calculated from the null hypothesis we want to test. Once the observed frequencies (O) and expected frequencies (E) have been determined, χ^2_{obs} could be calculated by:

[39] Following Galton's estimation, fingerprint evidence has been widely accepted as certain for more than a century and only until recently its reliability has become doubtful (Gigerenzer, 2002, pp.12-13).
[40] According to Stigler (1986), one incentive to Pearson's work on the generalized form of the probability curve had been to do better than Edgeworth had (p.338).
[41] The Greek letter chi (χ) is used because the distribution of this test statistic belongs to a group of his skew distributions that he has designated the chi family. He called it 'chi-square' because the test statistic, indeed, behaves like the square of chi (Salsburg, 2001, p.96).

$$\chi^2_{\text{obs}} = \sum_i \frac{(O_i - E_i)^2}{E}.$$

He has proved that the chi-square statistic has a probability distribution that is independent of the data used (Pearson, 1900). In other words, he is able to tabulate the probability distribution of this statistic and uses this set of tables for any test. This probability density distribution depends on a single parameter called 'degrees of freedom'. The following figure shows, for example, the probability density distributions of χ^2, for various values of degrees of freedom (v):

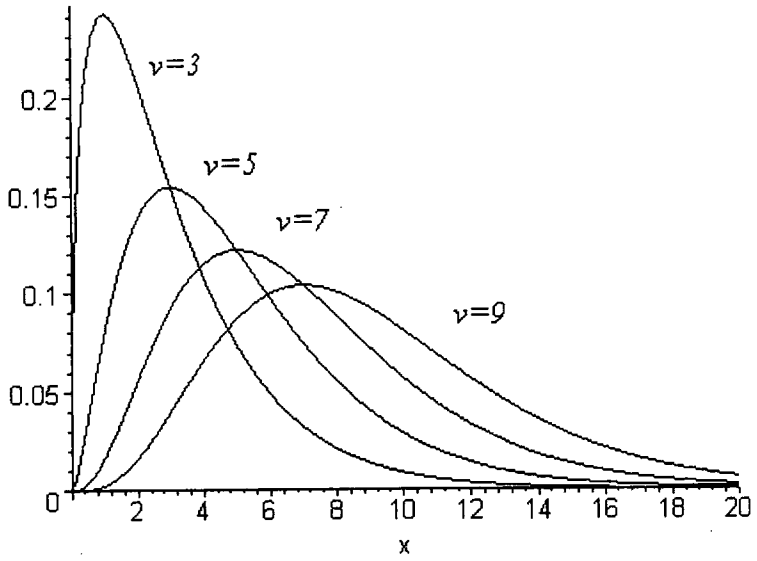


Figure 1 The probability density distributions of χ^2, for various values of degrees of freedom (v)

The null hypothesis is rejected if the χ^2_{obs} is greater than the χ^2_{α} (the χ^2 found from its probability density distribution for the given degree of freedom at the rejection

level α) [42]. With the advent of the chi-square test statistic, the level of rejections began to be standardized. According to Pearson (1900), 0.1 level was regarded as 'not very improbable' and 0.01 level as 'very improbable'. William Gosset, who used the pen name 'Student'[43] to publish the paper 'The probable error of the mean' in which the now famous t distribution for small samples is developed, alleged that a level of three times the probable error 'for most purposes would be considered significant' (Student, 1908, p.13). Two years after the publication of Student's (1908) article, Wood and Stratton recommended agricultural researchers to take "30 to 1 as the lowest odds which can be accepted as giving practical certainty that a difference is significant' (1910, p.433). Thirty to one odds corresponds to 1/31, i.e. $p = 0.032281$ or a mean difference of 3.17 probable errors. Other similar standards have been proposed. For instance, three times the probable error was suggested to be the accepted standard for the undoubted significance of an obtained difference between averages (McGaught, 1924). Three times the probable error is equal to 2.0235 times the standard deviations or $p = 0.043025$. Ronald A. Fisher is probably the first person who mentioned explicitly the use of p =.05 level as determining statistical significance:

> The value of which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (Fisher, 1973a, 44)[44]

[42] For example, if ν =5, α = 0.05, χ^2_α is about 11.071 which can be obtained from the table or software.
[43] For the history why Gosset used pen name in publishing papers, see, for example, Salsburg, 2001, pp.27-28.
[44] This paragraph also appeared in the first edition of the same book which published in 1925. Some have, however, regarded that the choice of the .05 was quite arbitrary and casual. For example, Cochran has expressed his view that 'Fisher sounds fairly casual about the choice of 5% for the significance level, as the words 'convenient' and 'prefers' have indicated' (p.17). We will return to this point later.

Cowles and Davis (1982) have argued that by the time Fisher published his first book on statistical methods, three times the probable error was a commonly adopted convention for determining statistical significance in different fields of scientific studies that employed statistical test[45]. It is thus suggested that Fisher rounded these figures to 0.05 to express the significance level in terms of standard deviation instead of probable error especially when there was a need to provide a broader base for general understanding by those who did not have sufficient training in statistics during the time when statistical analysis began to extend to the social sciences (Cowles & Davis, 1982, p.557). There were, however, a number of problems behind this method of SST and they will be discussed later.

2.3 Fisher's significance testing vs. Neyman-Pearson hypothesis testing

As we have seen in previous sections, SST can date back to the eighteenth century. But SST can hardly be said to be prevailing amongst researchers in different fields until the first half of the twentieth century during which the two rival approaches to SST were put forward by Ronald A. Fisher and by Jerzy Neyman in collaboration with Egon Pearson[46]. (Howson & Urbach, 1993; Lehmann, 1993).

[45] There were certainly researchers who did not accept the use of .05 level. For instance, in 1917 J.E. Coover expressed clearly that he would not accept a p value of .00476 as a decisive indication of some cause beyond chance (Coover, 1975, p.82).

[46] There were pronounced philosophical and methodological differences between these two approaches. Debate between these schools was thus inevitable. According to Lehmann (1993), the debate was mainly carried out by Fisher and Neyman as the collaboration between Neyman and Pearson had stopped before Neyman participated in the confrontations with Fisher.

In the first half of the twentieth century there was a tendency for the methodology of experimental procedure to shift from single-subject research that focused on experimental control and the a priori minimization of error to a focus on treatment group experimentation with comparison of aggregate means and the measurement of error after the fact (Danziger, 1990). This tendency provided an environment in which SST flourished. It was in this environment that Fisher (1925/1973a, 1935/1971) published his small-sample statistical procedures which in turn had an enormous impact on the research methodologies in the early twentieth century. The focus of Fisher's approach is to attempt to challenge and reject a null hypothesis of interest[47] (e.g., a new teaching method does not lead to a difference in learning outcomes) in a study. According to the null hypothesis, the observed data (e.g., difference in learning outcomes) are merely a result of random sampling. Given the truth of this null hypothesis, either the observed and more extreme data are very unlikely to happen (e.g., the probability is smaller than 0.05), which will lead to a rejection[48] of the null hypothesis, or not very unlikely to happen (say, not smaller than 0.05), which fails to provide sufficient evident for us to reject the null hypothesis.[49] The procedure, which is called a significance test, can be summarized as follows:

1. To identify the null hypothesis, H_0, which states that a sample comes from a hypothetical infinite population with a known sampling distribution

[47] Some (for example, Howson & Urbach, 1989; McClure & Suen, 1994) have argued that the focus of Fisher's approach is the research hypothesis of interest (e.g. a new teaching method leads to a difference in learning outcomes) instead of the null hypothesis. We will see later that it is not an appropriate interpretation of Fisher's approach.

[48] According to Fisher (1956/1973b, pp.45-47), rejection of a hypothesis involves both a policy decision to treat the hypothesis as if were false and the formation of an attitude of incredulity towards it. And regarding a rejected hypothesis as objectively incredible is not taking an irreversible decision, i.e., we should be prepared to be 'convinced by future evidence that appearances were deceptive and that a very remarkable and exceptional coincidence has taken place' (p.35).

[49] According to McClure and Suen (1994) the first case will constitute evidence for the corresponding research hypothesis and the second case cast doubt upon the research hypothesis. We will return to this point later .

2. To determine the appropriate test statistic[50] and figure out its distribution under the assumption that H_0 is true.

3. To calculate the value of the chosen test statistic from the observed data.

4. To calculate the probability, given the truth of H_0, of the statistic taking a value as or more extreme than the one calculated in step 3 (i.e., the p value).

5. To specify a significance level[51] α. When $p \leq \alpha$, the result is said to be significant at the α level and H_0 is said to be rejected at the α level.

The rationale for the significance testing is that if H_0 is true the probability of an outcome or more extreme result is small (say, less than 0.05). A logical consequence is the simple disjunction: 'Either the hypothesis is not true, or an exceptionally rare chance has occurred' (Fisher, 1960, p.8[52]). The outcome has occurred, thus discrediting the hypothesis.[53] According to Fisher, the smaller is the value of p, the greater inductive evidence against H_0 will be. That is to say, p value is a measure of evidence against the null hypothesis (Berger, 2002; Johnstone, 1986, 1987; Spielman, 1974). For example, suppose we suspect that a particular die is unfair in the sense that not all six outcomes in throwing the die are equally likely and the outcomes seem to favour the throwing of '6' rather than that of other numbers. If we get 9 '6's in 10 throws then, according to our

[50] See Cox & Hinkley, 1974, p.66 for a more formal definition of test statistic.

[51] Fisher (1935/1971) had suggested that, 'It is usual and convenient for experimenters to take 5 per cent, as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of fluctuations which chance causes have introduced into their experimental results' (p.13). But this suggestion has no implication that Fisher would support the 0.05 level so much that he did not willing to give up. Indeed, he had explicitly said, 'If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. Point), or one in a hundred (the 1 per cent. Point) (1926, p.504).

[52] Similar disjunction can be found in Fisher (1956/1973b): 'Either an exceptionally rare chance has occurred, or the theory of random distribution is not true' (p.42).

[53] Whether the logic behind the reasoning is fallacious will be discussed in Chapter 5.

intuition, our conjecture seems to be true. On the contrary, the conjecture seems to be false if we get only 2 '6's in 10 throws. How about 4 or 5 '6's in 10 throws? Our intuition is not always reliable in such intermediate cases. Fisher has thus promoted a more reliable and objective method for drawing inductive inference from the particular to the general, or from samples to populations – using probability values rather than 'eyeballing' the data.

In this example, we have to pose a null hypothesis H_0 which asserts that the probability of getting a '6' = 1/6 and the probability of getting a number other than '6' is 5/6. And we are going to see if we are able to refute it.[54] To test the hypothesis, the die is thrown a predetermined number of times, say 10, and we record the result. We then specify the results which the experiment could have produced. In this case, we choose the number of '6's obtained as the test statistic.[55] Then it is clear that the test statistic can take any integral values from 0 to 10. Assuming that H_0 is true, the probability of getting exactly r '6's when the die is thrown 10 times is $C_r^{10}\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{10-r}$. The probabilities are listed in the Table 1 and displayed graphically in the Figure 2:

[54] Our null hypothesis here should not be confused with the hypothesis that the die is fair (i.e. the probability of getting a 'n' in throwing the die = 1/6, where n = 1, 2, 3, ..., 6).

[55] Some may suggest that we should use the chi-square goodness-of-fit test here. But if our discussion here focuses only on the two outcomes: getting the number '6' and getting numbers other than '6', then the statistic adopt here still works. There are, of course, various random variables that can be regarded as test statistic, see, for example, Howson & Urbach, 1993. We will return to this point later.

Number of '6's (r)	Probability
0	0.16150558
1	0.32301117
2	0.29071005
3	0.15504536
4	0.05426588
5	0.01302381
6	0.00217064
7	0.00024807
8	0.00001861
9	0.00000083
10	0.00000002
	1.00000000 (TL)

Table 1 The probability of getting r '6's when the die is thrown 10 times

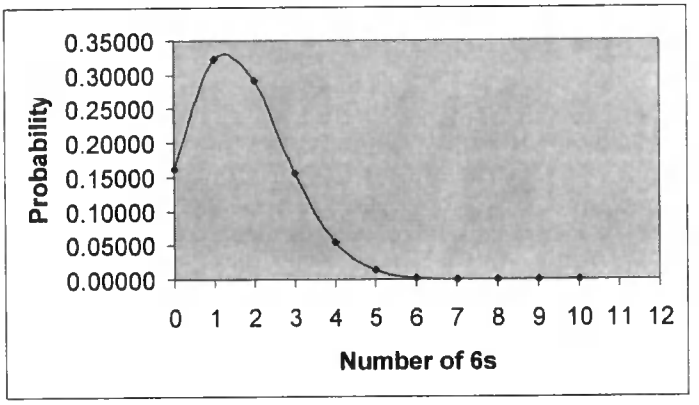


Figure 2 The probability of getting r '6's when the die is thrown 10 times

For the sake of illustration, suppose the die, after being thrown ten times, produced one '1', two '2's, one '3', one '4', and five '6's. From the table, we see that getting five '6''s occurs with probability less than 0.013 under the null hypothesis. The probability that the experiment would throw up any one of the results which have less or

equal probability to 0.013 is $p = 0.015$.[56] Since $p < 0.05$, the null hypothesis of this example is thus rejected at the 0.05 level.

In this simple example, the (binomial) distribution of the test statistic under the null hypothesis can be easily calculated (at least theoretically). But in practice, the situation is different. Consider a hypothetical but more realistic example, suppose we want to know whether male and female students have the same mean math score in the same examination. But it is impracticable for us to get all students' math score and we are only able to select a random sample of students from the two genders and get each student's math score. The null hypothesis in this case merely asserts that their means are the same but it does not specify the distributions of math score in the two populations. The statistical test cannot be performed with test statistics whose probability distributions under the null hypothesis are not known. What we can do is to make further assumptions. For instance, if the populations are normal, the *t*-statistic, developed by Student, implied by the data collected can be calculated.[57] We can establish, from the known probability distribution of *t*, how probable it is for a result expressed as a value of the *t*-statistic to happen. The procedure then goes as that of our previous case: if the probability of our result collected lies below our previously designated significance level, then the null hypothesis is rejected at that corresponding level.

We can see from these examples that the Fisher's approach has not made any comment about the acceptance of any hypothesis. In our first example, if the null hypothesis is rejected, we cannot decide which rival hypotheses should be accepted. If

[56] $p = 0.013 + 0.002 + 0.000 + 0.000 + 0.000 + 0.000 = 0.015$

[57] For detail of the calculations, see, for example, Hopkins, Hopkins, and Glass, 1996.

the null hypothesis is not rejected, the study is clearly inconclusive simply because sampling fluctuation still remains a viable rival explanation for the observed result. Indeed Fisher (1956/1973b) has alleged explicitly, 'A test of significance contains no criterion for "accepting" a hypothesis' (p.45).

In Fisher's approach, the region of low probability is concentrated in one or both of the tails of the probability distributions of outcomes. According to Howson and Urbach (1993), if we are merely interested in judging if the outcome of a test fell in a region of low probability, there seems to be no reason to prevent us from, say, choosing a narrow region in the centre of the bell-shaped distribution. Moreover, there are always a number of random variables that may be defined on any given outcome space. Fisher's approach leaves open which test statistic should be used. And not all random variables will lead to the same conclusion when they are put to use as the test statistic in a significance test. For example, suppose the statistic of the example of throwing a die is modified so that all the outcomes exhibiting zero, one '6's and two '6's, three, four and five '6's, six, seven and eight '6's, nine '6's and ten '6's are combined into four groups. The new groups may be arbitrarily renumbered from one to four, thus giving the following distribution of the new statistic *t'*.

Value of Statistic (*t'*)	Probability
1 (zero,one, or two '6's)	0.77522680
2 (three, four, or five '6's)	0.22233505
3 (six, seven, or eight '6's)	0.00243731
4 (nine or ten '6's)	0.00000084
	1.00000000 (TL)

Table 2 The probability distribution of the *t'* statistic

It is obvious that, with this new statistic, a result of getting five '6's is no longer significant at the level of 0.05 (the new p^* value is 0.222). Some may argue that the grouping of outcomes appears to be arbitrary. Unlike the statistic r (in our previous example) whose value '1' has a clear natural meaning in sentences like 'r = 1' (i.e. the number of '6's is 1), the statistic t' is *artificial* in the sense that the value of t' in sentences like 't' = 1' is void of natural meaning. But the notions of 'artificiality' and 'lack of natural meaning' are not precise enough for us to regard t' as an illegitimate statistic. As a result, we have to impose some more precise restriction on test statistic so as to give consistent result (Howson and Urbah, 1993).

In order to resolve these problems that undermine Fisher's approach[58], Jerzy Neyman and Egon Pearson (1928a, 1928b, 1933) proposed an alternative approach to SST – hypothesis testing, a term that they employed to distinguish it from Fisher's 'significance testing' (Howson and Urbach, 1989; Gill, 1999; Hubbard, 2003; Hubbard and Bayarri, 2003a). According to this Neyman-Pearson approach, the statistical hypothesis is tested relative to one or more other statistical hypotheses (they need not be labeled as 'null' or 'alternative'):

> ... in addition to H there must exist some other hypotheses, one of which may conceivably be true. Here, then, we come to the concept of the 'set of all admissible hypotheses' which is frequently denoted by the letter Ω. Naturally, Ω must contain H. Let $\overline{H}$ denote the complement, say $\Omega - H = \overline{H}$. It will be noticed that when speaking of a test of the hypothesis H, we really speak of its test

[58] Some of the problems will be discussed in later chapters.

'against the alternative $\overline{H}$.' This is quite important. The fact is that, unless the alternative $\overline{H}$ is specified, the problem of an optimal test of H is indeterminate. (Neyman, 1977, p.104)

Based on considerations regarding the decision criterion, sample size and effect size, this approach introduces the probabilities of committing two kinds of errors. The first type of error (Type I) is the error made when the null hypothesis (H_0) is rejected by the hypothesis testing if H_0 is true, whereas the second type of error (Type II) is the error made when the hypothesis testing fails to reject H_0 when it is false. The problem of hypothesis testing is the problem of selecting critical regions so as to control the rates of Type I and Type II errors. Neyman (1950) maintained that the Type I error is more important to avoid than the Type II error (p.265). Hence, we have to fix arbitrarily a small number α called 'the level of significance' prior to the collection and analysis of the data,[59] and to require that the probability of committing the Type I error does not exceed α. Then the final choice of the test is made so that the probability of accepting the null hypothesis when it is false (i.e. the Type II error) is minimized. The probabilities associated with Type I and Type II errors are usually symbolized by α and β respectively[60] and these two error probabilities will define a critical region for the chosen statistic. If a result falls into the critical region, then the alternative hypothesis is to be accepted and the null hypothesis rejected; on the contrary, the null is to be accepted and the alternative rejected.

[59] The reason that we have to fix the value of α prior to the collection and analysis of data will be discussed in Chapter 4.

[60] The same symbol 'α' is used to denote both the significance level in Fisher's significance testing and the probability associated with Type I error in Neyman-Pearson hypothesis testing. That may explain, at least partly, why many people blur the distinction between these two notions. We will return to this point in Chapter 4.

Consider our die again, suppose one has claimed that he had loaded the die with lead so that the probability of getting a '6' would be 1/2 and that of getting a number other than '6' would be 1/2 too. We shall label this hypothesis as H_1 and the former hypothesis (which asserts that the probability of getting a '6' = 1/6 and the probability of getting a number other than '6' is 5/6) as H_0. Using the number of '6's obtained as the test statistic, we get the following table and the corresponding graph:

Number of '6's	Probability (H_0)	Probability (H_1) [61]
0	0.16150558	0.00097656
1	0.32301117	0.00976563
2	0.29071005	0.04394531
3	0.15504536	0.11718750
4	0.05426588	0.20507813
5	0.01302381	0.24609375
6	0.00217064	0.20507813
7	0.00024807	0.11718750
8	0.00001861	0.04394531
9	0.00000083	0.00976563
10	0.00000002	0.00097656
	1.00000000 (TL)	1.00000000 (TL)

Table 3 The probability of getting r '6's when the die is thrown 10 times under two hypotheses

[61] Assuming that H_1 is true, the probability of getting exactly r '6's when the die is thrown ten times is given by : $C_r^{10}\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{10-r} = C_r^{10}\left(\frac{1}{2}\right)^{10}$

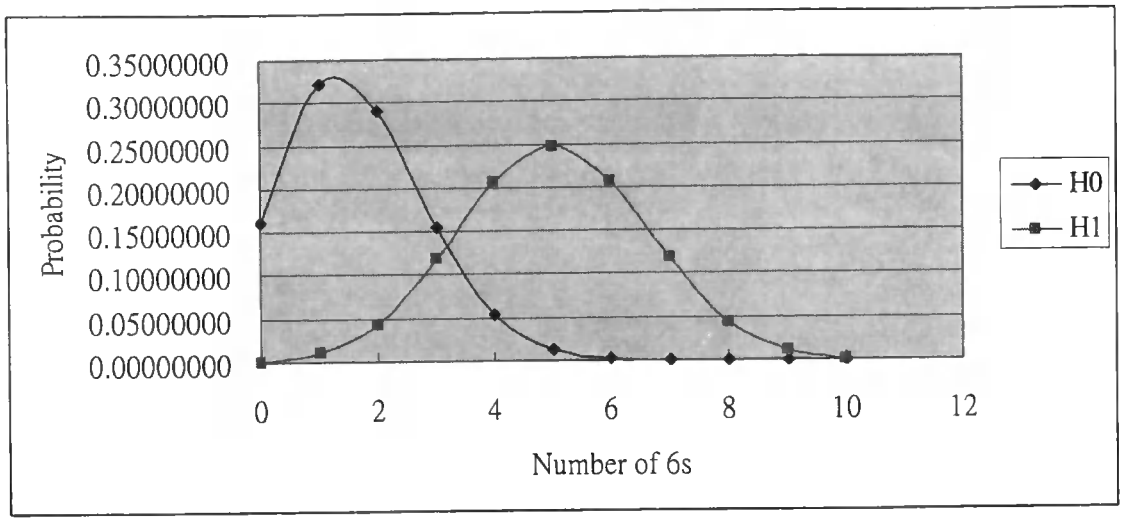


Figure 3 The probability of getting r '6's when the die is thrown 10 times under two hypotheses

If we accept the following rule: when the die is thrown ten times,

$$\begin{cases} H_0 \text{ is rejected if five or more '6's are obtained} \\ H_0 \text{ is accepted if four or less '6's are obtained} \end{cases}$$

we would reject H_0 given our previous outcome (i.e. getting five '6's). With this rule, the probability of rejecting H_0 when H_0 is true is 0.015, i.e. α = 0.015, whereas the probability of accepting H_0 when H_1 is true is 0.377.[62] If we assume further that exactly one of the two hypotheses must be true, then 0.337 is also the probability of accepting H_0 when H_0 is false, i.e. β = 0.337. The power of the test, which is regarded as a measure of the degree to which the test discriminates between the two hypotheses, is 1 – β. In this case, the power is 0.663, which is also the probability of rejecting H_0 when H_0 is false.

[62] 0.205 + 0.117 + 0.044 + 0.010 + 0.001 = 0.377

In the Neyman-Pearson approach, as long as the space of outcomes and its associated probability distribution are known, the critical region for a particular hypothesis, say H_0, and a rival hypothesis H_1 is uniquely determined for each significance level. It thus explains why we regard some particular region but not other as critical, and such a reason is lacking in the Fisher's approach (Howson and Urbach, 1993).

The distinctions between Fisher's approach and Neyman-Pearson approach are not made consistently clear in modern statistical writing and teaching and blurring the distinctions between measures of evidence (p value) in Fisher's significance testing and Type I error probability (α) in Neyman-Pearson hypothesis testing have been extensively reported (Hubbard, 2003). According to Royall's (1997) observation, textbooks on mathematical statistics tend to adopt Neyman-Pearson approach, while textbooks on statistical methodology tend to lean more towards Fisher's approach. The jargons are also not standard, and the same terms and symbols are often adopted in both approaches, thus blurring the differences between the two approaches. In fact, many researchers have implicitly employed some hybrid of these two approaches in their research studies (Thompson, 1996), and these two approaches have eventually evolved into the practice of the modern form of SST (Harlow, 1997). According to Huberty's (1993) study, in the 1930s, writers of statistics textbooks began to refer to Fisher's approach. The first book that discussed the two types of errors appeared in 1940. Later, in the 1950s, the two approaches began to be unified in textbooks but it did not gain any endorsement from any of the originators. The unified theory was accepted by the 1960s in a number of academic disciplines. According to Nix and Barnette (1998), SST had become the unified form that become so ubiquitous that over 90% of articles in major psychology

journals adopted SST as justifications for drawing their conclusions from the data collected.

Contrary to Fisher's approach, according to which no explicit complementary hypothesis to the null hypothesis is identified, Neyman-Pearson approach comprises two rival hypotheses, one is usually labelled as the null hypothesis (H_0)[63]. According to Fisher's hypothesis testing, the p value resulting from the model and the data is figured out as the strength of the evidence for rejecting the null hypothesis and there is no place for the notion of power of the test nor that of accepting alternate hypotheses. On the contrary, Neyman-Pearson hypothesis testing allows us to accept one hypothesis if another one is rejected on a predetermined α level and has a precise notion of power of the test. To some researchers, SST partially uses the Neyman-Pearson decision process but regards failing to reject the null hypothesis as a modest support for the null hypothesis (Gill, 1999). Fisher's approach defines the significance level afterwards as a function of data whereas Neyman and Pearson approach defines the significance level before getting the data. Some advocates of SST straddle these two approaches. It appears that the significance level is selected a priori, but in fact p values are sometimes used to evaluate the strength of evidence. Different researchers may give different accounts of the form of SST. As a consequence, there is a family of akin but different procedures all under the name of 'SST' (Erwin, 1998). Gigerenzer (1993) even argued that much of the controversy over SST is due to its hybridism and different interpretations. These points

[63] There are at least two types of null hypothesis. One is the point-null hypothesis (or the simple null hypothesis) which specify the same unique value for the population statistic or states that there is no difference between the two population statistics (e.g. $H_0 : \mu_E = \mu_C$). Another is the directional null hypothesis (or the composite null hypothesis) which specifies a different range of possible values for the population statistic or states an inequality between two population statistics. (e.g. $H_0 : \mu_E \geq \mu_C$). (Liu and Stone, 1999; Chow, 1996)

will become clearer when we discuss the distinctions between these two approaches in Chapter 4.

Chapter 3 Interpretations of probability

From the contemporary mathematical point of view, probability is not difficult to define – it is a countably additive set function on a σ-field, normalized with a total mass of one. This definition was first introduced by Andrei Nikolawvich Kolmogorov in 1933 (Kolmogorov, 1956).[64] There is not much controversy over the axiomatic system of the probability calculus[65]. Contemporary mathematicians, with this axiomatic system of probability theory, can state theorems with clarity and prove them with rigor[66]. To them, the numerical values of probabilities and the meanings attached to the primitive terms in its axioms or theorems are almost of no particular significance. But to researchers in practical fields, such as social sciences, the numerical values of probabilities and its meaning are what they strive for and merely an axiomatic presentation of probabilities is not useful. In the Section 2.2, we have discussed the prior probability of a hypothesis $P(H)$. For a specific hypothesis, it is either true or false. What do we mean by saying

[64] Here is a brief introduction of Kolmogrov's axiomatic system:
A probability space is a triple (Ω, F, P) where Ω is an arbitrary non-empty set, F is a σ-field of subsets of Ω (A collection of subsets of Ω is called a σ-field if it contains Ω itself and it is closed under complements [i.e., if $E \in F$, then $E^c \in F$] and closed under countable unions [i.e. if $E_i \in F$ for $i = 1, 2, \ldots$, then $\cup E_i \in F$]), and P is a measure on F (i.e. P is a a non-negative, real-valued function on Ω, such that P is countably additive: if $E_i \in F$ for $i = 1, 2, \ldots$, and the sets are pairwise disjoint, then $P(\cup E_i) = \sum P(E_i)$) such that $P(\Omega) = 1$. P is called probability measure or shortly probability.
A modern treatment could be found in elsewhere, for instance, Capiński and Kopp (1999, p.46) and Kolmogorov's original treatment could be found in Kolmogorov (1956, p.2). For a discussion of Kolmogorov's development of the measure theoretic probabilities in the twentieth century, see von Plato (1994, pp198-237).
[65] A main controversy over the Kolmogorov probability arises from the foundation of quantum mechanics. For instance, negative probabilities are postulated in almost all quantum models but only non-negative probabilities are allowed in Kolmogorov's probability theory. The controversy is beyond the scope of this thesis. For the discussion of this issue, see Atkinson, 1998; Brody, 1993; Khrennikov, 1999. There is also an alternative set of axioms for probabilitiy proposed by Fetzer (1981), but this non-standard probability theory is not consistent with the enormous body of theorems based on Kolmogorov axioms and is thus not well received by the mathematical community.
[66] Measure theory is necessary to be introduced if we want to study probability theory in a rigorous manner. For the explanations why it is needed, see, Rosenthal, 2000, pp.1-5.

that its prior probability is, for instance, 0.05? What objects in the world correspond to probabilities?

If we try to think about the precise meaning of the word 'probability', we may find that the word 'probability' is ambiguous. For instance, when we say that:

(3.1) The probability of getting a '1' in throwing the die is 2/3,

we are talking about the die. The sentence (3.1) is either true or false, regardless of our knowledge about the die. If it is true, it is because of how the objective world is (including how the die is loaded and probably how we throw the die, etc[67]). It is also possible for us to test (3.1). We could throw the die[68] for a number of times and observe the relative frequencies of the occurrences of '1'. If we get 200 '1's out of 600 throws, we will probably admit the truth of (3.1). But if we get 10 '1's out of 600 throws, we will be very dubious about (3.1).[69]

On the other hand, when we say that:

(3.2) Based on all the available evidence, the probability that the source of the international outbreak of Severe Acute Respiratory Syndrome (SARS) in March 2003 is a doctor from Guangzhou is 2/3 [70],

are we talking about the source of the international outbreak of SARS? The following sentence:

---

[67] The conditions can hardly be fully stated here. And we do not intend to delve into the distinctions between 'tossing' and 'throwing' here although it has been argued that there are much more problems arising from tossing or flipping a coin. For instance, Gelman and Nolan (2002) have argued that we can hardly bias a coin especially when we totally ignore how the coin is flipped or tossed.

[68] Of course we are assuming that the way we throw the die is the same (or almost the same) for each trial.

[69] We, however, will not say we could prove or falsify the probability statement, like (3.1). See Chapter 6 for further discussion.

[70] http://www.biomedcentral.com/news/20030320/09

(3.3) The source of the international outbreak of Severe Acute Respiratory Syndrome (SARS) in March 2003 is a doctor from Guangzhou,

is either true or false, regardless of our knowledge about the source of the international outbreak of SARS. Its truth depends solely on how the world is, especially what happened in March 2003. But if (3.2) is true, it is not merely because of how the world is, but also because of how well the available evidence supports (3.3). The truth of (3.2) depends also on the relation between the available evidence and the truth of (3.3). Unlike (3.1), (3.2) can no longer be tested by experiments since it seems to be nonsense to do repeated trials on the source of a particular outbreak of a particular disease at a particular time. It seems to suggest that the word 'probability' in (3.1) and (3.2) are of different meanings.

Could we a have a single theory that can explicate all of our ordinary usages of the word 'probability' as well as regulate it? Or do we need pluralist views of probability to accommodate different interpretations each of which is valid in a particular context? There are many views of probability to be found in the literature and they can be divided for our present purposes into five major categories.[71] We will discuss briefly the major tenet of each interpretation and leave the lengthy discussion of the objections and criticisms to the appendices so that we could discuss in turn the implications of the interpretations of probability for the analysis of SST.

[71] There are many other different classifications of theories of probabilities, see Weatherford (1982, pp.11-17) for a brief description.

3.1 The classical theory

The classical theory of probability was first introduced by the thinkers of the European Enlightenment such as James Bernoulli and Pierre Simon de Laplace. One of the Enlightenment's characteristic ideas is admiration for Newtonian mechanics of which a consequent is the belief in universal determinism. Universal determinism is the thesis that in our universe the past determines a unique future. Of course, many people could imagine or consistently describe that there are many ways in which the world might go on. But many of these conceivable futures might be physically impossible. For instance, given the laws of nature and the past[72], it is deterministic that either there must or must not be a lunar eclipse on 30 March 2003. Even though I could picture to myself what it would be like for there to be a lunar eclipse tonight, this imaginable future is in fact physically impossible.

In a deterministic universe (or system), probabilities are not inherent in nature and must be relative to human ignorance. For example, suppose in a certain circumstance, there are only n possible outcomes. According to universal determinism, one of these n outcomes (say, outcome O_1) must occur. If some persons know all relevant laws of nature and the actual past (or sufficient initial conditions), they could be able to predict the occurrence of O_1.[73] If we, as normal humans, do not know enough about the laws of nature or the initial conditions in a sufficient degree of precision[74], we may not be able to

[72] The laws of nature per se do not dictate when particular events like eclipses will occur. For example, given the Newton's laws of motion but without given sufficient initial conditions, it is possible that no unique consequence can be deduced.

[73] Such persons have come to be known as Laplace's demons. See Laplace (1825/1994, p.2).

[74] In some systems (not necessarily chaotic systems), a minuscule change in initial conditions can get greatly amplified by later events (Smith, 1998, p.1) So even if we know enough about the laws in these systems, we can still get tremendous error in predictions if we do not know the sufficient details of the initial conditions.

predict which of the *n* outcomes will occur. In this case, according to the classical theorists, we should have recourse to probabilities. As Laplace (1925/1994) put it, 'probability is relative in part to this ignorance and in part to our knowledge. Suppose we know that, of three or more events, one alone must occur, but that nothing leads us to believe that one of them will happen rather than the others. In this state of indecision, it is impossible for us to say anything with certainty about their occurrence. However, it is probable that one of these events chosen at will (or at random), will not occur, because there are several equally possible cases that exclude its occurrence, while only a single one favours it.' (pp.3-4).

In other words, according to the classical theory of probability, there is no objective chance or indeterminism. Probability is only a measure of our partial ignorance. If there are *n* possible outcomes in a particular situation, it is deterministic that only one of them must occur. Suppose we, unlike a Laplace's demon who perfectly knows all laws of nature and initial conditions so that he knows which outcome must occur, do not have any reason or evidence to expect that one of the *n* outcomes will occur rather than the others, there is a principle called 'the Principle of Indifference'[75] according to which we have to regard the *n* outcomes as equally possible. In these circumstances, although there is no genuine objective chance, we do have objective rules of assigning or generating probabilities and combining probabilities. The rule of assigning probabilities is as follows: suppose out of these *n* possible outcomes there are *m* of them are

[75] This is Keynes's (1921, p.41) terminology. Sometimes it is also called the 'Principle of Insufficient Reason' (first suggested by J. von Kries in 1886) (Keynes, 1921, pp.41-42; Howson & Urbach, 1993, p.52). In Appendix 8, we will expound this principle in detail.

favourable to the outcome O, then the probability of O is defined to be m / n.[76] This is the classical definition of probability based on equally possible cases. There are, however, a number of objections or criticisms that make the classical theory of probability discredited. For example, the term 'equiprobable' is required in the definition of probability and that renders the interpretation circular. Moreover, the classical theory yields a number of intractable inconsistencies or paradoxes and is unfit for explicating the use of probability in common usage.[77] In the next section, we will consider the logical theory of probability, which is the first theory of probability emerging in the twentieth century and the one very similar to the classical theory.

3.2 The logical theory

The logical theory[78] was developed in the early twentieth century by John M. Keynes, Rudolf Carnap[79] , Karl Popper and later by Harold Jeffreys (Gillies, 2000, p.25). Logical theory has three major tenets: First, probability is a logical relation between sentences or propositions. Second, the relation is completely determinable by the application of logic and the rules of probability to the two sentences. It follows that probability is not known by empirical means. Third, ascriptions of probability without being relative to certain evidence are either elliptical or nonsense.

[76] The rules of combining probabilities (or the probability calculus), such as the addition rule for mutually exclusive events, are basically mathematics in nature. They are almost the same for all different theories of probability and there is no need for us to discuss in depth here.

[77] See Appendix 8 for details.

[78] Logical theory of probability is sometimes called 'A priori theory of probability'. Some may consider logical theories as a sub-class of a priori theories and Keynes's theory should be regarded as a version of a priori theories instead of a logical theory (for instance, Weatherford, 1982, p.76), but these terms are not as precise as terms like Keynes's logical theory and Carnap's logical theory, which will be used in our following discussion.

[79] Carnap has indeed contended that there are two probability concepts: Probability$_1$ and Probability$_2$ (1945, pp.521-525; 1962, pp.19-51). It is his Probability$_1$ concept that corresponds to our logical theory discussed in this section.

Same as classical theorists, logical theorists regard all propositions as either true or false. Speaking of propositions as certain or probable is not because we think it so nor it is a characteristic of the propositions in themselves, rather it is merely an expression of relationship in which they stand to a *corpus* of knowledge, actual or hypothetical (Keynes, 1921, pp. 3-4). In other words, when it is said that a proposition C is probable, what is being asserted is that the proposition C bears a certain relation to another proposition, or a set of propositions, which may also be described as confirming, or supporting, or providing evidence for C (Ayer, 1973, p.188). In the case of deductive logic, the relation between the proposition (conclusion) and a set of propositions (premises) of a valid argument is *necessary* by which it means that if the premises are all true then the conclusion cannot be false. For example, consider the following argument:

(P1) All students who know how to multiply numbers will know how to add numbers;

(P2) Raymond is a student who knows how to multiply numbers;

Therefore, (C) Raymond knows how to add numbers.

The relation between the sentences (P1), (P2) and (C) is necessary – if (P1) and (P2) are true then it is certain that (C) is true. In other words, (C) is entailed by (or follows logically from) (P1) and (P2). But for induction, the relation between the conclusion and the premises is no longer necessary. For example, consider the following argument:

(P1') Six hundred students who know how to multiply numbers are observed and they all know how to add numbers;

(P2) Raymond is a student who knows how to multiply numbers;

Therefore, (C) Raymond knows how to add numbers.

In this case, even though (P1') and (P2) are true, (C) is possibly false. In other words, (P1') and (P2) do not entail (C). But we would tend to say that (P1') and (P2) do *partially* entail (C) as (P1') and (P2) certainly give some support for (C). If (P1') changes to:

(P1'') Six *million* students who know how to multiply numbers are observed and they all know how to add numbers;

the support for (C) seems to be greater. In other words, the degree of partial entailment from (P1') and (P2) to (C) is less than that from (P1'') and (P2) to (C), which is in turn less than that from (P1) and (P2) to (C). And probability is the degree of partial entailment (Keynes, 1921, 5-6). In Keynes's terminology, (C)/(P1)&(P2) > (C)/(P1'')& (P2) > (C)/(P1')&(P2)[80]. However, it is worth noting that in the case of partial entailment it should be the content of the evidence instead of the numbers of students in (P1') and (P1'') that matter to the support for (C). For example, consider:

(P3) Raymond studied under the same curriculum with the 6 hundred students in (P1')

(P3') Raymond studied under a curriculum that is different from that studied by the 6 million students in (P1'')

Then it could be likely that

(C)/(P1') & (P2) & (P3) > (C)/(P1'') & (P2) & (P3')

even though the number involved in (P1'') is much greater than that in (P1').

[80] According to Keynes (1921), the value of the symbol a/h, if it exists, represents the conditional probability of a with reference to the evidence of h (p.40).

One immediate consequence of this approach is that probabilities are all relative to given evidence[81]. At first sight, it seems that this interpretation of probabilities is not consistent with our ordinary use of the probability concept, as we often speak simply of the probability of some outcome. A standard reply is that it is because in making probability statements we seldom bother to specify the evidence on which we are relying and our probability statements are commonly elliptical (Keynes, 1921; Kneale, 1949). To logical theorists, ascriptions of probability without being relative to certain evidence are thus either elliptical or nonsense.

If probability is interpreted as the logical relation of propositions, how could we determine the numerical values of probabilities? To Keynes (1921), not all probabilities are measurable or they must have a numerical value. He contended further that some pairs of probabilities may not even be comparable (p34). Our knowledge of the logical relations between propositions, according to Keynes (1921), is based on our direct acquaintance with logical relations[82], which are a fundamental source of our knowledge, directly available to our intuition, and neither can nor should be referred to anything else as their source or justification (pp.12-14). The set of probabilities is thus not linearly ordered. In Keynes's own words:

[81] Formulation of probability with reference to the evidence is not exactly the same as the formulation in the form of a conditional clause. See Carnap (1962, pp.32-33) for a thorough discussion on this point.

[82] Keynes has adopted Russell's (1912) position on knowledge – some of our knowledge is obtained directly or 'by acquaintance' (such as those based on our immediate sense perception, for example, I know by acquaintance that the cover of the book in front me is black in colour) – which is called 'immediate knowledge', and another kind of knowledge is the general knowledge which is conjoined from immediate knowledge and a priori knowledge of the truths of logic and mathematics (e.g. our knowledge that Everest is the earth's highest mountain) – which is called 'knowledge by description.' See Grayling, 1996, pp.39-47 for a brief introduction of Russell's work on theory of knowledge.

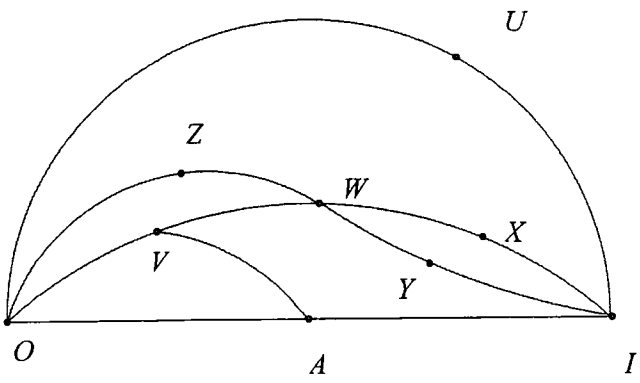


Figure 4 Keynes's conception of probabilities (Keynes, 1921, p.39)

'*O* represents impossibility, *I* certainty, and *A* a numerically measurable probability intermediate between *O* and *I*; *U*, *V*, *W*, *X*, *Y*, *Z* are non-numerical probabilities, of which, however, *V* is less than the numerical probability *A*, and is also less than *W*, *X*, and *Y*. *X* and *Y* are both greater than *W*, and greater than *V*, but are not comparable with one another, or with *A*. *V* and *Z* are both less than *W*, *X*, and *Y*, but are not comparable with one another, *U* is not quantitatively comparable with any of the probabilities *V*, *W*, *X*, *Y*, *Z*. Probabilities which are numerically comparable will all belong to one series, and the path of this series, which we may call the numerical path or strand, will be represented by *OAI*' (p.39).

Keynes's complicated conception of probability has at least two major difficulties. First, the presence of non-numerical probabilities will make the theory deviate from the simplicity and power of the mathematical theory of probability (Gillies, 2000, p.35). Second, for a specific probability, how could we determine whether it has numerical value or not? Keynes's (1921) answer is that numerical measurement is possible only when we can be given a number of equiprobable alternatives (p.41). This is clearly a

resort to the Principle of Indifference. As we have discussed in last section, the application of this Principle to assign initial probabilities has demonstrated to be unsuccessful. As a matter of fact, not all the advocates of logical theory of probability agree with Keynes in regarding probability as an unanalysable logical relation. Carnap is certainly an example.

Carnap's (1945; 1962) logical conception of probability (or called 'probability$_1$' to differentiate the one in terms of observed frequencies, or of the limits towards which they are supposed to tend – probability$_2$) is analyzable – probability$_1$ is a measure of the partial inclusion of the range of one sentence in that of another, which is of a purely logical nature (p.294). With the use of mathematical logic, Carnap is able to refine this conception to make probability$_1$ a rigorous mathematical relation between sentences in a restricted class of simple languages, which is called the Languages *L*. But Carnap's conception still suffers insurmountable difficulties (see Appendix 9). For example, a separation from experience or our empirical world will make logical theory not useful in its application in natural sciences or social sciences. And it is this reason that some have proposed another theory of probability that is supposed to be able to have wider applications in sciences – the frequency theory which is the topic of our next section.

3.3 The frequency theory

The frequency theory of probability was developed largely by Richard von Mises and Hans Reichenbach in the twentieth century. As a major opponent of logical theory, the frequency theory is the view that probability is the empirical relative frequency of

occurrence of some feature of the real world rather than the logical relation between sentences. Consider the following sentence that has been discussed before:

(3.1) The probability of getting a '1' in throwing the die is 2/3

We are, according to the frequency theory, making an empirical sentence about our objective world – the die and the way we throw the die (or if wish, the device for throwing). The probability in (3.1) is used to describe a particular physical property – the relative frequency of '1' in repeating throwing will approach 2/3 in the long run.

Contrary to the logical theory, the probability value in frequency theory is not relative to any evidence and is determined uniquely by the objective world. Furthermore, in frequency theory all probabilities are known a posteriori only, and it thus renders probability theory as one on par with science. As von Mises (1957) elaborated, the concept 'probability' in the frequency theory, like many other concepts in science[83], starts with the imprecise concept of ordinary language. But when we are going to construct a scientific theory it will be replaced by more precise notion. This more precise notion of probability is what the frequency theorists try to introduce by means of explicit definition.

Just like dynamics which is dealing with the motions of bodies and the forces which act on them, probability theory, to the frequency theorists, is dealing with the problems in which either the same event repeats itself again and again, or a great number of uniform elements that are involved at the same time. Contrasting to the subjective theory that will be discussed in the next section, probabilities in the frequency theory are

[83] He has used the concept of 'work' in mechanics as an example to illustrate this point (pp.5-6). In mechanics, work is defined as the scalar product of forces and displacement (or the line-integral of force) and this scientific definition has no associations with the use of 'work' in everyday matters (such as the work performed by the musician).

associated with collections of events or other elements that are considered to be objective and regardless of our beliefs or knowledge, just as masses or bodies in dynamics that are independent of the observers. Single events that do not share this feature will be beyond the scope of probability theory (such as the probability of winning a particular battle would have no place in the frequency theory). Probability theory is thus restricted to special problems as game of chance, social mass phenomena, and statistical treatment of mechanical and physical phenomena (von Mises, 1957, pp.10-11) More precisely, von Mises has proposed the term 'collective' to formalize this restriction of the applicability of probability theory. The term 'collective' is used by von Mises to denote a sequence of uniform events or processes which differ by certain observable attributes such as colours, numbers, or anything else (1957, p.12). For example, all the throws of a die made in the course of a game can form a collective and the number of points thrown can be the attribute; all the molecules in a given volume of gas can form a collective while the velocity of a single molecule can be its attribute; the whole class of students who were sitting for a particular public examination in 2003 is a collective while the grade of each student can be his/her attribute. According to von Mises (1957), the definition of probability is only concerned with 'the probability of encountering a certain attribute in a given collective' (p.12). With an empirical law of probability which states that 'experience has shown that in the game of dice, as in all the other mass phenomena which we have mentioned, the relative frequencies of certain attributes become more and more stable as the number of observations is increased' [84] (von Mises, 1957, p.12), the probability of the attribute considered within the given collective is defined as the limit to which the relative frequency of the observed attribute would tend.

[84] This empirical law is sometimes called 'the Law of Stability of Statistical Frequencies' (Keynes, 1921, p.336).

There are, however, regular cases where the relative frequencies converge towards definite limiting values, but where it is nevertheless inappropriate to talk about probability (von Mises, 1957, p.23). For example, consider a road at the side of which there are a succession of large milestones and small stones between the milestones at intervals of 1/10 of a mile. The attribute large milestone and the attribute small stone will have a limiting frequencies of 1/10 and 9/10 respectively. But this is not a genuine collective since the sequence of results is determined rather than random[85]. Hence, von Mises (1957) has to propose that the limiting values of the relative frequencies of the attributes must remain the same in all partial sequences which may be selected from the original one in an arbitrary way (pp.24-25)[86]. In a nutshell, the probability of an attribute is the limit of the relative frequency with which it appears in a random collective.

Like the classical theory and the logical theory, the frequency theory also suffers some intractable difficulties[87] that lead theorists to propose another objective theory of probability – the propensity theory which will be discussed in Section 3.5.

---

[85] A simple example to illustrate the point that probability and relative frequency are not the same is: suppose a coin tossing device could give 'H' and 'T' occurring in alternative orders. We toss for 28 times and get the sequence: T H T H T H T H T H T H T H T H T H T H T H T H T H T H. What is the probability that the next observation will be 'T'? In this case, the relative frequency of 'T' is 1/2 but we could predict with perfect certainty that a 'T' will occur in the next toss. This also explains why von Mises requires that a collective should have random outcomes as an essential characteristic of a collective.

[86] More precisely, the limiting value of the relative frequency of each attribute in a collective w is the same in any infinite subsequence of w which is determined by a place-selection. 'Place-selection' here means any effectively specifiable method of selecting indices of members of w, such that the decision to select or not the index i is allowed to depend at most on the first $i - 1$ attributes in w. The satisfaction of this condition is warranted by the other empirical law called 'the Law of Excluded Gambling Systems' (von Mises, 1957, pp.25-27; Gillies, 2000, pp.95-96) by which randomness and the failure of gambling systems are related – the authors of gambling systems (for example, a gambling system in roulette is something like: 'bet on black after a run of four consecutive reds') have all, sooner or later, had the sad experience of finding out that no system is able to improve their chances of winning in the long one. Von Mises's has not been able to give a precise definition of randomness or random sequence, and only recently there has been a definition of it that is free of contradiction. For reviews on this problem, see Martin-Löf, 1969; Gillies, 2000, pp.105-109.

[87] See Appendix 10 for details.

3.4 The subjective theory and Bayesianism

The subjective theory of probability, which was first developed independently by Frank P. Ramsey[88] and Bruno de Finetti[89], identifies probability with a person's degree of belief in a proposition. Unlike other interpretations of probability, this theory provides a straightforward interpretation of the probabilities of single events. For instance, when we identify probability as a person's degree of belief, sentences about single event like 'the probability that US will be involved in war with North Korea in the future (say, in 2007)' simply means a person's degree of belief in the outbreak of war between US and North Korea in 2007. Unlike frequency theorists, the subjectivists need not worry about the appropriateness of using the term 'probability' in this way, or to struggle for figuring out what the reference class is. But two immediate questions will be raised: Does it mean that, in the subjective interpretation of probability, we should no longer speak of *the* probability, but rather of David's probability, Mary's probability? If we would like to know the probability of E, we have to measure a person's degree of belief in E. But how could we make such a measurement and compare one's degree of belief with another's? To the first question, many subjectivists (advocates of the subjective theory of probability) will answer yes. There is no need, according to the subjective interpretation, to assume that the degrees of belief will be the same for all rational persons. To the second question, though there is a bold but unrealistic proposal that degrees of belief could be measured by a psycho-galvanometer (Ramsey, 1931, p.161), a betting scheme is regarded by most subjectivists including Ramsey and De Finetti as the most reasonable way to measure degrees of belief. For example, if we want to measure the degree of belief of David in a

[88] The theory was proposed in Ramsey's paper 'Truth and probability' which was first presented in a talk he gave to a philosophy club in Cambridge in 1926 and later published in Ramsey (1931).
[89] De Finetti first published his papers on subjective theory of probability in 1930. For the history of development of subjective theory by De Finetti and Ramsey, see Gillies, 2000, pp.50-51.

proposition E, we have to get David to agree to bet with us on E under the following conditions:

(1) David has to select a number r in the closed interval [0, 1] as the betting rate[90] on E,

(2) After David's selection, we select the stake $\$S$ and choose either one of the following option:

Option 1: David pays us $\$rS$ and he will get $\$S$ from us if E is true and get nothing if E is false.

Option 2: We pay David $\$rS$ and we will get $\$S$ from David if E is true and get nothing if E is false.

Suppose David knows all these conditions and accepts the betting. If he chooses 0.2 as his betting quotient, then 0.2 is taken to be his degree of belief in E. Three remarks must be made here. First, David has to make his selection of betting rate before knowing which option is selected. If he knows that option 1 is selected before selecting the betting rate, he would choose r as small as possible. On the other hand, if option 2 is selected before his selection he would choose r as large as possible. In neither case would r represent David's genuine degree of belief in E. Only under this betting scheme, David would try his best to select the betting rate as fair as possible if he doesn't want his opponents to impose disadvantages on him.

The second remark is about the stake. Although the stakes here are taken to be in money, any other tenable media will do. Indeed our subsequent discussion of the subjective theory will not be much affected if one insists to take the stakes to be in utility,

[90] Betting rate means the bet per stake and it is sometimes called 'betting quotient'.

or in other tenable media. Some worry that, if the bets are to be in money, the magnitude of the stake should not be too large or too small in compared with David's fortune. Large stake will impose an unbearable risk of bankruptcy on him and therefore induce him to select a smaller betting rate. Small stake will probably be a trifle to him and is thus unable to make him consider seriously about the bet. This problem can be resolved if the magnitude of the stake is chosen such that the amount of bet is small enough that it won't impose a risk of financial disaster on the person but large enough to make him ponder over the betting rate (Gillies, 2000).

We have seen how the subjectivists use the betting scheme to measure the degrees of belief. But can any degree of belief be regarded as probability? It seems that a person's degrees of belief could be quite arbitrary, how could we guarantee that the degrees of belief must satisfy Kolmogorov's axiomatic system of the probability calculus? The subjectivists will reply that not any degree of belief or betting rate is admissible. Only those coherent sets of betting rates will be admissible and they will satisfy the axioms of probability. A set of betting rates is said to be coherent if the opponent cannot choose a set of corresponding stakes and options such that the opponent will win whatever happens[91]. For example, suppose David bets on the following two propositions:

E_1: Mary will get a pass in the mathematics examination.

E_2: Mary will *NOT* get a pass in the mathematics examination.

If his betting rates on these two propositions are 3/5 and 3/4 respectively, his opponent could make a Dutch book against David by offering stakes and options as follows:

[91] If, on the contrary, the opponent of a person is able to choose stakes and options so that the opponent will win whatever happens then the opponent is said to have made a *Dutch book* against the person.

For E_1: The stake is \$1000 and the option is: David pays the opponent \$600 and he will get \$1000 from the opponent if E_1 is true and get nothing if E_1 is false.

For E_2: The stake is \$1000 and the option is: David pays the opponent \$750 and he will get \$1000 if E_2 is true and get nothing if E_2 is false.

If E_1 is true, then E_2 is false and David will lose \$350. If E_1 is false, then E_2 is true and again David will lose \$350. In other words, David will be a loser no matter what happens and his set of betting rates is thus said to be incoherent. On the other hand, if his betting rates on the two propositions E_1 and E_1 are 3/5 and 2/5 respectively[92], it can be shown that nobody would be able to make a Dutch book against him and this set of betting rates is said to be coherent. In fact there is a theorem called the Ramsey-De Finetti theorem (or the Dutch book theorem) which assures that a set of betting rates is coherent if and only if it satisfies the axioms of probability[93].

As a result, if degrees of belief can be represented by coherent betting rates then they will *ipso facto* satisfy the axioms of probability. It renders the subjective theory looking highly plausible in interpreting probabilities because it seems to be able to demonstrate why it is rational for us to accept the axioms of probability calculus. Furthermore, as we have discussed, subjective theory provides a straightforward interpretation of the probabilities of single events. Despite these promising features, subjective theory suffers serious difficulties. For examples, different subjects may have

[92] Indeed, in the present case, any pair of betting rates which add up to 1 will do.
[93] The proof can be found elsewhere, for instance, Hacking, 2001, pp.165-170; Gillies, 2000, pp.59-65.

different degrees of belief or betting rates on the same proposition, could we be sure that their betting schemes are all coherent? If no, does it imply that the incoherent subjects cannot 'do probability'? On the other hand, even though their betting schemes were all coherent, how could we interpret the objective statements like 'the probability of that a particular radioactive element will disintegrate in 3 years' that should be dependent on the objective world rather than how we believe it? The difficulties of subjective theory will be further discussed in Appendix 11 and it could be shown that the subjective probability derived from the Dutch Book Argument and the Representation theorem, which constitute two major arguments to support the subjective theory, is in fact the subjective estimate of objective chance, rather than subjective uncertainty of the mind assumed by subjective interpretation (Sun, 2003).

3.5 The Propensity theory

The propensity interpretation of probability was first proposed by Karl Popper (1957). The theory was then taken up by a number of writers, some developed and reformulated it, and some criticized it (Galavotti, 2005; Gillies, 2000; Mellor, 2005, Miller, 2002). Popper maintained that probabilities are completely objective features of the world but he did not agree with von Mises that objective probabilities for single events should not be admitted, since probabilities for single events are inevitable in quantum theory. Moreover, he has presented an argument against von Mises' frequency theory (Popper, 1957): Consider two dice, one is fair and another is loaded in such a way that in the long run '6' occurs about 1/4 of the throws. If we throw the loaded die for many times (say, 10^6 times) except the 100th throw in which the fair die is thrown, the relative frequency of '6' in the sequence of these throws will be approximately 1/4.

According to the frequency theory, the probability of getting '6' in the 100th throw, that is still a part of the collective for which P('6') = 1/4, should be 1/4. This result is, however, contrary to our intuition that its probability should be 1/6. Popper (1957, 1959) then suggested that even though probabilities might be said to be relative frequencies, the frequencies would depend on the experimental arrangement. If we repeat the experiment very often, the experimental arrangement is liable to produce a sequence with relative frequencies which depend on these particular generating conditions for the experiment and probabilities are dispositional properties of these conditions, i.e., propensities. In other words, according to the propensity theory, probability is a characteristic property of the experimental arrangement (or generating conditions of an experiment) rather than as a property of a sequence or collective[94].

If the probability of an outcome is construed as a measure of the inclination of the current state of affairs to realize that outcome (Miller, 1994, p.182), then probability can be ascribed to a single event, no matter it has been realized or is merely possible. This interpretation underpins an indeterministic picture of our world:

> 'The future is open: objectively *open*. Only the past is fixed; it has been actualized and so it has gone. The present can be described as the continuing process of the actualization of propensities; or more metaphorically, of the freezing or the crystallization of propensities. While the propensities actualize or realize themselves, they are continuing processes. When they have realized themselves, then *they are no longer real processes*' (Popper, 1990, p.19).

According to Miller (2002), this constitutes 'a factual hypothesis that our world is faced at any time with a range of possible ways forward, and that these possibilities may be

[94] In 1933, Kolmogorov had raised similar viewpoint in associating probabilities with generating conditions rather than sequences or collectives in the first edition of his *Foundations of the theory of probability* (Kolmogorov, 1957, pp.3-4).

differently weighted. It is a second hypothesis that these weights or propensities conform to the axioms of the calculus of probability. But if propensities really are probabilities, and are sufficiently well behaved, then a bridge between propensities and frequencies is provided by the laws of large numbers. Statements of propensity are in principle testable' (p.112). The propensity theory is thus not a metaphysical theory as some (e.g., Gillies, 2000; Galavotti, 2005) have interpreted. As stressed by Popper (1990), 'propensities should not be regarded as properties *inherent in an object*, ..., but that they should be regarded as *inherent in a situation*' (p.14). For example, the probability for the single event that Mr. William J. Clinton is still alive 40 years from today (1 September 2005) to occur, according to the propensity theory, is neither an intrinsic property of Mr. Clinton's present state of health (or his genetical make-up) nor the relative frequency of certain collectives (such as the life expectancy for US Presidents). The probability will indeed vary day by day[95] and will be affected by his activities or lifestyle (e.g., smoking), other people's activities (e.g., assassination), and the progress of medicine, etc. (Mellor, 1994, p.181; Popper, 1990, p.14).

A major problem for the propensity theory, first raised by Paul Humphreys (1985), is that: Could Bayes's theorem be applicable to propensities? As tendencies to produce certain outcomes, propensities have a temporal or causal asymmetry that appears to be inconsistent with the symmetry characterizing probabilities. Using Humphreys' (1985) own example, there could be a disposition for a glass window to shatter when struck by a heavy object but the window has no disposition to be hit by a rock when broken (p.558).

[95] For example, the propensity on 1 September 2005 of his survival for another 20 years, given that he will act as a matador in tomorrow's bullfighting will be different from the propensity on 1 September 2005 of the same event, given that he will give a speech in tomorrow's lecture.

But in probability calculus, once $P(A|C)$ has been defined its inverse probability $P(C|A)$ could be figured out by using Bayes's theorem $P(A \mid C) = \frac{P(C \mid A)P(A)}{P(C)}$. This inconsistency constitutes a criticism of the propensity theory and is now known as 'Humphreys' paradox' (Fetzer, 1981, p.283). Proponents of propensity theory have proposed different further interpretation of propensities to circumvent the problem.[96] For example, Gillies (2000) has argued that the adoption of the long-run version of propensity theory is able to solve the Humphreys' paradox. Gillies (2000, pp.130-134) uses an example to argue for his point: suppose that in a factory there are two machines producing Frisbees, Machine 1 produces 800 pieces per day with 1% defective and Machine 2 produces 200 pieces per day with 2% defective. At the end of each day a Frisbee is selected at random from 1000 produced by the two machines. Let D denote the event that the selected Frisbee is defective, M denotes that it was produced by machine 1, N denotes that it was produced by machine 2. The conditional probabilities $P(D \mid M)$ and $P(D \mid N)$ make perfect sense, that denote respectively the propensities of machine 1 and 2 to produce defective Frisbee. But what do we mean by '$P(M \mid D)$' or '$P(N \mid D)$'? Does the propensity of the defective Frisbee selected at random at the end of day to have been produced by machine 1 (or 2) make sense? First, by Bayes's theorem, the value of $P(M \mid D)$ is given:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

[96] There are however some proponents of propensity theory who do not attempt to solve this problem. For example, Fetzer (1981) has construed propensities as partial causes (i.e., $P(A|C)$ is a measure of the strength of the propensity of the conditions C at time t to produce an event A at some time later than t) and conceded that this interpretation makes senses in some cases only. This strategy is hardly appealing as it leaves the calculus of probability only partially interpreted.

$$= \frac{P(D \mid M)P(M)}{P(D \mid M)P(M) + P(D \mid N)P(N)}$$

$$= \frac{0.01 \times \frac{800}{1000}}{0.01 \times \frac{800}{1000} + 0.02 \times \frac{200}{1000}}$$

$$= \frac{2}{3}.$$

According to Gillies (2000), if propensities were interpreted as partial causes, this conditional probability would mean 'the drawing of a defective frisbee in the evening is a partial cause of weight 2/3 of its having been produced by machine 1 earlier in the day. Such a concept seems to be nonsense, because by the time the frisbee was selected, it would either definitely have been produced by machine 1 or definitely have not been produced by that machine' (p.131). By adopting the long-run propensity theory, Gillies (2000) first identifies the set of repeatable conditions specifying that the two machines produce their daily output of Frisbees (S). The conditional probability $P(M \mid D)$ is a short-hand of $P(M \mid D \wedge S)$ which means:

> Suppose we repeat S each day, but only note those days in which the frisbee selected is defective, then, relative to these conditions, there is a propensity that if they are instantiated a large number of time M will occur, i.e. the frisbee will have been produced by machine 1, with a frequency approximately equal to 2/3. (p.132)

This interpretation does, however, fail to tackle the problem of single event. For example, if the Frisbees produced by machine 1 is marked with a small '1' on its surface while those produced by machine 2 is marked with a small '2'. When a Frisbee is selected at the end of day, the small number will show that either it is made by machine 1 or made by machine 2. A propensity of 2/3 does not make sense for this single event.

Another attempt to solve Humphreys' paradox is being developed by David Miller (1994, 2002) and later Popper (1990) through the universe version of the single-case propensity theory. This version accentuates a larger framework than that consisting in the generating conditions. When we consider the conditional probability $P(C|A)$, Miller (1994, 2002) proposes to add temporal index to each of the occurrences (e.g., A and C) so that they could be described by a basic statement, i.e. a single statement equipped with coordinates of time and place. For example, we have to specify the time when an occurrence A is actualized, if it is actualized. Suppose that A_{t_1} denotes an occurrence A that is actualized, if it is actualized at all, at time t_1, and that C_{t_2} denotes an occurrence C that is actualized, if it is actualized at all, at time t_2. The propensity at time t for C to be actualized at time t_2 given that A is an occurrence that is actualized at time t_1 should be written as $P_t(C_{t_2} \mid A_{t_1})$. By considering the temporal sequence of t, t_1, t_2, we have a total of thirteen situations that could be grouped into 4 cases[97]:

(1) $t < t_1 < t_2$

This is the simplest case that the time of the conditioning occurrence A is earlier than that of the conditioned occurrence C. When we talk of the propensity at time t for C to be actualized at t_2 given that A is actualized at t_1, both A and C have not reached their time for actualization. The probability $P_t(C_{t_2} \mid A_{t_1})$ can take any value in the interval [0, 1]. Humphreys' paradox does not arises in this case.

[97] Our grouping here renders the four cases being mutually exclusive and exhaustive. See Miller, 2002, pp.113-114 for another form of grouping from which the cases generated overlap slightly.

(2) $t < t_2 \leq t_1$ [98]

In this case, the time of the conditioning occurrence A is not earlier than that of C. But when we talk of the propensity at time t for the actualization of C at t_2 given that A is actualized at t_1, the actualization of A at time t_1 is supposed to be given at t rather than t_2. Since the time t is earlier than t_2, the occurrence C has not been actualized when we consider the propensity at t and $P_t(C_{t_2} \mid A_{t_1})$ can still take any value in the interval [0, 1]. For instance, in our previous Frisbee example, if $P(M \mid D)$ really means $P_t(M_{t_2} \mid D_{t_1})$, where $t_2 < t_1$ as the defective Frisbee must be selected after its production by the machine and the time t (say, the time very early in the day such that the machines has not started operating) is earlier than the time of production and selection, talking of the probability $P_t(M_{t_2} \mid D_{t_1})$ at t has no implication that the Frisbee's defectiveness has any causal influence on its earlier production. What it means is the propensity for the present world at t to develop into a world in which the Frisbee selected in the evening will have been produced by machine 1 (in the afternoon, say), given that the present world at t is a world which will develop into one of the worlds in which the Frisbee selected in the evening is defective.

(3) $t_2 \leq t$ [99]

If the time for the realization of C has been passed at time t, either C has been already actualized or failed to be actualized. Hence, the value of $P_t(C_{t_2} \mid A_{t_1})$ could be only 1 or 0.

---

[98] This case consists of two situations: (i) $t < t_2 < t_1$ and (ii) $t < t_2 = t_1$.

[99] This case comprises of 8 situations: (i) $t_2 = t_1 = t$, (ii) $t_2 = t_1 < t$, (iii) $t_2 < t_1 = t$, (iv) $t_2 < t_1 < t$, (v) $t_2 < t < t_1$, (vi) $t_1 < t_2 = t$, (vii) $t = t_2 < t_1$, (viii) $t_1 < t_2 < t$.

(4) $t_1 \leq t < t_2$ [100]

In this case, the time for the realization of A has been passed at time t but the time for realization of C is later than t. If A has been already actualized at or before t, A is a necessity at any time at or after t. It means that at t, the absolute propensity for C to be actualized will be the same as the propensity for C to be actualized under the condition that A has been actualized, i.e. $P_t(C_{t_2} \mid A_{t_1}) = P_t(C_{t_2})$. On the other hand, if A has failed to be actualized at or before t_1, A would become an impossibility at or after t. According to the probability conditional theory (Jeffrey, 1964; Ellis, 1973; Adams, 1998), the probabilities of conditional sentences like '$A \rightarrow B$' are conditional probabilities (e.g. $P(A \rightarrow B) = P(B \mid A)$). Moreover, if A is false then the conditional 'if A then C' will be true no matter what the truth value of C is. Therefore, the value of $P_t(C_{t_2} \mid A_{t_1})$ is 1.[101]

From the above discussions, we see that the resolution of Humphreys' paradox rests on recognizing that conditional propensities is not a measure of the causal dependence of the conditioned occurrence upon the conditioning occurrence. Inverse conditional propensities, as McCurdy (1996) has pointed out, do not represent inverse dispositions and they are well-defined concepts and can take non-trivial values as demonstrated in the above analysis.

It should be clear from the preceding discussion that although all interpretations of probability suffer from its own difficulties, the propensity theory is the one that is most

[100] This case comprises of 2 situations: (i) $t_1 < t < t_2$, (ii) $t_1 = t < t_2$

[101] It is worth noting that the probability calculus does not hold for those involving conditional probabilities $P(B|A)$ where $P(A) = 0$. In other words, it is not true that $P(B|A) + P(\sim B|A) = 1$ when $P(A) = 0$. Hence, it will not follow from this interpretation that $1 = P_t(C_{t2}|A_{t1}) + P_t(\sim C_{t2}|A_{t1}) = 1 + 1 = 2$.

able to stand up to its objections. And we will see in the next section whether we could explicate the concept of 'the probability of a hypothesis' viewed in this light.

3.6 Implications for the probability of a hypothesis

As learnt from David Hume, no number of specific observations or deduction can establish the truth of a general hypothesis with certainty. Inductive inferences are logically invalid, i.e. even if all premises are true the conclusion is not necessarily true. The truth of the premises of the inductive argument leaves the truth of the conclusion uncertain. Some thus attribute to the induced hypothesis some degree of probability and the problem of induction will then consist in figuring out the probability of the hypothesis (H) in light of the observation (O), i.e. $P(H \mid O)$, which is considered to be the degree to which our certain knowledge of the observation justifies our hypothesis. According to them,

O confirms H iff $P(H \mid O) > P(H)$

O disconfirms H iff $P(H \mid O) < P(H)$

O is evidentially irrelevant to H iff $P(H \mid O) = P(H)$.

If the observation O is a logical consequence of the hypothesis H, then $P(O \mid H) = 1$. We have:

$$P(H \mid O) = \frac{1 \times P(H)}{P(O)}$$

$$= \frac{P(H)}{P(O)}$$

$$\geq P(H) \ (\because P(O) \leq 1)$$

In other words, when we can deduce an observation from a hypothesis, and the observation really comes to pass, the result will confirm H unless $P(O) = 1$. It is worth

noting that even though a hypothesis is confirmed by an observation, the hypothesis may still be very improbable in light of the observation. For example, in a bag there are 1000 dice of which all are normal except one whose six faces are all marked with '6'. Suppose a die is selected randomly from the bag and a '6' is obtained when it is thrown (O). Let H be the hypothesis that the die selected is the one with '6' marked on all six faces. It is clear that $P(O \mid H) = 1$. Although $P(H \mid O) = \frac{2}{335}$ which is nearly six times of $P(H) = \frac{1}{1000}$, it is still very improbable (in compared with $\sim H$) in light of observation. Hence, 'confirm' here merely means 'probability raising' and it has no implication that the confirmed hypothesis must possess a high probability worthy of belief.

In this example, even though we are not Bayesians or advocates of subjective interpretation of probabilities, it still makes sense for us to talk about the prior probability $P(H)$ and posterior probability $P(H \mid O)$ since the hypothesis (the die selected is the one with '6' marked on all six faces) is a possible outcome of a chance process. But many of the hypotheses in which we are interested (e.g., all normal children who reach age 8 could perform the spatial reasoning task T, all particles arise from the resonant oscillation modes of strings[102]) are by no means possible outcomes of a chance process. Consider string theory: although some have argued that it is merely a theoretical model, we could still in principle be able to find evidence to confirm it by searching for the many new particles it predicts (the particles that correspond to the many possible oscillations of the string) (Randall, 2005, p.294). It thus makes sense to say what string theory predicts about observations. But what is the probability that string theory is true? If we do not

[102] It is a basic hypothesis in string theory (Randall, 2005, p.283).

believe that there is a God who had chosen the laws that govern our universe by a chance process, could we still make sense of the idea that a hypothesis has a prior probability?

There are at least four rejoinders to this criticism. First, some advocates of frequency theory (e.g., Reichenbach, 1949) have attempted to extend the notion of relative frequency so as to include the probability of hypothesis. Nonetheless, according to Popper (1959/1980), all proposed ways in assigning a probability value to a hypothesis with reference to the relative frequency are shown to be unsuccessful. For example, suppose one counts all experimentally testable statements belonging to the hypothesis and define the probability of the hypothesis as the relative frequency of those statements that turn out to be true. In case the hypothesis is universal, it will make an infinite number of experimentally testable statements. Since the number of observations can only be finite, the relative frequency of those statements that turn out to be true will then always remain zero, which is certainly an absurd consequence. If there were only a finite number of testable statements, the definition would make bizarre result too – a falsifying observation will only produce a very small decrease in the probability of the hypothesis, which is inconsistent with the fact that a falsifying outcome will suffice to establish the falsity of the hypothesis. Moreover, if we consider the hypothesis as an element of a reference class, as we have discussed in Section 3.1, there is a difficulty of determining the reference class for the frequency theory. Leaving this problem aside (say, assuming that the reference class is the class of hypotheses proposed by all other researchers), frequency theorists have another problem engendered by talking about the probability of

hypothesis – the truth of many hypotheses can never be ascertained, even in principle.[103] It is thus impossible for us to talk about the relative frequency of true hypotheses in a particular reference class. On the other hand, if we define the probability of a hypothesis as the ratio of the non-refuted (instead of true) hypotheses to all hypotheses in the reference class, then it will lead to bizarre consequence: either (a) if the number of all hypotheses is infinite then the probability of the hypothesis (or any non-refuted hypothesis in the reference class) will be 1 as the number of refuted hypothesis must be finite; or (b) if there are only finite number of hypotheses in the reference class then for any refuted hypothesis in this reference class, its probability will not be zero.[104]

Second, it seems to be less obscure for the probability of a hypothesis in light of evidence. Suppose we have different rival hypotheses each of which has certain chance to produce the observation O, it seems to be legitimate for some interpretations of probabilities to attach an intelligible meaning to the conditional probability $P(H \mid O)$ – for example, in the long run of repeating some particular procedures (say, Neyman-Perason experimental setup in controlling Type I error), we could define the probability of the hypothesis in light of O as the relative frequency of true hypotheses given the occurrence of O. However, as all probabilities are conditional (i.e., relative to a reference class) according to frequency theory, the insuperable difficulties encountered in last paragraph still persist in this case.

[103] As Popper (1959/1980, p.259) has pointed out: If we were able to know the truth of a hypothesis, we had no point to talk about its probability.
[104] Cf. Popper, 1959/1980, pp.259-260, 316.

Third, some may argue that we seldom assess the truth of a single hypothesis. In case of two of more competing hypotheses, we could use Bayes's theorem to make comparisons amongst different hypotheses:

$$P(H_1 \mid O) = \frac{P(O \mid H_1)P(H_1)}{P(O)}$$

$$P(H_2 \mid O) = \frac{P(O \mid H_2)P(H_2)}{P(O)}$$

From these two equations, we have:

$$P(H_1 \mid O) > P(H_2 \mid O) \text{ iff } \quad P(O \mid H_1)P(H_1) > P(O \mid H_2)P(H_2)$$

The posterior probabilities of two competing hypotheses depend upon the prior probabilities and their likelihoods. In this case, advocates of classical theory[105] may argue that if we have no reason to assign these two hypotheses different probabilities we have to assign them the same probability and the inequality in RHS will then become $P(O \mid H_1) > P(O \mid H_2)$, which should be in principle easily to be computed. But assigning the same probability to the two hypotheses is based on some versions of the Principle of Indifference and in Section 3.1 we have argued in length that this Principle can hardly be tenable. Moreover, even though it is legitimate for us to assign the same prior probabilities to these hypotheses, we have still to make sense of the idea that a hypothesis has a prior probability.

The fourth rejoinder is to regard the prior probabilities as describing a researcher's degree of belief. The degree of the researcher's belief in string theory before observation will determine what his or her prior probability is. As we have argued in

[105] See, for example, Laplace's treatment (Section 2.4).

Section 3.4, the subjective theory suffers from serious difficulties that it can hardly provide a intelligible interpretation of probabilities in objective cases. For example, different people may have different degrees of belief in an objective hypothesis, how could we judge which is correct? And in Section 3.4 we have also argued that we cannot assume that the process of Bayesian conditionalization must be able to provide rational researchers a learning strategy so that whatever prior probabilities they assign to the hypothesis their posterior probabilities will converge towards the same value. Furthermore, as Sober (2002) has argued, if our hypothesis is about our objective world, the subjective degrees of prior belief in the hypothesis do not have much scientific standing – what we want to know when we read research reports or papers is the information about the phenomena under study, but not autobiographical remarks about the authors of the study.

One more point for the measure of confirmation of a hypothesis in light of observation: Some (e.g., Redhead, 1985) have suggested that we could use $\frac{P(H \mid O)}{P(H)}$ as a measure of the support of H by the observation. As we have discussed before, if O follows logically from H, this ratio will equal $\frac{1}{P(O)}$, which is independent of H. Consider $H_1 = H \wedge T$, where T is another hypothesis[106]. As H logically implies O, H_1 will imply O as well. It is clear that $\frac{P(H \mid O)}{P(H)} = \frac{P(H_1 \mid O)}{P(H_1)}$ and they are thus supported to exactly the same degree by O if the ratio is really a measure of the support of the hypothesis by the observation. This is certainly not acceptable. Hence, this construal of

[106] Cf. Gillies, 1986.

measure of support of hypothesis needs to be revised and we will return to this point in Chapter 6.

As an advocate for frequency theory, Ronald A. Fisher did not accept $P(H|O)$ as a meaningful notion. In order to assess the relative merits of rival hypotheses in light of observation, he developed the concept of likelihood (Edwards, 1992): The likelihood of the hypothesis given observation O (and a specific model) , $L(H \mid O)$, is proportional to $P(O \mid H)$, with the constant of proportionality being arbitrary. If we take the constant to be 1, then $L(H \mid O) = P(O \mid H)$. Figuring out $P(O \mid H)$ requires that probabilistic predictions about the observations can be derived from the hypothesis. In cases like tossing the coins or dice, the hypothesis itself can often be described as a probabilistic model,[107] in which probabilistic predictions about the outcomes can be derived directly from the hypothesis. However, not all hypotheses we are going to study would by themselves result in probabilistic predictions about the outcomes. In this connection, their assessment under likelihood requires further probabilistic assumptions, which are collectively called the 'model'. We will discuss this issue in Chapter 4.

Apart from Fisher's likelihood, degree of corroboration is another concept developed by Popper (1957/1980, 1983) for comparing rival hypotheses in light of empirical observation:

$$C(H \mid O) = \frac{P(O \mid H) - P(O)}{P(O \mid H) - P(O \wedge H) + P(O)}.$$ [108]

[107] For instance, the die is fair, or P ('i')=1/6, for all i = 1, ..., 6.
[108] For simplicity, here we ignore the background information in the formula.

According to Popper (1983), this definition of degree of corroboration will lead to some highly intuitive results (pp.241-243). First, we should note that the denominator of the fraction must be non-negative for the following reason: since all probabilities are not greater than 1 and not less than 0, we have

$$P(O \mid H) - P(O \wedge H) + P(O) = P(O \mid H) - P(O \mid H)P(H) + P(O)$$
$$= P(O \mid H)(1 - P(H)) + P(O) \geq 0$$

Therefore, the sign of degree of corroboration $C(H|O)$ will be completely determined by the sign of the numerator. If O supports H, i.e. O follows from H or $P(O|H) = 1$, then the numerator will be positive. And the degree of corroboration will then be positive. If O undermines H, i.e. non-O supports H or $P(\sim O|H) = 1$, then $P(O|H) = 1 - P(\sim O|H) = 0$. And the degree of corroboration $C(O|H)$ will thus be negative. If O is independent of H, i.e. $P(O|H) = P(O)$, then the numerator will be zero and the degree of corroboration will then be zero. Second, we can observe that in the above three cases $C(O|H)$ will not be greater than 1 and in fact the denominator is chosen merely because it can normalize the whole fraction. That is to say, the maximum value which $C(O|H)$ can reach is 1 if this normalization factor is chosen. Here is the reason:

Since $P(O \wedge H) \leq P(O)$, we have $C(H \mid O) \leq \dfrac{P(O \mid H) - P(O)}{P(O \mid H)} \leq 1 - \dfrac{P(O)}{P(O \mid H)} \leq 1.$

The maximum value 1 will be attained if and only if $P(O) = 0$ because:

$$\frac{P(O \mid H) - P(O)}{P(O \mid H) - P(O \wedge H) + P(O)} = 1$$

$$\Leftrightarrow P(O \mid H) - P(O) = P(O \mid H) - P(O \wedge H) + P(O)$$

$$\Leftrightarrow 2P(O) - P(O \wedge H) = 0$$

$$\Leftrightarrow 2P(O) - P(H \mid O)P(O) = 0$$

$$\Leftrightarrow P(O)(2 - P(H \mid O)) = 0$$

$$\Leftrightarrow P(O) = 0 \ (\because P(H \mid O) \neq 2)$$

Since $P(O) \geq P(O \wedge H) = P(O \mid H)P(H)$, we have $P(O \mid H) = 0$ or $P(H) = 0$ if $P(O) = 0$. That is to say, if $P(O) = 0$ and $P(O|H) = 1$, $C(H \mid O)$ will attain its maximum value 1 and in this case $P(H) = 0$. This result matches with our intuition that only an observation O which is extremely improbable but become highly probable in the presence of H can give H maximum degree of corroboration. Similarly, the minimum value which $C(O|H)$ can reach is –1. And this minimum value will be attained when $P(O \mid H) = 0$ for:

$$\begin{aligned} C(H \mid O) &= \frac{P(O \mid H) - P(O)}{P(O \mid H) + P(O \wedge H) + P(O)} \\ &= \frac{P(O \mid H) - P(O)}{P(O \mid H)(1 - P(H)) + P(O)} \\ &= \frac{0 - P(O)}{0(1 - P(H)) + P(O)} = -1 \end{aligned}$$

This result is desirable because O is an evidence which falsifies H when $P(O \mid H) = 0$. Popper's idea is similar to that of Fisher: neither attempt to assign probabilities to hypotheses.[109] It is worth noting that Popper did not think that $C(H|O)$ is the only way to measure the degree of corroboration and he had indeed proposed different formulae for measuring the degree of corroboration although he maintained that the above formula is the simplest and most lucid one (Popper, 1983, p.242) and the degree of corroboration defined above does not conform to probability calculus (Popper, 1983, p.243) and its implications will be discussed in Chapter 6.

[109] See de Queiroz, 2003; de Queiroz and Poe, 2001, 2003; Kluge, 1997, 2001; Sidall and Kluge, 1997 for the discussion (and debate between de Queiroz and Poe and Kluge) on the similarities between likelihood and collaboration.

We would like to make two concluding remarks before closing this chapter. First, for the hypotheses that are not the outcomes of chance processes, talking about their probabilities is so obscure that further explication is necessary. Unlike the truth of a single event that could one day, at least in principle, be recognized and its probability can thus be assigned meaningfully under the propensity interpretation, the truth of hypothesis could never be ascertained and talking about its probability is still unintelligible unless we are preparing to accept the subjective theory. But in this case, one has to answer how the difficulties encountered by the subjective theory could be resolved. Second, even if we were able to explicate the meaning of 'the probability of a hypothesis' satisfactorily, it would not imply that the judgment of the reasonableness of a hypothesis would be necessarily related to its probability. Whether probability is a notion that can be used to characterize the degree to which a hypothesis has stood up to tests is a problem that will be discussed in Chapter 5.

Chapter 4 The logical foundations of SST and misconceptions associated with SST

In this chapter, we will examine the logical foundations of SST and the fundamental notions, such as null hypothesis, p value, statistical significance and Type I, II errors so that we could clear up a way for further discussion of the arguments for and against SST in the next chapter.

4.1 The logic of hypothesis testing

Rather than give our examination of the logical foundations of SST straight away, we will first consider some preliminary discussions on the concept of hypothesis testing which might more naturally suggest themselves. First of all, a basic confirmation practice[110] involving hypothesis testing in science runs in this way. If we are going to test a hypothesis H , we have to deduce observation statements from H. By conducting an experiment or making an inspection, we would be able to determine if the observation statements are true or not. If all observation statements are found to be true, then H is said to be confirmed though we are not still sure if H is really true or not. On the contrary, if some of them are not true then we would be able to refute H conclusively. The logic behind the refutation[111] is as follows:

If H is true, then S is true.

S is not true.

Hence, H is NOT true.

[110] The inductive nature of confirmation will be discussed in Chapter 6.

[111] 'Refutation' and 'falsification' are used interchangeably in this thesis.

This is a valid argument form called *modus tollens* (MT)[112]. Its validity can be easily proved (e.g., we can use a truth table to demonstrate that the sentence $(H \rightarrow S) \wedge \sim S \rightarrow \sim H$ is a tautology).

The genuine situation is, however, more complicated. First, even though the hypothesis is simply a universal statement without referring to any unobservable entities or constructs such as 'all metals expand when heated' or 'all normal children who reach age 8 could perform the spatial reasoning task *T*', the hypothesis cannot be directly tested with the argument as shown above. For example, let *H* be the hypothesis that all normal children who reach age 8 could perform the spatial reasoning task *T*. We cannot test directly if all normal children who reach age 8 could really perform the task *T*. We have to include an additional condition *C* such as 'David is a normal child aged 8' so that *H* and *C* together, instead of *H* alone, could entail an observable statement: David is able to perform *T*. Second, a substantive hypothesis usually refers to a number of theoretical constructs, therefore it is the hypothesis in conjunction with auxiliary theories or hypotheses, rather than the hypothesis alone, could entail observational statements. For example, consider the hypothesis that concrete manipulatives are efficacious in learning mathematics (Clements and McMillen, 1996; Sowell, 1989; Suydam, 1986). It cannot be directly tested because 'concrete manipulatives' have to be specified and 'efficacious in learning mathematics' is hardly observable. Only when we add the auxiliary hypotheses such as 'computer manipulatives like the software *Shapes* (a software version of pattern

[112] See, for example, Copi and Cohen, 1998. *H* and *S* here can be substituted by the names of any two sentences.

blocks) are concrete manipulatives'[113], 'if a manipulative M is efficacious in learning mathematics then students who use M in their mathematics classes outperform those who do not'[114], and 'the students' performance could be measured by the students' scores in a particular mathematics test', the hypothesis could be putting into testing. For example, from the substantive hypothesis and the above auxiliary hypotheses, we could deduce that 'students using the software *Shapes* in their mathematics classes would get higher scores in a particular mathematics test than those who do not use *Shapes*', which can in principle be directly tested[115].

Accordingly, we have to revise the argument as follows:

If H is true and C is true, then S is true.

S is not true

Hence, H is NOT true or C is NOT true.

where C is a set of auxiliary hypotheses and specific initial conditions.

In other words, the premises of the argument consists of three components, the hypothesis H (the theory or hypothesis being tested), a set of auxiliary hypotheses and/or specific initial conditions C and the observation statement S. If what we expected to occur does not happen, i.e. S is false, then either H is false or the set of auxiliary hypotheses and specific initial conditions C is not true. We could still retain the hypothesis by blaming not all initial conditions having been met or not all auxiliary

[113] We could also regard 'computer manipulatives are concrete manipulative' as auxiliary hypothesis and 'the software *Shapes* (a software version of pattern blocks) is a particular concrete manipulative' as an initial condition. But there is no need for us to delve into such intricate distinction here.

[114] The auxiliary hypotheses mentioned here are hardly exhaustive. For example, we have to assume further that the two groups should have the same performance if there is no difference among their uses of *M*.

[115] We will elaborated this point in Section 4.5.

hypotheses are true. Consider the above examples, if David is able to perform the task, then H is said to be confirmed to a certain extent[116]. But if he is unable to perform the task T, could we refute H conclusively? Those cognitive psychologists advocating H may argue that although David is a child aged 8, he is not a *normal* kid at all. Hence the fact that David is unable to perform the task does not necessarily entail that H is false. Moreover, in our second example, even though there is a study which shows that students not using concrete manipulatives outperformed classes using manipulatives on a mathematics test, researchers like Clements (1999) could still retain the hypothesis that concrete manipulatives are efficacious in learning mathematics by rejecting the auxiliary hypothesis 'if a manipulative M is efficacious in learning mathematics then the students who use M in their mathematics classes outperform those who do not' – Clement indeed contents that the manipulative M is efficacious only when the students use it *properly* and in the study students sometimes learn to use manipulatives only in a rote manner, which can hardly, according to Clement, be regarded as a proper use and it thus does not fulfill his amended auxiliary hypothesis.

Another complication about hypothesis testing is that many hypotheses (or auxiliary hypotheses) do not take the form of exact and mathematically formulated laws[117]. For example, the hypothesis that concrete manipulatives are efficacious in learning mathematics (or the auxiliary hypotheses) is clearly not in the form of exact and mathematically formulated law. Furthermore, even though we have added the auxiliary

[116] This is a controversial point that we will discuss in later Chapters.

[117] Examples of exact and mathematical formulated laws include the law of gravitation, which states that the force between two objects is directly proportional to the product of their masses and inversely proportional to the square of the distance between them.

hypotheses as mentioned above, we still need procedures that could be used to compare the scores obtained by the group of students who use *Shapes* in their mathematics classes (G_1) and the group of students who don't (G_2) in a particular mathematics test. The common practice is something like this: the researcher selects at random a sample of students (say, 40 students) amongst all of the students in a particular form of a school. The researcher then randomly assigns the 40 students to two groups and makes effort to check that they are almost equivalent in their previous knowledge of mathematics and other known attributes that may affect their performance in the mathematics test. The same teacher then teaches a mathematical topic to both groups of students for a certain period of time. The teacher will strive for keeping the teaching and learning materials as well as the teaching strategies for the two groups identical except that students in G_1 use the software *Shapes* in the classes and the students in G_2 do not. In other words, the researcher has to ensure that all conditions except the use of *Shapes* should be kept constant in the two groups. At the end of the period, the same mathematics test is given to both groups. The researcher could thus gather the scores obtained by the two groups of students. Since there are many factors, in addition to the use of concrete manipulatives, that can affect a student's performance in the test, students in either group will not have the same score. The researcher has to use statistical means to determine if the hypothesis has to be confirmed or rejected in light of the collected data. And SST, which has been discussed in Chapter 2, is the usual statistical means adopted. Before discussing its logical foundation, we have first to disentangle the conceptual confusion over the Fisher's significance testing and Neyman-Pearson hypothesis testing.

4.2 Distinctions between Fisher's significance testing and Neyman-Pearson hypothesis testing

In Section 2.8, we have discussed the historical development of Fisher's significance testing and Neyman-Pearson hypothesis testing and mentioned that the prevalent SST is a hybrid of these two approaches. Many studies show that many students (Falk & Greenbaum, 1995) and even researchers (Oakes, 1986) have no real insight into the meaning of a SST result. And as many have reported (Goodman, 1999a, 1999b; Hubbard, 2004; Hubbard & Bayarri, 2003a, 2003b; Hubbard & Ryan, 2000), confusion over the reporting and interpretation of results of SST is widespread amongst researchers and many erroneously believe that SST is prescribed by a single coherent theory. We will in this section discuss in depth the fundamental differences between Fisher's views on significance testing and Neyman-Pearson's ideas on hypothesis testing and expose their incompatibility so that we could have a more clear picture about the other notions in SST.

First, the need of an alternative hypothesis critically distinguishes between Fisher's and Neyman-Pearson approaches. In Fisher's significance testing the researcher postulates only the null hypothesis. Explicit incorporation of competing hypotheses is due to Neyman and Pearson. Neyman (1952) has commented that 'if satisfactory tests are actually devised without explicit consideration of anything beyond the hypothesis tested, it is because the respective authors subconsciously take into consideration certain relevant circumstances, namely, the alternative hypothesis that may be true if the hypothesis tested is wrong' (p.44). Fisher certainly denied the need for an alternative hypothesis and was discontented with the Neyman-Pearson view that 'the purpose of the

test is to discriminate or "decide" between two or more hypotheses' (1956/1973b, pp. 45-46) and he maintained that such a purpose has greatly obscured the understanding of SST. Are these conflicting views arising from the difference in questions that they suppose the tests are purporting to address? Royal (1997) has suggested that Fisher's significance testing purports to measure evidence and addresses the question, 'How should I interpret these observations as evidence?' while Neyman-Pearson hypothesis testing aims at choosing between competing hypotheses and addresses the question, 'What should I do, now that I have this observation?' (Royall, 1997, p.64). There is probably difference between significance testing and hypothesis testing in their purposes but we will see in the next paragraph that the underlying principles behind the tests are also different.

Second, in Fisher's significance testing the researcher attempts to reject the null hypothesis by establishing the probability (p value) of obtaining the observed or more extreme outcomes under the assumption of the null hypothesis. A small value of p will imply that the observed outcome would be highly implausible if the null hypothesis were true. If the implausible outcome has indeed occurred, it will constitute evidence against the null hypothesis[118]. Neyman and Pearson did not agree that the mere occurrence of a rare or implausible outcome would be adequate for the rejection of the null hypothesis. They would not reject the null hypothesis, no matter how unlikely the observed outcome is, unless they could ascertain that there is a competing hypothesis under which the

[118] The logic behind this reasoning will be discussed in Chapter 5.

outcome would be more likely to occur.[119] This is perhaps the genuine reason why alternative hypothesis must be included in Neyman-Pearson hypothesis testing.

Third, Fisher's significance testing provides evidence against the null hypothesis in a single experiment or study whereas Neyman-Pearson hypothesis testing is not concerned with which individual hypothesis is true or false but attempts instead to control mistakes in the long run. In other words, hypothesis testing centres on the inductive behaviour[120] rather than the truth of a particular hypothesis. In Neyman and Pearson's own words:

> 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 ... Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong' (1933, p.290-291).

It is worth noting that the process of Fisher's significance testing is an inductive inference drawing conclusions about the null hypothesis (in a single study) from observation. Since the inductive inference is not as formal as the logical or mathematical inference, its process is fluid, non-quantifiable and includes combining the p value in some unspecified way with background information (Fisher, 1956/1973b; Goodman, 1999). On the contrary, the process of Neyman-Pearson testing is an apparently automatic way to bound the rate of mistaken conclusions in the long run. Its outcome is an action or behaviour: to

[119] In his later reflection, E.S. Pearson (1990) remarked that William S. Gosset had left him in a letter with the idea: 'The rational human mind did not discard a hypothesis unless it could conceive at least one plausible alternative hypothesis' (p.82).

[120] The term 'inductive behaviour' means the habit of humans (or other animals, such as Russell (1912/1980)'s chicken) to adjust their behaviour or actions to limited amount of observations, so as to avoid undesirable consequences, e.g., humans and some animals tend to take cover whenever dark clouds appear in the sky (Neyman, 1950; Hubbard, 2004). For Neyman's defence of his preference for inductive behaviour over inductive inference, see Neyman, 1950, pp. 1-2.

reject one hypothesis and accept the other. To reject a hypothesis H means only that the rule prescribes us to take an action A rather than action B. This does not imply that we have to believe that H is false. Similarly, to accept a hypothesis H means only to take an action B and this does not mean that we believe that H is true (Neyman, 1950, pp.259-260).

Apart from these three distinctions, there is one more distinction between these two approaches – the p value in Fisher significance testing and the Type I error rate α in Neyman-Pearson hypothesis testing. Since the interpretation of these notions have great implications for the later discussion, we will discuss this issue in the following section.

4.3 The interpretation of *p* values and Type I error rates

The p value, the significance level and Type I error rate α are extensively reported in SST, but misunderstanding of p values and significance levels by researchers can often be revealed in the literature especially when more and more researchers are able to access computers and computer programs for statistical analysis (Sackrowitz & Samuel-Cahn, 1999). Moreover, the use of these terms is quite confusing. For example, the p value has been referred to as '$P_{calculated}$' (Thompson, 1994), 'probability level', 'descriptive significance level', 'prob-value', and 'tail probability' (Williams, 1999). And the level of significance is represented by the symbol α, but sometimes by $P_{critical}$ (Thompson, 1994). Of course, the interpretation of these concepts creates greater problems.

As we have discussed before, Fisher's p value is a measure of the probability of the observed and more extreme outcomes assuming the truth of null hypothesis. Based

on Fisher's logic, it is an index measuring the strength of evidence against the null hypothesis: the smaller the p value, the greater the weight of the evidence is. For convenience, Fisher suggested to take 0.05 as a standard level of significance by which it means the researchers are prepared to ignore those results that fail to reach this standard (Fisher, 1935/1971, p.13). This 0.05 significance level is flexible[121], not pertaining to any pre-specified error rate, and has no long-run frequentist implication that, for example, the researcher 'allows himself to be deceived once in every twenty experiments' (Fisher, 1929, p.191). On the other hand, the significance level α in Neyman-Pearson hypothesis testing is the long-run relative frequency of Type I errors conditioned on the truth of the null hypothesis. The lower the value we set for α, the lower the probability of Type I error. Since there are many different possibilities for the alternative hypotheses (e.g. the difference of means could have many different values), the probability of a Type II error (β) is a variable which is a function of the true value of the parameter that is unknown. Usually we are interested in some particular values for this probability and compute this probability for the most unfavorable case. Power is the complement of β and is thus a measure of the chance of rejecting a null hypothesis when the null hypothesis is false. Like Type II error, power is also a variable which depends on the value of the parameter (e.g. the difference of means). In general, Type I and Type II errors are inversely related. The lower the value of α, the lower is the power of the test. In other words, the Type I error control used in hypothesis testing may lead to the lack of power in statistical research studies (Nix and Barnette, 1998). Low power will have a substantial impact on the ability to replicate traditional experimental studies based on hypothesis testing (Ottenbacher, 1996). Suppose a researcher does not strive for significance results, the

[121] For Fisher's remark on this point, see footnote 51 in Chapter 2.

likelihood of being published is severely diminished due to the publication bias that is present for statistically significant results (Nix and Barnette, 1998). As a consequence, a number of important studies could not be published because they are not good enough according to hypothesis testing. With the consideration of these constraints, the researcher has to figure out the critical region and the researcher should only state if the result fell in the critical region but not where it fell, as might be indicated by a p value. As Gigerenzer (1993, p.317) has explicated, the p value is a property of the data or, more accurately, a relation between a body of data and a theory and the significance level α (in Neyman-Pearson hypothesis testing) is a property of the test, not of the data.

Following Neyman and Pearson's tradition, many textbooks writers on hypothesis testing suggest that the significance level α must be specified or fixed prior to the collection of data[122] but only a few of them have tried to explain why we have to do this way. For example, Lind, Marchal and Mason (2002) only made the suggestion without any attempt to provide non-tautological explanation:

> We selected the significance level, .01 in this case, before setting up the decision rule and sampling the population. This is the appropriate strategy. The significance level should be set by the investigator, but it should be determined *before* gathering the sample evidence and not changed based on the sample evidence. (p.345)

Some simply call this approach 'fixed-level hypothesis testing' as if the name itself could explain why we should not simply report the actual probability found as the significance level (Daly et al, 1995, pp.314-315; Wild and Seber, 2000, p.390). And some merely

[122] In developing their theory of hypothesis testing, Neyman and Pearson were inspired by the French mathematician Emile Borel's suggestion that 'the criterion to test a hypothesis (a 'statistical hypothesis') using some observations must be selected not after the examination of the results of observation, but before' (Neyman, 1977, p.103). See also Neyman and Pearson, 1933, p.290.

state the rule and regard disobedience as a moral problem: 'Adjusting the level of alpha to achieve statistical significance after the study has been completed is unethical' (Drumm, 1995), without explaining why ethically researchers should not adjust the level of alpha to achieve statistical significance after the collection and analysis of data.

There are also writers who do not completely agree with this rule but their reasons are still obscure. For instance, Blaikie (2003) has alleged:

> Setting probability levels and deciding on the appropriate test before the data are collected, and certainly before the analysis is conducted, can be regarded as adding a degree of objectivity to the research, and as avoiding *post hoc* interpretations. However, if our ultimate aim is to test theoretical hypotheses,... Setting a rigid level of confidence in advance may not serve this purpose ...' (pp.182-183).

Indeed, there are textbook writers who suggest further that 'a reasonable procedure is for the researcher to adopt a significance level but also to report the actual probability found as a result of testing H_0, regardless of whether H_0 is rejected or not' (McCall, 2000, p.226). It seems that the actual probability in hypothesis testing could provide important information for the researcher, otherwise, there will not be some researchers 'who dislike the combination of arbitrary significance levels and dichotomous reject/do not reject decisions and prefer to report probabilities without decisions...' (McCall, 2000, p.226). Moreover, there are writers like Moore (1997) who maintain that: 'The P-value is more informative than a statement of significance, because it allows us to assess significance at any level we choose. For example, a result with P =0.03 is significant at the α = 0.05 level, but not significant at the α = 0.01 level' (Moore, 1997, p.490).

The genuine reason that the Type I error probability α must be specified or fixed prior to the analysis of data and cannot be adjusted after the examination of observed data is that: only by doing so we are able to control the long-run relative frequency of Type I error at a fixed level α. The reason could be explicated with an example. Consider two scenarios. In the first scenario, if all researchers in Group A agreed that a 0.05 significance level is specified before collection of data. Out of 10 000 true hypotheses, about 500 will be rejected at this 0.05 significance level. As a whole, the research reports in Group A would, in the long run, have a Type I error rate of 0.05. Similarly, if all researchers in Group B agreed that a 0.01 significance level is specified before collection of data, then the research reports in Group B would have a Type I error probability of 0.01.

In the second scenario, nothing changes except that researchers in Group C decide not to supply the significance level until the data are obtained. For instance, if the value of the test statistic exceeds the 0.01 critical value, the result will be reported as significant at 0.01 level; whereas the test statistic falls between the 0.01 and 0.05 levels, the result will be reported as significant at the 0.05 level. When the null hypothesis is rejected at the 0.01 level, it must be rejected also at the 0.05 level. The research reports in Group C do have a Type I error probability of 0.05 rather than 0.01 despite that some of them would claim that they have a Type I error probability of 0.01. It is thus misleading for some reports in Group C to state that they have a Type I error probability of 0.01.

As construed above, the significance level of hypothesis testing helps controlling the Type I errors in the long run. But what most researchers are eager to know is the

truth of their own hypotheses. If the p value in SST could on one hand limit our mistaken conclusion in the long run and on the other hand measure evidence and thus assess the truth of the null hypothesis from a single test, this hybrid will be much more attractive than significance testing or hypothesis testing, each of which can serve only one role. Is it really possible for the p value to represent both the strength of the evidence against the null hypothesis and at the same time the Type I error rate under the null hypothesis? The p value always lies exactly on the border of the tail region represented by the p value. Therefore, the tail region represented by a particular p value (e.g. $p = 0.04$) will not include future outcomes with smaller p values: for example, if we get an outcome with $p = 0.02$, we will report the p value as 0.02 rather than 0.04. This does not amount to the concept of error rate which requires that a result can be anywhere within the tail region. In other words, if we interpret the p value as the Type I error rate, it is akin to the second scenario discussed in the last paragraph that the genuine Type I error rate will be higher than the claimed p value. It is thus mistaken for us to use the p value to denote at the same time the evidence against the null hypothesis and the observed Type I error rate.

Carver (1978) has summarized the misinterpretations of statistical significance into three categories: '(a) the probability is .05 that the results are due to chance, or the probability is .95 that the results were not caused by chance; (b) the probability is .95 that the results will replicate, or we can be .95 percent confident that the results are reliable; and (c) the probability is .95 that the research hypothesis is true, or we can be .95 percent confident that our results are valid' (p.387). It is widely recognized that SST is a difficult subject to teach and learn and statistical literature shows evidence of misconceptions at all ages and all levels of expertise (Garfield & Ahlgren, 1988;

Williams, 1999). Haller and Krauss (2002) suggested that textbooks are one possible source of these misconceptions. For example, in Nunally (1975), eight interpretations[123] of a significant test result have been provided in three pages (pp.194-196) and they are regarded as different ways to say the same thing.[124] Despite textbooks, journals also have wrong interpretations of significance, explicitly given by, as Sedlmeier and Gigerenzer (1989) reported, authors and editors of journals of these journals. Moreover, Haller and Krauss (2002) found in a recent study that, despite publication of numerous articles on the misunderstandings about SST, little seems to have changed – 'Nearly 90% of the *scientific psychologists* perceive at least one of the false "meanings" of a *p*-value as true. However, our novel finding that even among the *methodology instructors* 80% share these misinterpretations is flabbergasting' (p.7). These misunderstandings will be discussed in turn.

In Neyman-Pearson hypothesis testing, the level of significance (or the Type I error rate) α is the probability of rejecting a null hypothesis H_0 provided H_0 is true. One common misinterpretation of the level of significance arises from the confusion of two conditional probabilities: $P(R_H \mid H_0)$ and $P(H_0 \mid R_H)$, where R_H denotes the outcome that the null hypothesis is rejected. The first conditional probability means the probability that the null hypothesis H_0 will be rejected given that H_0 is true; while the second means the probability that H_0 is true given that H_0 is rejected. The level of significance should be understood as the conditional probability $P(R_H \mid H_0)$. If we follow the frequency

[123] All of the interpretations are either obscure or even mistaken. For example, 'the improbability of observed results being due to error', 'the probability that an observed difference is real', 'the degree to which experimental results are taken "seriously"'.

[124] With respect to the textbook definitions of null hypotheses, a lack of consistency and comprehensiveness is observed in Truran's (1998) study.

theory of probability (Section 3.3), 'a level of significance of 0.05' means 'on average, 5 out of every 100 times the null hypothesis is true, we will reject it', instead of 'on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong'. The misinterpretation (c) reported by Carver (1978) belongs to this category. Moreover, Falk (1986) has reported that a large class of undergraduate students and 15 teachers and teaching assistants of statistics believed that α is the probability of being wrong when rejecting the null hypothesis at the significance level α. Similar findings have been found by Birnbaum (1982): school students who have studied statistical inference thought that the sentence, 'a significance level of 5% means that, on average, 5 times out of every 100 times we reject the null hypothesis we will be wrong' (p.24) sounded true. Oakes (1986) conducted a systematic probe of the meaning attached to a significant test result by 70 academic psychologists and it was found that the interpretation endorsed by most of the subjects was that the p value of the test conveys the probability of being wrong in rejecting the null hypothesis.[125]

Another misconception associated with the p value is the interpretation that the p value is the probability that the result (the occurrence of D) is due to chance or a p value of 0.05 means that the probability is .95 that the result was not caused by chance (Bakan, 1966). As Matthews (1999) has noted, what researchers are really interested in is the probability that the effect is just a fluke, given the obtained result. It is quite understandable that many researchers have hoped that the p value could give answer to

[125] Other similar results could be found in Falk and Greenbaum (1995) and Pollard and Richardson (1987).

this question they are really interested in.[126] But, as we have discussed in 4.2, p value is the probability of obtaining the result D or one more extreme when H_0 is true and there are no other factors influencing the result. In other words, a p value of 0.05 means the probability of obtaining results at least as impressive as those obtained is 0.05, assuming that mere chance is their true explanation. Merely from the given significant result, we can hardly make any inference about the causes of the result. For example, even if concrete manipulatives such as *Shapes* are not efficacious in learning mathematics, a significance difference between the students' performances could be observed if not all conditions apart from the use of concrete manipulatives have been kept constant in the two groups (say, the students who use *Shapes* have worked much harder in preparing the test than those who do not). It is thus incorrect to interpret p value as the conditional probability that the result is due to chance, given the data actually obtained.

Replication is a cornerstone of scientific research. Credibility will be much lower if results from a study cannot be reproduced. Another misconception is that, as Carver (1978) and Batanero (2000) have noted, the significance level indicates the probability of successful replication. For example, a significance level of 0.05 means that the probability is 0.95 that the results will replicate, or we can be 0.95 percent confident that the results are reliable. In Neyman-Pearson hypothesis testing, a significance level 0.05 means that in the long run it is anticipated that about 5 out of the 100 tests will be significant merely by chance given that the null hypothesis is true. This does not entail that, regardless of the truth of the null hypothesis, we would still get successful

[126] Even a researcher (e.g. Dunn, 2001) who taught others about the concept of statistical significance wrote in this way: 'statistical testing uses mathematical procedures to examine particular differences between groups to see if it is likely that the observed difference could have arisen by chance alone. If it is unlikely enough that the difference would have arisen by chance alone, the difference is "statistically significant.'

replication about 95 out the 100 tests. Unless our concern is the process of replication when the null hypothesis is true, it should be the statistical power rather than the Type I error rate will play a more vital role in determining the result replicability.[127]

On the other hand, does a small p value entail a higher probability of repeating a statistically significant result? As we have discussed before, the p value is a measure of the degree of conflict of the observed data with the null hypothesis Although it is not related to the rate of Type I error, it seems that we could argue in this way: the smaller the p value, the more evidence there is against the null hypothesis, rendering the null hypothesis more likely to be false. And the more likely the null hypothesis is to be false, the more likely that the significant results will replicate. We will discuss if this reasoning is really valid in next chapter.

A final point about the level of significance is the reasons why certain figures (like 0.05) of significance level have been frequently chosen. As we have discussed in Section 2.7, Ronald A. Fisher is probably the first person who explicitly suggests choosing a significance level of 0.05 as a convention to recognize significant results in experiment and some have explained why the value of 0.05 was selected by Fisher. But it is not a reason that is supported by mathematical theory. Indeed, the use of different levels of significance is a matter of convention. Research literature has shown that, in

[127] Suppose a null hypothesis is being tested at the 0.05 level of statistical significance, the probabilities that a true H_0 will be accepted twice and that a true H_0 is rejected twice in succession when using SST are respectively $0.95^2 = 0.9025$ and $0.05^2 = 0.0025$. Thus the overall probability of replication is 0.9025 + 0.0025 =0.905 when H_0 is true. This value does not, however, take into the consideration the effect of statistical power. If our purpose is to conduct a replication study to examine a genuine effect (say, concrete manipulatives are efficacious), then power instead of the value of p will become a critical factor in determining the success of any replication effort. See Ottenbacher, 1996 for detail of the impact of statistical power on the process of research replication and our discussions in the next chapter.

addition to 0.05, the commonly used levels of significance include 0.01 and 0.001. Skipper, Guenter, and Nass (1970) have suggested that adopting different levels of significance would enable the differentiation of research findings that will be published or not and remind the researchers to choose level of significance with full awareness of its implications for the problem under study. First, it is possible that a greater value of α, if interpreted as Type I error rate, might be preferable especially when the power is low and a Type II error is crucial. Second, in Fisher's significance testing, whether different significance levels α for p values do really reflect the corresponding degree of evidence is a problem that will be addressed in next chapter.

4.4 The logical foundation of Fisher's significance testing

As we have discussed in Section 2.3, the focus of SST is to attempt to challenge and refute the statistical hypothesis. Using the hypothesis that concrete manipulatives are efficacious in learning mathematics as an example, the procedures of Fisher's significance testing can be summarized as follows:

1. Assuming that at the end of the learning period the students in groups G_1 and G_2 come from two theoretical populations of students. And the mean test scores of the two theoretical populations of students in G_1 and G_2 are respectively μ_1 and μ_2. If the two theoretical populations are identical, then $\mu_1 - \mu_2 = 0$ which is identified as the null hypothesis H_0. According to our original hypothesis, what we expect to occur is that $\mu_1 - \mu_2 > 0$, which is described as the alternative hypothesis H_1.

2. Assuming that groups G_1 and G_2 are random samples chosen from the corresponding populations. Determine the appropriate test statistic[128] and figure out its distribution under the assumption of the truth of the null hypothesis. In the case of this example, it could be a *t*-distribution with 38 *df*.

3. From the observed data, we could calculate the means and the variances of the two samples and then figure out the unbiased estimates of the variances of the populations. Finally, we will get a calculated value of *t* from these data.

4. The probability of obtaining a *t* value as extreme or more extreme than the calculated *t* value when the null hypothesis is true is called the *p* value. If the *p* value is small (say, less than a significance level 0.5), then either the hypothesis is not true, or an exceptionally rare chance has occurred. In other words, if the null hypothesis is true, the results based merely on fluke are very unlikely. The observed data thus provides evidence against the null hypothesis. On the other hand, if the f the *p* value is not smaller than α and the null hypothesis H_0 is true, the discrepancies between the mean scores in the two groups could be explained by the chance fluctuation in sampling. As a result, the data does not provide evidence against the null hypothesis H_0.

The basic reasoning behind significance testing is: in order to decide if a null hypothesis H_0 should be rejected, we first assume that H_0 is true and see what consequence could be deduced. Suppose we could deduce from the truth of H_0 that it is very unlikely that the data *D* will occur (for example, a particular test statistic is very unlikely to fall in a certain region as hybrid model suggests). If in fact *D* does not occur

[128] For a formal definition of test statistic, see Cox and Hinkley, 1974.

(for example, the test statistic does not fall in the region), this result is expected under the assumption of the truth of H_0. We thus have to retain H_0 in this case. On the contrary, if in fact D does really occur, then this result is astonishing if H_0 is true and therefore we seem to have good reason to reject H_0. The logic behind the whole argument could be summarized as follows:

> If H_0 is true, then the probability that the data D will occur is very small.
>
> The data D occurs.
>
> Hence, the probability that H_0 is true is very small.

Is this argument valid? For a certain null hypothesis H_0, it must be either true or false, what does it mean by 'the probability that H_0 is true'? These questions would be addressed in the next chapter.

4.5 The concepts of refutation and rejection

Similar to what we have discussed in Section 4.1, the following argument form (MT) is valid:

> If H_0 is true, then the data D will not occur.
>
> The data D occurs
>
> Hence, H_0 is NOT true.

If the two premises are true, the conclusion must be true. In other words, if we could deduce from the truth of H_0 that the data D will not occur and it is found that D does occur, then H_0 is said to be refuted. 'Refutation of the null hypothesis H_0 by D' means that H_0 is shown by the evidence or observation (D occurs) to be false. Of course, as we

have discussed in Section 4.1, such a situation is over-simplified and the intrusion of auxiliary hypotheses and initial conditions is inevitable in most scientific research. However, even in this case, the revised argument:

If H_0 is true and C is true, then the data D will not occur.

The data D occurs.

Hence, H_0 is NOT true or C is NOT true.

would still imply that H_0 and C are refuted by D although we are not certain which one (or both) is being refuted.

It is worthwhile for us to note that in the argument behind SST what we could *at most* conclude is that the probability that H_0 is true is very small[129]. It is clear that when the probability of an event is very small (even zero), it does not entail that the event is impossible[130]. Hence, based on the conclusion that the probability that H_0 is true is very small, we reject H_0; it does not mean that H_0 has been refuted or shown to be false. In other words, the rejection of H_0 does not imply that H_0 has shown to be false. It is at most a conclusion that H_0 is very unlikely to be true. We will return to this point in Chapter 6.

4.6 The concepts of different types of hypotheses

Apart from the p values and significance levels, 'null hypothesis' is another concept that is usually misunderstood. Confusions between the roles of the null and alternative hypotheses and between the statistical alternative hypothesis and the research hypothesis could be found in some research papers (Chow, 1996; Batanero, 2000). As

[129] Whether we could really reach this conclusion will be discussed in Chapter 5.
[130] For example, when we randomly select a number from the open interval (0, 1), the probability that it is rational is 0 although it is perfectly possible that it could really be a rational number.

we have discussed in Section 4.1, what we are really interested in is the substantive hypothesis (which, in our example, is concrete manipulatives are efficacious in learning mathematics) . It is however not feasible for us to test this hypothesis directly. We have to deduce observable implication from the substantive hypothesis and some additional auxiliary hypotheses (and initial conditions). The outcome, such as 'the students who use computer manipulatives in mathematics classes could outperform the students who do not' is called 'research hypothesis'. The research hypothesis is usually still not specific enough for putting to test. It is thus required to construct an experimental hypothesis, such as 'the students who use the software *Shapes* in their geometry classes would get higher scores in a particular geometry test than those who do not use *Shapes*'. An experiment will then be conducted to test the experimental hypothesis. For example, in Section 4.1, we have discussed in details what specific procedures have to be followed under the experimental context. And in order to determine if the students in group G_1 will get higher scores in a particular geometry test than those in G_2, we have to conduct statistical analysis which requires a specification of the null hypothesis (H_0: $\mu_1 - \mu_2 = 0$) and the alternative statistical hypothesis (H_1: $\mu_1 - \mu_2 > 0$).

With this hierarchy in mind, we would be clear that SST is merely a means to test the statistical hypothesis. Suppose in an experiment we are able to reject a null hypothesis with a significance level of 0.01. Even though the logic behind SST is sound (we will discuss this point in Chapter 5), it does not imply that we are at the same level of certainty to reject the experimental hypothesis. Because experimental control (for instance, if all relevant factors that could influence students' performance have been kept constant) and theoretical considerations from the field under study (for instance, if the

test constitutes a reliable and valid measure for students' performance) would play an indispensable role in assessing the backward deduction from the rejection of statistical hypothesis to that of experimental hypothesis. It is also this reason that Chow (1996) who has argued that many of the criticisms against SST are misdirected, as they refer not to the statistical process but to other parts of the inferential procedures.[131]

In SST, we attempt to reject the null hypothesis so that we can gain evidence of the alternative hypothesis. It is in this connection that the null hypothesis, as originally suggested by Fisher, means the hypothesis that is going to be nullified (Cohen, 1994). In this general sense, the null hypothesis does not necessarily refer to the hypothesis of no difference or relationship. For example, 'the difference of the means of two populations is 3 marks (i.e. $\mu_1 - \mu_2 = 3$)' could be our null hypothesis. But now the word 'null' in 'null hypothesis' is usually regarded a synonym of 'nil' or 'zero'. Construed in this way, the null hypotheses of no difference are sometimes called the 'nil hypotheses' so as to distinguish it from other null hypotheses (Cohen, 1994). Some have argued that unless in the most rare of instances the nil hypothesis must be able to be rejected (Bakan, 1966; Cohen, 1994; Meehl, 1967). Certainly, it is extremely rare to find two identical cases of anything in our world. If the measure is fine enough, any observed objects will differ on whatever variable we choose to measure. For example, could we expect to find two lakes with identical amount of water? Even if a machine is designed to produce identical bottles of distilled water, we hardly expect that it can produce two bottles of water that will have exactly the same amount of water. Even if we fail to measure the difference, we can always resort to a more accurate way of measurement. But as Hagan (1997, 1998)

[131] This point will be further discussed in Chapter 5.

has argued, the null hypothesis is a statement about the population rather than the sample. According to him, samples drawn from the sample population will always differ but the population will be exactly identical to itself. Moreover, the nil hypothesis can always be rejected at some sample size. According to APA(1994), reports of effect sizes are recommended because sample size largely drives rejection of the null hypothesis. Thompson (1998) thus concludes that 'statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we've been too lazy to drag in enough participants' (p.799). We will leave the question whether these criticisms to the concept of nil hypothesis are tenable to the next chapter.

4.7 Statistical significance and practical significance

Although many authors have made note of the fact that statistical significance has nothing to do with practical importance, it is still not unusual for researchers to herald the significance level as being synonymous with importance (Glaser, 1999; Thompson, 1996, 1997). In practice, we identify the hypothesis of interest with the alternative hypothesis which usually says nothing about the exact magnitude of the difference between the population means (e.g. $\mu_1 - \mu_2 > 0$). Statistical significance is thus not necessarily informative about the practical significance of the data. For example, if the mean scores of the two groups of students G_1 and G_2 are 60 and 59 out of 100 marks. A statistically significant difference may still be demonstrated if the sample size is large enough. However, if a difference in score of 1 mark is educationally significant is another matter. In other words, statistically significant difference does not imply the difference is significantly big.

Besides, in an experiment, if a difference between the experimental and control group is demonstrated to be statistically significant, it does not entail that the difference must be due to the treatment differences. For example, if the mean scores of the two groups of students G_1 and G_2 are shown to be significantly different. There is always possibility that the difference is due to factors other than the use of computer manipulatives. We have to look carefully at the study design to see if it is really possible that something other than the use of computer manipulatives that might explain the difference in scores.

Chapter 5 Arguments for and against SST

As noted in previous chapters, there is abundant literature on the controversy of SST. As early as 1931, Ralph W. Tyler cautioned against the uncritical use of SST and pointed out that 'differences which are statistically significant are not always socially important. The corollary is also true: differences which are not shown to be statistically significant may nevertheless be socially significant' (1931, pp. 116-117). In fact, as reported by Pearce (1992), criticisms of SST began immediately with Fisher's introduction of it in 1925. Joseph Berkson (1942)[132], as a physician and a practitioner who has frequently applied SST, has noticed the logical problem of SST[133] and raised a number of queries about its use[134]. In 1957 Lancelot Hogben's book-length critique of SST appeared. This book was regarded by Morrison and Henkel (1970) as 'a systematic and damaging attack on various probability practices in research' (p.3). These criticisms were however not heeded by most researchers. In fact, in the middle of the 20th century SST was still widely used and regarded as correct by the vast majority of researchers with good credentials although there were fragmentary queries about SST, in addition to Berskson (1938, 1941, 1942) and Hogben (1957)[135]. In this connection, many

[132] Berkson (1938; 1941) had begun his criticisms on SST when he discussed chi-square tests and tests of departure from normality in his two earlier papers. Based on his experience on applying the chi-square test, he has observed that 'when the numbers in the data are quite large, the *P*'s tend to come out small...If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large – for instance, on the order of 200 000 – the chi-square *P* will be small beyond any usual limit of significance' (1938, p.526).

[133] For example, Berkson (1942) has argued that the argument behind SST is 'basically illogical' by considering the symbolic form: 'It says "If *A* is true, *B* will happen sometimes; therefore if *B* has been found to happen, *A* can be considered disproved." There is no logical warrant for considering an event known to occur in a given hypothesis, even if infrequently, as disproving the hypothesis' (p.326). Similar query has been raised in Berkson (1938, pp.530-531). We will return to this argument in later sections.

[134] For Fisher's comment on Berkson's criticism of SST and Berkson's response, see respectively Fisher (1943) and Berkson (1943).

[135] For instance, see Geary, 1947; Selvin, 1957; Yates, 1951. There were also several papers criticizing the abuses of the one-tailed test of significance, see Eysenck, 1960; Goldfried, 1959 for details.

researchers in the forties and fifties of the 20th century did not take the criticisms of SST seriously (Spielman, 1974).

In the sixties of the 20th century, more criticisms of SST had been accumulated[136] and this led Clark (1963) saying that 'the shortcomings in the methodology of statistical hypothesis testing used in educational and psychological research have been emphasized repeatedly in recent behavioral science and statistical literature' (p.455) and Bakan (1966) claiming that his arguments against SST is hardly original and 'in a certain sense, what "everybody knows," and that 'to say it "out loud" is...to assume the role of the child who pointed out that the emperor was really outfitted only in his underwear' (p.423). The first systematic collection of papers on the controversy over the use of SST, edited by Denton E. Morrison and Ramon E. Henkel, was published in 1970. Inside the book a number of papers criticizing SST published prior to the seventies have been included. Having said that, the situation did not change much for the subsequent decades as Cohen said twenty eight years later after the publication of Bakan's paper, 'this naked emperor has been shamelessly running around for a long time' (1994, p.997).

Carver (1978) has identified a number of misinterpretations of SST and 15 years later he reported that the practices were hardly changed (Carver, 1993). Despite no big change in the practices, widespread criticism of SST appeared in the psychological literature (Gigerenzer and Murray, 1987; Lunt and Livingstone, 1989; Gigerenzer, 1993; Cohen, 1994; Dracup, 1995). In response to the criticism the American Psychological Association (APA) set up a task force on investigating statistical inference and a report

[136] See, for example, Binder, 1963; Grant, 1962; Lubin, 1962; McNemar, 1960; Rozeboom, 1960; Salvage, 1957.

was then published (Wilkinson & Task Force on Statistical Inference, 1999) suggesting a pluralistic approach with respect to different research methodologies, and a flexible approach to data analysis that emphasized the value of exploratory data analysis and graphical methods. Statistics was regarded by the Task Force as a means for communicating research findings and no statistical tests were suggested to be banned from journals although it was agreed that perhaps too much had been made of statistical testing.

Views on SST are diverse. Some argue that SST should be banned (e.g. Carver, 1978, 1993; Falk and Greenbaum, 1995; Hunter, 1997; Schmidt, 1996; Schmidt and Hunter, 1997; Shrout, 1997) but some defend that SST is elegant and useful (Abelson, 1997a; 1997b; Chow, 1996, 1998a, 1998b; Hagen, 1997, 1998; Rindskopf, 1997). There are also researchers who hold relatively moderate views. For example, Harris (1997) argued that SST could be very useful if we abandon two-valued logic, Cohen (1990, 1994) and Thompson (1996, 1999a, 1999b, 1999c, 1999d) argued that SST could still be meaningful and useful if more thought is given to formulating meaningful hypotheses at the front end of the research process. In this chapter, we will examine the major arguments against SST and critically analyze the rebuttal to these objections.

5.1 Is the null hypothesis always able to be rejected?

A number of decades ago, Berkson (1938, 1942, 1943), Bakan (1966), Edward et al. (1963), Meehl (1967) and Nunnally (1960) all have argued that the null hypothesis of no difference (or called the 'nil hypothesis') must be able to be rejected in behavioural research. As Murphy (1990) has noted, statistics textbooks and discussions of statistics

in the research literature often state or imply that the null hypothesis is never true or credible. For example, Twaite and Monroe (1979) claimed, in discussing the null hypothesis, that the researcher 'never believes that the null hypothesis is true' (p.230). As we have mentioned in Section 4.6, the reasons are twofold. First, finding two identical cases of anything in our world is extremely rare. As Bakan (1966) has put it, 'a glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature' (p.426). Tukey (1991) alleged further that asking 'For any A and B, are the effects of A and B different?' is foolish since they 'are always different – in some decimal place'(p.100).[137]

Second, the nil hypothesis can always be rejected at some sufficient large sample size. Nunnally (1960) has pointed out 'if the null hypothesis is not rejected, it usually is because the *N* is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data' (p.643)[138]. *The Publication Manual of the American Psychological Association* (APA, 1994, p.18) also admits that sample size can drive rejection of the null hypothesis and thus recommends reports of effect size[139].

[137] For other researchers who also claimed that the null hypothesis cannot possibly be true, see, for example, Grant, 1962; Meehl, 1967; Murphy, 1990; Weitzman, 1984. Some, such as Anderson, et al. (2000) and Johnson (1995), even claimed that nearly all null hypotheses are false on a priori grounds.

[138] Nunnally (1960) has provided evidence for this contention: After a factor analysis of the results obtained in a study of public opinion consisting of 700 subjects, he calculated the correlation coefficients of the factors with age, sex, income and a number of other variables and found that nearly all correlations were significant, including those that made little sense (p.643). Similar evidences could be found in Bakan, 1966, p.425; Cohen, 1994, p.1000; Meehl, 1967, p.109; 1990, p.212. See also Streiner, 2003 for his analogy: 'sample size is like magnification with a microscope: the smaller the object that's being observed, the more magnification we need.' (p.761).

[139] Nevertheless, as suggested by numerous empirical studies of papers published since 1994 in different fields (such as psychology and education), merely 'recommending' has not greatly influenced reporting practice (Thompson, 1998, p.799).

If these two reasons were sound, all null hypotheses must be false or could always be rejected. Controlling the Type I error would be insignificant and rejection of null hypotheses would then become a trivial exercise. And these would constitute a major criticism of SST (Kirk, 1996). Before discussing whether the first reason is really justified, we have to look into the logic behind it. There are infinite numbers of real numbers in any closed interval of real numbers, say [–20, 20]. If we randomly select a real number from the interval, the probability of selecting any particular real number, such as 0, will be zero.[140] According to Frick (1995a), it is this mathematical fact that renders some thinking that the probability of zero difference between two effects is zero, as zero difference is merely one point in the interval of infinite possible differences between two effects. For example, Murphy (1990) explicitly stated that 'the null hypothesis represents a zero effect, taken to an infinite level of precision, and the alternative hypothesis includes everything else. This type of null hypothesis is clearly not very credible; the prior probability of a hypothesis such as this should indeed be zero' (p.404). Despite the misconception that an event having a zero probability implies the event being impossible to happen,[141] this mathematical fact can hardly be applicable to the case of hypothesis unless the assumption, random selection, has been met. If the selection of a number is not totally random, for example, a few special numbers can be allocated a nonzero probability before spreading the remaining probability over the

[140] It is worth noting that the probability here is exactly zero, not 'essentially zero' as asserted by Frick (1995a). Indeed, the probability of selecting a finite subset or a countable infinite subset of the interval will also be zero. Furthermore, Frick (1995a) was false in stating that 'only an *interval* of real numbers can receive a nonzero probability' (p.133). It is a well-known mathematical fact that the set of irrational numbers in [-20, 20], for instance, is not an interval but it could still receive a nonzero probability, 1.

[141] As we have mentioned in Section 4.5, it is true that if an event E is impossible to happen, then $P(E) = 0$. But its converse is false. For example, it is possible to randomly select a rational number from [-20, 20] though the probability of a rational number being selected is exactly 0. It seems that in their exchanges both Frick (1995a, 1995b) and Edgell (1995) did not note this point. Frick (1995a, 1995b) used to consider whether a hypothesis would be possible to be true, without knowing that the probability of the hypothesis being possible to be true could still be zero.

remaining real numbers, the probability of selecting any one of the special numbers will certainly not be zero. Zero is one of these special numbers.[142] It is thus not a priori truth that the probability of a nil null hypothesis must be zero.

Moreover, if we merely consider whether a null hypothesis is not possible to be true we may leave the issue of zero probability aside. When the null hypothesis is about manipulation of merely one variable, it is possible for the variable to have nil effect and therefore it is possible for the nil null hypothesis, one variable has nil effect on another variable, to be true[143]. For example, suppose that we want to investigate whether the first digit of a student's Identity Card number would have an effect on the student's performance in mathematics, we see no reason why in this case nil effect is not a possible hypothesis. In the case where manipulation involves two or more variables, the situation is more complicated. For example, suppose we are going to compare the effects of the Abacus and Mindabacus Course and Kumon Mathematics Course, two prevalent courses in Hong Kong, on students' ability in problem solving. Since it is possible for both courses to have nil effect on students' ability, the null hypothesis that there is no difference between their effects is possibly true. However, suppose further that Kumon Mathematics Course provides merely drilling of routine problems and causes a decline of students' ability in problem solving, which involves mainly novel problems. The null hypothesis would be true only if the Abacus and Mindabacus Course has exactly the same detrimental effect. If the effect is measurable on a scale of real numbers, it is very

[142] Frick (1995a) has used the hypothesis about the probability of the average height a species can jump as an example to illustrate this point: although the probability of the average height humans can jump being *exactly* 0.5 m could be zero, the probability of the average height oaks can jump being exactly 0 m need not be zero.

[143] Unless we assume radical views, like what Hays (1981) has asserted – 'there is surely nothing on earth that is completely independent of anything else. The strength of association may approach zero, but it should seldom or never be exactly zero' (p.293).

unlikely that the two real numbers are exactly identical. That may explain why some researchers, such as Bakan (1966), Cohen (1990), Grant (1962), Murphy (1990), and Weitzman (1984), have claimed that the null hypothesis cannot possibly be true. As a result, it is not true that the nil null hypothesis cannot possibly be true. Whether a nil null hypothesis can be true depends on the content of the hypothesis – whether it involves manipulation of one variable and in the case of complex variables whether it is known that at least one variable has an effect. In other words, the first reason for claiming that the null hypothesis is generally false is not tenable. How about the second reason?

Researchers, such as Bakan (1966), Nunnally (1960) and Meehl (1967), have presented different examples to demonstrate their contention that a large enough sample size will always result in rejection of the null hypothesis. But William Oakes (1975) has argued that these examples are all involving the self-selected-groups (SSG) design[144] and their contention is not supported by evidence resulting from the use of a true experimental design[145]. In an SSG design, groups are selected on the basis of a difference on one subject variable, i.e., the independent variable. The groups are then compared on various other subject variables, i.e. the dependent variables. Since the independent variable is itself casually determined by some combination of other (extraneous) variables, it is highly probable that these extraneous variables would exert their influence not only on the independent variable but also the dependent variables. If it is the case, a comparison of the groups with respect to any of those dependent variables whose

[144] A research study is said to be utilizing the self-selected-groups design whenever the level of the independent variable is determined for each subject by the subject's own characteristics (W. Oakes, 1975, p.267).

[145] A true experimental design involves a manipulated independent variable of which the levels are assigned at will to subjects by the experimenter.

determining variables overlap with those determining the level of the independent variable will find the groups differing with respect to that particular dependent variable. It follows that, no matter how small the difference in the dependent variable values for the independent variable groups is, the null hypothesis will be rejected if the sample size is large enough.

From above discussion, we could note that a large enough sample size could result in rejection of the null hypothesis only if there is some fixed, nonzero amount of difference in the dependent variable values for the independent variable groups (Oakes, 1975). But does this fixed difference necessarily occur in all experimental designs? In a true experimental design, the subjects are assigned at random to different groups. Subsequently, different levels of the independent variable are induced in subjects by the experimenter's manipulations according to the groups the subjects are belonging to. The random procedure will ensure that there is no relationship between the basis for assignment of groups and measure of any dependent variable. In other words, before manipulation, the probability of the groups differing on any variable is independent of the sample size. If the independent variable does indeed have no effect upon the dependent variable, the probability of rejecting the null hypothesis will not increase with an increase in the sample size. It is thus possible that, as elaborated by Oakes (1975), after getting 'almost significant' results in conducting a true experiment increasing the sample size would have the effect wash out as non-significant with the larger sample size.

Accordingly, not all null hypotheses must be false on a priori grounds or they must be able to be rejected with sufficiently large sample size even though many SST

reported in the research literature[146] are associated with null hypotheses that are not plausible, as Kline (2005) expected. What we have to be cautious about is that the null hypothesis should not be a straw man of which the rejection could hardly advance science.

5.2 Is statistical significance difference not necessarily an important difference?

According to Kirk (2001), researchers are interested in answering three basic questions when examining the relationships between variables. First, is an observed effect real or should it be attributed to chance? Second, if the effect is real, how large is it? Third, is the effect large enough to be useful? SST could only tell us the probability of obtaining the effect or a more extreme effect if the null hypothesis is true. If the null hypothesis is a nil hypothesis, SST could be used to address the first question. But, in this case, SST can hardly provide any information about the magnitude of the effect or whether the effect is important or not. As Tyler (1931) has pointed out, a statistically significant difference is not necessarily an important difference, and a difference that is not statistically significant may be an important difference. Since then this warning has been repeated many times (Carver, 1978). For example, Gold (1969) alleged that 'statistical significance is only a necessary but not sufficient criterion of importance' (p.44).

As we have discussed in previous chapters, it is certainly mistaken to interpret the *p* value, Type I error, or the level of significance as a measure of effect magnitude[147].

[146] For example, after reviewing the null hypotheses tested in several hundred empirical studies published from 1978 to 1998 in *Ecology* and the *Journal of Wildlife Management*, two prominent environment sciences journals, Anderson et a. (2000) found that there are overwhelming occurrence of false null hypotheses in their samples of articles.

[147] It is unfortunate that outcomes with lower *p* values are still sometimes interpreted as having stronger effects than those with higher *p* values. Oakes (1986) has found that not only students but also researchers

But that SST is unable to measure effect magnitude does not constitute a serious challenge to SST, especially when SST is purported to address only the first question. It is worth noting that the first question is more fundamental than the other two. Without knowing whether the observed effect is genuine, it is unjustified to ask if the effect is large or large enough to be useful. It is thus inapt to conclude that the rejection of a null hypothesis is not very informative.[148] Moreover, the null hypothesis is not necessarily nil. In case when we really want to test whether certain effect is present, we could set up a particular hypothesis which states that such an effect does not exist. A rejection of this null hypothesis could be an indication of the existence of this particular effect magnitude although whether it would really succeed in providing this information will be discussed in later sections.

5.3 Is SST indispensable?

Some believe that without SST, we are not able to tell when a finding is worth interpretation and it is less possible for different researchers to arrive at the same decision or conclusion from data (see, for example, Davis, 1958). And to some researchers SST does serve a useful purpose – addressing the question, 'is an observed effect real or should it be attributed to chance?'. For example, Mulaik, Raju and Harshman (1997) asserted that 'We cannot get rid of significance tests because they provide us with the criteria by which provisionally to distinguish results due to chance variation from results that represent systematic effects in data available to us' (p.81). It is certainly pleasing if

in psychology would overestimate the size of the effect based on a significance level from 0.05 to 0.1. Mittag and Thompson (2000) have, however, found in the AERA survey that respondents strongly disagreed that *p* values directly measure study effect size. For discussion of the difference between these two studies, see Gliner, Leech, & Morgan, 2002.

[148] Many have made such conclusion. See, for example, Carver, 1978; Oakes, 1986; Kirk, 1996; Shulman, 1970.

we have a statistical procedure that could be used to judge whether an observed effect is real or merely attributed to chance though some[149] are very doubtful whether it is possible to have such procedures. However, even though such procedures exist, it does not imply that SST is the one that can perform the feat.

We will first discuss the Neyman-Pearson hypothesis testing approach and leave the discussion of Fisher's significance testing to the next section. According to a number of studies,[150] the average power of SST in research literature is between 0.4 and 0.6. With a power within this range, Schmidt and Hunter (1997) argued that about half of all tests in a research literature will be non-significant. This is certainly mistaken. What it means should be: amongst the false null hypotheses about half of them will not be rejected. Nevertheless, they were correct in concluding that coin flipping would in many cases provide a higher level of accuracy than SST.

One may argue that this is not a defect of SST, but a problem of low statistical power. Using large enough sample sizes would be able to ensure high power, but at the same time, as argued by Schmidt and Hunter (1997), the effect sizes or relations examined in most research are small enough that power of even 0.8 requires more subjects than are often feasible to obtain. And it will render most studies being impossible to be conducted in reality.

[149] See, for example, Schmidt & Hunter, 1997.
[150] For instance, Cohen, 1988; Schmidt, 1996; Schmidt, Hunter, & Urry, 1976; Sedlmeier & Gigerenzer, 1989.

5.4 Does *p* value provide purported evidence against the null hypothesis?

The logic of Fisher's significance testing is that from the truth of the null hypothesis we can deduce that the probability of certain event E is very small. Hence, either the null hypothesis is false or rare event E occurs. If the event E really occurs, it will provides evidence against the null hypothesis. If it is the case, the following argument will work as well: for any outcome E of tossing a coin for 10 times,

If the coin is fair, the probability of E will be very small (0.5^{10}).

E really occurs

Hence, the die is not fair.

This argument still runs no matter what the outcomes are and no matter how fair the coin is. Fisher's p value is NOT the probability of the observed outcome under the assumption that the null hypothesis is true, as the probabilities of the outcomes more extreme than the observed one are also included in calculation of the p value. We will show in this section that using the tail region to represent a result that is actually on its border will make the case against the null hypothesis appear much stronger than it in fact is. Consider a simple example: assuming that the null hypothesis and the alternative hypothesis are as follows:

H_0: the distribution is the normal distribution $N(0, \sigma^2)$

H_1: the distribution is the normal distribution $N(d, \sigma^2)$, where $d > 0$.

Suppose that the test is a two-tailed test with known direction of the effect and the alternative hypothesis is the one against which the hypothesis testing has 90% power and the predetermined Type I error rate is 0.05, the probabilities 0.9 and 0.025 are represented by the areas of the two regions as shown in Figure 5:

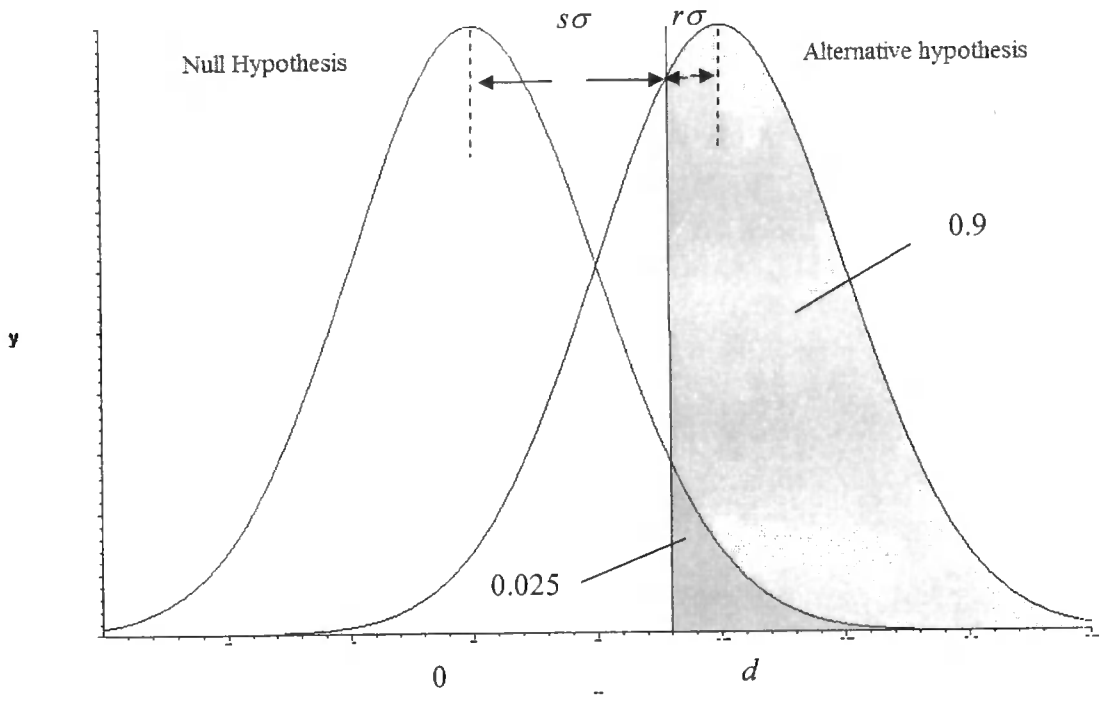


Figure 5 The graph showing the curves of the normal probability density functions for outcomes under the null hypothesis and the alternative hypothesis (Not in scale)

Let the vertical line be the line corresponding to Type I error rate 0.05. That is to say, it will cut the curve for the null hypothesis (the left curve in Figure 5) such that the area of the right tail region is 0.05/2 = 0.025. If the vertical line is at a distance $s\sigma$ from the mean of the distribution under the null hypothesis, where σ is the standard deviation of the two distributions under the two hypotheses, then it is clear that $P(s \leq z) = 0.025$.[151] Since the hypothesis testing is assumed to have 90% power when the predetermined Type I error rate is 0.05, the vertical line will cut the curve for the alternative hypothesis (the right curve in Figure 5) such that the area of the right tail region is 0.9. Suppose the line is at a

[151] With the use of Maple 8, we can find that the value of s is approximately 1.959964.

distance $r\sigma$ from the mean of the distribution under the alternative hypothesis, then $P(z \geq -r) = 0.9$. Since normal curves are symmetrical, we have $P(z \leq r) = 0.9$.[152] It is clear that the distance between the two vertices of the curves is $(s + r)\sigma$.[153] If an outcome with p = 0.02 has been observed, the probability associated with one of the two tail regions is 0.01. Using the likelihood ratio (or Bayes factor)[154] as a measure of the relative evidential support given by the data to the two hypotheses, we could compare the relative evidential supports given by the precise outcome (p = 0.02) and the imprecise outcome ($p \leq 0.02$) corresponding to the tail regions of the two curves.

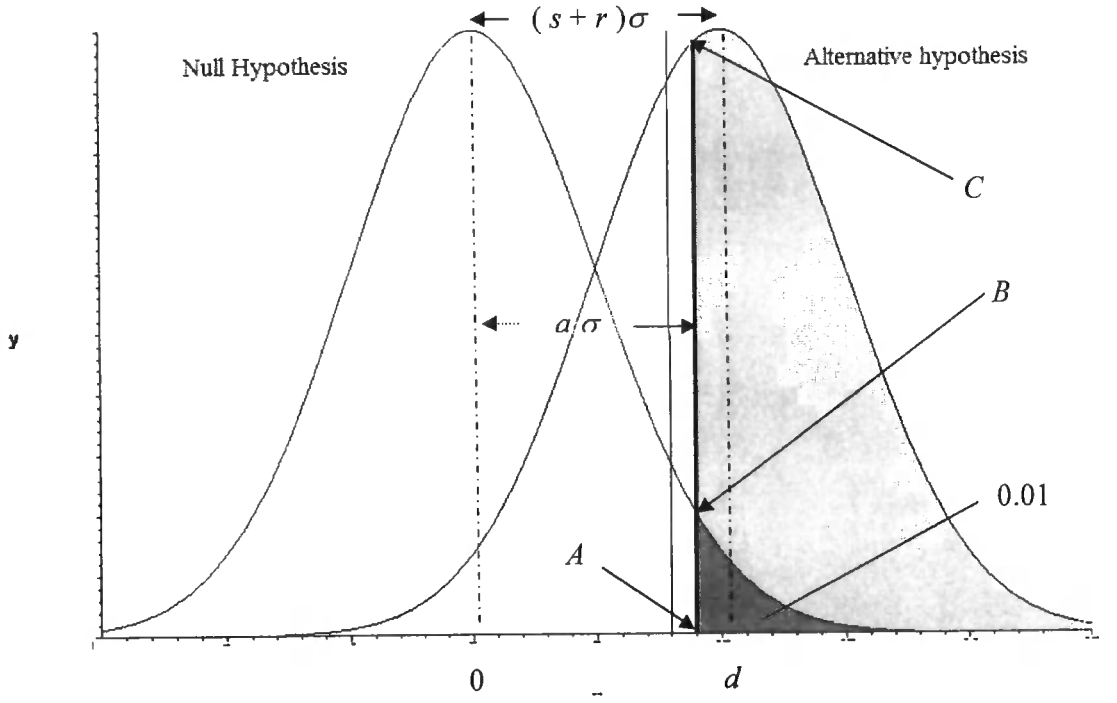


Figure 6 The graph showing the likelihood ratios associated with the precise p value (p = 0.02) and the imprecise p value ($p \leq 0.02$) (Not in scale)

[152] The value of r is approximately 1.281551.
[153] The value of $r + s$ is approximately 3.241515. It means that the value of d is approximately 3.241515σ.
[154] See Section 2.2.

The likelihood ratio associated with the precise p value (i.e. $p = 0.02$) is measured by the ratio of curve heights at the observed data , i.e. $\frac{AB}{AC}$ as shown in Figure 6. It is clear that the length AB corresponds to $\phi(a)$, where ϕ is the probability density function of the standard normal distribution[155]. On the other hand, the length of AC corresponds to $\phi(s+r-a)$ [156] as the distance between the vertical line AC and the mean of the distribution under the alternative hypothesis is $(s + r - a)\sigma$. Therefore, the likelihood ratio of the null hypothesis to the alternative hypothesis of a difference associated with 90% power, $\alpha = 0.05$, under the experimental result with the precise p value 0.02 is given by:

$$\frac{\phi(a)}{\phi(r+s-a)} \approx \frac{0.026652}{0.262448} \approx 0.1016 .$$

On the other hand, the likelihood ratio associated with the imprecise p values ($p \leq 0.02$) is measured by ratio of the small shaded area to the total shaded area, as shown in Figure 6. Obviously, the small shaded area is $P(z \geq a) = 1 - \Phi(a) = 0.01$.[157] The total shaded area is given by $P(z \geq -(r+s-a)) = P(z \leq r+s-a) = \Phi(r+s-a)$. Therefore, the likelihood ratio of the null hypothesis to the alternative hypothesis of a difference associated with 90% power, $\alpha = 0.05$, under the experimental result with the imprecise p value ($p \leq 0.02$) is given by:

$$\frac{P(z \geq a)}{P(z \leq r+s-a)} = \frac{1-\Phi(a)}{\Phi(r+s-a)} = \frac{0.01}{\Phi(r+s-a)} \approx \frac{0.01}{0.819948} = 0.0122.$$

---

[155] By solving the equation $\Phi(a) = 1 - 0.01$, where Φ is standard normal cumulative distribution function, we could find that the value of a is approximately 2.326348. Therefore, $AB \approx \phi(2.326348) \approx 0.026652$.

[156] $AC = \phi(s + r - a) \approx \phi(3.241515 - 2.326348) \approx 0.262448$

[157] See footnote 155.

As we have discussed in Section 2.2, the smaller the value of the likelihood ratio (of the null hypothesis to the alternative hypothesis), the greater the evidence is said to be against the null hypothesis (or in favour of the alternative). The above results thus indicate that in this particular case the actual outcome provides much less relative evidential support against the null hypothesis than that of the outcomes corresponding to the tail region. The ratio of these two likelihood ratios u is $0.1016/0.0122 \approx 8.33$. It means that the likelihood ratio for the precise p value (i.e. the ratio of heights of the two probability densities at the observed data) is about 8.33 times less evidence against the null hypothesis than does the likelihood ratio for the imprecise p value (i.e. the ratio of areas of the probability density functions beyond the observed data).

From Figure 7 in which the ratio of the two likelihood ratios (u) is plotted against different p values (the power and the Type I error rate remain the same), we can see that the ratio is always greater than 1 (the horizontal line $u = 1$ is also plotted) in the range from 0.0001 to 0.1.[158] The results that the ratio is greater than 1 are obtained for different powers and Type I error rates. See Figure 8 in which u is plotted against different p values and predetermined Type I errors (the power = 0.9) and Figure 9 for different p value and different powers (the Type I errors = 0.05).

[158] The graphs are plotted with Maple 8. See Appendix 12 for the Maple Worksheets.

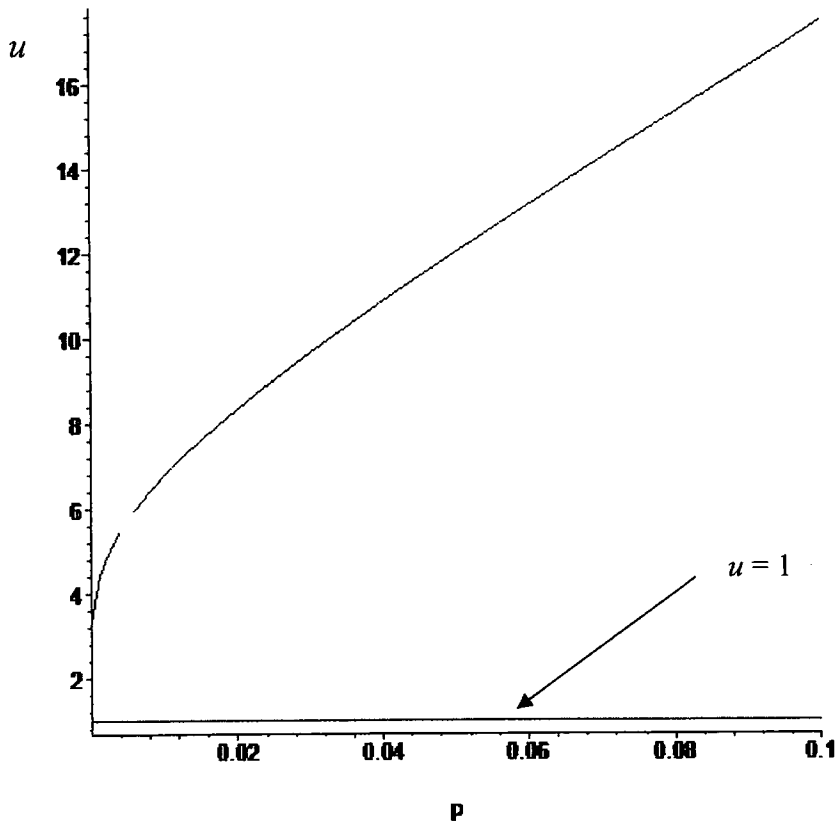


Figure 7 The graph showing the ratio of the two likelihood ratios (u) under different p value (from 0.0001 to 0.1)

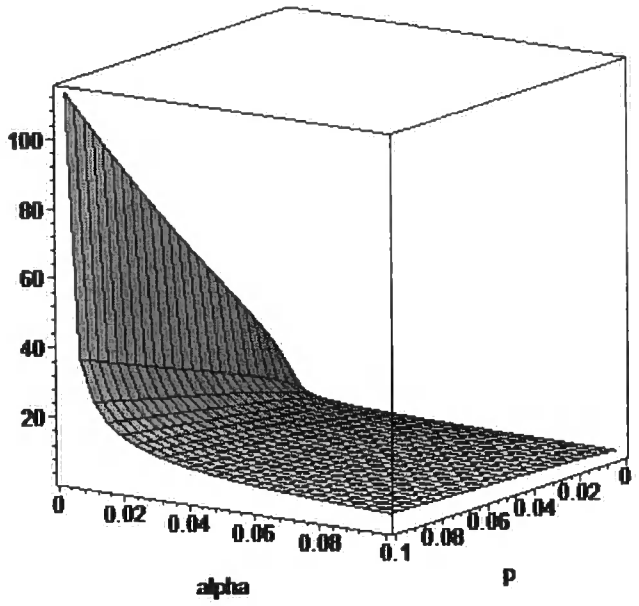


Figure 8 The graph showing the ratio of the two likelihood ratios under different p values (from 0.001 to 0.1) and Type I error rates α (from 0.001 to 0.1).

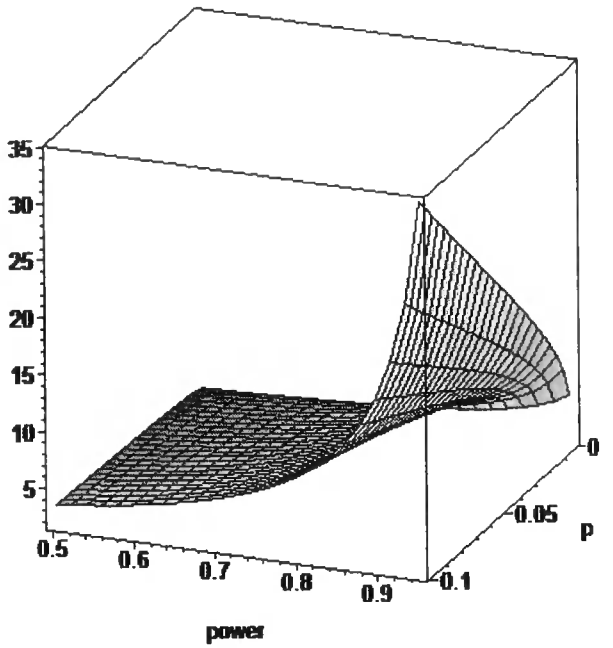


Figure 9 The graph showing the ratio of the two likelihood ratios under different p values (from 0.001 to 0.1) and power (from 0.5 to 0.95).

As we have noted above, different alternative hypotheses will produce different likelihoods for the same null hypothesis. Can likelihood ratios be used in lieu of p values? In fact, for a given null hypothesis and observed data, there is an alternative hypothesis which will result in a maximum likelihood – the one whose mean is equal to the observed estimate (see Figure 10).

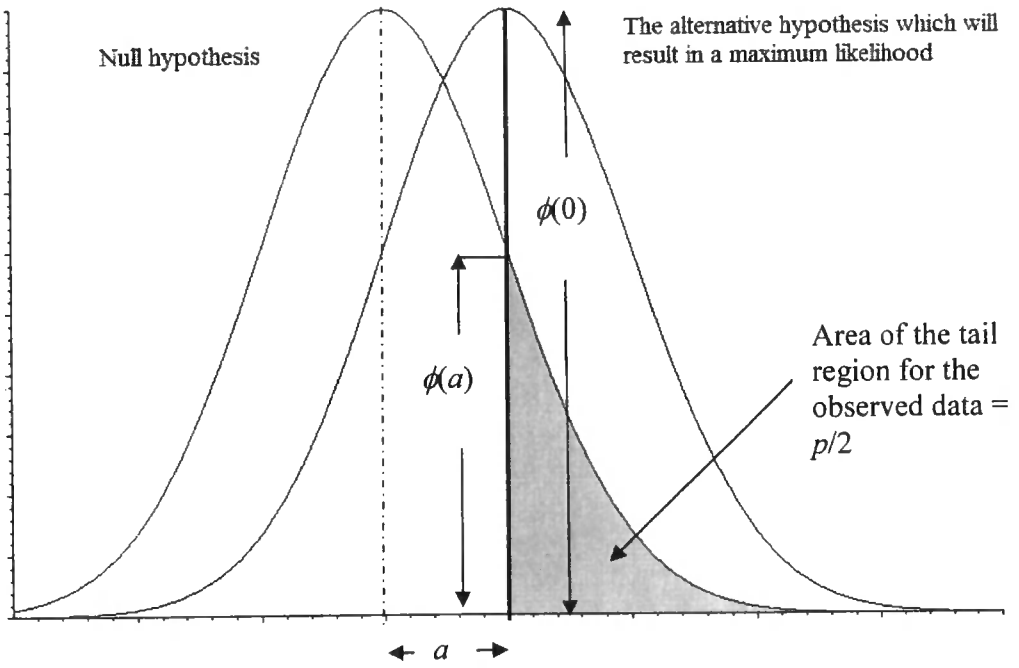


Figure 10 The graph showing the null hypothesis and the alternative hypothesis which will result in a maximum likelihood for the observed data

This so-called 'standardized likelihood' will represent the greatest degree of evidence against the null hypothesis. It can be easily shown that in this case the ratio of these two likelihood ratios u for the outcome ($z = a$) becomes

$$\frac{\phi(a)}{p\phi(0)}, \text{ where } p/2 = 1 - \Phi(a)$$

For example, for the outcome with $p = 0.05$ occurs, the ratio of the two standardized likelihood ratios is $\frac{\phi(1.96)}{0.05\phi(0)} \approx 2.93$. A plot of the ratio against different p values is shown in Figure 11:

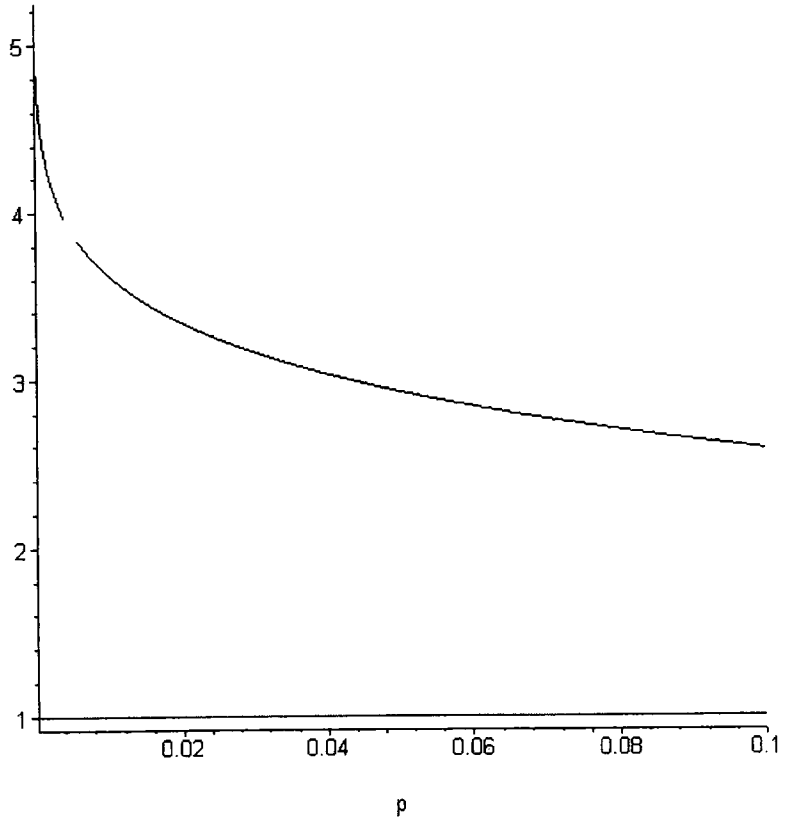


Figure 11 The graph showing the ratio of the two standardized likelihood ratios under different p values (from 0.001 to 0.1).

We can observe from what we have discussed above the likelihood ratio for the precise p value (i.e. the ratio of heights of the two probability densities at the observed data) provides much less evidence in support of the null hypothesis than does the likelihood ratio for the imprecise p value (i.e. the ratio of areas of the probability density functions beyond the observed data). The use of tail region to represent a result that is actually on the border will thus overstate the evidence against the null hypothesis.

5.5 The logical fallacy behind hypothesis testing

Many supporters of SST would agree that SST have been misused or misinterpreted by some researchers but the test itself is not inherently misguided or flawed[159]. We will see in turn whether they are correct. Suppose that a pregnant woman of age 38 who is worried about her baby having a serious disease (say, Down's syndrome) consults a doctor. The doctor recommends the woman to receive a prenatal amniocentesis test[160]. According to the doctor, amniocentesis is a very effective test to determine whether the baby has Down's syndrome or not. The test is conducted and the result is negative. The doctor then comforts the woman by reporting the negative result and assuring that, if the baby has Down's syndrome the result would very probably have been positive. In this case it seems that we should conclude with high probability that the baby does not have Down's syndrome. On the other hand, if the test result is unfortunately positive it seems that we should conclude with high probability that the baby does have Down's syndrome. The inference, that seems to be prevalent in our daily life and straightforwardly valid, is indeed the heart of the argument of SST – the assumption that H_0 is true will lead to an improbable result D. If D really occurs it will show that the probability that H_0 is true is very small, and therefore leads to the rejection of the assumption that H_0 is true. This argument is similar to a proof strategy commonly used in mathematics or mathematical logic – reductio ad absurdum[161], also known as

[159] See, for instance, Abelson, 1997a; Cortina & Dunlap, 1997; Frick, 1996; Hagen, 1997.

[160] Amniocentesis is the process in which a needle is passed through the mother's lower abdomen into the amniotic cavity inside the uterus so as to withdraw some amniotic fluid from around baby. The amniotic fluid can then be used for testing for certain conditions or birth defects in the baby.

[161] Strictly speaking, reductio ad absurdum is a derived rule of inference (derived from the axioms and the rule of inference – modus ponens) or a metatheorem, i.e. a theorem about an axiomatic system (for the distinction between a theorem and a metatheorem, see Hunter, 1971, p.11). For example, in the formal axiomatic theory L for the propositional logic (see Mendelson (1997, pp.35-36) for details of L), reductio ad absurdum can be expressed as:

'if $\Gamma, p \vdash_L q \wedge \sim q$ then $\Gamma \vdash_L \sim p$' or 'if $\Gamma, p \vdash_L \sim p$ then $\Gamma \vdash_L \sim p$'.

'indirect proof' or 'proof by contradiction'[162]. Reductio ad absurdum is a type of valid logical argument where we assume a premise for the sake of argument, arrive logically at an absurd result (such as a contradiction), and then conclude the original assumption must have been false and thus has to be rejected, since it gives us this absurd result. For example, in elementary number theory we prove that there is no greatest prime number, by assuming that there is a greatest prime number and proceeding to prove that such an assumption will lead to an absurdity – there is a number[163] that is both prime and not prime, and therefore we have to reject the original assumption that there is a greatest prime number.

Some, including textbook writers, have indeed made an analogy between SST and argument by contradiction (for instance, Freedman, Pisani, and Purves, 1998; Harshbarger, 1977; Reeves & Brewer, 1980) and some explicitly claim that the logic behind SST is 'an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must therefore be rejected. ..... The null hypothesis is creating absurdities, and should be rejected. In general, the smaller the observed significance level, the more you want to reject the null. The phrase "reject the null" emphasizes the point that with a test of significance, the argument is by contradiction' (Freedman, Pisani, and Purves, 1998, p.482). Though most of them do not really maintain that the null hypothesis rejected in SST is definitely false, they do believe

---

Here, '$\Gamma, p \vdash_L q$' means that q could be derived in the formal axiomatic system L from Γ and p. It is, of course in a stringent sense, mistaken to characterize reductio ad absurdum as tantamount to the theorems like '$(p \rightarrow (q \wedge \sim q)) \rightarrow \sim p$' or '$(p \rightarrow \sim p) \rightarrow \sim p$' (such conflation could be found in some elementary logic textbooks, for instance, Tymoczko and Henle, 1995, p.94).

[162] See Quine (1982, p.194) for the historical note of reductio ad absurdum (or indirect proof).

[163] This number is $p_1 \ldots p_{n-1} p_n + 1$ where p_n is the greatest prime number and $p_1, \ldots, p_{n-1}$ are the rest of all prime numbers that are less than p_n. The details of the proof could be found in any standard textbooks, such as the one coauthored by the author - Man, Leung , and Ng, 1997, p.127.

that the null hypothesis has been rendered improbable and that explains why it has to be rejected. Despite the strong resemblance between these two argument forms – reductio ad absurdum is a valid argument form (or a metatheorem for an axiomatic system for the propositional or predicate logic) [164] but its probabilistic counterpart in SST[165] is not.

Consider the example of amniocentesis again: suppose the amniocentesis is very effective in the sense that the probability of a positive test result given the baby is being affected by Down's syndrome, and the probability of a negative test result given the baby is normal (not being affected by Down's syndrome) are both 0.995 (denoted by 'Γ') (Pauker and Pauker, 1979). Assuming that the baby is normal (denoted by '~DS') we would have a high probability (0.995) that the test result is negative (denoted by 'neg'). In other words, we have

$$\Gamma, \sim DS \vdash_{prob} \text{neg}.$$

The result is in fact positive. As $\sim\text{neg} \vdash_{prob} \sim\text{neg}$, we have

$$\Gamma, \sim DS, \sim\text{neg} \vdash_{prob} \text{neg} \wedge \sim\text{neg}.$$

---

[164] The proof of reductio ad absurdum is simple. For example, in the formal axiomatic theory L for the propositional logic, reductio ad absurdum could be proved as follows:

(1)	$\Gamma, p \vdash_L q \wedge \sim q$	
(2)	$\Gamma \vdash_L p \rightarrow q \wedge \sim q$	(1), deduction theorem
(3)	$\Gamma \vdash_L \sim(q \wedge \sim q) \rightarrow \sim p$	(2), law of the contrapositive
(4)	$\Gamma \vdash_L \sim(q \wedge \sim q)$	law of contradiction
(5)	$\Gamma \vdash_L \sim p$	(3), (4), modus ponens (QED)

[165] The probabilistic counterpart of reductio ad absurdum can be expressed as:

'if $\Gamma, p \vdash_{prob} q \wedge \sim q$ then $\Gamma \vdash_{prob} \sim p$'' or 'if $\Gamma, p \vdash_{prob} \sim p$ then $\Gamma \vdash_{prob} \sim p$''

where $A_1, A_2, \ldots, A_n \vdash_{prob} B$ if and only if it is logically impossible for the premises $A_1, A_2, \ldots, A_n$ all to be true and the conclusion B to have low probability (cf Sainsbury, 1991, p.105).

According to the probabilistic counterpart of reductio ad absurdum, we should conclude (from Γ and ~neg) with high probability that the baby does have Down's syndrome, i.e.

$$\Gamma, \sim neg \vdash_{prob} DS.$$

But is it the case? Suppose for women of age 38 the incidence of live born infants with Down's syndrome is 1/800. Then the situation could be represented in the following figure 9:

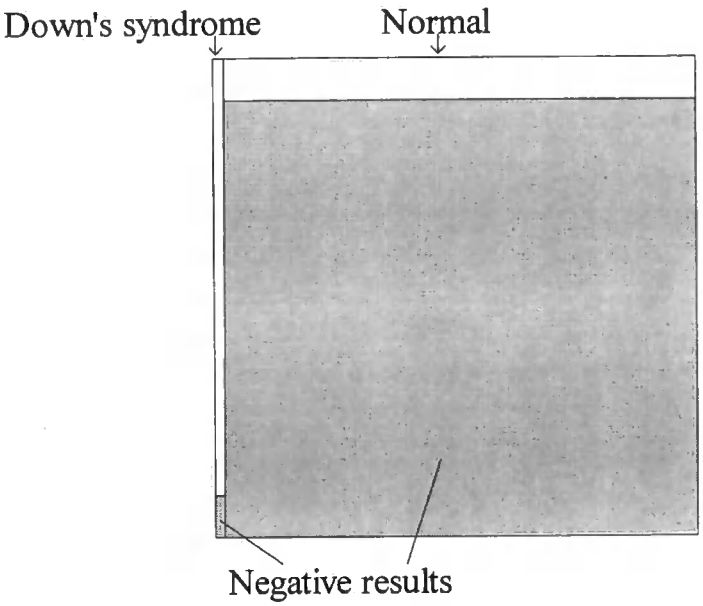


Figure 12 Down's syndrome Situation 1

The whole square represents the proposition that the infants are borne by women of age 38. The left and the right rectangles represent respectively the propositions that the infants are affected and the infants are normal; shaded regions represent the proposition that negative amniocentesis results were obtained. The areas of the regions in the diagrams represent the probabilities of the propositions they correspond to (though not

drawn in exact scale)[166]. For example, the ratio of the area of the left rectangle to the area of the right rectangle should be 1:800. From the figure, it is clear that the probability of an affected infant (left rectangle), given a positive test result (white region) is not very high[167]. In other words, even though 'Γ, ~DS, ~neg $\vdash_{prob}$ neg ∧ ~neg' is true 'Γ, ~neg $\vdash_{prob}$ DS' is false. This example suffices to show that the probabilistic counterpart of reductio ad absurdum is not a valid rule of inference (or argument form).

From the above analysis, we could observe that the fallacy occurs because we ignore the low base rate of DS in the particular population. If for women of age 38 the incidence of live born infants with Down's syndrome is 400/800 instead of 1/800 then the situation becomes:

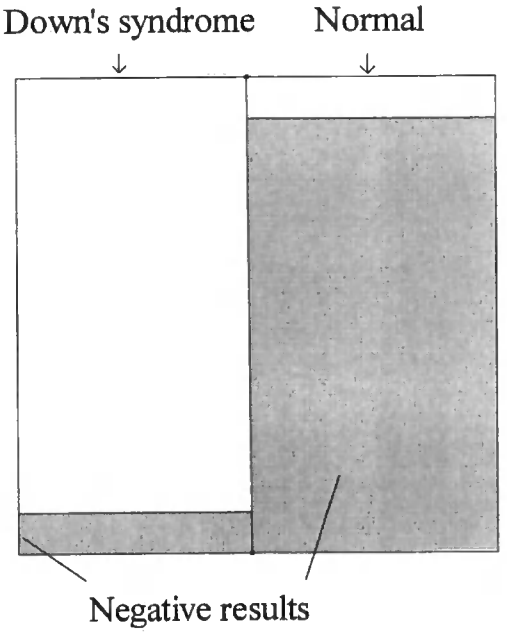


Figure 13 Down's syndrome Situation 2

[166] See Adams (1998, pp. 11-19) for details of how Venn diagram could be used to represent probabilities of compound propositions.

[167] Simple calculation will reveal that the probability is only 0.995×1/(0.995×1 + 0.005×799) = 0.199.

Now the probability of an affected infant (left rectangle), given a positive test result (white region) is very high (0.995) and 'Γ, ~neg $\vdash_{prob}$ DS' is thus no longer false. In fact, hundreds of studies conducted on the use of base rates in probability judgment tasks revealed that base rates were universally ignored (Tversky & Kahneman, 1982; Chistensen-Szalanski & Bushyhead, 1981). Not only laymen who have little or no knowledge on probability theory commit such error in probability judgment tasks, psychologists and even trained statisticians may ignore base rate probabilities unconsciously[168].

The logic behind the argument of SST could also be analyzed from two other perspectives. First, the argument could be phrased as:

> If H_0 is true, then the probability that the data D[169] will occur is very small.
>
> The data D occurs
>
> ---
>
> Hence, the probability that H_0 is true is very small.

For example, in our previous example of amniocentesis, the argument is:

> If a baby is normal, then the probability that the test result is positive is very small.
>
> The test result is positive
>
> ---
>
> Hence, the probability that the baby is normal is very small.

[168] The author (Wu, Ng, & Sze, 2003) has conducted a research project (in 2001-2002) to study if the student teachers majoring in mathematics education in the Hong Kong Institute of Education would still commit the base-rate fallacy and the result is consistent with others (such as Hacking, 2001).

[169] The data D here could mean an outcome in the rejection region of the null hypothesis H_0, rather than the point numerical value of the statistic obtained.

The argument is of the form:

If p is true, then the probability that q is true is very small.

q is true

Hence, the probability that p is true is very small.

Or, more precisely,

For any x, if $P(x)$ is true, then the probability that $Q(x)$ is true is very small.

$Q(a)$ is true

Hence, the probability that $P(a)$ is true is very small.

According to Salmon (2005), this form of inference is extremely important as it seems to represent a common form of inference in statistics. Giere (1979, p.97) even characterizes this form as the standard form that the good inductive arguments in scientific reasoning almost always have. This form looks very similar to *modus tollens* (MT)[170], a valid argument form which can be expressed as follows:

If p is true, then q is not true.

q is true.

Hence, p is not true.

[170] We have mentioned this form in Section 4.1. Here p and q can be substituted by the names of any two sentences. In the formal axiomatic theory L for the propositional logic, modus tollens is a metatheorem that could also be expressed as:

$$p \rightarrow \sim q, q \vdash_L \sim p$$

Or

For any x, if $P(x)$ is true, then $Q(x)$ is not true.

$Q(a)$ is true.

Hence, $P(a)$ is not true.

Notwithstanding their high similarity, the probabilistic modus tollens is not a valid argument form (or a metatheorem). Here is a counterexample showing its invalidity[171]:

If n is a non-negative integer less than 1000, then the probability that n is zero is very small.

n happens to be 0.

Hence, the probability that n is an integer less than 1000 is very small.

The first premise is true as the probability that a non-negative integer less than 1000 is zero is 1/1000, a very small number. If n happens to be 0, n must be a non-negative integer less than 1000. Hence, the probability that n is a non-negative integer less than 1000 is 1, which is certainly not very small. It thus shows that the probabilistic modus tollens is not valid. In order to illustrate this case with the use of Venn diagram. Let us consider another similar but finite case:

(P_1) If a person is working in Hong Kong, then the probability that the person is working at the University of Hong Kong (HKU) is small.

(P_2) Sheena is working at HKU.

(C) Hence, the probability that Sheena is working in Hong Kong is very small.

[171] See Cohen (1994), Falk and Greenbaum (1995), and Falk (1998) for similar examples. Moreover, our previous amniocentesis example could serve the same purpose.

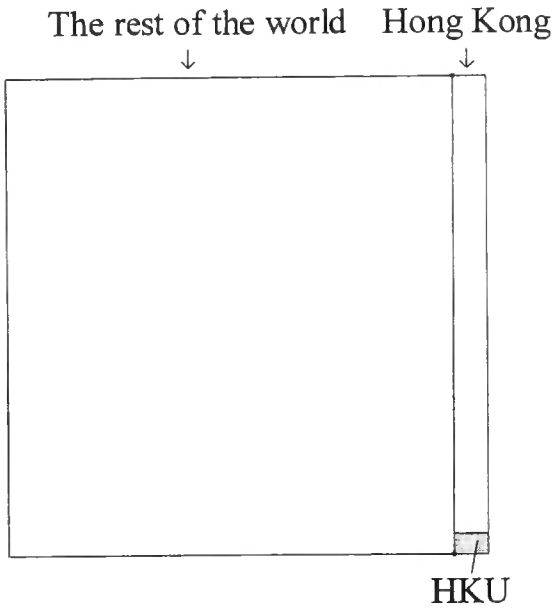


Figure 14 Probability of working at HKU

According to the ratio of number of staff members of HKU to the total number of employers in Hong Kong, the first premise (P_1) is clearly true. Thus the ratio of the areas of the shaded region to the right rectangle is small. Since all persons who are working at HKU must be working in Hong Kong, the whole shaded region in Figure 11 lies inside the right rectangle. Therefore, if Sheena is working at HKU then Sheena has to be working in Hong Kong. And the probability that Sheena is working in Hong Kong is thus 1 and it can hardly be very small. The conclusion (C) cannot be concluded from these two premises. Hence, this argument form is invalid.

Although the arguments against the probabilistic MT appear to be very persuasive, there are still some objections to the counterexamples. First, some[172] may argue that the consequent of the conditional[173] (P_1) is true in and of itself – the probability that a person

[172] For instance, Cortina and Dunlar (1997).

[173] A conditional is any sentence of the form 'if p then q' where p and q are called respectively the antecedent and the consequent of the conditional. (Blackburn, 1994, p.73)

is working at HKU is very small. As a result, almost any sentence could be used as antecedent of the conditional (P_1) and that its truth would not affect the truth of the whole conditional. If a person is working at HKU then that person has to be working in Hong Kong, and it is because of these two aspects of the particular example chosen that the probabilistic MT breaks down. Cortina and Dunlar (1997) have argued that there are a number of instances in which the probabilistic use of MT is correct. For example, they have cited the following as more representative of psychology: (1) If Sample A were from some specific population of 'normals,' the sample A probably would not be 50% schizophrenia. (2) Sample A comprises 50% schizophrenic individuals; therefore, (3) Sample A is probably not from the 'normal' distribution.

Cortina and Dunlar (1997) are mistaken on at least three points. First, for the conditional discussed above, the consequent is not true in and of itself. For example, if the antecedent happens to be 'a person is working at HKU', then neither the consequent nor the conditional will be true. Second, when we want to prove that a universal sentence is false (e.g., all prime numbers are odd), we will deliberately choose a counterexample (e.g., the prime number 2 is not odd). It is ridiculous for a critic to challenge, 'why don't you choose 'the prime number 11 is odd' as the example?' In order to prove that an argument form is invalid (i.e., not all arguments in this form will have true conclusions when their premises are true)[174], what we should do is to choose an argument of this form which has true premises but false conclusion as the counterexample and it doesn't matter whether or not there are many other arguments of this form whose premises and

[174] In other words, if an argument form is valid, then all arguments of this form must have true conclusion if their premises are true.

conclusion happen to be true.[175] Third, the argument cited by Cortina and Dunlar (1997) could in fact still be invalid. Consider an extreme but possible case: suppose there are only two populations in our world: (1) 'Normal' population which comprises 95% non-Schizophrenic and 5% Schizophrenic; (2) 'Super' Population which comprises 100% non-Schizophrenia. It is obvious that the sample A which comprises 50% schizophrenic is still probably from the 'normal' population.

The last perspective on the logic behind the reasoning of SST is about conditional probabilities. In SST, what we try to deduce is that when we obtain data D, such that $P(D \mid H_0)$ is low, $P(H_0 \mid D)$ is also rendered sufficiently low to guarantee rejection of H_0. But it is quite clear that the two conditional probabilities $P(D \mid H_0)$ and $P(H_0 \mid D)$ are completely different even though there are reports that many people still believe that these two conditional probabilities are equal[176] (Bakan, 1966; Birnbaum, 1982; Diamond & Forrester, 1983). Furthermore, even though one may be aware that the two conditional probabilities are in general unequal one may still think that a low value of $P(D \mid H_0)$ will warrant a low value of $P(H_0 \mid D)$. The reasoning is based on the probabilistic analogue of contraposition:

Same as reduction ad absurdum and MT, contraposition is also a valid argument form (or metatheorem)[177]:

'$A \rightarrow B$' if and only if '$\sim B \rightarrow \sim A$'

[175] Baril and Cannon (1995) have made a similar mistake in criticizing Cohen's (1994) selection of examples. See also Cohen's (1995) reply.

[176] For further discussion of this fallacy of equating two inverse conditional probabilities, see Down , 1988; Eddy, 1982.

[177] Sometime, contraposition could be used to refer the theorem '$A \rightarrow B \leftrightarrow \sim B \rightarrow \sim A$'. These two representations could be shown to be equivalent with the use of deduction theorem.

According to the probability conditional theory (Jeffrey, 1964; Ellis, 1973; Adams, 1998), the probabilities of conditional sentences like '$A \rightarrow B$' are conditional probabilities (e.g. $P(A \rightarrow B) = P(B \mid A)$). Hence,

$$P(B \mid A) = 1 \text{ if and only if } P(\sim A \mid\sim B) = 1$$

It is this relation that renders some inferring a high value of $P(\sim A \mid\sim B)$ from a high value of $P(B \mid A)$ or vice versa (i.e. $P(B \mid A) \approx 1$ if and only if $P(\sim A \mid\sim B) \approx 1$), and this inference is what we call 'the probabilistic analogue of contraposition' (P-contraposition). As a low value of $P(D \mid H_0)$ implies a high value of $P(\sim D \mid H_0)$ and a high value $P(\sim H_0 \mid D)$ implies a low value of $P(H_0 \mid D)$ (Rule*)[178], some could deduce from the probabilistic analogue of contraposition that a low value of $P(H_0 \mid D)$ would result from a low value of $P(D \mid H_0)$:

a low value of $P(D \mid H_0)$

$\Rightarrow$ a high value of $P(\sim D \mid H_0)$ (Rule*)

$\Rightarrow$ a high value of $P(\sim H_0 \mid D)$ (P-contraposition)

$\Rightarrow$ a low value of $P(H_0 \mid D)$ (Rule*)

That may explain why some think that a low value of $P(D \mid H_0)$ will warrant a low value of $P(H_0 \mid D)$ even if they are aware that the two conditional probabilities are unequal. However, similar to the probabilistic counterpart of reductio ad absurdum or MT, the probabilistic analogue of contraposition is also invalid. Our previous examples could demonstrate this point. Nevertheless, the inference from a high value of

[178] Since $P(D|H_0) + P(\sim D|H_0) = 1$, a low value of $P(D|H_0)$ will imply a high value of $P(\sim D|H_0)$. Similarly, $P(H_0|D) + P(\sim H_0|D) = 1$, a high value of $P(\sim H_0|D)$ will entail a low value of $P(H_0|D)$.

$P(\sim D \mid H_0)$ to a high value of $P(\sim H_0 \mid D)$ could become valid if an additional condition has been met. Consider the following figure:

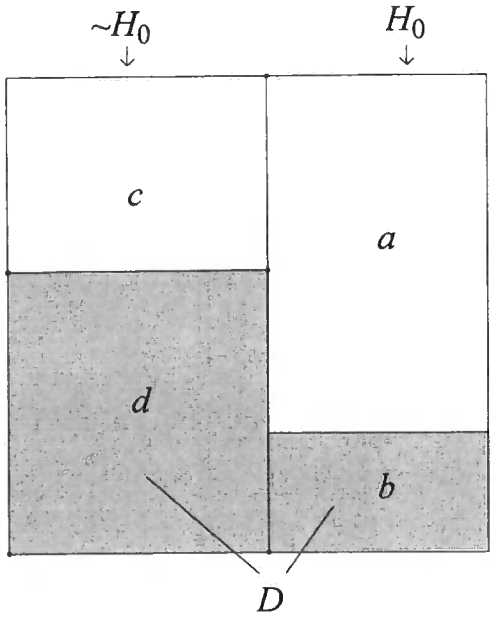


Figure 15 Conditional probabilities

Let a, b, c, d be the areas of the corresponding regions. In other words, a, b, c, d denote the probabilities of $H_0 \wedge \sim D$, $H_0 \wedge D$, $\sim H_0 \wedge \sim D$, and $\sim H_0 \wedge D$ respectively and their sum of these probabilities equals 1.

$$P(\sim D \mid H_0) \leq P(\sim H_0 \mid D) \text{ iff } \frac{a}{a+b} \leq \frac{d}{b+d}$$

$$\text{iff } ab + ad \leq ad + bd$$

$$\text{iff } a \leq d \quad (\text{provided } b \neq 0)^{179}$$

$$\text{iff } P(H_0 \wedge \sim D) \leq P(\sim H_0 \wedge D).$$

Also, $a \leq d$ iff $a + b \leq b + d$

$$\text{iff } P(H_0) \leq P(D),$$

and $a \leq d$ iff $2a + b + c \leq a + b + c + d$

[179] If $b = 0$ (and $a \neq 0$, $d \neq 0$), $P(\sim D \mid H_0) = P(\sim H_0 \mid D) = 1$ and the inequality must hold.

$$\text{iff} \quad (a+b)+(a+c) \leq 1$$

$$(\because a+b+c+d=1)$$

$$\text{iff} \quad P(H_0)+P(\sim D) \leq 1.$$

In other words, if $b \neq 0$ then the following 4 conditions are equivalent:

$P(\sim D \mid H_0) \leq P(\sim H_0 \mid D)$, $P(H_0 \wedge \sim D) \leq P(\sim H_0 \wedge D)$, $P(H_0) \leq P(D)$,

$P(H_0)+P(\sim D) \leq 1$.

That is, if $P(H_0) \leq P(D)$, $P(H_0 \wedge \sim D) \leq P(\sim H_0 \wedge D)$, or $P(H_0)+P(\sim D) \leq 1$ [180] then the probabilistic contrapositive $P(\sim H_0 \mid D)$ is greater than or equal to the probability of the original conditional $P(\sim D \mid H_0)$ and thus the inference from a high value of $P(\sim D \mid H_0)$ to a high value of $P(\sim H_0 \mid D)$ becomes valid. In any situation in which we would like to apply the probabilistic analogue of contraposition we have to consider if the condition has been met. The above analysis may help us resolve the paradox of confirmation proposed by Carl G. Hempel (1945a; 1945b)[181]. But when doing a realistic SST, how could we check if the conditions, say, $P(H_0) \leq P(D)$, have been met? We cannot simply assume that $P(H_0) \leq P(D)$ since is tantamount to assuming that $P(\sim D \mid H_0) \leq P(\sim H_0 \mid D)$, and it is exactly the point at issue.

[180] Reichenbach (1976, pp.129-133) has shown similar result.

[181] It is also called 'Hempel's paradox' or 'Raven paradox', which was proposed to illustrate a problem where confirmation theory violates intuition (or more precisely, a shortcoming of Nicod's criterion) – When we want to test the theory that all ravens are black, we have to go out and examine ravens and see if they are all black. But 'all ravens are black' are logically equivalent to 'whatever is not black is not a raven'. If a black raven is a confirming evidence for 'all ravens are black' then a white thing that is found to be non-raven (such as a shirt) should also be a confirming evidence for 'all ravens are black' too, which does clearly violate our intuition. For the discussion of how Hempel's paradox would vanish from a probabilistic point of view, see Reichenbach, 1976; Salmon, 2005.

Hagen (1997) also raised objections to the attack on the logic of SST. He argues that Cohen (1994) has tried to find a way to relate the probability of H_0 to 'countable', 'empirically based' relative frequencies and Cohen's effort led him to define H_0 and H_1 in ways that the SST does not allow. In our response to Cortina and Dunlar (1997), we have clearly shown that merely insisting that the hypotheses must be about the population is not able to save the logic of SST. Hagen (1997) also defends the logic of the SST by arguing that arguments can be reasonable and defensible even when they are not logically valid in a formal sense. It will be quite interesting to see how one could prove this point. But Hagen (1997) fails to do so. He merely gives an example to illustrate this point:

> If you contract AIDS, you will probably die of some opportunistic infection within 10 years.
> You did contract AIDS.
> You will probably die of some opportunistic infection within 10 years.
> This probabilistic argument is not formally logical because one could accept the premises but still reject the conclusion. The argument is, however, quite reasonable and defensible based on data (p. 22).

Our rejoinder is twofold. First, this argument is not probabilistic in the sense that the modal term 'probably' attached to the sentence plays no role in determining the validity of the argument. Second, this argument is indeed an instance of modus ponens which is certainly a valid argument form in propositional logic. That explains why this argument looks 'quite reasonable and defensible based on data' and no person could reasonably accept the two premises but reject the conclusion.

As Falk (1998) has argued, 'the faulty belief that a statistically significant result makes H_0 improbable and deserving rejection, is central to the NHST reasoning. If this belief is erroneous, the whole structure collapse: rejecting a hypothesis whose posterior probability is moderate or high is unacceptable' (p. 798).

5.6 Does probability provide an appropriate measure of the plausibility of a hypothesis

We have discussed the logic behind SST in the last section but there are, however, two intertwined assumptions that we have not yet challenged: (1) a hypothesis with high probability is something good or something we have to aim at; (2) the low probability of a hypothesis provides us a good reason to reject the hypothesis. As we have discussed in Chapter 3, talking of the probability of the hypothesis that is not an outcome of a chance process is indeed unintelligible and the attempt to identify the probability of a hypothesis with the probability of events is doomed to failure. But for the sake of argument, suppose we were able to assign probability value to a hypothesis in the same way as we assign it to an event. Should we then take it for granted that a hypothesis with high probability is really what we ought to aim at when we conduct research? First, if a hypothesis achieves its highest probability, i.e. 1, what it means is that H is a necessary truth. But only a hypothesis void of empirical content (e.g., the hypotheses which consist of solely analytic sentences such as 'all mathematics high achievers are high achievers in mathematics') could be a necessary truth. It is clear that no one will regard a hypothesis void of empirical content as a useful hypothesis that we ought to aim at in conducting empirical research. Second, testability is a virtue of hypotheses that we ought to aim at. A more testable hypothesis is one which can be better tested. If a hypothesis, going to one

extreme, is never susceptible of test (either verification or falsification), then it can hardly be regarded as scientific. Generally speaking, a hypothesis with more empirical content is more susceptible of falsification than the one with less empirical content. For example, 'all swans are white and all ravens are black' has more empirical content than 'all swans are white'. Every instance that can be used to falsify the second sentence (e.g., the existence of a white swan) can falsify the first sentence but not the vice versa (e.g., the existence of a non-black raven can falsify the first but not the second sentence). The first sentence is thus more susceptible of falsification than the second one. As a result, the greater the testability or the more empirical content the hypothesis possesses, the more improbable the hypothesis will be, The intuition that a hypothesis with high probability must be something good or something we have to aim at when conducting empirical research is thus not justified unless 'probability' here has another meaning – one that does not conform to the probability calculus. In fact, as argued by Popper (1983, p.225), we have to distinguish between the probability of a hypothesis with respect to its chance and the *probability* of a hypothesis with respect to its test. What we have already discussed is the first usage.

For the second usage of *probability*, a hypothesis will be considered to be more *probable* (or to be having a higher *probability*) if it can stand up to more severe tests (or it has some other virtues such as it is more simple and has greater explanatory power and fewer assumptions). It is trivial that under this interpretation a hypothesis with high *probability* must be something good or something we have to aim at and on the other hand the low *probability* of a hypothesis is a good reason for us to reject the hypothesis. In order to avoid confusion, we follow Popper's (1983) suggestion in using the term

'degree of corroboration'[182] to characterize the degree to which a hypothesis has been tested and leaving the use of 'probability' only for those concepts that satisfy the probability calculus (p.277). Can degree of corroboration be treated in terms of probability calculus? If a hypothesis H_1 implies (or is said to be stronger than) another hypothesis H_2, every refutation of H_2 will necessarily be a refutation of H_1, but not vice versa. It is thus impossible for the degree of corroboration of H_2 to be greater than that H_1. But logically H_1 will be more improbable than H_2.[183] Hence, even though corroboration could be treated in terms of probability calculus, it would be more closely related to the improbability of a hypothesis than to its probability (Popper, 1983, p.231).

As we have discussed in 3.6, Popper (1957/1980, 1983) has suggested a formula for the measure of the degree of corroboration. But he had also stressed many times that what really concerned him was not the way we define the degree of corroboration. He did not believe that such a numerical evaluation of this degree would have any practical significance, such as contribution to science.[184] There is thus no need for us to delve into its details here.

In conclusion, talking of probability of a hypothesis is itself questionable, unless we are willing to adopt a subjective interpretation of probability. But even so, a hypothesis with high probability is not what we should aim at when doing scientific research, as a hypothesis void of empirical content would achieve its highest probability,

[182] Popper had indeed used another label 'degree of confirmation' in his earlier writings. See Popper, 1983, pp.228-230 for the reasons why he later changed the term into 'degree of corroboration'.
[183] Here we still assume that we can talk about the probability of a hypothesis. It is as if we are talking about the probability of an event E, and in that case, if E_1 implies E_2, the probability of E_1 will not be greater than that of E_2.
[184] See Popper, 1983, p.221, 233, 254.

i.e. 1. 'Probability' might have another usage when we talk about the probability of a hypothesis. For this usage, 'the *probability* of a hypothesis' is tantamount to 'the degree of corroboration of the hypothesis'. If *probability* could be expressed in terms of probability, it will be more closely related to the improbability of hypothesis than to its probability. That is to say, using the concept 'probability' that satisfies the probability calculus to describe a hypothesis is absurd. A hypothesis with high probability might have little empirical content or it is less corroborated, and thus it cannot be something good or something we have to aim at when conducting scientific research. So far, we have seen why SST, or other statistical tests which enable one to decide between hypotheses on the basis of the probability that the hypotheses are true, has gone astray. We will discuss in the next chapter if there are any alternatives to SST that are on the right track.

Chapter 6 Alternatives to SST

Many have noted the problems of SST though not all of them genuinely understand the limitations and difficulties that we have discussed in the last two Chapters. As we have discussed in Chapter 5, the use of SST has been criticized in an unfair way. For example, if we aim at using SST to address only the question of whether an observed difference is produced by chance, then asking how to supplement SST so as to address the questions, such as 'How large is the effect?' or 'Is the effect large enough to be useful?', is not a problem that we have to bother with here.[185] We will focus here on whether there are any better alternatives to SST so that they could be used to tackle the problems that SST tries to solve, but fails. Nevertheless, there is a great number of sources criticizing the over-reliance on SST and advocating that in reporting outcomes from research more widespread use of confidence intervals (CIs), amongst other statistical techniques, would improve research communication (for example, Cohen, 1994; Cumming and Finch, 2001; Harlow, Muliak, & Steiger, 1997; May, 2003; Wilkinson & Task Force on Statistical Inference, 1997). We will see in the first section whether CIs are really able to tackle the problems encountered in using SST. In such a volume as this, we are not able to discuss all suggested alternatives to SST, such as the effect sizes and Bayesian statistics. We, however, hope that our discussion of CIs could shed light on the tenet that all other suggested alternatives that are inductive in nature would suffer the same limitation and in the last section we will suggest how a methodological framework for doing quantitative research could resolve this problem.

[185] For how to supplement SST, see, for example, Kirk, 1996, in which he has proposed 40 different ways to measure the effect magnitude.

6.1 Is the confidence interval an alternative to SST?

In response to the challenges to SST, APA Publication Manual has suggested that 'Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended' (APA, 2001, p.22). Cumming and Finch (2001) have offered four cogent reasons to advocate the use of CIs, namely that "they are (a) readily interpretable, (b) are linked to familiar statistical significance tests, (c) can encourage meta-analytic thinking, and (d) give information about precision' (p.532). We will see if they really constitute reasons for regarding CIs as an alternative to SST.

As we have discussed in the last two chapters, researchers have severe misconceptions about SST. Could CIs be free from miscomprehension? Are they really readily interpretable. Brandstätter (1999) has alleged that CIs are easier to understand than SST.[186] We can, however, hardly be optimistic, at least in Hong Kong. Even the Hong Kong official curriculum documents recommending for use in schools, for example, *Syllabuses for Secondary Schools – Applied Mathematics (Advanced Supplementary Level)*, contains statements like: 'Confidence interval for the mean of a normal population with known variance: (Notes on teaching) In general, teachers should point out that an interval estimate of an unknown population parameter (e.g. the mean μ) is a random interval constructed so that it has a given probability of including the parameter'

[186] Fidler and Cumming (2005) also maintained that there are benefits of teaching inference via CI, rather than SST even though they noted that a number of misconceptions about CIs had been reported in different studies.

(Curriculum Development Council, 1998, p.64)[187], a big blunder in interpreting CIs. Unquestionably, the curriculum statement is true to point out that it is the interval, but not the population parameter of interest, that is random.[188] But it is false to assert that a particular CI has a given probability of including the parameter purported to be estimated by the CI. As shown in the following figure, 100 different 95% confidence intervals constructed from different data sets drawn from the same population (with mean equals zero) were generated with the use of Maple 8:

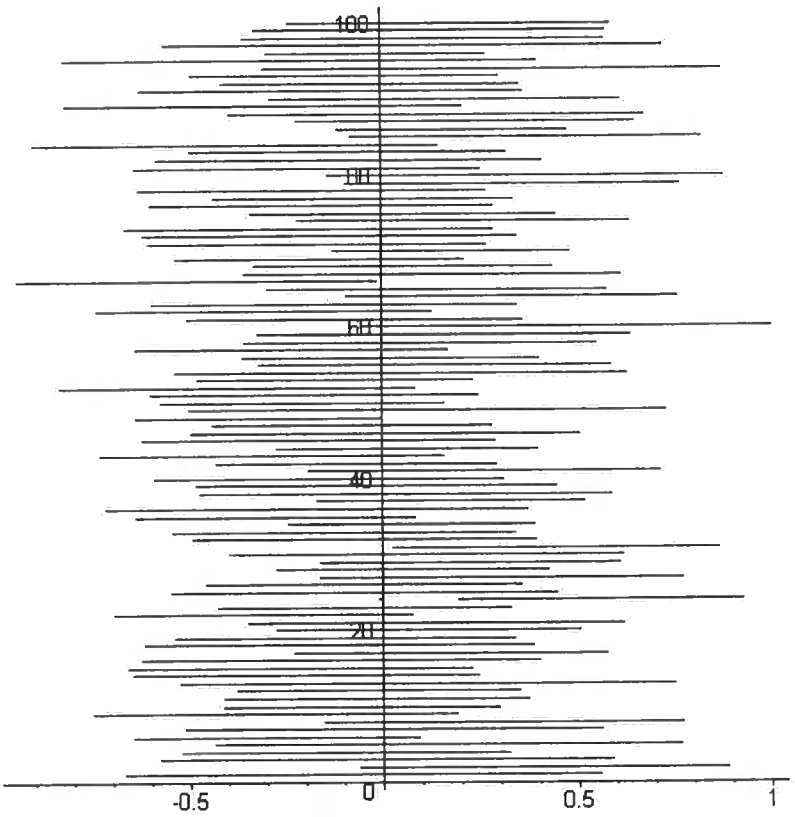


Figure 16 100 different 95% confidence intervals constructed from different data sets drawn from the same population (with mean equals zero)

[187] The same statement could be found in another Mathematics Syllabus: Curriculum Development Council, 1992, p.151.

[188] The interval is random in the sense that the sample mean is random. Of course we are not saying that based on the particular sample at hand the CI would vary although this point is, as reported by Blume and Royall (2003), often confused by students in introductory statistics classes.

It is noteworthy that a few of the CIs in the figure do not overlap. If a 95% CI did really mean that there is a 95% probability that the true value of the parameter being estimated lies inside the interval, the probability that the parameter lies inside any two non-overlapping (or mutually exclusive) CIs would be $2 \times 0.95 = 1.9$, which is greater than 1. This is certainly an absurd result, and on that account, a 95% CI does not mean that there is 95% that the true value of the parameter being estimated falls inside the CI.

Besides, there is a more profound reason why the above interpretation is mistaken. We use a standard CI problem as an example. Suppose we are dealing with a random sample of size n less than 30 from a normal distribution with the mean μ and variance σ^2, then[189]:

$T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$, where $\overline{X}$ and S^2 are the mean and the variance of the random sample,

has the t distribution with n -1 degrees of freedom. Using $t_{\alpha,\nu}$ to denote the value such that the area to its right under the curve of the t distribution with ν degrees of freedom is equal to α, i.e., $P(T \geq t_{\alpha,\nu}) = \alpha$, we have

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha .$$

From these two equations, we get a (1 - α)100% CI for the mean μ of the population:

$$\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},$$

[189] This is a standard result that can be found elsewhere. For example, Miller, & Miller, 1999, p.285.

Here $\bar{x}$ and s^2 are respectively the mean and variance of a particular sample, and they should not be confused with the random variables $\overline{X}$ and S^2.[190] There are at least four interpretations of this CI constructed from the random sample:

$$(CI_1) \quad P(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} \mid \text{population mean} = \mu) = 1 - \alpha$$

$$(CI_2) \quad P(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \mid \text{population mean} = \mu) = 1 - \alpha$$

$$(CI_3) \quad P(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}, \text{where population mean} = \mu) = 1 - \alpha$$

$$(CI_4) \quad P(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \text{where population mean} = \mu) = 1 - \alpha$$

The difference between (CI_1) and (CI_3), or between (CI_2) and (CI_4), is similar to the difference between $P(A \mid B)$ and $P(A \text{ and } B)$. In the process of getting the above inequality for the CI, the calculation is conditional on the value of population mean. That is to say, Figure 13 could not be drawn without first deciding what the true value of the population mean is. Hence, what we really assert is (CI_1) or (CI_2) rather than (CI_3) or (CI_4) when we say that the (1 - α)100% CIs have a probability of (1 - α) of containing the mean μ of the population (Sober, 2005).

[190] For discussion of this confusion, see Schield, 1997.

On the other hand, probability statements are referentially opaque.[191] For example, it is known that for a fair die the probability that the result of the next throw is an odd number is 0.5. Even if the result of the next throw is 3, we cannot substitute '3' for 'the result of the next throw' in the sentence 'the result of the next throw is an odd number', without affecting the truth of the probability statement, because the probability that 3 is an odd number is certainly not 0.5. Accordingly, we cannot conclude from the truth of (CI_1) to the truth of (CI_2) although (CI_2) is an instantiation of (CI_1). What (CI_1) describes is that if we apply this statistical procedure of construction to different data sets drawn from the same population with mean μ,, we can expect to construct CIs that in the long run about (1 - α)100% of them will contain μ. When this procedure is applied to a single data set (CI_2), the particular CI constructed either includes μ, or it doesn't.[192] All that we can be said of this particular CI is that it is constructed by a procedure that is described by (CI_1). The probability is not really about this particular CI for the CI either contains μ. (with probability one) or it does not (with probability 0). The misconception that (CI_2) is true is very prevalent. Even those who are discussing the problems associated with SST and CIs would still commit this mistake. For example, in a response to Cohen's (1994) classic article against SST, Frick (1995c) endorsed the interpretation

[191] Truth about a given object is not usually influenced by the way of referring to it. In other words, if the terms A and B have the same reference then term A can be substituted for term B in any sentence in which B occurs, without affecting the truth-value of the sentence. For example, as 'Hesperus' and 'Phosphorus' refer to the same planet, one can substitute 'Phosphorus', without loss of truth, for 'Hesperus' in the sentence 'Hesperus is the planet Venus'. Quine (1980) has, however, argued that in some contexts, which are called 'referentially opaque', intersubstitutivity does not necessarily occur without loss of truth (pp. 139-159). For example, that 'David knows that Phosphorus is Phosphorus' is true does not imply that 'David knows that Phosphorus is Hesperus' is true.

[192] Brewer (1985) has argued that damage does not necessarily occur when researchers regard the parameter μ as a variable except that with this treatment must come the philosophy in which it makes sense to do so. In particular, if probabilities are construed as degrees of belief, we can say, for instance, we are 95% sure that the parameter falls inside a particular confidence interval, in which one starts with prior distribution of the parameter and finds a posterior probability distribution, which is the conditional probability distribution of the parameter given the data. This Bayesian interpretation of CIs is philosophically controversial and we will not delve into its detail here.

that 'the 95% confidence interval is an interval within which the true value is 95% probable to fall' (p.1102).

In a survey of 180 undergraduate psychology students, they displayed misconceptions about both the definition of a CI [193], and how aspects of CI relate to each other[194] (Fidler & Cumming, 2005). Besides, even world-leading researchers have been found in different studies (Cumming, Williams & Fidler, 2004) to have a range of serious misconceptions about CIs. That may explain why Smithson (2003) claimed that 'perhaps the most obvious difficulty with confidence intervals lies in how we interpret what the confidence statement means' (p.16), and why Thompson (1987) concluded that '...both significance tests and CIs are subject to misinterpretation. The issue in choosing between the two is therefore not whether one is immune from interpretation but rather which of the two is more useful to a thoughtful reader' (p.191). Furthermore, by identifying different misconceptions surrounding CIs by textbook writers, Brewer (1985) found that these misconceptions appeared to be related to the interpretations of CIs and their relation to SST. Besides, some even argued that the logic of hypothesis testing is relatively uncomplicated and uncontroversial compared to that of CIs (Simon, unpublished). Therefore, CIs are not really readily interpretable and being free of miscomprehension is not a reason that we have to replace SST by CIs.

In the last Chapter, we have argued that the logic behind SST is fallacious. Could CIs escape the logic of SST and thus offer a defensible procedure for making a dichotomous decision on the null hypothesis. Being regarded as a means of estimating

[193] For example, many of them failed to realize the inferential nature of CIs.
[194] Only 16% of them know how the CI width is related to the sample size.

the value of a parameter, CIs can indeed be used to either reject or retain the null hypothesis which specifies the value of the parameter. To test the null hypothesis, we could determine whether the value of the parameter specified in the null hypothesis lies within the CI. If the value of the parameter lies outside the CI, then the null hypothesis could be rejected at the level of confidence. On the contrary, if it lies inside the CI, then the null could not be rejected at that level. It appears that we could now be able to test a hypothesis concerning a parameter in a more direct way – a CI could inform us the range in which we should locate the parameter rather than the range in which we should not locate it. Is it really so? First, we are indeed not sure whether the CI calculated from the observed data does or does not contain the value of the parameter specified in the null hypothesis. As we have discussed above, we cannot even assign a probability to the statement about the particular CI. All we could know is that in the long run about 95% of the CIs would contain the parameter. Hence, it is incorrect to say that a particular CI could inform us the range in which we could locate the parameter. Second, in SST we begin with a null hypothesis which specifies the value of a single parameter. From this parameter, we develop, say, a sampling distribution against which the observed sample statistic is compared. But for CIs, we start from the observed sample statistic, and establish a CI against which we test an infinite number of parameters (Hagan, 1997). All hypotheses with specified parameters that are outside the CI would be rejected. The logic invoked here is, as illustrated by Hagan (1997): 'The probability that a population with "this" parameter produced "this" datum from which "this" confidence interval was constructed is very low. Therefore, we reject the idea that the datum came from such a population.' (p.22). In this regard, the logic is the same as that behind SST and it is a

plain mistake to regard the CI estimation as the inverse operation to SST.[195] CIs can thus hardly escape the logical fallacy, discussed in the last Chapter. We may agree with McGrath (1998) that testing the null hypothesis may not be the primary reason for some who prefer CIs, or Thompson (1998) that we are doing little more than SST if we do mindlessly interpret a CI with reference to whether the CI subsumes zero[196]. But these cannot sway our conclusion that CIs are no better than SST for the purpose of testing null hypothesis.

Moreover, when comparing multiple means, researchers are sometimes recommended to compare the results from CIs and decide if the intervals overlap.[197] The difference is judged significant when there is no overlap, and not significant when there is overlap. This procedure can, however, lead to mistaken conclusions. Schenker and Gentleman (2001) have shown that even though CIs overlap, there could be a statistically significant different between the means. In other words, a rejection of the null hypothesis by this overlap method entails rejection by the familiar SST, but not vice versa. Therefore, the overlap method is more conservative and less powerful than the standard method and the use of CI in comparing multiple means is thus not as useful as some have expected.

There are of course many suggested virtues of CIs. For example, according to McGrath (1998), the basic question under SST is 'based on this sample, what is our best

[195] It is worth noting that such mistake really exists. For example, Nicholls (2000) explicitly alleged that 'The reporting of confidence intervals would allow readers to address the question "Given these data and the correlation calculated with them, what is the probability that H_0 is true?" rather than "Given that H_0 is true, what is the probability of finding a correlation this strong(or stronger)?"' (p.984)
[196] Whether in reality people are conforming to their 'wishes' is an issue that we will discuss below.
[197] Abundant examples could be found in Schenker & Gentlemen, 2001.

guess about whether or not ρ equals 0?', but CIs allow for a more interesting question: 'based on this sample, what is our best guess about the value of ρ?'. And, perhaps more important, many have claimed that CIs are able to give more information about the effect size[198], and thus useful for conducting meta-analysis[199], and about replication[200] and precision[201]. But as we have argued in previous Chapters, SST does not purport to provide the information on these areas. Hence, even though CIs could provide information on these areas, the use of CIs in these aspects serve merely as a supplement to SST. There is thus no need for us to delve here into the details of CIs. Moreover, some of the claims about the usefulness of CIs are dubious. For example, Matthews (1998) has argued that 'CIs share many of the same problems of interpretation as P-values. Most importantly, they also share an inability to take into account the plausibility of the hypothesis under test. As such, 95 per cent confidence intervals are also prone to exaggerate both the size and the "significance" of intrinsically implausible effects.' And we may note, in passing, that some other objections to the use of CIs have been raised[202].

Moreover, even though the integrity of CIs had not been threatened by these arguments, it would still be dubious whether CIs being used merely as an algorithm for making inference from data could really serve as a supplement to SST. In fact, CIs have

[198] The CIs about effect sizes are not the same as the CIs about means. See, for example, Cumming & Finch, 2001; Smithson, 2003; Thompson, 2002, for details.
[199] For how CIs facililitate meta-analytic thinking or its role in meta-analysis, see Sim & Reid, 1999; Thompson, 2002.
[200] See Cumming, Williams & Filder, 2004.
[201] To some researchers (e.g., Cumming & Finch, 2001, Montori, et al., 2004), CI width could reflect a number of aspects of precision of a study, including the amount of variability in the population, the sample size, sampling error, and the amount of error in the dependent variable.
[202] For example, see Seidenfeld, 1979 for criticism leveled at the Neyman-Pearson theory of CIs and Mayo, 1981 for rejoinder to this criticism; see Walley, 1991 (section 7.5) for arguments against the use of CIs and see Smithson, 2003 for defense. For a more recent criticism of CIs, see Pawitan, 2001, section 5.10.

been recommended as a supplement in many different fields for decades[203] and the changes of editorial policy in journals appeared to be quite effective, but reporting of CIs does not guarantee that they will be used to interpret the data. Different studies on the effect of editorial policy on statistical practice show that the compliance was superficial. For example, Savitz, Tolo and Poole (1994) found that although 70% of articles in the *American Journal of Epidemiology* reported CIs, the inferences were based merely on the location of the null value with respect to the bounds of the CI, i.e. CIs were still used as a means to test hypotheses. And Fidler, et al. (2004) also found that the authors presented CIs and not p values seldom use CIs to justify their interpretation of the data, as suggested by the title of their paper 'Editors can lead researchers to confidence intervals, but can't make them think'. In other words, even though researchers know that SST has many insurmountable problems and accept the recommendation that CIs could serve as a supplement to SST, they either still use SST with an attachment of a report of CIs or simply use CIs as an alternative to SST, without knowing that similar insurmountable problems would be encountered when using CIs to test the null hypothesis. We will discuss in the next section why such phenomena happen and see if it is possible for us to have an algorithm other than SST or CIs for testing hypothesis.

6.2 Any other algorithms for making inference from data?

As noted in the previous section, there are formulae for constructing CIs for the purpose of testing hypotheses or estimating the parameter, such as means, specified in the null hypothesis. The logic behind CIs is as simple as that behind SST. Not much extra

[203] Over twenty journals in education and psychology now require CIs or effect size reporting (Thompson, 2002). See also *Publication manual of the American Psychological Association (5th Ed.)* (APA, 2001), *Publication Manual, Memory & Cognition* (Loftus, 1993), *Journal of Consulting and Clinical Psychology* (Kendall, 1997).

effort is required if CIs are merely used as an alternative to SST for testing the null hypothesis. In this connection, a deal could easily be made if researchers are merely requested to report CIs or simply to switch from SST to CIs in testing the null hypothesis. On the other hand, when CIs are used as a supplement to SST, for example, for measuring effect sizes, researchers have to confront at least two daunting technical difficulties: the use of noncentral distribution, which are less familiar to many researchers, and the lack of a generic and ready-made formula for computing CIs for effect sizes (Thompson, 2002). It is thus understandable why people prefer using CIs to test the null hypothesis, or to estimate the parameter specified in the null, rather than using it to measure effect sizes. That explains why using SST or CIs to test the null hypothesis could still be so prevalent, especially when many researchers are still conceiving these practices as an algorithmic method for conducting research – what they thought they have to do is to design an experiment and to collect data and there is no need for them to bother about the inferences at all as there are well-established algorithms or automatic rules for making inferences from the data.

Looking for an algorithm or an automatic rule for making inferences from the data might be a natural quest to many researchers. Since both SST and CIs encountered intractable problems in testing the null hypotheses, are there any other statistical tests that can be free from these problems in providing algorithms for testing the null hypotheses? In what follows we will consider three different cases and see if they could shed light on this question.

First, undoubtedly the calculations involved in SST or CIs are algorithmic, at least in the sense that abundant computer software are now available for researchers to perform these calculations after inputting the data. But it is also important to see, behind the computations, if there are any assumptions of which violation would give incorrect result. Indeed, before we conduct a test we have to check if all requirements have been met. For example, when we conduct a SST to compare groups (say, to test if the group of students who were taught by a teacher making use of a new teaching strategy could perform better than a normal group in a standard mathematics test), we could compute the standardized sample mean on the assumption that the sampling distribution is normal (in case the population variance is known) and follow the standard procedures of SST as discussed before. And it is the central limit theorem that gives a remarkable result about the sampling distribution of the sample mean so that we can make inferences from data without bothering about the distribution of the population. Many textbook writers reassure the readers that in one-sample *z*-test the sampling distribution of the means from very non-normal populations will become normal as the sample size increases.[204] But the question that how many observations must be used when computing the sample means so that the normal distribution will be a fairly good approximation to the sampling distribution of the means is often ignored by the researchers. Some recent writers have made good use of computer algebra system to answer this question empirically.[205] Let us consider a uniform distribution which is usually a standard illustration of the central limit theorem. We use Maple 8 to generate 10000 random numbers from [0, 1], simulating a uniform distribution U (0, 1). Figure 17 shows the histogram of this simulated population:

[204] See, for example, Hopkins, Hopkins, & Glass, 1996, pp.158-164; Moore & McCabe, 2006, pp.362-364.
[205] See, for example, Karian and Tanis, 1999, pp.106-114; Braselton, 2003, pp.427-437.

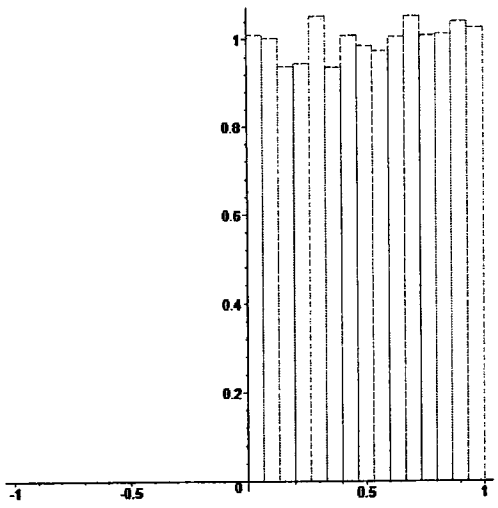
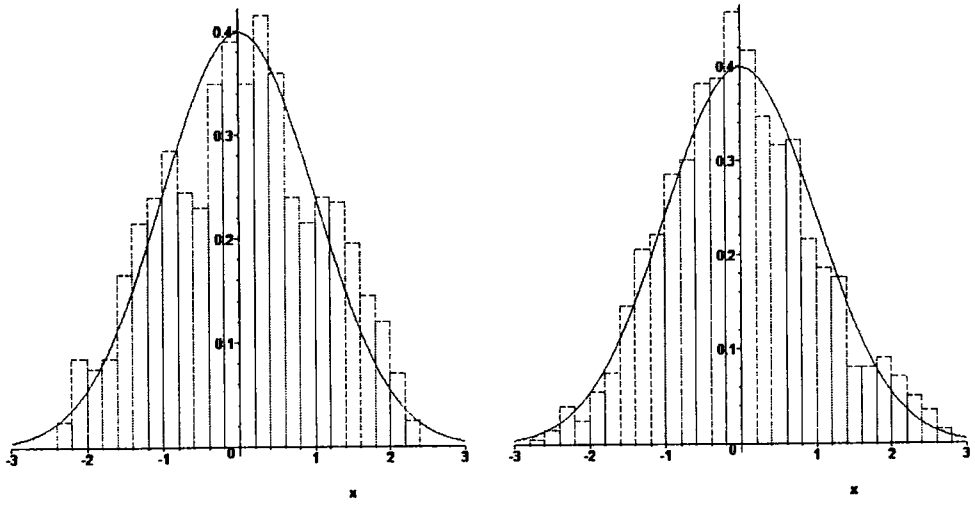


Figure 17 Simulation of the uniform distribution $U(0, 1)$

The standardized sample mean (or called 'one-sample z statistic) is calculated assuming the population mean and population variance are known. The effect of the sample size 2 on the distribution of the sample means is shown in Figure 18. If we increase the sample size to 20, there will be better agreement between the plot of sampling distribution of the means and the normal curve.



n =2 n =20

Figure 18 The histograms of the standardized sample means when n =2 (L) and n =20 (R) and the corresponding normal curve for the parent uniform population $U(0,1)$

Consider an exponential distribution with probability density function $f(x) = e^{-x}$, where $0 < x < \infty$. We repeat our computer experiment (also assuming that the population mean and population variance are known) and the histogram of this simulated exponential population is shown in Figure 19 while the histograms of sampling distribution of the means when n =2 and n =20 and the corresponding normal curves are shown in Figure 20.

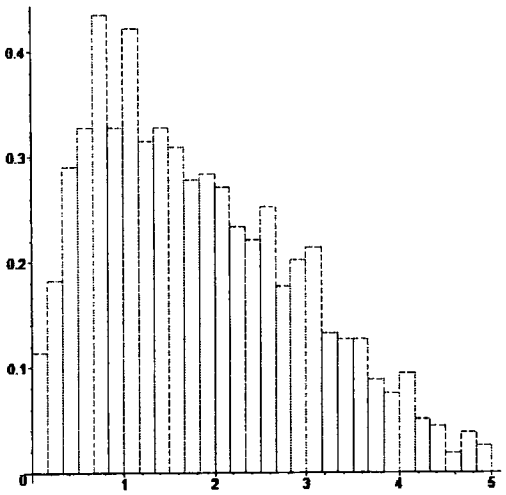
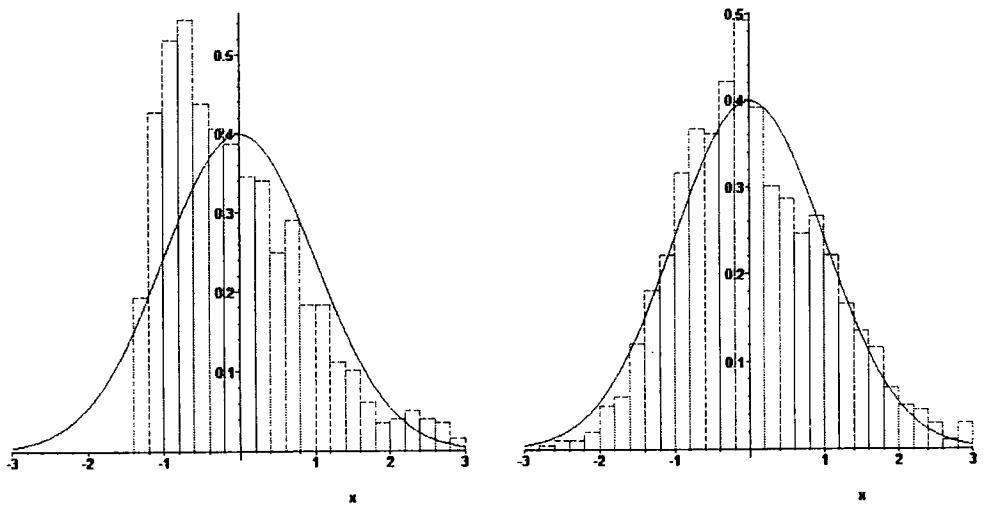


Figure 19 Simulation of the exponential distribution with probability density function $f(x) = e^{-x}$



n =2 n =20

Figure 20 The histograms of standardized sample means when n =2 (L) and n =20 (R) and the corresponding normal curve for the parent population with probability density function $f(x) = e^{-x}$

In these two examples, we start with distributions whose curves do not look like a normal curve but plots of sampling means are approximately normal when each mean is based on only 20 or even 2 values. Some may thus think that in general the central limit theorem applies with small sample sizes. But is it the case? Consider the lognormal distribution[206] which has a skewed and heavy-tailed probability curve as shown in Figure 21.

[206] It is biological science which first introduced lognormal distribution especially when exponential growth is combined with further symmetrical variation. But the distribution has much more applications nowadays. For a general discussion of lognormal distributions and how it is used across different areas of science, see Limpert, Stahel, and Abbt, 2001.

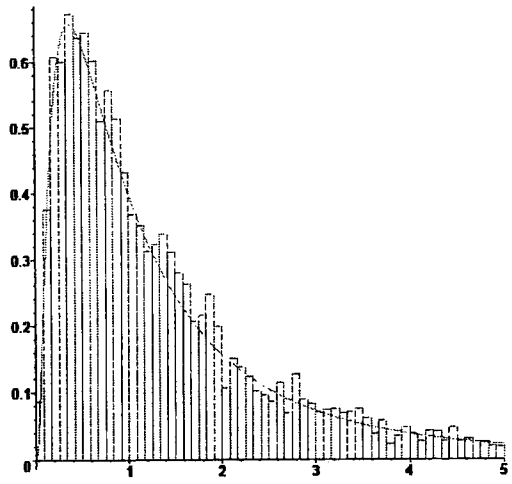


Figure 21 The histogram of a lognormal distribution

Same as before, 20 observations are used to compute each sample mean and the plot of the means is shown in Figure 22.

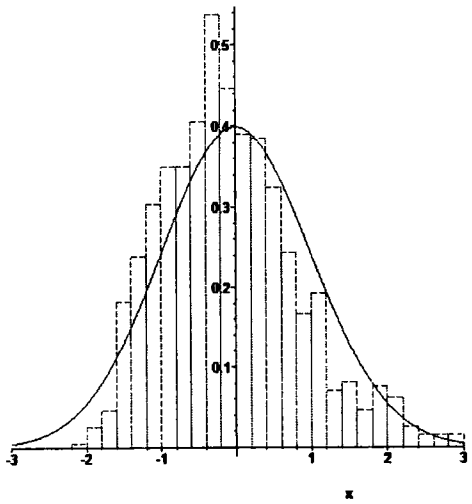
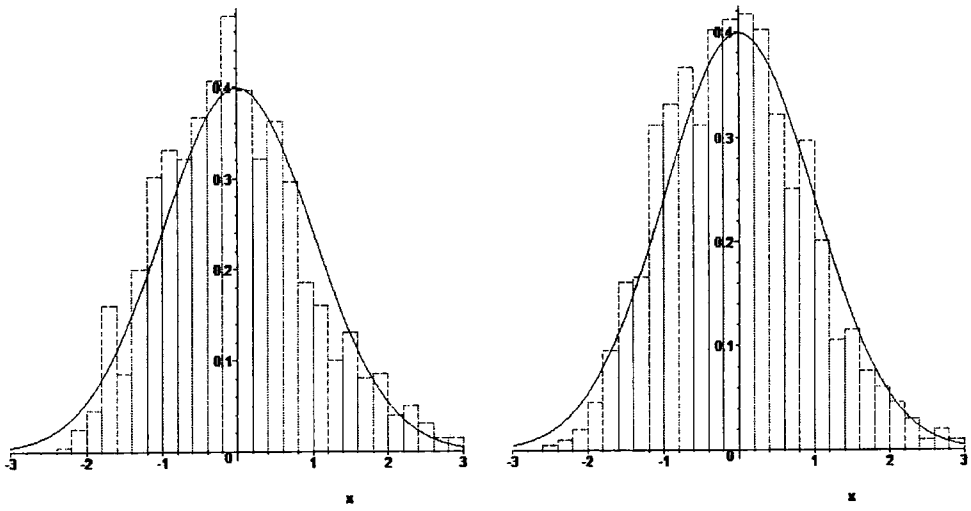


Figure 22 The histogram of standardized sample means when n =20 and the corresponding normal curve for the lognormal distribution

This time the plot of the means is poorly approximated by a normal curve, particularly in the left tail, as indicated in Figure 22. If we increase the sample size to 40, the approximation is better though as a whole it still remains poor. And the plot of means looks better when the sample size increases to 80 (see Figure 23 for n =40 and n = 80). As a result, we could get good approximation with 20 observations in some cases but there are cases where more observations are needed.



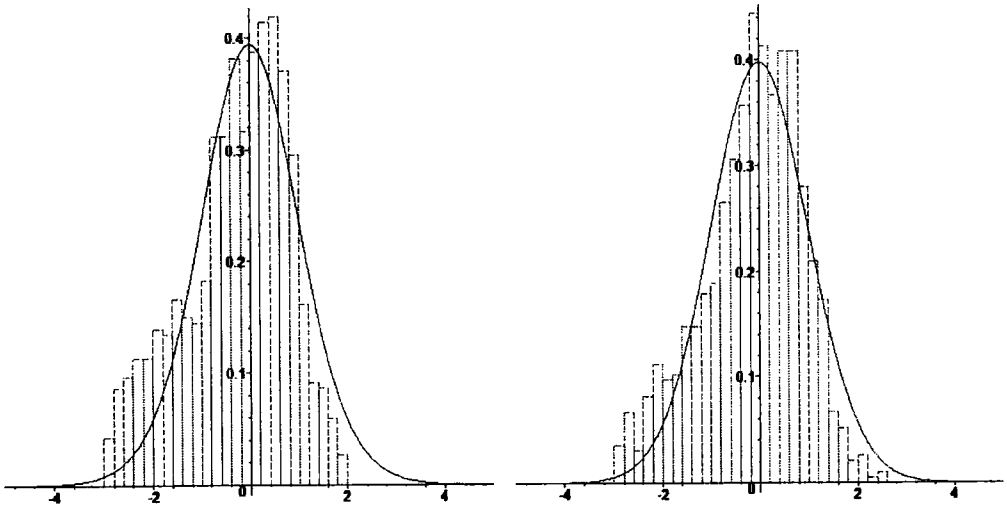
$n = 40$ $n = 80$

Figure 23 The histograms of standardized sample means when n = 40 (L) and n =80 (R) and the corresponding normal curve for the lognormal distribution

Not only z statistic but also t statistic (which is supposed to have Student's t distribution when the population variance is not known, that has been mentioned in Section 2.2) is sensitive to the population from the sample is drawn.[207] Figure 24 shows the plot of sample t statistics for samples with size 20 drawing randomly from the lognormal distribution and the graph of the corresponding t distribution (with 19 degrees

[207] For further discussion on the practical problems with Student's t, see Wilcox, 2001, pp.67-91.

of freedom). It is obvious that the actual distribution of sample t statistics is skewed and deviated significantly from the t distribution if the sample size is 20 (the approximation is better but still not so good when the sample size increases to 80)



$n = 20$ $n = 80$

Figure 24 The histograms of the sample t statistics taken from a lognormal distribution when $n = 20$ and $n = 80$ and the curve for t (19) distribution

Since the left tail of the actual sampling distribution is skewed and its left tail is much thicker than the t distribution, the actual probability coverage will be substantially less if we compute what we claim is a 0.95 CI for the mean. We use Maple 8 to select 5000 random samples (with size 20) from the lognormal distribution and calculated the ratio of the samples which really lie inside the 95% CI (i.e. $[-a , a]$, where a = 2.093024054, computed from the t (19) distribution). It is found that only 86.4% of samples are covered by the CI. Similar errors would occur when we estimate Type I and Type II errors. Hence, even though we suppose the problems surrounding SST did not exist, the computation of CIs would be inaccurate and the control over the Type I error would be poor if the samples with insufficient sample size are taken from populations

with lognormal distribution. In other words, not only SST but also other inferences from data (such as computing CIs) would be sensitive to violations of assumptions. And checking whether the assumptions have been violated is certainly not an algorithmic process.

Second, in many cases when researchers use SST to compare groups, the genuine null hypothesis they are going to test is that the groups come from the same population, i.e. the probability curves associated with the groups are identical. For instance, the observed result that renders us rejecting a null hypothesis about the means with Student's t is certainly an indication that the probability curves of the groups differ in some manner. However, is the converse also true? Now suppose, after conducting a SST, we have to accept the null hypothesis that the means of the two groups are equal. Could we thus conclude that the probability curves associated with the groups are identical?. The answer is yes if we assume that the means must differ if the probability curves differ (Wilcox, 2001, p.87). However, this assumption is plainly false. For example, we can use Maple 8 to draw two different probability curves that have equal means (1.66) and variances (4.66) (see Figure 25).

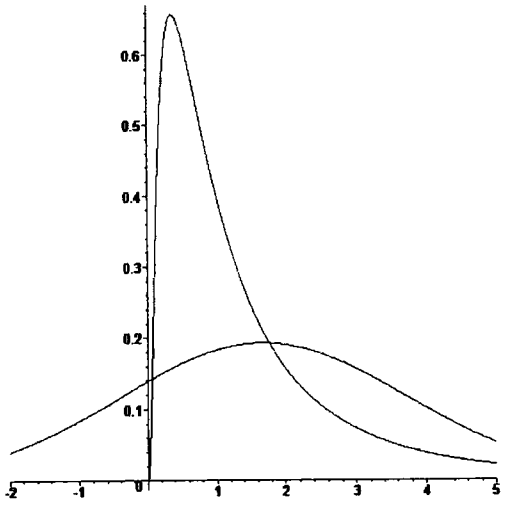


Figure 25 The probability curves of a lognormal and a normal distribution that have equal means (1.66) and variances (4.32)

Hence, mere a compliance with the algorithm is not enough for making correct inference. In this case, maybe a graph of the data from these two samples could help in avoiding this trivial mistake.

Third, there are many cases in which very different data can give rise to the same level of achieved significance, the same correlation, and the same regression equation (Anscombe, 1973; Macdonald, 2002). For example, consider the four data sets of ordered pairs given in the following Table 4:

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
4	4.27	4	3.1	4	5.39	8	6.58
5	5.67	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.13	8	7.04
10	8.04	10	9.14	10	7.45	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.75	8	6.88
14	9.96	14	8.1	14	8.84	19	12.5

Table 4 Four data sets of ordered pairs

In each data set, there are 14 ordered pairs. They all give rise to the same means of the x's (9.0) and y's (7.5), the same regression line ($y = 0.5x + 3$), the same correlation coefficient R^2 (0.67), and the same level of significance (0.02). But we could see from the their graphs (as shown in Figure 26 – 29) that not all of the data sets conform to the theoretical descriptions, such as the result of regression analysis. Figure 26 shows that there is probably a linear relationship between the variables and it conforms to what most of us anticipate when a regression analysis is performed. But in the second data set the relationship between the variables is perfectly nonlinear and it is dubious whether there is any random variation (y seems to have a perfectly smooth curved relation, such as quadratic, with x). The third data set shows a perfect linear relationship between the variables except for one outlier. The regression line obtained is thus deviated from the one that perfectly fits the points excluding the outlier. For the fourth data set, the data are quite consistent with the assumptions of a linear relationship – homogeneity of variance and a normally distributed error term even though all the variation on x comes from only one point.

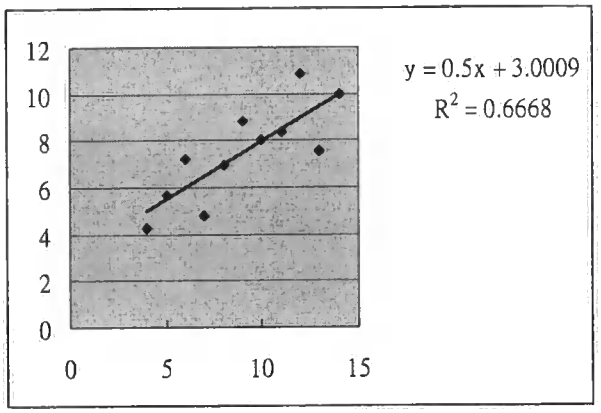


Figure 26 Graph of the first data set

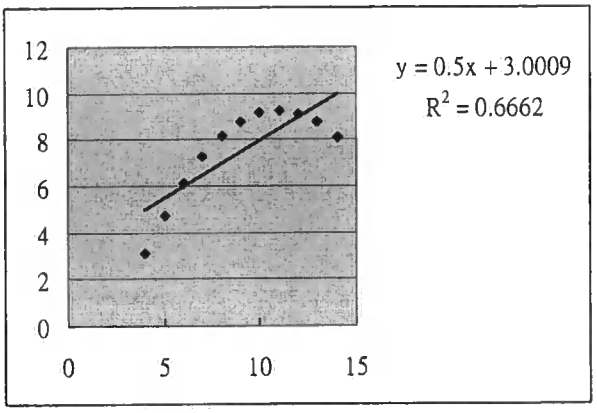


Figure 27 Graph of the second data set

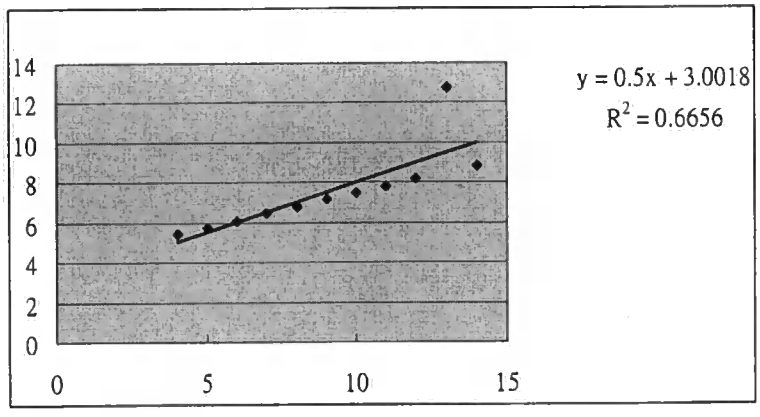


Figure 28 Graph of the third data set

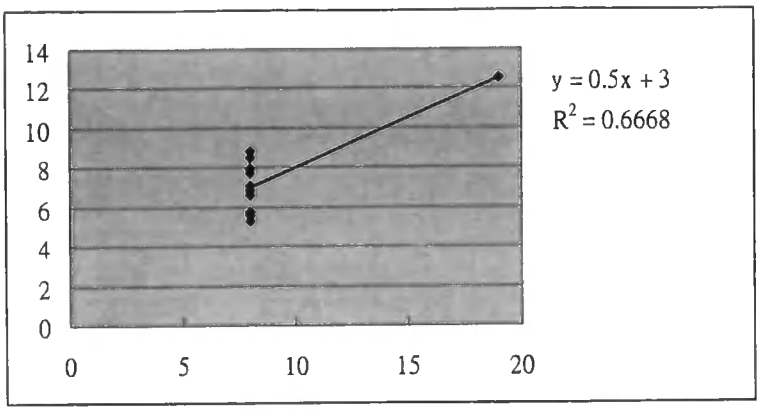


Figure 29 Graph of the fourth data set

This case shows that sometimes statistical tests are not sufficient for us to decide between the models on the probability that the models are true, and that data could undermine a model and that this could influence our confidence in the results of the statistical tests. For example, the graph of Figure 27 suggests that the linear model for the second data set is misspecified although, following the algorithms for finding level of achieved significance, correlation coefficient and regression equation, all data sets give rise to the same results. It is worth noting that in this case, unlike our first example in which we challenge if all assumptions have been met, data could cast doubt on the significance test but without challenging on the assumptions on which it is based. Our discussion here is not only applicable to regression analysis but also to, for example, analyses of variance (ANOVA) and *t* tests (Macdonald, 2002). Hence, whenever any statistical test is applied we must be cautious that unexpected patterns of results can affect the conclusions that should be drawn from it.

From the discussion of these three cases, we could see that having only a compliance with the algorithm or rigid rules dictated by SST, CIs or any other statistical

theory is not sufficient for drawing a good inference from data. We should be cautious about the assumptions of the tests. Sometimes the violation of these assumptions could lead to incorrect results. Also, particular consideration should be given to the data, which could sometimes be able to strengthen or undermine the hypothesis. The list could be extended to a considerable length (Macdonald, 2002). In a nutshell, it is not reasonable for us to expect the existence of a statistical test providing algorithms or rigid rules by conforming to which all problems about testing hypotheses could be solved.

We have now come to a situation where both SST and CIs have insurmountable problems in testing hypotheses, and there are no other statistical tests that could provide algorithms to enable one to decide between hypotheses merely on the basis of these algorithms. It seems that searching for a panacea that could replace SST for testing hypotheses is doomed to failure. Then, what practices should we adopt in testing hypotheses? We will address this question in the next section.

6.3 Induction versus falsification

As Macdonald (2004) has argued, all forms of statistical inference, of which SST is certainly one, involve the following characteristics: observations are made, a probability is computed, and a conclusion is drawn. Making statistical inferences involves reasoning on the basis of incomplete information and does not lead to necessary conclusions. And it is this reason that probability comes into play. Besides, it was argued that statistical reasoning is a sort of induction based on an inference with some missing premises, i.e. enthymemes (e.g., Macdonald, 2004). But induction is not deductively valid, that is to say, the truth of the premises cannot guarantee the truth of

the conclusion. Could we justify induction and statistical reasoning, with the help of enthymemes? Certainly, it is always possible to construe induction as an enthymeme. But it does not help much. Induction is still logically invalid. Because if enthymeme does work in this way, all invalid arguments would be turned into valid arguments[208]. For example, we know that the following argument is invalid:

David is a mathematics teacher

Hence, David is an expert in Sudoku

We could add the sentence 'Any mathematics teacher is an expert in Sudoku', or simply, 'David is an expert in Sudoku' as a missing premise, then the argument will become pretty valid. Hence, enthymeme cannot help in turning induction into valid inferences unless we are prepared to accept that there are no invalid arguments at all.

We have concluded from the preceding discussion that both SST and CIs have insurmountable problems in testing hypotheses, and there appear to be no other statistical tests that could provide algorithms to enable one to decide between hypotheses merely on the basis of these algorithms. Do these problems arise from the inductive nature of statistical tests? If so, could scientific research be conducted without induction? We will discuss these questions in what follows.

In science, the hypotheses usually consist of universal statements. Unlike singular statements such as 'David is a mathematics teacher', universal statements cannot be proved or conclusively confirmed. For example, 'all ravens are black' cannot be established with absolute certainty no matter how many black ravens have been

[208] For further discussion on this issue, see Musgrave, 1999a, 1999b.

observed. It is thus commonly believed that the hypotheses in science cannot be conclusively confirmed[209]. What researchers can do is to make use of induction, say,

The raven observed at t_1 is black

The raven observed at t_2 is black.

....

The raven observed at t_n is black

Hence, all ravens are black.

This inference looks perfectly sensible, although it, unlike deduction, is not logically watertight, for the truth of all premises does not guarantee the truth of the conclusion. No matter how big n is, it is still possible that the next observed raven is not black. This sort of inferences is known as 'inductive inferences' in which we move from premises about singular statements, such as those express the results of observations or experiments, to a universal statement. A received view, at least before Karl Popper, is that science is characterized by the use of induction. And it is Popper who has argued forcefully that this view is mistaken.

Although inductive inferences are not deductive, they nonetheless seem to be a perfectly legitimate way of forming our beliefs about the world. For instance, we believe that all children under 2 years old are unable to solve differential equations because all children under 2 years old that we have observed are unable to do so; the sun will rise tomorrow because the sun has risen every day up until now. The list can go on indefinitely. Could we justify these seemingly ubiquitous inductive inferences? To this question David Hume gave a negative answer. Admitting that we use induction all the

[209] For fallabilists, all knowledge is fallible, that is to say, not only scientific theory but also all other factual statements cannot be established with absolute certainty.'

time, in daily-life as well as in science, Hume maintained that our use of induction is merely a matter of brute animal habit (Hume, 1739/1990). It appears that when making inductive inferences, we are based on a principle called the 'uniformity of nature' which states that the course of nature continues uniformly over time and space. Is this principle true? Or, could it be justified? It is obvious that this principle is neither a purely logical truth nor a priori truth, like 'if Dodo is a black cat then Dodo is black', since a world violating this principle is conceivable. But can it be justified empirically? Suppose up to this moment this principle is true for all instances we have examined, this still cannot constitute a justification to the principle since the argument from all these examined instances to the truth of this principle is itself based on this very principle – it is begging the question. Hume thus concluded that induction is not justifiable. If we accept that to reason in an unjustifiable way is irrational and induction is unjustifiable, then we must conclude that to reason with the use of induction is irrational. But in science, or in daily-life, it seems that we reason, and must reason, with the use of induction. Does it imply that we are, and must be irrational?

Popper (1957/1980) agreed that the various difficulties of induction are insurmountable, but he argued that science could make no appeal to induction. His basic idea is that: confronted with an experience-transcending hypothesis *H* in science, we could either try to justify *H*, give reasons to it, show that it is true or highly probable, or try to criticize it, give reasons against it, show that it is false (Musgrave, 2004). Being experience-transcending, *H* cannot be justified with only experience. According to what we have discussed, invoking unjustifiable induction to justify *H* is itself

unjustifiable, and to show that *H* is highly probable also does not work. Hence, it leaves only the latter option to us.

For the latter option, there are two possible outcomes. First, we succeed in falsifying the hypothesis, that is to say, the hypothesis is shown to be false[210]. The hypothesis will then be expelled from the body of science. The falsification, as discussed in Chapter 4, is based completely on a valid logical argument form, *modus tollens*. As no invalid or unjustifiable inductive inferences are required in falsifying *H*, the falsification could sustain the attack from Hume's inductive skepticism. Second, we fail in falsifying the hypothesis, i.e. *H* stands up to our effort to reject it. What does it mean to have passed these tests? It is the answer to this question that constitutes the main difference between Popper's falsificationism and justificationism[211] (Miller, 1980, 1994). For justificationists, only the hypotheses that have passed the tests, or have been confirmed, would be admitted to the realm of scientific knowledge. But for Popper's falsificationism, all hypotheses that are falsifiable[212] would be admitted. A test of a hypothesis is a serious attempt to falsify it. If a hypothesis stands up to our effort in falsifying it, it does not make any significant difference to the status of the hypothesis. Although some may regard it as corroborated, it merely means that the hypothesis could still remain in the realm of scientific knowledge, temporarily and perhaps it will be falsified in next test. But a hypothesis's passing a test does not mean that it has been

[210] The genuine situation is more complicated. See Section 4.1 and the following discussion.
[211] Justificationism is premised on the tenet that it is only reasonable to believe what has been justified (shown to be true, or probable).
[212] In a nutshell, a hypothesis is said to be falsifiable if and only if there are conditions under which the hypothesis will be rejected, i.e. it can make some definite predications that are capable of being tested against experience. It is falsifiability that Popper proposed it as a criterion of demarcation between science and pseudoscience. For more technical detail, see Popper, 1983, pp.xix-xxv. It should be stressed that the term 'falsifiable' is not used in the sense that the hypothesis can definitively or conclusively or demonstrably be falsified.

justified, confirmed nor proved.[213] There has been much discussion over Popper's falsificationism in the past decades. For the purpose of clarification, it is worthy to note the following five points about Popper's idea on falsificationism.

First, some may wonder if we never invoke induction, how could the knowledge of our world beyond the content of our observations be obtained? For example, suppose we know that all metals are electrical conductors. This piece of knowledge can neither be observed, as it is impossible for us to test all metals, nor inferred from what we observed, if no inductive inference is allowed. Our reply is that it is true that empirical knowledge cannot be obtained in these ways, but it does not imply that we have to do with induction. Because empirical knowledge could come from conjectures or guesswork. In fact, most great theories or hypotheses in sciences are not an accumulation of observations. For example, Newton's first law of motion states that: An object moving in a straight line will continue moving in a straight line, unless acted on by an outside force. It seems to be contrary to our observations. On one hand, we cannot directly observe force; and on the other hand, what we usually observe about a moving object in a straight line is that it will sooner or later come to a rest, rather than continue moving. This law of motion is indeed a bold guesswork. Of course, we are not saying that observations have no place in making conjectures. Sometimes it is the observation that inspires researchers to make their guesswork. But it does not mean that there is a necessary linkage between observation and the bringing up of the

[213] Musgrave (2004) would, however, argue that *H*'s being able to stand up to our effort to criticize it will constitute a good reason to believe *H*, though it is not a reason for *H* itself (p.24). According to Musgrave (2004), 'reason for believing *H*' and 'reason for *H*' are completely different. See Miller, 2002/2006 for comments on this view. It is too involved a subject to be treated here in detail.

hypothesis. Inductive inferences do not have a role in putting hypotheses into science. The only 'entrance' requirement for a hypothesis into science is its falsifiability.

Second, one of the distinguishing features of scientific theories or hypotheses is that they do have predictive import. It is also an important factor that we treasure science, instead of pseudo-science. However, does having predictive import imply that some form of ampliative inference[214] must be incorporated? Popper (1974) said 'no'. Science has predictive import simply because the hypotheses or theories consist of universal statements. When a hypothesis states that 'all metals are electrical conductors', what it asserts is about all metals including those which are tested in the future. There is no need for us to invoke any ampliative inference for predictive import.

Third, like all fallabilists, Popper believed all factual statements could be mistaken and thus no certain truth could be achieved. But unlike other fallabilists, Popper maintained that it is pointless for us to target at making our knowledge more certain, more reliable[215] or more probable.[216] It is truth, but not probable truth or reliable truth, which we should try to achieve in science (Miller, 1980). It is this contention that engenders researchers to put their effort in falsification rather than justification. What researchers should do is to make strenuous effort to put any hypothesis to test, as severe as possible, to ensure all false hypotheses could one day be

[214] An inference is ampliative when its conclusion contains more information than its premises.
[215] According to reliabilists, knowledge is true belief generated by a reliable method or process. Induction on its own is not reliable, as it often leads to false beliefs. But combining with other methods, induction could become more reliable (Goldman, 1986).
[216] The notion 'probable truth' is quite obscure. See Miller, 1980, p.114 for further discussion.

shown to be false. As we only aim at truth, we need not bother about the measures of certainty, truthfulness, or reliability of the hypotheses that have passed all tests. No inductive inference is thus needed to keep the hypotheses in the body of science.

However, there seems to be a difference amongst hypotheses – we have more confidence in the hypotheses which have already been passed many severe tests than that not yet been tested. But it is merely a sort of psychological difference. A hypothesis in which no matter how confident we feel could not ensure it will never be falsified by empirical evidence. For example, we had pretty high confidence in Newton's classical mechanics for many years but there was still one day we found that the motion of Mercury's perihelion, the point at which Mercury passes closest to the sun, did not behave as predicted by Newton's theory.[217]

It is noteworthy that the prediction made by science is merely deduced from the hypotheses which have not yet been falsified[218]. It should not be regarded as one that has been justified by the observation. But this is not a big problem for, as shown by Hume, no ampliative or inductive inferences could lead to justified predictions either.

Fourth, if a hypothesis has been falsified, how could we be certain that it will be falsified again in the future, if induction is not allowed? (Hesse, 1974; Warnock, 1960) Before answering this question, we have to make a distinction first. When a

[217] Scientists had tried to fix this problem with many auxiliary hypotheses. For example, some had proposed that there was dust between the Sun and Mercury. But none of them were consistent with other observations, for instance, no dust could be found when the region between Mercury and the Sun was carefully scrutinized. It was later that Einstein's General Theory of Relativity could predict the orbit of Mercury with an astounding accuracy. For further discussion, see Torretti, 1999, pp.246n, 299, 417-419.
[218] Strictly speaking, auxiliary hypotheses and/or specific initial conditions are needed, in additional to the hypotheses. See Chapter 4 for details.

hypothesis is said to be falsified, we will use the word '**FALSIFIED**' with bold capital letters to denote that the hypothesis has in fact been falsified, and use the word '*falsified*' with italic letters to denote that the hypothesis was mistakenly falsified by the researchers. In other words, when researchers claim that a hypothesis has been falsified, either the hypothesis is **FALSIFIED**, in this case the hypothesis is really false though the researchers might never be absolutely certain of it, or it is *falsified*, in this case the hypothesis is true and some mistakes must have been made in the process of falsification. Let us return to the question. Hesse's (1974) argued that 'one past falsification of a generalization does not imply that the generalization is false in *future* instances. To assume that it will be falsified in similar circumstances is to make an inductive assumption, and without this assumption there is no reason why we should not continue to reply upon all falsified generalization' (p.95). If 'falsification' here means **FALSIFICATION**, then one past **FALSIFICATION** will imply that the generalization or the hypothesis is false, and false forever. It is deductive logic, rather than induction, that will guarantee that it will be false in future instances. However, some, in particular fallibilists, might argue that all factual statements cannot be established with absolute certainty and we would thus never be absolutely certain that whether the falsified hypothesis is **FALSIFIED** or *falsified*. It is true that we may not be absolutely certain if the falsification is **FALSIFICATION** or *falsification*, but it does not mean a hypothesis could not be falsified until we are absolutely certain of it. If we never risk falsifying a hypothesis that might be true, then no hypothesis would be falsified at all. In fact, Popper has stated clearly that anything like conclusive proof to settle an empirical question, such as finding a conclusive practical experimental proof of falsity, does not exist (Popper, 1983, p.xxii). That is to say, a hypothesis could be

falsified even though we are not sure if it is **FALSIFIED** or not. What we must bear in mind is that if a hypothesis is falsified by a test, it does not provide evidence that it must fail a repetition of the test although it does give reasons for believing that the hypothesis is false (Popper, 1974, p.1043). Likewise, unless we smuggle induction in somewhere, the fact that a hypothesis has passed a test, no matter how stringent, does not provide any evidence that it will pass a repetition of the test.

Despite saying this, we are not alleging that we have to repeat the same test again and again if the hypothesis has already passed the test. As Popper (1963/1989) has elaborated, we have to use our background knowledge in searching for a counterexample, for we always try to falsify the hypothesis with the most risky predications. If a theory has passed many such tests, then, 'owing to the incorporation of the results of our tests into our background knowledge, there may be, after a time, no places left where (in light of our new background knowledge) counterexamples can be expected to occur with a high probability. But this means that the degree of severity of our test declines. This is also the reason why an often repeated test will no longer be considered as significant or a severe...' (p.240). To put it in another way, we have to repeat a test only because we want to check if there were any mistakes that had been made in conducting the test, or if the observation is indeed one that can be repeated. But, in doing so, what we are really testing is not the substantive hypothesis itself. This thus explains why a repeated test is less severe.

Fifth, there are some hypotheses that have passed a number of strenuous tests. Our practical decisions are made on the basis of these best-tested hypotheses, not because

the passing of tests constitutes a reason for the hypothesis nor it could raise its probability of being true, but 'because these decisions (and the proposals from which they emanate) stand up best to criticism and rational comparison with other proposals about our practical decision' (Miller, 1980, p.127). It is worth noting that the proposal that stands up best to criticism is not necessarily the one that we have reason to suppose to be successful. Moreover, as Miller (1980) pointed out, 'the best we can do in the way of criticism is to deploy all the theoretical knowledge that we have at our disposal' (p.128). What we have to make use of in criticism is what we know now, not what we will know when our decision is implemented.

We have discussed a number of important differences between induction and falsification. In the next section we will then discuss how Popper's perspectives on falsificationism could shed light on the problems being encountered by SST.

6.4 SST and falsificationism

SST, no matter it is Fisher's significance testing, Neyman-Person hypothesis testing or their hybrid, is about statistical hypotheses. How is it possible to falsify a statistical hypothesis which comprises probability statements? Popper (1957/1980) stated very clearly that '*probability statements will not be falsifiable*' (pp.189-190). To see why, let us consider a very simple probability statement about a coin: the probability of getting a head is 1/2. Unlike the singular statement 'the weight of the coin is 8 g', which is clearly falsifiable, the hypothesis 'the probability of getting a head is 1/2' is consistent with all possible outcomes. No matter how long we toss the coin, say, for n times, and whatever number of heads we observe, e.g., m, the outcome will always have

a non-zero probability: $C_m^n(\frac{1}{2})^n$. Even though in 1000 tosses we get 1000 tails, we still cannot falsify the probability statement. If falsifiability is a criterion of demarcation between science and non-science, it would seem that statistical hypotheses could not be regarded as scientific.

In order to address this question, Popper (1957/1980) proposed the notion of methodological falsifiability, with the development of alternative accounts of randomness, convergence, and the formal theory of probability. Popper (1957/1980) first gave a modified definition of randomness on top of the Frequency theory of von Mises and then developed his own definition for what it is for a finite sequence to be random.[219] The claim that a sequence, whatever its distribution is, is random can be falsified if its initial behaviour, rather than just its eventual behaviour, is sufficiently remote from the ideal randomness. For example, even though the probability for a fair coin to have 1000 tails in 1000 tosses is greater than zero, this result is so unrepresentative or physically impossible that it can occur only within a much more extensive context, say, the sequence of 1000 tails will occur when there are 2^{1000} or more games, in each game the coin is tossed 1000 times. But it is physically impossible for a coin to be tossed for such a large number of times[220]. Hence, if we really obtain such a result, the probabilistic hypothesis that the coin is fair could still be practically falsified even though we are not certain if it has been **FALSIFIED** or *falsified*. In this

[219] This idea had been developed, by G.J. Chaitin, A.N. Kolmogorov, P. Martin-Löf, R. Solovay, into the information-theoretic approach, according to which a sequence is random if it is incompressible (or irreducible). For discussion on different definitions of randomness, see Chatin, 2001, pp.111-127.

[220] On one hand, no physical coin could be tossed for so many times without smashing into dust; and on the other our present universe does not have enough time to generate so many tosses: If we toss the coin once per second, it will take us about 3.4×10^{296} years to run 2^{1000} games. But our universe is only about 1.5×10^{15} years old! There are certainly cases that are not so clear-cut. In these cases, the rule will be decided by the researchers, but it will always be subject to criticism. We will return to this point later.

example, the practical falsification is based on some kind of methodological rule, namely, that the basic statement about the sequence of heads and tails would be accepted as a consequence of the statistical hypothesis if and only if the sequence is representative[221] in a physically possible context of tossing the fair coin.

The methodological rule that cases of extreme non-probability be excluded resembles the rule of rejection in SST. That's why some have thought that SST is a falsifying rule in Popper's falsificationism. For example, Keuth (2005) stated that 'As is common practice in statistics, a region of acceptance and a corresponding region of rejection is defined. Occurrences that are very unlikely in view of the hypothesis fall within the region of rejection. If a test results in such a sample, then the hypothesis is rejected for the present; otherwise, it is accepted for the present. The definition of a region of acceptance and a region of rejection functions as a methodological rule...We need not follow Popper's considerations any further, for they add nothing to what we know from statistics textbooks' (p.176). The first explicit falsifying rule for probability statement was proposed by Gillies (1971). This falsifying rule was endorsed by Popper (1957/1980, p.191), who regarded Gillies's proposal as 'a most important contribution to the problem of the falsifiability of probabilistic or statistical theories and falsifying statistical tests' (p.419). The falsifying rule indeed agrees with the procedures of SST (Gillies, 2000, pp.147-148).[222] If so, are there any differences between SST and falsificationism? We will discuss this problem in turn.

[221] 'Representative' here is not an objective notion. We will soon return to this point.
[222] See also Mayo, 1996 for adopting some forms of SST as methodological rules for falsification of statistical hypotheses.

First, the methodological rule, according to Popper (1957/1980), is 'a rule, for instance, which might demand that the agreement between basic statements and the probability estimate should conform to some minimum standard.' (p.204). The dividing line drawn by the methodological rule decrees that only reasonably representative segments (or reasonably 'fair samples') are permitted, while atypical or non-representative segments are forbidden. This rule is for statistical hypotheses, but similar rules are already present for error treatment in deterministic hypotheses. For example, suppose we are testing a deterministic hypothesis H, from which we could deduce that a particular measurable quantity x would have a value of x_0 under some conditions. By measuring x under these conditions, we would be able to check if its value does amount to x_0. In principle, H would be falsified if the value of x does not equal x_0. But in fact we will not expect, even though H is known to be true, that the value of x will be exactly the same as x_0, for errors and uncertainties in making measurements are almost always inevitable.[223] Hence, although the value of x does not exactly equal x_0, H should not be regarded as falsified if it is still sufficiently near x_0. Only if the value of x differs from x_0 to a certain extent, H would be regarded as falsified. In other words, we may specify an interval I, usually symmetrical about x_0, say $[x_0 - \delta, x_0 + \delta]$,[224] such that we will regard H as falsified if the value of x does not lie inside I, and maintain H otherwise. Hence, practically, the methodological rule is required not only by the falsification of statistical hypotheses but also by that of deterministic hypotheses.

[223] There are many potential sources of error and uncertainty in making measurements. For example, errors in making measurements include human error, calibration and systematic error, etc. See Gott & Duggan, 2003.

[224] This is manifest in particular when the experimental errors are non-systematic, and the measurements can be considered as independent trials such that the error in the measurement X, i.e. to $x - x_0$, is a random variable whose distribution is symmetrical about 0.

Second, according to falsificationism, when we put a statistical hypothesis to a test, we have to formulate a rejection region for the hypothesis. The rejection region here may be no different from that in STT. For example, the rejection region could be formulated in such a way that the probability of getting outcomes in this region is, say, 0.05 assuming the truth of the hypothesis. When the observed outcome really lies inside the rejection region, we will regard the hypothesis as falsified, and no more than that. That is to say, falsification, aiming at expelling the false hypotheses from the body of science, has no implication that when one hypothesis has been falsified another hypothesis[225], like the alternative hypothesis in Neyman-Pearson hypothesis testing, would thus turn out to be accepted. Besides, according to falsificationism, there are only two possible outcomes when we put a hypothesis to a test. Either the hypothesis is falsified or not. Contrary to Fisher's significance testing, in which the p value is regarded as a measure of degree of the evidence against the hypothesis, there is no place for the p value in falsificationism. Furthermore, as with the falsification of deterministic hypotheses, the falsification of statistical hypotheses could be mistaken and it is always possible for a falsified hypothesis to pass the repeated test. However, if the repeated tests are independent and the hypothesis is true, the probability for the hypothesis to fail twice would be much smaller. On the other hand, if the statistical hypothesis passes the test, we simply maintain the hypothesis until it fails in another test. Passes in tests would not

[225] Of course, when H is false its negation $\sim H$ will be true. But the negation itself may not be an interesting hypothesis that we are looking for or it is even not falsifiable. For example, researchers in economics proposed the hypothesis called 'the law of demand' – for all goods, the lower the price the more quantity will one demand – to explain human's behaviour. If it were falsified, the negation would be: for some goods, it is not true that the lower the price the more quantity will one demand. Since it does not specify which goods would not obey the law of demand, when the price of a good drops, the quantity demanded, no matter it increases, decreases or remains unchanged, would always comply with this negation of the law of demand. That is to say, this new hypothesis will turn out to be practically unfalsifiable. And as we have discussed before, only falsifiable hypotheses, at least practically, would be allowed in science.

constitute any evidence for the truth of the hypothesis. Researchers should continue to make strenuous effort to falsify the hypothesis with another more severe test.

Third, what falsificationists are really concerned with is the hypothesis per se, rather than the Type I error rate in the long run or the truth of the alternative hypothesis. In other words, even if Neyman-Pearson hypothesis testing, contrary to what we have argued before, had succeeded in grouping all different hypotheses into a single batch and estimating that 5%, say, of them would be falsely rejected, this Type I error rate is not what falsificationists are bothered about, for it cannot tell if a particular hypothesis is falsified or not. Also, in Neyman-Pearson hypothesis testing, the rejection of the null hypothesis is tantamount to the acceptance of the alternative hypothesis. But for falsificationists, if we are really interested in the alternative hypothesis, we have to put the alternative hypothesis directly to test.

Fourth, one may note that when we discuss the methodological rules for statistical hypotheses and deterministic hypotheses, we encountered some notions such as 'representative in a physically possible context' or 'an interval I which is symmetrical about x_0' that seem to be quite arbitrary. These notions, like the rejection region, are postulated for practical purpose. For example, when testing Newton's laws, physicists might accept certain measurement error of distance. But later when A. Einstein postulated the theory of special relativity, the same standard might no longer work for detection of the time dilations for particles derived from the theory of special relativity. In order to falsify Newton's law, physicists have to put the theory to test under more stringent requirements. It is thus not meaningful for asking for an objective and one-

size-fit-all falsifying rule. If a hypothesis passes a test with a rejection region, it is always legitimate for other researchers to criticize it and test it again with a more stringent rejection region. That is to say, the rejection region is a factor that determines the test-severity rather than a predetermined fixed region that could be used for all practical purposes. Moreover, Type I error and Type II error rates are important notions in Neyman-Pearson hypothesis testing. But for falsificationists, all judgments are conjectural. No matter how we set the significance level, the long term Type I or Type II error rates, we cannot get falsification done once and for all. That is to say, falsificationists are always prepared to revise the previous decisions they have made about what has been falsified and what has not. They would only commit an error in tentatively maintaining a false hypothesis or the error in tentatively rejecting a true hypothesis.

Fifth, in SST, we sometimes need not postulate all details of the hypothesis before testing, especially when SST is regarded as an inferential process from sample to population. For example, when we conduct a one-sample *t* test to compare the mean of a sample to a population, the population variance is unspecified before observation for its value is estimated from the sample. In most research situations researchers use SST, as Aron, Aron and Coups (2005) mentioned, when they do not know the population mean, plus, they usually have not one set but two sets of data (p.228). No matter the *t* test is for dependent means or independent means, the means and the variances of the populations would not be pre-specified. But for falsificationists, it does not matter if we know the population mean. What we have to do is to postulate the details of the hypothesis before putting it to test. If we do not specify the details of the hypothesis

clearly, we can hardly make continuous effort in putting the *same* hypothesis to tests. For example, if we want to test the hypothesis which claims that the Secondary Four students who get a special lesson in using Dynamic Geometry Software would perform better in geometry, we select a group of students to receive the special lesson. Before and after the special lesson, the students are tested with their ability in solving problems in geometry. In this repeated-measures design, a standard *t* test for paired samples[226] would be performed to test the null hypothesis: students' scores in the test (which measures their ability in solving problems in geometry) do not change from before to after the special lesson. Since there are many factors that could influence a student's score in the test, we would not expect that a student's difference score (after – before) must be zero[227] even though the null hypothesis is true. According to SST, we assume that if the null hypothesis is true then the population of difference scores will be normally distributed with a mean of zero. And we have to reject the null if the sample mean of the difference scores deviates much from zero. But it is noteworthy that whether the deviation is great or not depends on the dispersion of the distribution of the sample means, which in turn depends on the sample size and the dispersion of the distribution of the population difference scores. But in SST the latter is not specified in our null hypothesis until the sample mean has been obtained. We use the variance of the sample difference scores and the sample size to estimate the variance of the population difference scores. The estimated variances could be very different from one test to another. That is to say, even if we repeat the test again, the null hypotheses are different from one test to another.

[226] It is also known as '*t* test for dependent means', '*t* test for correlated means', or '*t* test for matched samples'.
[227] And of course the mean of the students' difference scores would not necessarily be zero too.

Besides, for the purpose of SST, the null hypothesis is merely a kind of straw man. What the researchers are really concerned with is the alternative hypothesis, which is in our present example that students' scores would increase from before to after the special lesson. As we have discussed in previous chapters, the rejection of null hypothesis on one hand cannot imply the truth of the alternative hypothesis and on the other cannot imply that the statistically significant difference is practically significant. For falsificationists, they have to specify the hypothesis with sufficient details, for example, the students' difference scores would be a normal distribution with a mean = 10 marks and standard deviation = 3 marks before putting the hypothesis to test. Of course, we have to specify a rejection region, say $\bar{x} < 10 - 2 \times 3$ or $\bar{x} > 10 + 2 \times 3$, to enable us to falsify the hypothesis practically. Here, one could choose a narrower rejection region to make the test less severe when the study is at the exploratory stage and make the test more stringent by broadening the rejection region in later stages.[228] Nevertheless, one may notice that the hypothesis would be easily falsified if it is stated in such a precise form. But it is not the fault of our approach. Only adopting the most severe tests could prevent us from becoming entrenched in false hypotheses camouflaged with their vague specifications.

Sixth, falsificationism could enable us to criticize others' hypotheses in a way that SST could not. Say, suppose some researchers report that the above null hypothesis about the effect of special lessons has been rejected by SST and then conclude that the students' scores would increase from before to after the special lesson. There are certainly a number of ways that the study could go wrong. For example, the tests might

[228] In exploratory studies, one could practically allow the mean to have a range of values and thus narrow the rejection region.

be invalid or not reliable; there might be researcher biases, such as halo effect[229]. Nevertheless, the best way for one who queries the conclusion is to repeat the test. But here comes the problem. As we have discussed in Chapter 5, this sort of null hypothesis can hardly be nil for there are numerous factors, other than the special lesson, that could intervene to affect the scores. The sample size could influence the probability of rejecting the null in this experimental design. Does it imply that the researchers who are skeptical about the rejection of the null have to keep the same sample size in their repeated test? Perhaps at least they have to give justification why they have reduced the sample size if it leads to result that the hypothesis is no longer being rejected. What is more, different sample sizes would result in different estimated population variances. The null hypotheses addressed by these researchers are not genuinely identical. It is thus hard for one to make any conclusion if one is rejected but another is not.

For falsificationists, they do not have the problems arising from the sample size. For one thing, they have to specify the variance of the population in the hypothesis before conducting the test. That is to say, the variance of the population would not be affected by the sample variance. Next, even though the hypothesis is easily to be falsified, it does not imply that we have to accept another hypothesis. Hence, we would not have the problems, as discussed in Section 5.1, confronted by SST, e.g., we might accept the truth of the alternative hypothesis simply by conducting a SST on the corresponding null hypothesis with a large sample size. Moreover, suppose some falsificationists have tried to put the hypothesis, with a specified rejection region, to test. If the hypothesis passes the test, all researchers skeptical of this result could put the

[229] For more other possible errors, see, for example, Onwuegbuzie & Daniel, 2003.

identical hypothesis to a repeated test. There will never be a conclusive test for a hypothesis. That is to say, hypotheses are always open to criticism. If the researchers are still disturbed by the hypothesis even though it has passed in repeated tests, they could put the hypothesis to more severe tests, for example, by improving the accuracy of measurement for a deterministic hypothesis or controlling the intervening variables for a statistical hypothesis so that they could have another rejection region to accept less extreme results as evidence for falsification.

More important, SST is used to direct researchers in deciding whether the difference has been produced by chance, sampling in our case, or by the treatment, the special lesson. However, suppose all intractable problems surrounding SST had disappeared, SST could at most inform researchers if the difference is due to chance. If the difference is not due to chance, SST will not tell us whether this difference is due to the factors other than the treatment. For falsificationists, even if the hypothesis passes in repeated tests, they have to propose more substantive hypotheses to test if the difference is due to the treatment. For example, if the difference is really due to the treatment instead of halo effect, one testable implication is that the difference scores would remain unchanged when we change other conditions, say the teacher involving in the special lesson, or the time between the two tests. Then we could change some of the conditions and test the hypothesis again. In this continuing process of criticism, we could render the hypothesis more precise and thus more easily refutable.

On the other hand, if the hypothesis fails in the test. Those who still think that the hypothesis is true could put it to a repeated test. If the rejection region is selected in

such a way that if the hypothesis is true there will be a probability of 0.05 that the sample means will fall in this region. Since they are testing the same hypotheses, falsifying a true hypothesis twice would be much less probable (0.05×0.05) if the tests are independent and have the same rejection region.

From the preceding discussion, we could see that there are a number of differences between SST and falsificationism. And it should now be clear that how falsificationism could contribute to solving the intractable problems encountered by SST. In the next section we will discuss how falsificationism could influence the conduct of education research.

6.5 Implications of falsificationism for education research

Popper's revolutionary approach to scientific method, in particular, and epistemology, in general, is also known as 'critical rationalism' which conceives human knowledge as consisting of bold guesses or conjectures. Criticism, according to Popper, should be applied in all fields of human experience, wherever there are problems for which we are interested in figuring out workable solutions. And educational discourse is certainly not an exception. There are many ways that critical rationalism could relate to education, for example, critical rationalism and the values underlying education, impacts of critical rationalism on learning and teaching practices, implications of falsificationism for the conduct of education research.[230] And for the purpose of our discussion, only the last issue will be addressed here. Moreover, we will use examples from the research on

[230] For a recent introductory overview of critical rationalism and educational discourse, see Zecha, 1999.

mathematics education for illustration all over this section although the major discussion below might apply equally to research in other educational contexts.

As we have mentioned in Chapter 1, a lot of money is spent per each year on education research (Pring, 2000). But it is dubious if the money is well spent (Hillage, 1998; Tooley and Darby, 1998). Education research is continually being criticized for its poor quality (Levin and O'Donnell, 1999) and its criticisms have been emerging for a number of years. The reputation of education research is said to be awful (Kaestle, 1993). One reason that could explain why such criticisms exist is that there are many conflicting theories and viewpoints. All are claimed by their proponents to be inspired by data or observations (Phillips, 1999, p.176). Indeed, if we count all observable implications of a hypothesis as evidence for the hypothesis, the hypothesis will always have evidence in its favour. There is, for example, evidence for the hypothesis that older and returning students would work harder in pre-calculus classes,[231] if we could find a returning student A who is older and work harder in a pre-calculus class than another student B. But at the same time, we might also find an evidence for another hypothesis that older and returning students would not work harder in pre-calculus classes, e.g. by identifying a returning student C who is older but does not work harder in a pre-calculus class than another student D. These two hypotheses are, however, contraries.[232] That is to say, the evidence for one hypothesis constitutes a falsification for another. If both evidences are true, the two hypotheses are indeed false even though each of them would have its own

[231] This question could probably be posed in a better way. But it is not a point that we have to bother with here. For discussion on posing research questions, see, for example, McKnight, Magid, Murphy, & McKnight, 2000, pp. 21-23.

[232] Two sentences are said to be contraries if and only if they cannot both be true but can both be false.

confirming evidence. This example shows that merely seeking for confirming evidence cannot effectively detect false hypotheses.

We know from preceding discussion that only by making continuous and strenuous effort to put the hypothesis to tests could prevent us from becoming entrenched in falsity. But falsificationism is still not prevalent, at least, in research on mathematics education. For example, in a recent conference on mathematics education held in Hong Kong, there were 10 theses and 22 research papers presented and later published in the *Proceedings of Conference on Mathematics Education 2005*. Many of them did not mention any hypotheses or theories in their theses or papers, for example, Leung & Park, 2005; Wong, 2005; Zhang, 2005. And those who had mentioned some hypotheses or theories, for instance, Brown, 2005; Cai, 2005, Mok, 2005, at the most, reported that they had found some evidences for the hypotheses but none of them had explained how the hypotheses could be falsified and certainly did not show any effort in falsifying the hypotheses. For example, both Mok (2005) and Huang and Mok (2005) explicitly mentioned the Variation theory[233] in their research reports. But when they tried to investigate the learning and teaching of mathematics in Hong Kong and Shanghai, they never asked what observations would constitute a falsification of the theory. It seemed

[233] According to the Variation theory, developed by Marton in collaboration with different researchers (Marton, et al., 2004), learning is a process in which we want learners to develop a certain capability or a certain way of seeing or experiencing. Experience of variation is an essential experience for discernment and to discern certain feature of something is necessary for seeing that object in a certain way. Hence, experiencing variation is significant for learning. They further argue that paying attention to what varies and what is invariant in a learning situation is crucial. Furthermore, learning always involves an object of learning. What is more important is how the teacher structures the lessons so that it is possible for the object of learning to come to the fore of the students' awareness, which is called the enacted object of learning. (Huang & Mok, 2005; Mok, 2005)

that all of the lessons, no matter what varies and what is invariant in a learning situation, would always conform to the Variation theory. Therefore, when they noted that, compared with Hong Kong mathematics teachers, Shanghai mathematics teachers practiced more with implicit variation in their lessons, they simply observed 'via the lens of the Variation theory' (Mok, 2005, p.25), without paying any attention to the problem whether the learners' outcomes are inconsistent with the Variation theory or not.

For falsificationists, the first thing to do in conducting a research on mathematics education is not only to propose a hypothesis that they hope is true. They also have to formulate the hypothesis precisely so that they could specify the conditions under which the hypothesis could be falsified. As Schoenfeld (2000) has noted, questions like 'Do students learn as much mathematics in large classes as in small classes?' could hardly be answered in the abstract. But Schoenfeld's criticism on these unanswerable questions seems to go astray. When addressing this question, according to Schoenfeld (2000), one must immediately ask, 'what counts as mathematics? How much weight will be placed, for instance, on problem solving, on modeling, or on the ability to communicate mathematically?' (p.642). He thought that a researcher has to know what to look for and what to take as evidence of it before being able to determine whether it is there. This saying is not mistaken. For example, without knowing the meaning of 'better', how could we determine if one approach in teaching algebra is better than another? It does make sense to require a researcher to know what to look for and what to take as evidence for the hypothesis. However, in reality this requirement is not difficult to meet. And, more important, this is not enough to discern falsifiable hypotheses from non-falsifiable. For example, one major tenet in Cai (2005) is that 'high quality mathematics instruction

should not only provide students with the opportunity to learn important mathematics and participate actively in the processes of constructing knowledge; it also should provide a setting for students to explain and justify their thinking and challenge the explanations of their peers and teachers.' (p.51, 67). What does 'high quality mathematics instruction' mean? Cai (2005) interpreted a high quality mathematics education as one conducted with effective classroom instruction. And an effective classroom instruction comprises of four critical features: students' learning goals, instructional tasks, classroom discourse, and the role of teachers (Cai, 2005, 50-53). So far so good! It seemed to meet Schoenfeld's requirement. But when looking into the details of classroom discourse, we could find that classroom discourse does indeed refer to the ways of representing, thinking, talking, and agreeing and disagreeing that teachers and students use to engage in instructional tasks (Cai, 2005, p.62) A desirable discourse in mathematics teaching is explicated by two teaching episodes, adopted from Thompson et al. (1994). And Cai (2005) praised the one which has provided a setting for students to reason and reflect on their reasoning (p.66). In other words, although Cai seemed to know what to look for, his major tenet, ignoring the details, is something no more than that high quality mathematics instruction, which has to provide students X, should provide students X. That is a tautological statement that no observation could falsify it. Moreover, even though the major tenet were not in the above form, the statement 'E should provide students Y' itself is a value judgment instead of a factual judgment. No matter A does in fact provide students Y or does not, 'E should provide students Y' could still be true. Cai's major tenet is thus not falsifiable even though it satisfies Schoenfeld's requirement.

As we have discussed, after proposing a falsifiable hypothesis, we have to carry out a study that is designed as a test which aims to detect if the hypothesis is erroneous. And the details of how falsificationism sheds light on tackling problems encountered by SST have been discussed in the last section. Although SST might give many people an impression that it is supported by sophisticated mathematics (Menon, 1993), SST is not so widely adopted in research on mathematics education. In a recent study of the use of statistical procedures in mathematics education research conducted by the present author (Ng & Wu, 2003), all quantitative articles, i.e. those involved collecting data and use of statistical analyses, published from 1994 to 2003 in the *Journal for Research in Mathematics Education* and *Educational Studies in Mathematics* were examined. It was found that only a few percentages of studies had adopted inferential statistical analysis. For example, the percentage of adopting *t* test is shown below:

	Journal for Research in Mathematics Education	Educational Studies in Mathematics
1994 - 1998	8.67 % (n = 173)	3.19 % (n = 188)
1999 - 2003	3.37 % (n = 178)	1.54 % (n = 195)

Table 5 The percentage of adopting *t* test

The percentages for adopting ANOVA, MANOVA, ANCOVA, Chi-squared test, *F*-test, correlation coefficients, and effect sizes are even smaller. The most favorable statistical tools adopted were for descriptive purposes:

	Journal for Research in Mathematics Education	Educational Studies in Mathematics
1994 - 1998	12.72 % (n = 173)	24.47 % (n = 188)
1999 - 2003	14. 04 % (n = 178)	20.51 % (n = 195)

Table 6 The percentage of descriptive statistics

The misconceptions of SST and related fallacies are also prevalent in the research on mathematics education where SST has been used, even though some may think that researchers on mathematics education should be much less susceptible to making mistakes in statistical reasoning. For example, Menon (1999) remarked that there was a researcher, having 25 years of teaching experience in mathematics and a good statistical background, who conflated the two conditional probabilities $P(H_0 \mid D)$ and $P(D \mid H_0)$ in a report on the effectiveness of writing-to-learn approach to learning mathematics (p.8). We have discussed the problems of SST and how falsificationism could shed light on tackling the problems encountered by SST. There is no point to repeat here. One last point we have to make here is that although there are many different types of research on mathematics education and what we have discussed in this chapter is focused on the quantitative research which involves the use of SST, it should not be assumed that other types of research could entirely insensitive to the implications of falsificationism, especially the points we have made in this section[234]. Phillips (1999), for example, has indeed discussed how falsificationism could enhance qualitative research.

In conclusion, if our purposes for conducting research on mathematics education are, as Schoenfeld (2000) suggested, to understand the nature of mathematical thinking,

[234] Indeed the research studies in the examples we have already discussed in this section are not quantitative.

teaching and learning and to use such understandings to improve mathematics instruction, our study has to be in touch with the empirical world and science would then step in. And only a science of mathematics education would be of use to the advance our understanding of the nature of mathematical thinking, teaching and learning and to education practitioners, including policymakers and teachers, in improving mathematics instruction. The science of education is not necessarily concerned with quantitative research. In light of Popper's falsificationism, the only requirement for a hypothesis or theory to be regarded as scientific is its falsifiability. It does not matter whether it is qualitative, quantitative, or statistical, in nature. Sometimes researchers are busy with the collection of data but without understanding the roles of data or observations. First, we should not expect that a great theory or hypothesis would automatically come out once we collect sufficient data. Second, no matter how many favorable observations we make, our hypothesis can never be conclusively confirmed or verified. The most important role of data is to falsify the hypothesis. If our hypothesis is statistical in nature, we could adopt some statistical falsifying rule to render the statistical hypotheses practically falsifiable. No matter the hypothesis is deterministic or statistical, its truth will never be certain. SST would not be able to guarantee the truth of hypotheses and any attempt in searching algorithmic rules for justifying hypotheses are doomed to be failures. The science of mathematics education is best construed as a process of criticisms. And, as Swan (2003) noted, 'the value of science both as a means of advancing knowledge, and of helping us to improve practice, lies only – but not insignificantly – in the method of criticism' (p.265). So long as we strenuously test those beliefs that we are using to guide our educational practices, eventually we will be able to detect our current faults, for example, we will one day eliminate some false judgment in regarding some teaching

practices as effective. Although our concern here has been with the science of mathematics education, it does not entail that we attach no value to other research. For example, the study on the aims of mathematics education is important,[235] even though it is primarily concerned with the value judgment (Ng, 1999), which can hardly be studied with the scientific methods. More important, there is clearly a place for our present study even though it is not scientific in nature. Science of mathematics education is thus not a panacea. It is only the most effective means at our disposal in addressing empirical questions.

[235] The problems addressed there are sometimes very important. For example, answer to the question 'why we require all secondary school students to study mathematics' could have very great implications for many people.

Chapter 7 Conclusion

SST has a long history. The first paper on the formal test of statistical hypothesis was written by John Arbuthnott about 300 years ago. But the basic logic underlying his arguments is still prevalent in the current practice of SST – a hypothesis is rejected because the observed data would be very unlikely if the hypothesis is true. The inference is so true that many of us are tempted to accept it. Much of the technical details of Arbuthnott's arguments had been challenged and advanced by his contemporaries and later commentators, as we have examined in Chapter 2. But this inference is an exception. Even nowadays there are many researchers who still regard this inference as valid.

Probability is inevitable in SST, which can be revealed in the above long-lasting inference. This inference is based on one important premise – the observed data is unlikely if the hypothesis is true. On one hand probability is a best measure of how unlikely the observed data are. That explains why, as we noted in Chapter 2, the probability theory, in particular Bernoulli's limit theorem and Bayes's theorem, has played an important role to the advancement of SST in its early stage of development. And on the other, 'probability of a hypothesis' is contained in a hidden premise of the inference – a hypothesis has to be rejected if it has a very low probability to be true. Apart from this inference, 'probability of a hypothesis' is also invoked in resolving the problem of induction. But what do we mean by 'probability of a hypothesis' if the hypothesis is not a possible outcome of a chance process? It is this question that led us to examine different interpretations of probability in Chapter 3.

After a brief discussion of different interpretations of probability, we came to the conclusion that the classical theory, the logical theory, and the frequency theory all suffer from fatal difficulties. To the subjective theory and the propensity theory, stiff objections are inescapable. However, it was concluded that the propensity theory is the one that are most able to stand up to its objections (Section 3.5). We certainly did not intend to settle all controversial issues surrounding the interpretations of probabilities for it is such a big task that requires at least another project. Moreover, we have noted that even granted that it would be the subjective theory, instead of the propensity theory, that could be most able to stand up to the harsh criticisms, our major conclusions about the role of SST in education research would still be tenable.

Nevertheless, we have come to three conclusions. First, if we do not accept the subjective theory of probability, talking about the probability of a hypothesis that is not the outcome of chance processes is unintelligible (Section 3.6). On the other hand, if we are prepared to accept the subjective theory of probability, we have to explain how the difficulties encountered by the subjective theory could be resolved, as we have done for the propensity theory. Second, suppose we insist on assigning probability value to a hypothesis in the same way as we assign it to an event, then we have to accept that it is the hypothesis with low probability, rather than high probability, that we should aim at when conducting scientific research (Section 5.6). That is, however, contrary to the hidden premise in SST – a hypothesis has to be rejected if it has a very low probability for being true. Third, there is another usage of 'probability' when talking about the probability of a hypothesis. For this usage, a hypothesis will be regarded as more *probable* if it can stand up to more severe tests, or it has some other virtues that we think

a good hypothesis should possess. That is to say, the low *probability* of a hypothesis could now constitute a good reason for us to reject it. But in this case the *probability* will not conform to the probability calculus (Section 5.6). Since the word 'probability' in the premise 'the probability that the data will occur is very small if the null hypothesis is true' refers to the one that conforms to the probability calculus, we cannot infer from this premise to the conclusion 'the probability that the null hypothesis is very small', where 'probability' has another meaning . Otherwise, no matter the argument form is valid or not, we would commit the fallacy of equivocation as we have put forward an argument where the word 'probability' changes from one meaning in the premise to another in the conclusion.

For those who believe that we should adopt the subjective theory or those who still insist on assigning probability value to a hypothesis in the same way as we assign it to an event, a detailed examination of the validity of the inference – a hypothesis is rejected because the observed data would be very unlikely if the hypothesis is true – is needed although in fact many people, including both the advocates and even critics of SST, either ignore this issue or simply regard the inference as valid when discussing the role of SST in conducting research. We thus devoted a rather long section to address this problem. With a detailed examination of the inference from different perspectives, we argued decisively that the inference is fallacious. Moreover, by making contrast with certain valid argument form and rule of inference, we could see why so many people have committed the fallacy.

As we have discussed in Chapter 2, SST is not a single concept that emerged all at once. There is indeed a cluster of concepts associated with SST and they were developed gradually since the publication of Arbuthnott's paper. We have also discussed how the two modern approaches of SST, Fisher's significance testing and Neyman-Pearson hypothesis testing, as well as their hybrid were evolved in light of the historical background discussed in previous sections. Despite the distinctions between these two approaches, they are also different from Arbuthnott's original idea. Moreover, not all advocates of SST would agree that we must invoke the concept of probability of a hypothesis in conducting SST. Hence, in addition to the above critical analysis, we have to examine whether the two approaches of SST and their hybrid could avoid invoking the concept of the probability of a hypothesis and become free of the insurmountable problems. But before that, we had to look more closely at the important notions and common misconceptions associated with SST. And it was our task in Chapter 4.

In Chapter 4, we first explained the logic of hypothesis testing in general so that we could see why SST would be invoked in conducting research. Distinctions between Fisher's significance testing and Neyman-Perason hypothesis testing were thoroughly discussed (Section 4.2). Only by doing so we could disentangle the subtle differences between the two most important notions in the modern versions of SST – p values and Type I error probability (Section 4.3) and understand why so many discussions on SST have gone astray.

To sum up, in Fisher's significance testing, the researcher attempts to reject the null hypothesis by establishing the probability (p value) of obtaining the observed or

more extreme outcomes under the assumption of the null hypothesis. The p value is a measure of the strength of evidence against the null hypothesis. The smaller the value of p, the greater the weight of the evidence is. There is no place for an alternative hypothesis or the control of the long run Type I or Type II error rates. It is interesting to note that Fisher himself simply regarded the inference as inductive, fluid, non-quantifiable and did not explicate the logic underlying the significance testing. Fisher and his advocates are thus confronted with two difficulties. First, what does it mean by 'evidence against the null hypothesis'? And how could it be measured? Second, suppose we know the answers to these questions, but why is the probability (p value) of obtaining the observed or more extreme outcomes under the assumption of the null hypothesis be used as measure of the strength of evidence against the null hypothesis? There are two possible answers to these questions. One invokes the probability of the null hypothesis, e.g. the evidence against a null hypothesis is the one that renders the hypothesis less probable, or the smaller the value of p, the more improbable the null hypothesis will be true. The underlying reasoning is, however, fallacious (Section 5.4 & 5.5). Another invokes the likelihood ratio associated with the observed or more extreme data but not the probability of a hypothesis. But in this case, we have to ignore Fisher's original idea that no alternative hypothesis is required in significance testing. This significance testing would encounter another problem, i.e. it would lead to misleading results as we have shown that the use of tail region to represent a result that is actually on the border would overstate the evidence against the null hypothesis (Section 5.4).

In contrast to Fisher's significance testing, Neyman-Pearson hypotheses testing does not aim at the truth of a single hypothesis and does not regard the mere occurrence

of rare outcome as adequate evidence for the rejection of the null hypothesis. What it is concerned with is to control the rate of mistaken conclusions in the long run. That's why the concept of Type I errors and Type II errors were developed in Neyman-Pearson's approach. It is important to note that the level of significance must be specified or fixed prior to the analysis of data. But many people have ignored this problem and even some suggested to reporting the observed p value as the exact significance level. We have provided a novel explanation in Section 4.3 for why such a blur of distinctions between the p value and the significance level would lead to undesirable consequences. And it also explains why the hybrid of these approaches, i.e. using the p value to limit our mistakes in the long run and at the same time to assess the truth of the null hypothesis, will not render SST to be more tenable. Although Neyman-Pearson hypothesis testing does not involve the concept of the probability of a hypothesis, it does have some other serious problems that can hardly be resolved. The most important one is that to reject a hypothesis at a certain Type I error rate means only that the rule, prescribing us to take an action, will ensure that in the long run we shall not be too often wrong. But it cannot address the researchers' genuine concern, i.e. the truth of the hypothesis they are interested in. What is more, the report of the calculated p values as the significance level will distort the overall long term Type I error rate (Section 4.3). Hence, an individual researcher cannot make the hypothesis testing itself more stringent merely by choosing a smaller significance level. The Neyman-Pearson hypothesis testing thus cannot provide an effective means for rejecting false hypotheses.

We have also discussed some widely held misconceptions about SST, such as the conflation between the two conditional probabilities, i.e. probability of rejecting a null

provided it is true and the probability that the null is true provided it has been rejected, level of significance being an indicator of the probability of successful replication, null hypotheses being tantamount to nil hypotheses, and the conflation between statistical significance and practical significance. On these grounds, we could on one hand see the limitation of SST and on the other judge whether some generally received criticisms on SST are really fair. First, it is not true that all null hypotheses must be false on a priori grounds or they must be able to be rejected with sufficiently large sample size, for the truth is dependent on whether it is self-selected groups design or a true experimental design (Section 5.1). Second, if a null hypothesis is not a nil hypothesis, a rejection of the null hypothesis could be an indication of an important difference (Section 5.2). More important, if what the researchers are interested in is a nil null hypothesis, for example, they aim at testing whether the difference is produced by chance or by the treatment, then the questions like 'How large is the effect?' or 'Is the effect large enough to be useful?' are not the questions that they have to bother with. Third, the average power of SST in research literature was found to be between 0.4 and 0.6. That is to say, about half of the tests for false null hypotheses are non-significant. But this is not an intractable problem for large enough sample sizes could be able to ensure high power (Section 5.3).

It should be concluded, from what has been discussed above, that although some criticisms on SST are not really fair, SST has insurmountable problems that could misguide the research paradigm. Many critics of SST have thus proposed different alternatives to SST. It must be noted that even though different reasons have been proposed for the use of CIs, we found, after a detailed examination of these reasons, that CIs have many problems too. And more important, the logic underlying CIs is the same

as that underlying SST. CIs are no better than SST for the purpose of testing null hypothesis (Section 6.1). Moreover, an algorithmic method for conducting research, i.e. what we have to do is to design an experiment and to collect data and there are ready-made automatic rule for making inferences from data, is certainly a tempting idea. But, as we have argued in Section 6.2, it is unreasonable to expect the existence of a statistical test that could provide algorithms or rigid rules by conforming to which all problems about testing hypotheses could be solved.

In proposing alternatives to SST, many critics of SST ignore one deeper reason for the failure of SST – the inductive reasoning underlying SST. Not only SST but also other tests invoking induction would fail in making valid inference from data to the hypothesis. This point becomes more apparent when making a contrast between induction and falsificationism in Section 6.3. According to the traditional view, advocated by many falsificationists, SST is a methodological rule for falsifying statistical hypotheses. We have, however, argued that there are indeed subtle differences between the methodological rules in falsificationism for falsifying statistical hypothesis and SST. And in light of these discussions, we could be clear how falsificationism could eschew the disadvantages of SST and other similar statistical inductive inferences (Section 6.4).

Education research is continually being criticized for its quality. Teachers often complain that the research output does not really matter to them or is futile to their practices. There are many reasons for that. One possible reason is that there are many conflicting theories and viewpoints in education research. And at least to this issue

falsificationism could make its distinctive contribution. The contribution is two-fold. First, falsificationism requires the hypotheses to be falsifiable if the hypotheses are about our empirical world. The researchers have to formulate the hypothesis precisely such that the conditions under which the hypotheses could be falsified are clearly stated. By doing so, the tautological or metaphysical theories, void of empirical content, will be expelled from the body of educational theories. Second, it is not difficult to find evidence for a hypothesis even if the hypothesis is false. It is falsification but not confirmation that could effectively discern the false theories.

The contribution is not only to the so-called 'quantitative research'. Whenever the hypothesis is about our empirical world, no matter it is quantitative or qualitative, falsificationism could be invoked. That's why the examples we adopted for illustration in Section 6.5 are not restricted to only the quantitative studies. Moreover, although most examples in this thesis were drawn from the research on mathematics education, the major discussion might apply equally to research in other educational contexts. Finally, it is interesting to note that Carver (1993) has subjected the published data of Michelson and Morley, which were used in 1887 to test the hypothesis that light travels through a medium called ether, with a simple analysis of variance. Historically, as we know, Michelson and Morely had concluded without using SST that light travels the same speed no matter what direction it is traveling. Based on their result, special relativity had been developed later by Albert Einstein. Would Michelson, Morely and Einstein change their minds if they had known that Carver (1993) could find statistical significance associated with the direction the light is traveling ($p < .001$)?

We do really hope that there could be research paradigms that would bring education research into a more prosperous situation in which teachers and other educational professionals would find the research output really matters to their practices. SST, as it should be clear now, is certainly not one of them, unless we prepare to include '*educational ether*' in our theories.

Appendix 1 A Proof of Arbuthnott's first argument

Here is a proof of Arbuthnott's first argument in modern terminology: suppose that the probabilities of getting M and F are both 0.5, the probability of getting $n/2$ M's by throwing n two-sided dice (where n is even) is:

$$C_{\frac{n}{2}}^{n}\left(\frac{1}{2}\right)^{n}=\frac{n!}{((n/2)!)^{2}2^{n}}\ .$$

By using Stirling's formula[236], we can easily show that its limit will be zero as n approaches infinity:

$$\lim_{n\to\infty}\frac{n!}{((n/2)!)^{2}2^{n}} = \lim_{n\to\infty}\frac{\sqrt{2n\pi}(\frac{n}{e})^{n}}{\left(\sqrt{2(\frac{n}{2})\pi}(\frac{n}{2e})^{\frac{n}{2}}\right)^{2}2^{n}}$$

$$= \sqrt{\frac{2}{\pi}}\lim_{n\to\infty}\frac{1}{\sqrt{n}}$$

$$= 0$$

But does it imply that 'it is very improbable that the outcomes would never reach as far as the extremities'? Although there is always a positive probability of getting an extreme outcome, the probability will rapidly diminish as n increases. That is to say, if we use S_n to denote the number of M's we get in throwing n two-sided dice. What we have shown above can be written as

$$\lim_{n\to\infty}P(S_n=\frac{n}{2})=0\,.$$

[236] According to Karl Pearson (1924), it was Abraham de Moivre who first gave the expansion of factorials and what James Stirling had indeed contributed is the determination of the constant term in the expression ($\sqrt{2\pi}$) (see also David, 1962/1998, pp.173-177). In this connection, maybe this formula or theorem should have been called 'de Moivre-Stiring theorem' as Pearson (1924) has proposed.

Some may think that the probability that S_n is near $n/2$ (or in a certain interval centred at $n/2$) will not diminish so rapidly as n increases (or its limiting value will no longer be zero). But this intuition is plainly false as shown below: for any positive number a,

$$\lim_{n\to\infty}\left(P(\frac{n}{2}-a\le S_n\le\frac{n}{2}+a)\right) = \lim_{n\to\infty}(\sum_{k=\frac{n}{2}-a}^{\frac{n}{2}+a} C_k^n\frac{1}{2^n})$$

$$= \lim_{n\to\infty}\left(\frac{2\sqrt{2}a}{\sqrt{\pi n}}\right)$$

$$= 0,$$

i.e., $\lim_{n\to\infty} P(-a\le S_n-\frac{n}{2}\le a)=0$.

It is an amazing but disappointing result: no finite interval $[-a, a]$ can be used to trap for the probabilities associated with $S_n-\frac{n}{2}$ as n approaches infinity. But, on the other hand, the weak law of large numbers tells us, as n approaches infinity, the probability that the average number of getting M's that deviates from 1/2 by more than any pre-assigned positive number ε will also tend to zero,

i.e. $\lim_{n\to\infty} P\left(\left|\frac{S_n}{n}-\frac{1}{2}\right|\ge\varepsilon\right)=0$, for any positive number ε .

Or in its equivalent form:

$$\lim_{n\to\infty} P\left(-\varepsilon\le\frac{S_n}{n}-\frac{1}{2}\le\varepsilon\right)=1,\ \text{for any positive number } \varepsilon .$$

Appendix 2 The Christenings in London, 1629-1710

Year	Males	Females	Year	Males	Females	Year	Males	Females
1629	5218	4683	1657	3396	3289	1685	7484	7246
1630	4858	4457	1658	3157	3013	1686	7575	7119
1631	4422	4102	1659	3209	2781	1687	7737	7214
1632	4994	4590	1660	3724	3247	1688	7487	7101
1633	5158	4839	1661	4748	4107	1689	7604	7167
1634	5035	4820	1662	5216	4803	1690	7909	7302
1635	5106	4928	1663	5411	4881	1691	7662	7392
1636	4917	4605	1664	6041	5681	1692	7602	7316
1637	4703	4457	1665	5114	4858	1693	7676	7483
1638	5359	4952	1666	4678	4319	1694	6985	6647
1639	5366	4784	1667	5616	5322	1695	7263	6713
1640	5518	5332	1668	6073	5560	1696	7632	7229
1641	5470	5200	1669	6506	5829	1697	8062	7767
1642	5460	4910	1670	6278	5719	1698	8426	7626
1643	4793	4617	1671	6449	6061	1699	7911	7452
1644	4107	3997	1672	6443	6120	1700	7578	7061
1645	4047	3919	1673	6073	5822	1701	8102	7514
1646	3768	3395	1674	6113	5738	1702	8031	7656
1647	3796	3536	1675	6058	5717	1703	7765	7683
1648	3363	3181	1676	6552	5847	1704	6113	5738
1649	3079	2746	1677	6423	6203	1705	8366	7779
1650	2890	2722	1678	6568	6033	1706	7952	7417
1651	3231	2840	1679	6247	6041	1707	8379	7687
1652	3220	2908	1680	6548	6299	1708	8239	7623
1653	3196	2959	1681	6822	6533	1709	7840	7380
1654	3441	3179	1682	6909	6744	1710	7640	7288
1655	3655	3349	1683	7577	7158			
1656	3668	3382	1684	7575	7127			

Table A2.1 The christenings in London, 1629 – 1710 (Arbuthnott, 1710, pp.189-190):

Appendix 3 Willem 'sGravesande's argument

Willem 'sGravesande's idea was that the probability of the observed data, given H_0, should be the probability of observing, in each of the 82 consecutive years, the number of male births falling between two specific values, defined by reference to the extremes of the observed data. His calculations were shown below[237]:

1. Since the annual total number of births (assuming to be the number of christenings) varied, he assumed that the yearly number of births was constant and its value was simply the average number of births over the 82 years.

   i.e. the yearly number of births $= \frac{938223}{82} = 11442.$

2. He then picked up the two years with the most extreme values for the ratio of male births: the total number of births.

Year	M/Total	Year	M/Total	Year	M/Total	Year	M/Total
1703	0.502654	1702	0.511953	1704	0.515821	1630	0.521524
1682	0.506043	1694	0.512397	1633	0.515955	1707	0.521536
1693	0.506366	1641	0.512652	1636	0.516383	1655	0.521845
1644	0.506787	1665	0.512836	1706	0.517405	1668	0.522049
1657	0.508003	1672	0.512855	1687	0.51749	1670	0.523297
1645	0.508034	1688	0.51323	1700	0.517658	1698	0.524919
1685	0.508079	1637	0.513428	1647	0.51773	1652	0.525457
1679	0.508382	1667	0.513439	1705	0.518179	1663	0.525748
1640	0.508571	1696	0.513559	1631	0.518771	1646	0.526037
1677	0.508712	1648	0.513906	1701	0.518827	1642	0.526519
1635	0.50887	1683	0.514218	1653	0.519253	1629	0.527017
1691	0.508968	1675	0.51448	1708	0.519417	1669	0.527442
1697	0.509318	1689	0.514792	1695	0.519677	1676	0.52843
1643	0.509352	1699	0.514938	1638	0.519736	1649	0.528584
1692	0.509586	1650	0.514968	1654	0.519789	1639	0.52867
1680	0.509691	1709	0.515112	1666	0.519951	1651	0.532202
1673	0.510551	1684	0.515236	1690	0.519953	1660	0.534213

[237] For details of 'sGravesande's calculation, see Hald, 1990, pp.275-285; Shoesmith, 1987, 133-146. The table here is generated with the use of Excel from Arbuthnott's original set of data.

1681	0.51082	1664	0.515356	1656	0.520284	1659	0.535726
1634	0.510908	1671	0.515508	1662	0.520611	1661	0.536194
1658	0.511669	1686	0.515517	1632	0.521077		
1710	0.51179	1674	0.515821	1678	0.521228		

Table A3.1 The ratio of male christenings to the total number of christenings in London, 1629 – 1710:

We can observe from the Table A3.1 that in 1703 the ratio was the least (0.502654) and in 1661 the ratio is the greatest. (0.536194).

Hence, the lowest 'observed' number of male births

$$= 0.502654 \times 11442 = 5751.$$

Similarly, the greatest 'observed' number of male births = 6135.

3. By using a binomial expansion with n = 11442 and p = ½, we can calculate the probability of observing the number of male births within the limits 5751 and 6135, in each of 82 successive 'average' years (with n = 11442). With the use of Maple 8, we get the following result[238]:

```
> sum(binomial(11442,k)*(0.5^11442),k=5751..6135);
```

.2906222800

```
> %^82;
```

$.9839437208\ 10^{-44}$

$$\text{i.e., } \sum_{k=5751}^{6135} C_k^{11442} \left(\frac{1}{2}\right)^{11442} = 0.291$$

[238] In the times where computer algebra systems were not available, 'sGravesande was required to do the summation longhand and he had invented certain minor sophisticated methods to cut down the calculations involved. See, Pearson, 1978, pp.301-2, for details.

And the probability of the observed outcomes over 82 years will thus become 0.98×10^{-44}, that is much less than the one by Arbuthnott's calculation: 2.07×10^{-25}.

Appendix 4 Nicholas Bernoulli's counterargument

Nicholas Bernoulli has proposed an argument against Arbuthnott's second argument. What Bernoulli challenges is the assumption of $p = 1/2$ for the probability of a male birth, on which both Arbuthnott and 'sGravesande had founded their logic. According to Bernoulli, this assumption was too restrictive an interpretation of 'chance' (Pearson, 1978, pp.161-162; Shoesmith, 1985, pp. 256-259; 1987, p.142). In his letter to de Montmort, he took the probability of a male birth to be 18/35. By adopting a model similar to 'sGravesande's, Bernoulli selected a constant value of 14000 for the total number of christenings (n) so that we needed not to use different binomial series for each of the 82 years. He then transformed the numbers of male and female christenings in each year in proportional to this constant value of n. As $p = 18/35$, the expected value of rescaled male christenings would be $14000 \times \frac{18}{35} = 7200$. The table of rescaled figures of christenings is as shown below:

Year	Males	Females	Year	Males	Females	Year	Males	Females
1629	7378	6622	1657	7112	6888	1685	7113	6887
1630	7301	6699	1658	7163	6837	1686	7217	6783
1631	7263	6737	1659	7500	6500	1687	7245	6755
1632	7295	6705	1660	7479	6521	1688	7185	6815
1633	7223	6777	1661	7507	6493	1689	7207	6793
1634	7153	6847	1662	7289	6711	1690	7279	6721
1635	7124	6876	1663	7360	6640	1691	7126	6874
1636	7229	6771	1664	7215	6785	1692	7134	6866
1637	7188	6812	1665	7180	6820	1693	7089	6911
1638	7276	6724	1666	7279	6721	1694	7174	6826
1639	7401	6599	1667	7188	6812	1695	7275	6725
1640	7120	6880	1668	7309	6691	1696	7190	6810
1641	7177	6823	1669	7384	6616	1697	7130	6870
1642	7371	6629	1670	7326	6674	1698	7349	6651
1643	7131	6869	1671	7217	6783	1699	7209	6791
1644	7095	6905	1672	7180	6820	1700	7247	6753

1645	7112	6888	1673	7148	6852	1701	7264	6736
1646	7365	6635	1674	7222	6778	1702	7167	6833
1647	7248	6752	1675	7203	6797	1703	7037	6963
1648	7195	6805	1676	7398	6602	1704	7222	6778
1649	7400	6600	1677	7122	6878	1705	7255	6745
1650	7210	6790	1678	7297	6703	1706	7244	6756
1651	7451	6549	1679	7117	6883	1707	7302	6698
1652	7356	6644	1680	7136	6864	1708	7272	6728
1653	7270	6730	1681	7151	6849	1709	7212	6788
1654	7277	6723	1682	7085	6915	1710	7165	6835
1655	7306	6694	1683	7199	6801			
1656	7284	6716	1684	7213	6787			

Table A4.1 Christenings in London (rescaled with $n = 14000$), 1629 – 1710

From this table, we note that the lowest value of the rescaled male christenings was 7037 (in 1703), which was 163 less than the expected value of 7200; and the greatest value was 7507 (in 1661), which was 307 greater than 7200. The next step was to calculate the probability that the rescaled male christenings would in any one year differ by no more than 163, either way, from the expected value of 7200:

$$\sum_{k=7037}^{7363} C_k^{14000}\left(\frac{18}{35}\right)^k\left(\frac{17}{35}\right)^{14000-k}$$

With the use of Maple 8, we get:

```
> p:=18/35:sum(binomial(14000,k)*(p^k)*(1-p)^(14000-
k),k=7037..7363);

> .9943058428
```

i.e., the probability is 0.9943.

Then we also notice that during the 82 years there are 11 deviations greater than 163:

Year	Males
1646	7365
1642	7371
1629	7378
1669	7384
1676	7398

1649	7400
1639	7401
1651	7451
1660	7479
1659	7500
1661	7507

That means the probability that the rescaled male christenings falling outside [7037, 7363] no more than 11 times in 82 years (and within them at least 71 times) is:

$$\sum_{k=71}^{82} C_k^{82} q^k (1-q)^k \text{ , where } q = 0.9943.$$

Using Maple 8, we have:

```
> sum(binomial(82,k)*(q^k)*((1-q)^(82-k)),k=71..82);
                    .9999999997
```

Such a high probability indicates that a chance mechanism could give rise to the regularity and consistency that Arbuthnott has observed.

Appendix 5 James Bernoulli's limit theorem

With the use of modern notation, the limit theorem presented by James Bernoulli can be put in this way:

Suppose a trial has $t = r + s$ equally likely outcomes, where r and s are two positive integers, of which r are favourable,[239] and let $p = \frac{r}{r+s}$. If s_n is the number of favourable outcomes in a series of n independent trials, then for any positive real number c, there exists an integer $N(r, s, c)$ such that

$$\mathrm{P}\left(\left|\frac{s_n}{n} - p\right| \le \frac{1}{t}\right) > \frac{c}{c+1}, \text{ for all } n \ge N.$$

Apart from proving the existence of the integer N, Bernoulli also shows how its value could be figured out: $N = \max(m_1 t + \frac{rt(m_1 - 1)}{s+1}, \quad m_2 t + \frac{st(m_2 - 1)}{r+1})$, where m_1 and m_2 are the smallest positive integers satisfying $m_1 \ge \frac{\ln(c(r-1))}{\ln(\frac{s+1}{s})}$ and $m_2 \ge \frac{\ln(c(s-1))}{\ln(\frac{r+1}{r})}$ respectively.[240]

After the proof of this theorem, Bernoulli gave one (and only one) example. We will use this example to illustrate the implications of this theorem: consider an urn with white and black pebbles in the ratio 3:2. We take out one pebble after another (with replacement) and observe how often a white pebble is withdrawn. If drawing a white

[239] It is also called, in Bernoulli's terminology, 'fecund events' or 'fertile events'.

[240] See Adams (1974) and Stigler (1986) for the outline of the proof in modern notation. A complete proof could be found in Hald (1990, pp.260-262). In modern texts, a much simpler proof can be obtained from Chebychev's inequality (see, for example, Hogg & Craig, 1995, pp.68-69). But Bernoulli's treatment has its own advantages over the modern one (for details, see Stigler, 1986, pp.66-69).

pebble is considered to be a favourable outcome, then $r : s = 3 : 2$. The values r and s could be chosen with some latitude (such as $r = 15$ and $s = 10$), depending on what limits for the relative frequency of the favourable outcomes we want to have. In Bernoulli's example, the values r and s were chosen to be 30 and 20, which meant that the limits for s_n/n were given by:

$$\left|\frac{s_n}{n} - \frac{3}{5}\right| \le \frac{1}{50}, \quad \text{or} \quad \frac{29}{50} \le \frac{s_n}{n} \le \frac{31}{50}.$$

He chose the value of c to be 1000, which meant that a 'moral certainty'[241] of 1000/1001 for this inequality to hold. With the use of the result shown above, he got m_1=211, m_2=301 and N = max(24728, 25550) = 25550.

As Bernoulli said in a letter to Leibniz, most of us know by some instinct of nature per se and by no previous instruction that the greater number of drawings, the surer we know about the proportion of white pebbles in the urn (Gigerenzer et al., 1989, p.30; Stigler, 1986, p.65). But the common sense does not suffice to show how many drawings warrant what degree of certainty. One of the great achievements by Bernoulli is the commencement of the journey toward a mathematical quantification of uncertainty. From the above calculation, we are now sure that when we make 25550 or more drawings, we will anticipate with a probability greater than 1000/1001 that the proportion of drawings in which a white pebble is taken out will differ from the true proportion of white pebbles in the urn (i.e. 3/5) by less than 0.02. Bernoulli's estimation of the number of drawings (N) was, of course, not the sharpest one. The number of trials 25550 was a

[241] Bernoulli applied 'moral certain' to an event whose probability nearly equals the whole certainty, so that a morally certain event cannot perceived not to happen. For Bernoulli's own elaboration, see Hald, 1990, pp.248-249.

very large number to Bernoulli especially when he intended to extend the result to solving practical problems. The number of stars listed in Flamsteed's catalogue (1725) in Bernoulli's time, as Stigler (1986, p.77) noted, was merely 3000 and the entire population of Basel was smaller than 25550. The number 25550 was thus at that time more than astronomical. This might explain why he seemed reluctant to publish his work though he had proved the theorem fifteen years before his death. He got such a large number partly because of his insistence on the moral certainty warranted by odds of 1000 to 1 which is much higher than the now common standard of certainty 19 to 1 (i.e. 0.95). Even though he had relaxed his standard to 'immoral certainty', say 19 to 1, the number of trials required would still be quite large – 15715. A number of people had thus tried to sharpen Bernoulli's result and reduce the number of drawings to a more practicable size. For example, Nicholas Bernoulli proved a theorem in 1713 by which the value of N was found to be 8400, and Abraham de Moivre reduced it further to 6500 from a theorem published twenty years after the two proofs of the two Bernoullis. (Hald, 1990, 267-274). Indeed, we could use Maple 8 to find that $N = 6450$ suffices for the required probability (but $N = 6449$ won't do).

Bernoulli's example acts as a model to investigate human mortality, weather or other important practical phenomena where the causes are hidden and the enumeration of equally likely cases is impossible. But in these real problems, the proportion of balls in the urn corresponds to the hidden causes and it is usually fixed but unknown. If the proportion of white pebbles in the urn is unknown, could we still use Bernoulli's result to estimate the proportion and determine how accuracy this estimate is from the number of favourable outcomes in a series of n independent drawings, s_n? At first sight, the answer

is affirmative – it seems that we could use Bernoulli's result to compute an interval, which can be taken as narrow as we wish[242], so that we may expect with a high probability that the true proportion will lie inside this interval, provided sufficiently large number of drawings, n, are made: it is obvious that the inequality

$$\left|\frac{s_n}{n} - p\right| \leq \frac{1}{t}$$

or the statement

$$s_n \in J(p,n), \text{ where } J(p,n) \text{ denotes the interval } [np - \frac{n}{t}, np + \frac{n}{t}]$$

is equivalent to the inequality

$$\frac{s_n}{n} - \frac{1}{t} \leq p \leq \frac{s_n}{n} + \frac{1}{t} \ ,$$

or the statement

$$p \in I(s_n,n) \ , \text{where } I(s_n,n) \text{ denotes the interval } [\frac{s_n}{n} - \frac{1}{t}, \frac{s_n}{n} + \frac{1}{t}] .$$

Then given the value of c, according to the Bernoulli's result, we seem to be able to ascertain that if the number of drawings n is N or more we may expect with a probability greater than $c/(c + 1)$ that the value of p will lie inside the interval $I(s_n,n)$.

A closer look into the calculations will reveal that the above argument is mistaken. Since N is a value which depends on the true proportion of white pebbles in the urn, we are still unable to know how many drawings are required to guarantee that we would have a probability greater than $c/(c + 1)$ that the value of p will lie inside the

[242] The length of the interval is given by $2/t$ in this example and we assume t being constant (=1/50) throughout the following discussion.

interval $I(s_n, n)$. With the modern notation of conditional probabilities, we could clearly distinguish the probability that Bernoulli has demonstrated how to compute:

$$P(s_n \in J(a,n) \mid p = a) \ (= P(a \in I(s_n, n) \mid p = a) \ )$$

and the one that we want to know when the true proportion is unknown:

$$P(p \in I(b,n) \mid s_n = b) \ (= P(b \in J(p,n) \mid s_n = b) \ )$$

.

The first probability reveals, when given the true proportion of white pebbles in the urn p, how likely it is that observed relative frequency (s_n / n) will approximate that probability to any desired degree of precision. What we really want to figure out in solving practical real problems is, however, the second probability which reveals, when given the observed relative frequency, how likely is it to approximate the true proportion.[243] It is apparent that Bernoulli himself had not conflated these two probabilities in his writings and there is not much evidence that he had pretended to a satisfactory answer to the problem of estimating the true proportion and determining the accuracy of this estimate. Maybe his failure to compute the second probability is another reason for his reluctance to publish *Ars conjectandi* (David, 1962, p.133; Hald, 1990, p.263; Glymour, 1992, p.196). The first systematic attempt to compute the second probability have to wait until 1763, half a century after the publication of *Ars conjectandi*, and this is our topic in the next Appendix.

Although Bernoulli's theorem cannot be used to estimate the true proportion and determine the accuracy of this estimate, it has been argued that Bernoulli's result does

[243] The second probability is also called 'inverse probability'. We will note, after the discussion of different interpretations of probability in the next chapter, that the first probability can be given any interpretation but the second one is only open to a limited kinds of interpretations.

begin to justify SST (Baird, 1981, p.50). For example, from the above discussion, we know that if an urn contains white and black pebbles in the ratio 3:2, in sufficiently large number of drawings (say, $n = 6450$) we will expect with a high probability (> 1000/1001) that the number of drawings in which a white pebble is taken out will lie between 3741 and 3999. If we find that, in 25550 drawings, there are 4010 drawings in which a white pebble is taken out; some may thus regard it as good evidence to reject the hypothesis that the true proportion of white pebbles in the urn is 3/5. The logic underlying this reasoning can be seen in subsequent development of SST.

Appendix 6 Bayes's theorem and SST

Here we would like to make two preliminary remarks regarding Bayes's theorem and SST. First, if an urn contains white and black pebbles in the ratio 3:2, in sufficiently large number of drawings (say, n = 6450) we will expect with a high probability (> 1000/1001) that the number of drawings in which a white pebble is taken out will lie between 3741 and 3999, i.e. $s_{6450} \in [3741,3999]$.[244] That means the following probability will be greater than 1000/1001:

$$P(E \mid p = 0.6),$$

where 'E' denotes the event $s_{6450} \in [3741,3999]$. Since $P(E \mid p = 0.6) + P(\sim E \mid p = 0.6) = 1$, it thus follows that the following probability will not be greater than 1/1001:

$$P(\sim E \mid p = 0.6),$$

where '~E' denotes that the event $s_{6450} \notin [3741,3999]$.

Suppose in 6450 drawings the number of drawings in which a white pebble is taken out is 4010 (i.e. $s_{6450} = 4010$), could we thus conclude that it is highly unlikely that the true proportion of white pebbles in the urn will be 3/5? Before invoking Bayes's theorem to express the posterior probability of the hypothesis that $p = 0.6$, we should first make clear one point about the relationship between the event $s_{6450} = 4010$ and ~E (i.e. $s_{6450} \notin [3741,3999]$). It is clearly that $s_{6450} = 4010$ implies ~E, but the converse does not hold. Hence, in general, $P(p = 0.6 \mid \sim E)$ and $P(p = 0.6 \mid s_{6450} = 4010)$ will be different.

[244] Here 's_n' denotes the number of favourable outcomes in a series of n independent trials. For details, see Appendix 5.

Which one should be adopted in answering our question in this very paragraph? I will argue a little bit later that only $P(p = 0.6 \mid s_{6450} = 4010)$ will make sense in the calculation of the posterior probabilities of hypotheses.

It can be easily shown that if A implies B then $P(A) \leq P(B)$[245]. Moreover, if A implies B then, for any proposition X, $A \wedge X$ implies $B \wedge X$, and hence $P(A \wedge X) \leq P(B \wedge X)$. By Bayes's theorem, it is certain that $P(A \mid X) \leq P(B \mid X)$, but the order of the sizes of the two probabilities – $P(X \mid A)$ and $P(X \mid B)$ – is still indeterminate[246]. In our present case, since $s_{6450} = 4010$ implies $\sim E$, we thus have $P(s_{6450} = 4010) \leq P(\sim E)$ and $P(s_{6450} = 4010 \mid p = 0.6) \leq P(\sim E \mid p = 0.6)$. With the use of Bayes's theorem, we have

$$P(p = 0.6 \mid s_{6450} = 4010) = \frac{P(s_{6450} = 4010 \mid p = 0.6)P(p = 0.6)}{P(s_{6450} = 4010)}.$$

Since the probability $P(\sim E \mid p = 0.6)$ is a very small number ($\leq 1/1001$), the probability $P(s_{6450} = 4010 \mid p = 0.6)$ will be much smaller. In fact,

$$P(s_{6450} = 4010 \mid p = 0.6) = C_{4010}^{6450} \times 0.6^{4010} \times 0.4^{2440} \approx 1.751243 \times 10^{-5}.$$

However, we cannot conclude, without making any assumptions about the prior probabilities $P(p = 0.6)$ and $P(s_{6450} = 4010)$, that $P(p = 0.6 \mid s_{6450} = 4010)$ will also be very small.

---

[245] See any standard text on probability logic, e.g. Adams (1998, p.32).

[246] For example, suppose a number w is selected randomly from {1, 2, 5}, assuming that each number is equally likely to be chosen. Let A : '$w \in \{1\}$', B : '$w \in \{1, 2\}$', X: '$w \in \{1, 5\}$', Y: '$w \in \{2, 5\}$'. We have A implies B, $P(X \mid A) = 1$, $P(X \mid B) = ½$, hence $P(X \mid A) \geq P(X \mid B)$. On the other hand, $P(Y \mid A) = 0$, $P(Y \mid B) = ½$, hence $P(Y \mid A) \leq P(Y \mid B)$.

Consider this example: suppose there are two urns, one of which (called 'U_1') contains 30 white and 20 black pebbles and the other urn (called 'U_2') contains 10 white and 40 black pebbles[247]. We first select one urn from U_1 and U_2, assuming that each urn is equally likely to be chosen. Without knowing which urn has been chosen, we take out one pebble after another (with replacement). We observe that in 6450 drawings the number of drawings in which a white pebble is taken out is 4010, which is NOT in the interval [3741, 3999]. Given these conditions, we could compute the probability that the urn selected is U_1 (in our example, the urn selected is U_1 if and only if p =0.6):

$$
\begin{aligned}
& P(p = 0.6 \mid s_{6450} = 4010) \\
= & \frac{P(s_{6450} = 4010 \mid p = 0.6) \times P(p = 0.6)}{P(s_{6450} = 4010 \mid p = 0.6) \times P(p = 0.6) + P(s_{6450} = 4010 \mid p = 0.2) \times P(p = 0.2)} \\
= & \frac{1.751243 \times 10^{-5} \times 0.5}{1.75124 \times 10^{-5} \times 0.5 + 3.164495 \times 10^{-1184} \times 0.5} \\
\approx & \quad 1 ,
\end{aligned}
$$

where p is the proportion of white pebbles in the selected urn.

It is thus clearly that we should conclude with a very high probability that the proportion of white pebbles in the selected urn is 0.6 (i.e. the urn is U_1) even though $P(s_{6450} = 4010 \mid p = 0.6)$ is very small. Let us consider a more conspicuous example which is the same as before except only that the urn U_2 now contains merely black pebbles. If in 6450 drawings the number of drawings in which a white pebble is taken out is 4010 then, without any calculation, we can conclude with absolute certainty that

[247] We will refer to this example by 'two-urns example' in later discussion.

the urn must not be U_2 because at least one white pebble is in the urn.[248] From this example, we could also see that why we should not use $P(p = 0.6 \mid\sim E)$ for computing the posterior probability of the hypothesis. Using Bayes's theorem to compute $P(p = 0.6 \mid\sim E)$ and $P(p = 0 \mid\sim E)$, we get respectively 9.945174×10^{-4} and 0.999005 and it seems to imply that we should expect that the urn selected is U_2 which is, however, inconsistent to the fact that the urn must NOT be U_2.

Second, in a section discussing about the relation between Bayes's theorem and SST, Baird (1981) has argued that 'if the prior probability for h is not very small then the likelihood $P(e \mid h)$ is the most directly relevant factor to the posterior probability $P(h \mid e)$. But this is the essence of the logic underlying SST: If the probability of e given h is low then probability of h given e is likewise low' (p.51). As shown in the above two-urns example, the prior probability for the hypothesis that the urn selected is U_1 is ½, which is denoted by $p = 0.6$, is certainly not very small; and the probability of the evidence, which is $s_{6450} = 4010$ in our example, given the hypothesis is $P(s_{6450} = 4010 \mid p = 0.6)$ which is certainly low ($\approx 1.751243\times10^{-5}$). But the probability of the hypothesis given the evidence, i.e. $P(p = 0.6 \mid s_{6450} = 4010)$, can hardly be regarded as 'likewise low' as Baird asserted. This sort of mistaken view is not rare as we will see in later chapters.

[248] If we insist on the use of Bayes's theorem, we get:
$P(p=0.6|s_{6450} = 4010) = 1.75124\times10^{-5}\times0.5/(\ 1.75124\times10^{-5}\times0.5+0\times0.5) = 1.$

Appendix 7 The nature of normal distribution

The normal distribution is fundamental to most of the modern statistical tests developed in the last century and its influence on social sciences is tremendous as we will see later.[249] The discovery of the normal distribution can be traced primarily to Abraham de Moivre who published in 1738 the first edition of *The doctrine of chances or a method of calculating the probabilities of events in play*, which may be regarded as a gambler's manual, in its revised third edition he continued the work of Nicholas Bernoulli and demonstrated a method of approximating the sum of a very large number of binomial terms in $(A+B)^n$.[250] As n increases, the number of terms in the expansion also increases and the graph of the distribution will begin to resemble a smooth curve, a bell-shaped symmetrical curve. For example, if a fair coin is tossed for n times, the probability of getting x heads (where $0 \le x \le n$) is

$$C_x^n \times 0.5^n .$$

In the following figure, the histograms of these probabilities are plotted for n = 10, 30, 100 and 1000. In each case, a curve given by the equation is also plotted:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}},$$

[249] About the normal distribution Bryan Morgan, a mathematician historian, has made a vivid remark, 'as characteristic of statistics as the hexagon is of organic chemistry or the parabola of ballistics' (1972, p.168).
[250] For a detailed illustration of de Moivre's normal approximation to the binomial distribution, see Hald, 1990, pp.468-508.

where $\mu = n \times 0.5$ and $\sigma = \sqrt{n \times 0.5 \times 0.5}$. This curve is now commonly known by the name that Karl Pearson has put it – the normal curve[251].

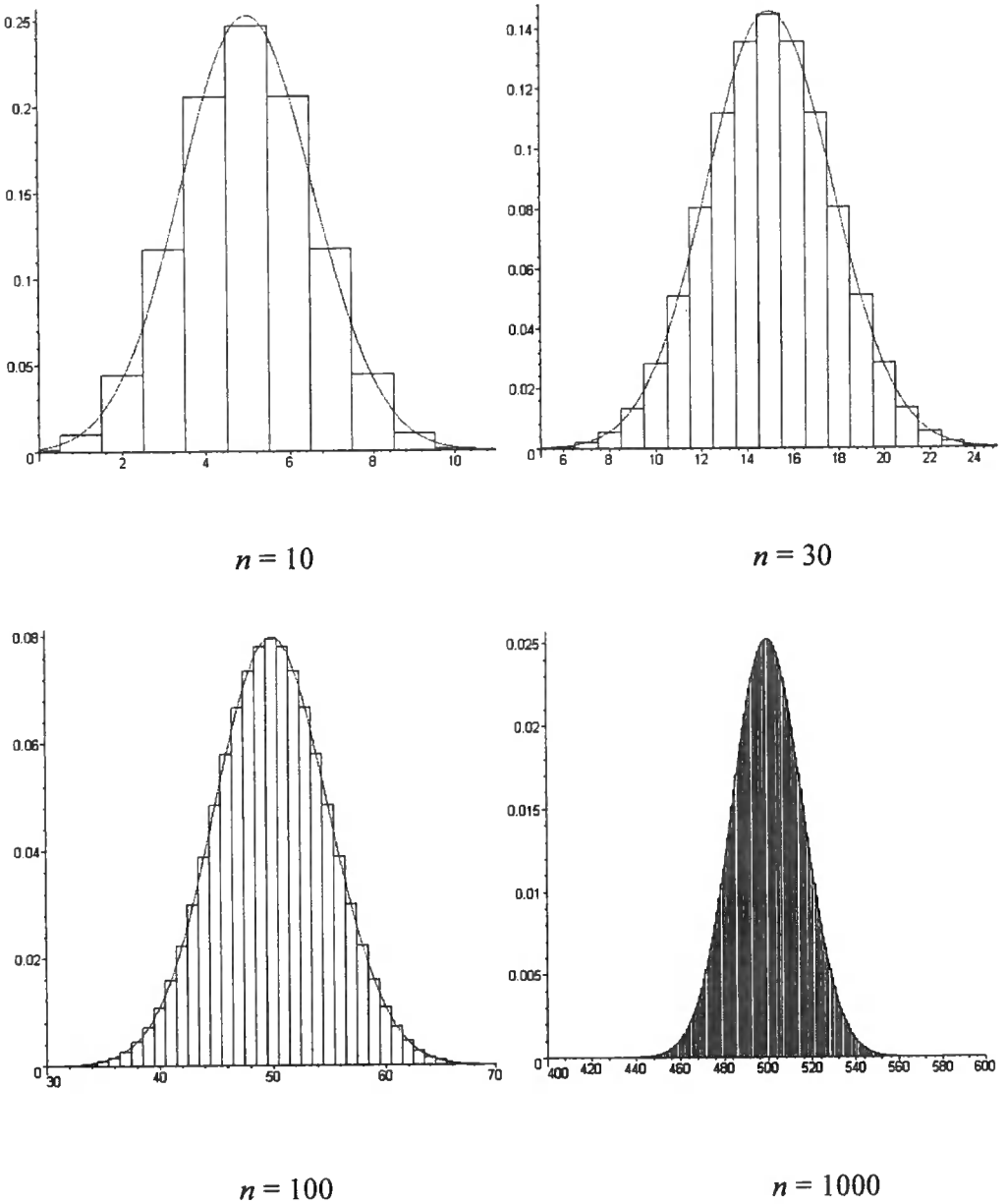


Figure A7.1 The histograms of the binomial (n, p =.5) probabilities superimposed with the corresponding normal curve for different values of n

[251] Karl Pearson is certainly not the first to use the term 'normal curve' and the first use of this term is still controversial (Cowles, 2001, p.10; David, 2001, pp.210-211; Tankard, 1984, pp.24-25). Recently, Kruskal and Stigler (1997) have argued that Charles S. Peirce should be credited with the first use of the term.

It is clearly that as n becomes larger and larger, the shape of the binomial probabilities becomes more and more like a normal curve.

The credit for discovering the normal curve is sometimes given to Carl Friedrich Gauss[252] and it can be revealed from the name 'Gaussian distribution'[253] which is often used to refer to the normal distribution. One of the earlier applications of the normal distribution outside of gaming is the assessment of errors in astronomical observations (Cowles, 2001, pp.10-12). It is Laplace's work, and contributions by many others, that interpreted the normal curve as the law of error[254] by which it means that measurements should follow the normal curve. An error, according to Laplace, is the resultant of a large number of sources of error of which one may affect the result in one direction or the opposite. Like every outcome of tossing a coin, every source of error may come up 'H' or 'T', if we use 'H' to denote the one that affect the result in one direction and 'T' the opposite. Gauss assumed explicitly that the two types of errors are equal likely, and thus derived that the distribution governing error is a binomial distribution of which the law of error is the limiting case when the number of sources of error tends to infinity.

Throughout the 1800's the normal distribution was used to describe a number of different phenomena. And the normal distribution was regarded by many researchers as a natural law. For examples, Adolphe Quetelet has provided masses of data (e.g.

---

[252] The credit also goes to Laplace, Charles S. Peirce and Wilhelm Lexis. According to Kruskal and Stigler (1997, p.86), such multiplicity of naming is conspicuous and it probably suggests that a prevailing contemporaneous evolving conceptual understanding of populations of people, of measurements and of their similarities in the 1870s

[253] In Germany, Gauss's portrait has been put on their 10 Deutschmark bill and a normal curve with equation is printed to the left of his portrait. See:
http://www.willamette.edu/~mjaneba/help/normalcurve.html

[254] It is also called 'the normal law of error', 'the normal law', or 'the Gauss-Laplace law of frequency of error' (Tankard, 1984, p.24).

measurements of the chest girths of 5738 Scottish soldiers[255]) that he claimed follow the normal distribution. (Boring, 1920, pp.10-11). These results had greatly impressed Francis Galton (1889, p.66)[256] who argued that the evidence, collected from various measurements (such as heights[257], span of arms, breathing capacity), for the law of error was in fact more than justified (p.56). But, as a law of nature, the law of error could hardly be a priori, analytic, or necessary[258]. It is always possible for the law to be false (i.e. its negation is always a possibility) or to be falsified. If a law cannot be falsified[259] no matter what experimental data we have, then it can hardly be used to provide factual information about our world. In its history of development, there are, however, researchers who have tried to give an a priori proof of this law. For instance, in 1850 Friedrich Wilhelm Herschel, a German-born British astronomer who discovered the planet Uranus, has argued in a way similar to Laplace that the probability of an error depends merely on its magnitude and not on its direction, and positive and negative errors

[255] It was this set of data that made Quetelet to become the first person to apply normal distribution to human data (Wild and Seber, 2000, p.238).

[256] A facsimile of the full text in PDF format could be downloaded at: http://www.mugu.com/galton/books/natural-inheritance/ .

[257] There are subcategories for certain types of measurements. Say, for height, it was subdivided into standing, standing without shoes, sitting, sitting from seat of chair. (Galton, 1889, p.201)

[258] There is a trio of distinctions that we will find useful in later discussion. They are the epistemological distinction between a priori and a posteriori, the metaphysical distinction between necessity and contingency, and the semantical distinction between analytic and synthetic truth. Roughly speaking, a truth is known a priori if it can be, in principle, known independently of experience of how things are in the world (e.g. a white swan is white, a father is the male parent); whereas a truth known a posteriori is one which can only be known on the basis on empirical investigation (e.g. snow is white, Pierre Simon Laplace died in 1827). An analytic sentence is one which is true or false merely in virtue of the meanings of the words used to make it and the grammatical rules governing their combination (e.g. an oculist is an eye doctor, a triangle has three sides) whereas a synthetic sentence is one whose truth could NOT be determined merely by the meanings of the words used to make it and the grammatical rules governing their combination (e.g. all dogs have kidneys, pure water boils at 100°C under normal conditions). A necessary truth is that whose denial will yield an impossibility, or it is true in all possible worlds (e.g. If all humans are featherless and Fisher is human, then Fisher is featherless; the sum of any two prime numbers that are greater than 2 is an even number) whereas the denial of a contingent truth is possible, or neither it nor its negation is necessary (e.g. All humans are featherless, there are nine planets in our solar system). For details, see any texts on philosophical logic, e.g., Grayling, 1997, pp.33-87; Wolfram, 1989, pp.80-128.

[259] A well-established law of nature is the one that can be falsified but has not yet been falsified by many observations. See Chapter 6 for further discussion.

are equally probable. Suppose we drop a ball from a given height and intend to make it fall on a given mark, errors in all horizontal directions are equally probable and that in perpendicular directions are independent. Hence, according to him, the law of error must necessarily be general and apply alike in all cases since the causes of error are supposed alike unknown (Baird, 1981, pp.62-63). But our question is: why must the positive and negative error be equally probable? Only our ignorance of the sources of error cannot warrant this equality. Suppose we are given a coin about which we know nothing, we cannot deduce from our ignorance of its fairness to the conclusion that the distribution of the outcomes will follow a normal curve especially when the number of tosses is not very large. For example, consider tossing a biased coin (probabilities of getting heads and tails are respectively 0.1 and 0.9) for 30 times. In the following figure, the histogram of probabilities are plotted and it is superimposed by the corresponding normal curve given by the equation:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $\mu = n \times 0.1$ and $\sigma = \sqrt{n \times 0.1 \times 0.9}$.

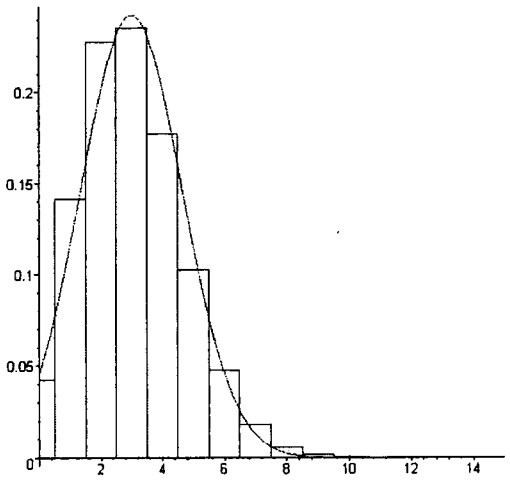


Figure A7.2 The histogram of the binomial (n = 30, p =.1) probabilities superimposed with the corresponding normal curve

From this figure, we can observe clearly that the histogram is skewed. Of course, it is now a well-known mathematical theorem that when n tends to infinity, the binomial distribution, no matter how p is different from 0.5, will approach the standard normal distribution[260] It can be illustrated by the biased coin example: the coin is now tossed for 500 times instead of 30 and the histogram of probabilities becomes symmetrical and looks more or less the same as the corresponding normal curve as shown in the following figure:

[260] For the theorem and its proof, see, for example, Miller & Miller, 1999, pp. 223-224. For an informal treatment to a normal approximation to a skewed distribution, and to the limit of a sum of a large number of random variables, see Hamming, 1991, pp.317-323.

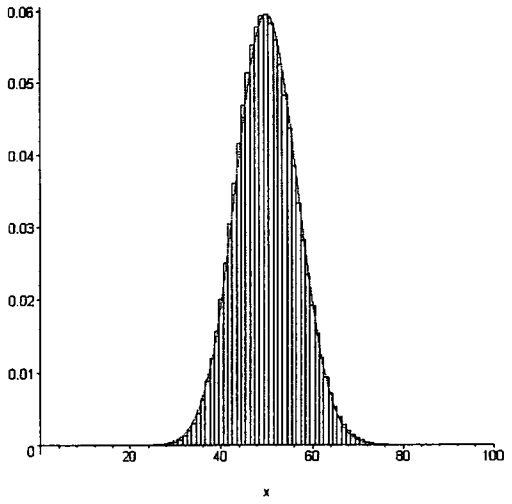


Figure A7.3 The histogram of the binomial (n = 500, p =.1) probabilities superimposed with the corresponding normal curve

At the time of Quetelet, many researchers, certainly including Quetelet himself, believed that all naturally occurring distributions of properly collected and sorted data would follow a normal curve, and that if there were any failure to exhibit such a shape, it was merely an evidence of ignoring some factors that could influence the results, or that the group under consideration is non-homogeneous. For instance, in an examination of the heights of 100 000 French conscripts, Quetelet found a discrepancy between observed and predicted values – out of 100 000 men there are 28 620 men who were of less heights than 5'2'', but there should only be 26 345 from the normal law. He explained this discrepancy by suggesting that there were Frenchmen who had stooped so low as to avoid military service (Boring, 1920, p.11). In 1863, Adolphe Bertillon, following Quetelet's study on the heights of French, found that the distribution of heights of 9 002 young men measured between 1851 and 1860 in the department of Doubs in France did not exhibit the usual symmetrical shape with a single modal value. What Bertillon found was a curve with two modal values. He then suggested that the inhabitants of Doubs must

consist of two human types and the suggestion was confirmed later by his colleague Lagneau (Stigler, 1986, pp.215-218).

The price for saving the normal law from the threats of falsification is high. First, the value of a scientific theory or a law of nature is that it is falsifiable. If we always regard the evidence of falsification as exceptions, the law will be irrefutable and it will become useless in explanation or prediction (Popper, 1963, pp.33-39). Second, if we always allow the possibility that the group under consideration is non-homogeneous then the occurrence of normal distribution will become an outcome entirely caused by the choices of researchers. For example, consider the distribution of the measurement of weights. As far as we know, no researchers have tried to survey the weights of all living things in the world. We hope it is very obvious that its distribution will not exhibit any shape like a normal curve (thinking about the weights of viruses and elephants, and their numbers). Perhaps some will argue that we should restrict our group to a smaller one. How about mammals? Mice and whales are mammals. The group seems to be too large. How about the species of Homo sapiens? Would different races (the pygmy of Equatorial Africa vs. the Aborigines of Australia), different age groups (babies and adults), different genders, different professions (sumo players[261] vs. jockeys) be regarded as different groups? The rule of thumb seems to be: narrow the group once you find that the distribution does not resemble the normal curve. At last, we would probably get a group that fits the normal law. But the normal curve will be, as Simon (1968, p.436) has put it, *made* by the researcher, rather than *met* by him (p.436).

[261] Sumo is the traditional national sport of Japan, usually played by very fat people.

According to Boring (1920, p.14), in 1894 Karl Pearson still gave evidence of being influenced by the sanctity of the normal law, but in 1900 his faith towards the sanctity was gone. Normal law is finally regarded as merely one of a series of important distributions. Of course, normal distribution still plays an important role in statistics and in particular SST, but for the reasons that are different from that believed by Quetelet or his contemporaries. First, it is a mathematical fact that if a random sample is taken from a normal distribution, then the distribution of various important functions of the observations in the sample can be derived explicitly and will themselves have simple forms[262]. As a result, for the mathematical simplicity and convenience, many researchers will tactically assume the normality of the distribution from which a random sample is drawn. Second, it is the central limit theorem that makes the normal law more prevailing. One of the implications of this theorem is that if a large random sample is taken from a distribution, no matter it is itself normal or not, the sampling distribution will be approximately normal. These two reasons are highly controversial. For the first reason, mathematical simplicity and convenience can hardly be reason for justifying that the distribution from which a random sample is drawn is normal. For the second, it is indisputable that the theorem has been proved. But the theorem itself does not state how large the random sample could warrant the normality of the sampling distribution and we will elaborate this point in Chapter 6.

[262] For the mathematical part, see DeGroot & Schervish, 2002, pp.268-280.

Appendix 8 Arguments against the classical theory

First, if we want to define probability as the ratio of the number of favourable outcomes to that of all possible outcomes, we have to assume that all possible outcomes are equiprobable.[263] But what does it mean by 'equiprobable'? If it is construed as meaning 'having an equal probability', then the concept of 'probability' enters into the very definition of 'probability' and vicious circularity thus arises. In order to avoid this circularity, some have tried to attach another meaning to 'equiprobable'[264] – the equiprobable outcomes have to meet two requirements: (1) equiprobable outcomes are those which are on the same logical level and can be subdivided in the same ways[265], and (2) there is no reason or evidence to support that one of the outcomes will occur rather than the others. The first requirement per se cannot warrant the outcomes to be equiprobable. For example, the six outcomes of a loaded die are on the same logical level but they are hardly equiprobable. The second requirement is what we have introduced – the Principle of Indifference. And it is this very Principle that makes the classical theory of probability to be discredited by many people.

The criticisms of the Principle of Indifference constitute the second objection to the classical theory of probability. The trouble with this principle is two-fold. One trouble is that the principle itself is not a priori means to warrant the outcomes to be equiprobable. It is commonly believe that the Principle of Indifference is applicable

[263] See, for instance, Laplace, 1825/1994, p.6.
[264] For examples, see Laplace, 1825/1994; Weatherford, 1982.
[265] In the example of throwing a die, the outcome of getting a '1' and that of getting '3' are of the same logical level since they can be subdivided in the same way – not being able to be subdivided; and getting '3, 4, 5, or 6' and getting '1 or 2' are not since the first case can be subdivided into 4 sub-cases but the second cannot. We will return to the discussion of this requirement in later paragraphs. See Keynes, 1921, p.60.

when the outcomes are equally undecided. But if indecision is to refer to one's psychological state, then different people might have different judgment and one's indecision could be totally mistaken. For example, a secondary school student might not be able to decide about whether a rational number or an irrational number is more likely to be selected from the interval [0, 1] and thus regards the two possible outcomes – rational or irrational – as equiprobable. The student is wrong and since the probabilities for selecting a rational and an irrational number are in fact 0 and 1 respectively. Hence, 'equally undecided' must not be referring to an individual's indecision.

The name of 'Principle of Indifference' might suggest that the Principle is only applicable to the outcomes amongst which there is no difference. But 'no difference' here must not mean that there is no difference in every aspect of the outcomes, otherwise the outcomes would become identical – there would be only one outcome[266]. Accepting that the outcomes are not exactly the same in every way, nor sharing all their qualities, we may suggest that the Principle could be applicable if and only the differences between the outcomes are not relevant to the problem. 'Not relevant to the problem' here means that the differences do not constitute any reason or evidence to support that one of the outcomes will occur rather than the others. But whether a certain piece of evidence is relevant to the occurrence of an outcome cannot in general be determined in a priori way. For example, in a car racing, we know that the colours of two cars are different. But we usually do not consider this piece of knowledge as relevant to their probability of winning – we will still apply the Principle of Indifference in this case unless we could find other

[266] There is a thesis called Leibniz's thesis of the identity of indiscernibles which states that no two things can be exactly the same in every way, sharing all their qualities (including the numerical-identity-with quality) (Kirwan, 1995, p.390).

known relevant differences, such as the differences in the size of engines or skill levels of the drivers, between the two cars. Another example is the fairness of die. Given a die, could we be sure that its six possible outcomes are equiprobable? Is the fairness of die a piece of a priori knowledge? The answer is certainly 'no'. We cannot distinguish a fair die from a loaded one without any empirical knowledge about the dice. One may argue that if the die is made of a perfect geometrical cube with perfectly homogeneous material, then we could establish in a logical sense that the die is fair and all possible outcomes are thus equiprobable. But how could we know that the die is a perfect cube? How could we know that the cube is made of homogeneous material? Or how could we know that different numbers marked on different faces will not affect its probability of landing as some primitive tribes might think that numbers could have some magical effects on physical world[267]? All of these are the questions that cannot be answered without a resort to empirical knowledge about the die. Hence, the Principle of Indifference is by no means theoretically fundamental and is hardly an a priori tool for us to assign probabilities to various outcomes.

Another trouble is that it yields a number of inconsistencies or paradoxes. Some of them can be eliminated by modifying the Principle of Indifference. For example, consider the paradox which is called the book paradox (Keynes, 1921, pp.43-44): Suppose we are about to borrow a book from a library, we have never seen the book before and we only know that its cover is mono-colour but have no idea what the colour is. It seems that we have no more reason or evidence to anticipate that the cover is red than not. Applying the Principle of Indifference, we have the probability that the cover is

[267] See von Mises, 1957, pp.72-73 for a discussion of this point.

red (i.e. P(red cover)) and P(not-red cover) are both 1/2. We could apply the same reasoning to black and blue, so we have P(black cover) = P(blue cover) = 1/2. Then P(red or black or blue cover) = 3/2 > 1, which is a contradiction. This paradox can be easily resolved if we introduce a further requirement that has been stated previously – equiprobable outcomes are those which are on the same logical level and can be subdivided in the same ways. The outcome that the book is non-red can be subdivided into black and not-(red or black), but the outcome that the book is red cannot be subdivided in the same way. Hence the Principle of Indifference is no longer legitimately applicable in the cases of red and not-red covers.

This suggestion to resolve the book paradox cannot, however, be used to eliminate the following paradox of specific volume and density[268]: Suppose we know that the specific volume[269] of a substance has a uniform probability density in the interval [1, 3]. By the Principle of Indifference, it is just as likely to be between 1 and 2 as between 2 and 3, so we have:

$$P\,(1 \leq \text{specific volume} \leq 2) = P\,(2 \leq \text{specific volume} \leq 3) = 1/2.$$

Since the specific volume of a substance lies between 1 and 3, its density is between 1/3 and 1. Applying the Principle of Indifference to equal intervals again, we have:

$$P\,(1/3 \leq \text{density} \leq 2/3) = P\,(2/3 \leq \text{density} \leq 1) = 1/2.$$

From these values, we have:

$$P\,(1.5 \leq \text{specific volume} \leq 3) = P\,(1 \leq \text{specific volume} \leq 1.5) = 1/2.$$

[268] Keynes (1921, p.45) attributes this paradox to von Kries. For the discussion of this paradox, see also Gillies, 2000, p.38; Weatherford, 1982, pp.56-57.

[269] Specific volume of a substance is defined as the volume divided by the mass, which is also the inverse of its density.

Combining with the first results of the specific volume, we have P (1.5 ≤ specific volume ≤ 2) = 0, which is contradictory to our assumption that is specific volume has a uniform probability density in the interval [1, 3].

Apart from this paradox, other paradoxes also arise in the application of the Principle of Indifference in the problems involving continuous parameters or geometrical probability. For the problems involving continuous parameters, paradoxes can be produced by the following method: consider a continuous parameter θ which takes values in a closed interval [a , b] , construct a bijective mapping f on [a , b] such that f is continuous and $f(\theta) \in [f(a), f(b)]$ if and only if $\theta \in [a, b]$[270]. If we have no reason to expect that θ is at one point of the interval [a , b] rather than another, we can apply the Principle of Indifference to give θ a uniform probability density in the interval [a , b]. Similarly, we have no reason to expect that $f(\theta)$ is at one point of the interval [$f(a)$, $f(b)$] than another and we apply the same Principle to give $f(\theta)$ a uniform probability in [$f(a)$, $f(b)$]. In general, the probabilities based on θ will be different from those based on $f(\theta)$ and thus inconsistence will arise. Let us use a simple example to illustrate this method of generating paradox. Suppose a square has been drawn on a paper and all we know is that the length of its side θ (in cm, say) lies in [1 , 4]. Consider its area A (in cm^2) which is a function of its length: $A = f(\theta) = \theta^2$. This function is clearly continuous and $f(\theta) \in [f(1), f(4)] = [1,16]$ if and only if $\theta \in [1,4]$. Since we have no reason to expect that θ is at one point of the interval [1 , 4] rather than another and A is at one point of the interval [1 , 16] rather than another, we an apply the Principle of Indifference

[270] In the paradox of specific volume and density, the parameter θ is the specific volume in the interval [1, 3] and the function is $1/\theta$.

to give each of θ and A a uniform probability density in the interval [1, 4] and [1, 16] respectively. When we try to figure out the probability that the area of the square is less than 9 cm^2, we get 2/3 based on θ but 1/2 based on A. These two answers cannot both be correct but we have no way to tell one is correct and another is wrong, paradox thus arises.

For the geometrical paradoxes, the most famous one is perhaps Bertrand's Paradox: given a fixed circle and select a random chord, what is the probability that this random chord is longer than the side of the equilateral triangle inscribed in the circle? It is now well-known that by using the Principle of Indifference in three plausible ways we could obtain three different answers: 1/2 (by considering the location of the mid-point of the chord along the length of the diameter that bisects the chord), 1/3 (by considering the angle between the chord and the tangent at one end-point of the chord), and 1/4 (by considering the area of the concentric circle which contains the centre of the chord).[271] Different considerations will give different answer to the same question. Similar difficulty appears in the needle problem. Gilles (2000) has indeed asserted, without further explication, that the needle problem raised by Buffon in 1733 is the earliest paradox arising from the Principle of Indifference. Buffon's needle problem itself is, however, not a well-known paradox. Many people regard this problem as a standard textbook problem without taking notice of the existence of paradoxes.[272] The problem goes in this way:

[271] These three solutions could be found in elsewhere. See for examples, Gillies, 2000, pp.38-41; Michalewicz and Fogel, 2000, pp.31-33. There is no need for us to go into its detail here.

[272] For example, in his now classic book on challenging problems in elementary probability theory, Mosteller (1965) has included this problem and give one solution to this problem. It seems that he does not know that this problem could have other solutions as what we will explicate below.

A table of infinite expanse has inscribed on it a set of parallel lines spaced $2a$ units apart. A needle of length $2l$, where $l < a$, is twirled and tossed on the table. What is the probability that when it comes to rest it crosses a line? (Mosteller, 1965, p.14).

Similar to Bertrand's Paradox, this problem can be solved by applying the Principle of Indifference in three plausible ways. The first way is what Mosteller (1965, pp.86-87) and many other writers have taken into consideration: the distance of the center of the needle from its nearest parallel line and the angle between the needle and the nearest parallel line. As shown in Figure A8.1, let P be the centre of the needle, x be the distance of P from its nearest parallel line, and θ be the angle between the needle and the line.

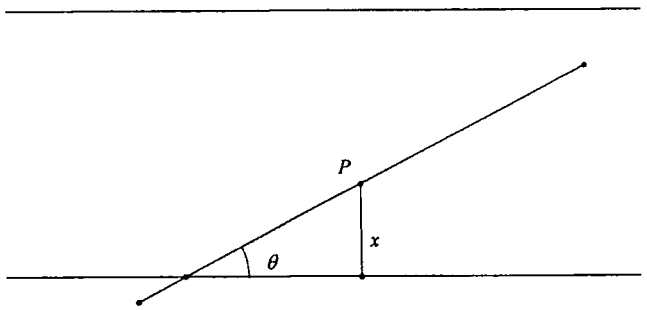


Figure A8.1

Then we have $0 \le x \le a$ and $0 \le \theta \le \pi$. The line will cross one of the parallels if and only if $x \le a \sin\theta$, i.e. if and only if the point (x, θ) lies inside the shaded region g_1 as shown in the Figure A8.2.

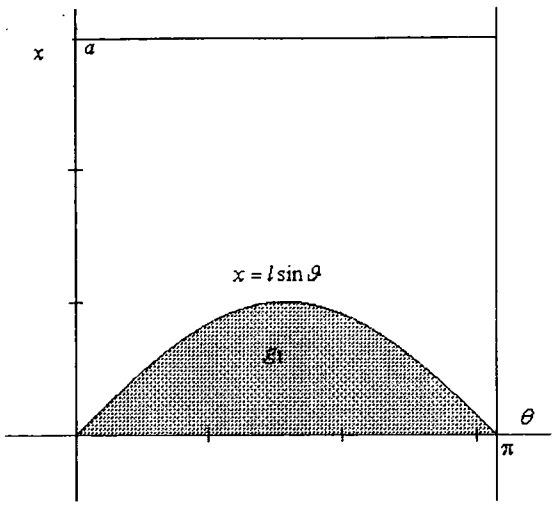


Figure A8.2

Suppose P is equally likely to fall anywhere between the parallels, then the probability that the needle crosses a line is

$$\frac{\text{Area of } g_1}{\text{Area of the rectangle bounded by } 0 \le x \le a \text{ and } 0 \le \theta \le \pi}$$

$$= \frac{\int_0^{\pi} l \sin\theta \, d\theta}{a\pi} = \frac{2l}{\pi a}.$$

For the second solution, let us consider the distance of the center of the needle from its nearest parallel line and the length of its projection on the nearest parallel line. As shown in the Figure A8.3, let P be the centre of the needle, x be the distance of P from its nearest parallel line, and $2y$ be the length of the projection.

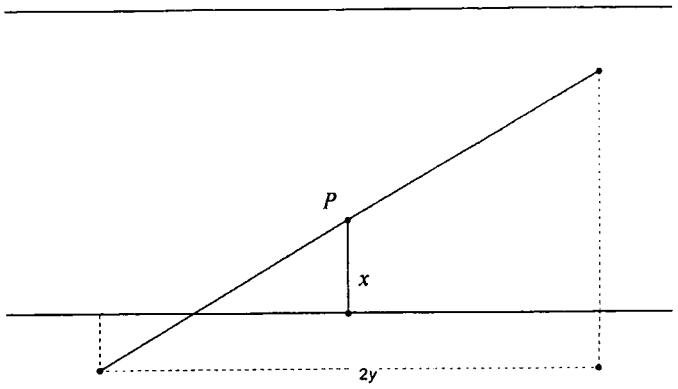


Figure A8.3

Then we have $0 \le x \le a$ and $0 \le y \le l$. The line will cross one of the parallels if and only if $l^2 - y^2 \ge x^2$, i.e. if and only if the point (x, y) lies inside the shaded region g_2 as shown in Figure A8.4:

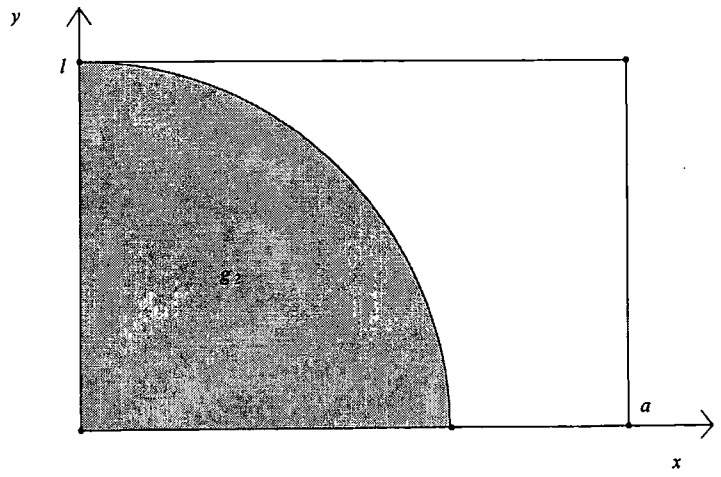


Figure A8.4

The probability that the needle crosses a line is thus:

$$\frac{\text{The area of the shaded region } g_2}{\text{The area of the rectangle bounded by } 0 \le x \le a \text{ and } 0 \le y \le l}$$

$$= \frac{\frac{1}{4}\pi l^2}{la} = \frac{\pi l}{4a}$$

For our third way of calculation, construct a perpendicular line to the parallel lines and along this perpendicular select one direction as positive. Using this directed perpendicular line and the closest parallel line as the y-axis and the x-axis respectively, we let y_1 and y_2 be the y-coordinates of the two end-points of the needle as shown in Figure A8.5.

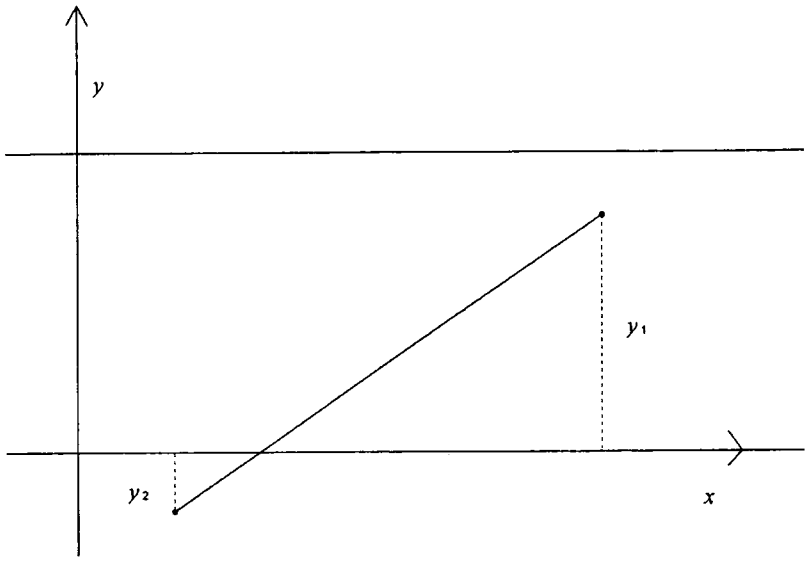


Figure A8.5

Then we have: $|y_1 - y_2| \le 2l$ and $|y_1 + y_2| \le 2a$. The line will cross one of the parallels if and only if $y_1 y_2 \le 0$, i.e. if and only if the point (y_1 , y_2) lies inside the shaded region g_3 as shown in Figure A8.6:

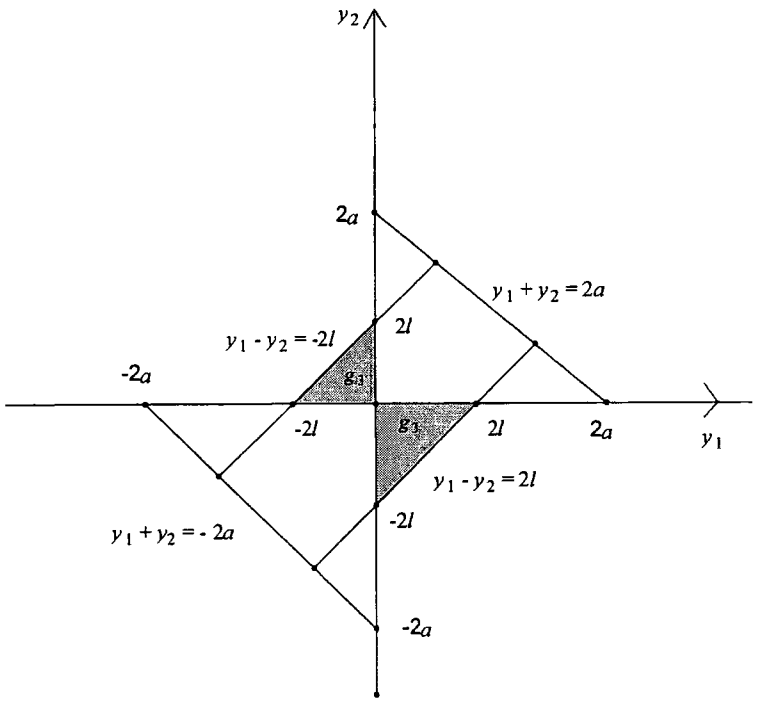


Figure A8.6

The probability that the needle crosses a line is thus:

$$\frac{\text{The area of the shaded region } g_3}{\text{The area of the rectangle bounded by } |y_1 - y_2| = 2l \text{ and } |y_1 + y_2| = 2a}$$

$$= \frac{4l^2}{2\sqrt{2}a \times 2\sqrt{2}l} = \frac{l}{2a}.$$

The reason why different approaches give different solutions lies in the fact that we have applied the Principle of Indifference to give a uniform probability density to three different random variables – i.e. the three order-pairs $(x\ ,\ \theta)$, $(x\ ,\ y)$ and $(y_1\ ,\ y_2)$ respectively. Assigning a uniform probability density to these three random variables are thus three incompatible assumptions. We are now able to conclude that the Principle of Indifference per se cannot be used to fix the probability, at least in the problems of involving continuous parameters or geometrical probability, as some classical theorists have alleged.

The third objection to the classical theory of probability is about its domain of application. In addition to the problems involving continuous parameters or geometrical probability in which the application of the Principle of Indifference could give rise to inconsistencies, the classical theory of probability breaks down in a number of cases. The first case is the sort of probability problems in mathematics that lead to irrational numbers as probabilities (for example, the probability that two natural numbers selected at random are relatively prime is $6/\pi^2$)[273]. Since an irrational number cannot be expressed as a ratio of two integers, the classical theory, according to which probability is defined as the ratio of the number of favourable outcomes to that of all possible outcomes, will thus fail to give an account of irrational probability. Another case is the situation in which the outcomes are empirically not equiprobable. For example, consider the loaded die which we have discussed before. Suppose we have tested the die empirically by throwing it for a large number of times and observed the number of occurrence of '1' (say, we get 200 '1's out of 600 throws), we then assert the following sentence:

(3.1) the probability of getting a '1' in throwing a loaded die is 2/3,

This sentence does, however, not make sense according to the classical theory. Because the six outcomes are no longer equiprobable even though they are on the same logical level (they all cannot be subdivided). Some may think that this is not a serious problem – we could restrict our use of: the term 'probability' in the cases where all possible outcomes are equiprobable. Despite the fact that this will leave the classical theory with very little to work on, it still gives rise to another problem – how could we know that the outcomes are equiprobable or not? We know that the classical theory is working well on

[273] See, for instance, Wells, 1986, p. 28.

a fair die. But what does it mean by 'fair'? This term cannot be defined with resort to a vicious circularity: a die is said to be fair if and only if the classical theory is working well on it. Otherwise, 'the classical theory is working well on a fair die' merely expresses an analytic truth 'the classical theory is working well on a die on which the classical theory is working well'. Since we have no a priori way to judge whether they are equiprobable or not, we cannot know in a particular case whether the term 'probability' can be legitimately used without performing empirical tests. But even a long series of experiments or empirical tests, no matter how long it is, is still unable to establish a conclusive proof that the die is fair. Tossing a fair die 6000 times, the number of occurrences of '1' could be 1000, or any number between 0 and 6000. Although we could apply Bernoulli's limit theorem to ascertain a high probability that the die is fair, the application itself would resort to the concept of 'probability' again as we have discussed in Section 2.2. Hence, according to the classical theory, we cannot practically talk about 'probability' in most cases. Any use of 'probability' that is not based upon equiprobable outcomes, including most of casual use such as the single event 'the probability that Brazil will win the World Cup 2006 is greater than 0.2', will be deviated from the classical theory. It thus certainly renders the classical theory unfit for explicating the use of probability in common usage.

Appendix 9 Logical theory and its difficulties

We will here use a simple example to illustrate Carnap's logical theory of probability. Consider the simple language L^2_3, which is a language that consists of 2 independent one-place predicates (M and N) and 3 individual constants (a, b, and c)[274]. By affirming or denying each property of each individual, a sentence[275] which completely describes a state of the world is called a state-description (Carnap, 1962, p.71). Each state-description could be regarded as characterizing a possible world. For examples, the following five state-descriptions will be characterizing five different possible worlds:

Z_{111111} : Ma & Mb & Mc & Na & Nb & Nc

Z_{011111} : $\sim Ma$ & Mb & Mc & Na & Nb & Nc

Z_{101000} : Ma & $\sim Mb$ & Mc & $\sim Na$ & $\sim Nb$ & $\sim Nc$

Z_{001111} : $\sim Ma$ & $\sim Mb$ & Mc & Na & Nb & Nc

Z_{100111} : Ma & $\sim Mb$ & $\sim Mc$ & Na & Nb & Nc

For our simple language L^2_3, there will be a total of 64 different possible worlds each of them will be characterized by one state-descriptions. If h is a sentence of L^2_3, its range R is the class of all state-descriptions in which h holds. For example, if h denotes 'Mc & Na & Nb & Nc', then its range is { Z_{111111}, Z_{011111} , Z_{101111} , Z_{001111}}. The degree of confirmation, q, of a hypothesis, h, on evidence, e, will be written as $c(h,e) = q$. Carnap represents the unconditional probability of h by $m(h)$, where m is the measure function associated with c. Therefore $c(h,e) = \frac{m(e \,\&\, h)}{m(e)}$. For the measure which assigns the

[274] Usual logical connectives are also included.

[275] Or a class of sentences if the language is L^{π}_{∞} (i.e. a language with π independent one-place predicates and infinite number of independent individuals).

state-descriptions the equal weights, Carnap writes the measure function as $m^\dagger$ and the corresponding confirmation function based on $m^\dagger$ is $c^\dagger$. In our example, if h denotes '*Mc* & *Na* & *Nb* & *Nc*' and e denotes '*Mb* & *Mc* & *Na* & *Nb* & *Nc*' since the measure of h is the number of state-descriptions in which it is true divided by the total number of state-descriptions, we have:

$$m^\dagger(h) = \frac{4}{64} = \frac{1}{16},$$

$$m^\dagger(e) = \frac{2}{64} = \frac{1}{32},$$

and $$m^\dagger(e \,\&\, h) = \frac{2}{64} = \frac{1}{32}.$$

Hence, $$c^\dagger(h, e) = \frac{m^\dagger(e \,\&\, h)}{m^\dagger(e)} = \frac{\frac{1}{32}}{\frac{1}{32}} = 1.$$

The value of this confirmation function is 1, which is reasonable since e logically entails h. This $c^\dagger$ function is, however, not wholly suitable as a foundation for a system of quantitative inductive logic. For example, if h denotes '*Mc*' and e denotes '*Ma* & *Mb*', following similar calculation, we have:

$c^\dagger(h, e) = \frac{1}{2}$, which is the same as $c^\dagger(h, \Gamma)$ where Γ denotes a tautology.

It means that even if we know that two more individuals a and b will satisfy the predicate M, the probability that the third individual c will satisfy the predicate M will remain unchanged. This $c^\dagger$ function thus fails to account for learning from experience[276]. That's

[276] We are not saying that the function $c^\dagger$ fails in accounting for learning from experience in all cases. For example, if h denotes '*Mb* & *Mc*' and e denotes '*Ma* & *Mb*', then $c^\dagger(h, e) = 1/2$ which is greater than $c^\dagger(h, \Gamma) = 1/4$, that means, in this case, it can demonstrate how we learn from experience.

why Carnap has constructed another confirmation function c^* to replace this $c^\dagger$ function (Howson and Urbach, 1993, pp.65-66; Weatherford, 1982, pp. 87-90).

The confirmation function c^* makes all structure-descriptions rather than all state-descriptions equiprobable. A structure-description corresponding to a world Z_i in L^π_N, j, is the disjunction of all worlds Z which are isomorphic[277] to Z_i arranged in lexicographical order (Carnap, 1962, p.116). For example, consider our previous example – the language L^2_3, there will be 20 different structure-description:

$S_{00} = Z_{000000}$

$S_{01} = Z_{000001} \vee Z_{000010} \vee Z_{000100}$

$S_{02} = Z_{000011} \vee Z_{000101} \vee Z_{000110}$

$S_{03} = Z_{000111}$

$S_{10} = Z_{100000} \vee Z_{010000} \vee Z_{001000}$

$S_{11} = Z_{100100} \vee Z_{010010} \vee Z_{001001}$

$S_{11'} = Z_{100010} \vee Z_{100001} \vee Z_{010100} \vee Z_{010001} \vee Z_{001010} \vee Z_{001100}$

$S_{12} = Z_{100011} \vee Z_{010101} \vee Z_{001110}$

$S_{12'} = Z_{100110} \vee Z_{100101} \vee Z_{010110} \vee Z_{010010} \vee Z_{001101} \vee Z_{001011}$

$S_{13} = Z_{100111} \vee Z_{010111} \vee Z_{001111}$

$S_{20} = Z_{110000} \vee Z_{101000} \vee Z_{011000}$

$S_{21} = Z_{110001} \vee Z_{101010} \vee Z_{011100}$

$S_{21'} = Z_{110010} \vee Z_{110100} \vee Z_{101001} \vee Z_{101100} \vee Z_{011001} \vee Z_{011010}$

[277] Two worlds are said to be isomorphic if and only if one can be derived from the other by merely exchanging some individuals for others by means of a one-to-one mapping. For example, Z_{001111} and Z_{100111} in our previous example are isomorphic, but Z_{111111} and Z_{011111} are not.

$S_{22} = Z_{110110} \vee Z_{101101} \vee Z_{011011}$

$S_{22'} = Z_{110101} \vee Z_{110011} \vee Z_{101011} \vee Z_{101110} \vee Z_{011101} \vee Z_{011110}$

$S_{23} = Z_{110111} \vee Z_{101111} \vee Z_{011111}$

$S_{30} = Z_{111000}$

$S_{31} = Z_{111001} \vee Z_{111010} \vee Z_{111100}$

$S_{32} = Z_{111011} \vee Z_{111101} \vee Z_{111011}$

$S_{33} = Z_{111111}$

If we treat each of structure-descriptions as equiprobable, then the measure of each structure-descriptions = 1/20. In other words,

$$m^*(S_{00}) = m^*(S_{01}) = \cdots = \frac{1}{20}.$$

Inside each structure-description, the weight will be distributed equally between all state-descriptions. For example, $m^*(Z_{000000}) = 1/20$ but $m^*(Z_{000001}) = \frac{1}{3} \times \frac{1}{20} = \frac{1}{60}$ because there are 3 state-descriptions in S_{01}. Using this new confirmation function to evaluate $c^*(h, e)$ and $c^*(h, \Gamma)$ again, where h denotes '*Mc & Na & Nb & Nc*' and e denotes '*Mb & Mc & Na & Nb & Nc*', we have $c^*(h, e) = 1$, which is reasonable as e logically entails h and the result is the same as that calculated by $c^\dagger$. Consider another example, if h denotes '*Mc*' and e denotes '*Ma & Mb*', we have $h \wedge e$ is *Ma & Mb & Mc*, and

$$m^*(h \wedge e) = m^*(S_{30}) + m^*(S_{31}) + m^*(S_{32}) + m^*(S_{33}) = \frac{4}{20} = \frac{1}{5}$$

$$m^*(e) = m^*(S_{30}) + m^*(S_{31}) + m^*(S_{32}) + m^*(S_{33}) + \frac{1}{3}(m^*(S_{20}) + m^*(S_{21}) + m^*(S_{21'}) + m^*(S_{22}) + m^*(S_{22'}) + m^*(S_{23}))$$

$$= \frac{4}{20} + \frac{1}{3} \times \frac{6}{20} = \frac{3}{10}$$

Hence, $c^*(h, e) = \frac{m^*(h \wedge e)}{m^*(e)} = \frac{\frac{1}{5}}{\frac{3}{10}} = \frac{2}{3}$.

And $c^*(h, \Gamma) = \frac{1}{3}(m^*(S_{10}) + m^*(S_{11}) + m^*(S_{11'}) + m^*(S_{12}) + m^*(S_{12'}) + m^*(S_{13}))$

$$+ \frac{2}{3}(m^*(S_{20}) + m^*(S_{21}) + m^*(S_{21'}) + m^*(S_{22}) + m^*(S_{22'}) + m^*(S_{23}))$$

$$+ m^*(S_{30}) + m^*(S_{31}) + m^*(S_{32}) + m^*(S_{33})$$

$$= \frac{1}{3} \times \frac{6}{20} + \frac{2}{3} \times \frac{6}{20} + \frac{4}{20} = \frac{1}{2}$$

Now our evidence could increase the value of the confirmation function c^* from 1/2 to 2/3, which is successful in reflecting how learning takes place from experience.

From these lengthy calculations, it seems to show how the value of probability$_1$, as a degree of confirmation, can be determined in a simple language completely by a priori means. Unlike Keynes who still wants to retain the Principle of Indifference as the only legitimate source of numerical initial probabilities, Carnap has explicitly rejected the Principle of Indifference[278]. But a closer look into Carnap's calculations will reveal that the spirit of the Principle of Indifference has indeed revived in his theory though in another form. When Carnap assigns probabilities to each structure-descriptions, he has declared that each is equiprobable. Furthermore, within each structure-description, he has assumed that probability is distributed equally between all state-descriptions. As we have argued before, one major objection to the Principle of Indifference is that different

[278] In his words, 'the classical theory claims to give a definition for probability, based on the concept of equipossible cases. The only rule given for the application of the latter concept is the principle of indifference, since we know today that this principle leads to a contradiction, there is in fact no definition for the concept of equipossibility' (Carnap, 1962, p.343)

applications of the Principle of Indifference depending upon different specifications of the alternatives will result in different values for the probability. We have already shown that Carnap's $c^{\dagger}$ and c^{*} functions would result in different values for the probability$_1$ of the same hypothesis, based on the same evidence.[279] If they are regarded as merely two abstract mathematical models, then it sounds perfectly legitimate for Carnap to make any assignment of probabilities. But when we try to apply the models to practical problems in our world, we have to determine which model will fit our world. And in this case we have still resorted to empirical means to make our judgment or evaluation[280]. As a result, the probability measures based on Carnap's method of distributing probabilities do not possess a genuine logical status.

Another objection to the logical theory of probability is that it cannot explain why our estimate of probability judgments should be in any way affected by the amount of evidence on which they are based (Ayer, 1973, pp.188-198). Consider an event: David will get a pass in the mathematics examination (h), its probability, as a measure of logical relation between evidence and conclusion, must be relative to certain evidence. But there are a number of relevant evidences that can be taken into account. Some are relevant to h, for examples, e_1 that David has got very high mark in a recent mathematics test; e_2 that David's father is a professional mathematician. But how about the evidences like: e_3 that David's mother has been confirmed to be a patient of Severe Acute

[279]These two conditional probability functions are merely the two that are regarded by Carnap as the most simple and natural ones. There is indeed a continuum of other probability functions defined on simple languages, each corresponding to a real-valued non-negative parameter λ. See Howson and Urbach, 1993, pp.66-72 for a brief discussion on how Carnap and his followers (such as J. Hintikka) have developed these functions.

[280] Carnap himself has indeed suggested that the parameter λ specifying the corresponding conditional probability function could be evaluated by a calibration process which consists in comparing the class of predictions assigned x percent probability with the frequency with which those predictions were true (Howson and Urbach, 1993, pp.70-71).

Respiratory Syndrome (SARS) three days before David's mathematics examination? Whether it is relevant to David's performance or is able to change the probability of h may in turn depend on the other evidences. For example, if we know that e_4 that David has been living with his mother until she was admitted into hospital two days before the examination, and some other evidences regarding the spread of SARS,..., etc, then e_3 together with e_4 will constitute relevant evidences that we should take it into account. On the contrary, if what we know is e_5 that David has not any contacts with his mother for 2 months rather than e_4, then e_3 together with e_5 will probably not relevant at all. It seems that we should take all evidences into account especially when certain piece of evidence per se is not sufficient for us to determine whether it is relevant or not. Of course it is practically impossible for us to gather all evidences[281], what we could do might only be to gather the total evidence available to us. But why should we have to take as evidence the total evidence available to us? From the point of logical theory, unless we had made logical mistakes, any one of these probabilities cannot be regarded as more, or less, correct than another one. For example, suppose we correctly arrive at $P(h,e_1) = p_1$ and $P(h,e_1 \wedge e_2 \wedge e_3 \wedge e_4) = p_2$, both results will be necessarily true and there is no ground for us to assert that one is more 'accurate' or 'correct' than the other. It is clearly unable to explain why in making a probability judgment we should base it on as many evidences as available. Another consequence of logical theory is that probabilities would never be refuted by experience. For example, based on only a piece of evidence e about a die (e.g. merely by observing its appearance), we figure out that the probability of getting a '1' in rolling this die once is 1/6, say. But later we throw it for 600 times and find that we get

[281] There is of course another serious practical problem that the logical theorists have to face: how could we determine the probability of h on $e_1 = P(h,e_1)$, $P(h,e_2)$, $P(h,e_3)$,...., etc? But for the sake of argument here, let us suppose the task of calculating all these probabilities could be accomplished.

600 '1's (evidence e^*). This new additional piece of evidence will certainly refute our estimation that the probability of getting a '1' is 1/6. Logical theorists would, however, reply that there is indeed no refutation at all, what we get are indeed $P('1', e) = 1/6$ and $P('1', e \wedge e^*) = 1$, and both of them are true. It thus show that the logical theory is separated from experience or our empirical world.

Appendix 10 Objections to the frequency theory

A main objection to the frequency theory is that there are many situations where we use probability but in which nothing like an empirical collective can be defined. These narrow limits of its applicability have, however, been well acknowledged by von Mises, as we have mentioned before. Von Mises (1957) has made the point clearly: 'From the complex ideas which are colloquially covered by the word 'probability', we must remove all those that remain outside the theory we are endeavouring to formulate. I shall therefore begin with a preliminary delimitation of our concept of probability' (pp.8-9). To von Mises, the frequency interpretation of probability will certainly not applicable to single events, for examples, 'the probability that David will get a pass in the mathematics examination is 0.8', 'the probability that US will be involved in war with North Korea at some time in the future (say, in 2007) is greater than 0.1', or 'the probability that Gaius Julius Caesar has visited Britain is less than 0.5'[282]. Critics to the frequency theory allege, however, that these sentences are absolutely meaningful and any theory of probability that cannot give an interpretation to them is certainly inadequate.

We have two rejoinders to this objection. First, although von Mises has argued that there is no such thing as the probability of a single event, it does not imply that no frequency theorists think that frequency theory can also be applicable to single event. Indeed, Reichenbach (1949) has attempted to give a frequency interpretation to single events by means of a *posit* (pp.372-378). According to Reichenbach, a posit is a sentence

[282] Von Mises has, however, argued that problems such as the probable reliability of witnesses and the correctness of judicial verdicts lie more or less on the boundary of the region which he has included in his treatment (1957, p.9).

with which we deal as true, although its truth value is unknown. A sentence about the probability of a single event is regarded as an 'elliptic mode of speech' which acquires a fictitious meaning by a transfer of meaning from the general to the particular single event. Second, von Mises's attempt could indeed provide a way out of the so-called 'the reference class problem'. As a critic against frequency theory, Hájek (1997) takes this example to illustrate the reference class problem: his probability of dying by age 60 (pp.74-75). What he wants is an unconditional probability[283]. But, according to von Mises's (1957) frequency theory, the notion of probability must be relativized – it is only the notion of probability in a given collective which is unambiguous (p.20). There are, however, many reference classes that Hájek can be placed in. For instances, the class of all living things, the class of all humans, the class of all male philosophers, the class of all fans of Woody Allen... We agree with Hájek that each of these reference classes will have its own associated relative frequency for death by age of 60. But we don't agree with him that the event (his death by age 60) has more than one probability. In fact, either Hájek will die by age 60 or not die by age 60. We will know the answer one day, say one hundred years later. It is totally legitimate for us to regard our common usage of 'the probability of his death by age 60' is really elliptical for a relativized probability sentence 'the probability of his death by age 60 on the condition that he belongs to the intersection of classes $C_1 \cap C_2 \cap \cdots C_n$. Here the classes are all relevant classes that are available to us and they are certainly not exhaustive. Indeed, if we could consider all relevant classes that he can be placed in, then he will be the only element in their intersection. We could theoretically be certain that he would either die or not die by age

[283] Hájek (1997) has made a very good distinction between a conditional probability of the form $P(B \mid A)$ and a relativized probability of the form $P_A(B)$: the former but not the latter presupposes that $P(A)$ is well-defined (p.86).

60 and the probability would then be either 1 or 0. Hence, when we talk about the probability of a single event, it is only an elliptical way to talk about the probability of that event relativized to certain known reference classes[284]. If for a certain single event we cannot obtain any reference class in which the event can be placed, then talking about its probability could be absolutely nonsense.

Another cluster of related objections to the frequency theory is about the notion of limiting relative frequency. These objections, that are interrelated, can be divided into three major categories[285] for the sake of discussion:

1. Epistemological problem: limiting relative frequencies cannot be known, at least for certainty;
2. Ontological problem: limiting relative frequencies do not exist in our natural world;
3. Methodological problem: sentences or theories with limiting relative frequencies is not refutable nor confirmable.

For the first problem, it has been argued that the limit relative frequency can hardly be known even for a simple problem like tossing a coin. According to the frequency theory, the probability of getting head is defined as the limiting relative frequency of heads in an infinite sequence of tosses. But no matter how many times the coin has been tossed, the relative frequency of heads is still not the limiting relative

[284] For the cases in which the reference classes or empirical collective can be identified, initial probabilities can be directly ascertained through induction by enumeration, at least in principle. This is contrary to the classical theorists' method of counting equiprobable cases according to the Principle of Indifference. That also explains why von Mises (1957) has argued that we can hardly say anything about probabilities when we are ignorant (pp.75-80). For a more detailed analysis of a posteriori establishment of a probability metric, see Reichenbach, 1949, pp.359-366.

[285] See Weatherford, 1982, pp.199-200 for a similar division of objections.

frequency of heads in an infinite sequence. The probability is thus never known to us with 100% of certainty and accuracy. It is true that we are indeed unable to toss the coin for an infinite number of times and thus unable to get an absolutely exact value for the probability of getting head. But some will argue that many of the physical quantities that could only be measured empirically are being unable to be known in this sense. Consider the mass of our earth, what we are able to measure is only an approximation of it. No matter how technologies have been advanced, we still cannot get its exact value. There is however one major difference between two cases. Though we cannot measure many of the physical quantities, such as the mass of the electron, with 100% of certainty and accuracy, its value measured with today's technology will be better than the one we measure 20 years ago and it is likely accurate to within our tolerable limit. A reasonable and finite number of measurements can establish the value of a normal physical quantities so convincingly that we have to explain where a previous error has been made if we want to make significant revision. But it is clearly not the case for limiting relative frequency. No matter how many times a die has been tossed and how stable the relative frequency of the occurrence of '1' is, it is always possible that its value will change dramatically without necessarily requiring an acknowledgment of any previous error.

When some assert that all students aged 18 or above could solve quadratic equations with 1 variable, we could either refute it by getting one student who is aged 18 or above but could not solve the equations, or confirm it by getting students who fulfill the requirements. But how about the sentences or theories with limiting relative frequencies? Consider the sentence 'the probability of getting '1' by throwing this die is 1/3'. If we throw this die for 1000 times and it is found that no '1' occur, we still could

not refute the sentence. Since the sentence only asserts that the limit relative frequency of '1' is 1/3, it is always compatible that the limit is 1/3 but no '1' occurs in the 1000 throws. Hence, it is quite clear that these sentences can hardly be refuted conclusively.

As a result, critics to frequency theory will query the existence of limiting relative frequency – it cannot share many properties possessed by other physical quantities simply because it is fictitious. Moreover, consider the case of tossing a coin, the probability of getting head is defined as the limiting relative frequency of heads in an infinite sequence of tosses. Suppose we toss the coin for 1000 times and get the relative frequency of heads, this sequence is merely an estimation for the limiting relative frequency in an infinite sequence. The infinite sequence itself has no physical existence. Unlike the moon that has a mass no matter we measure it or not, the coin would not have any sequence of heads if it had not been tossed. Could we say that a coin that is destroyed immediately after it has been made (so that it has not been tossed even for once) has no probability of getting head? Moreover, no matter how many times the coin had been tossed, the coin can never been tossed for infinite number of time. There is no such an infinite sequence of physical events (even though the coin's physical characteristics would not change a little when tossed for a large number of times), the characteristics of coin cannot explain why heads have probability 1/2. It is these difficulties that render the proposal of another objective theory of probability – the propensity theory which will be discussed in Section 3.5.

Appendix 11 Objections to the subjective theory

There are certainly cases in which probabilities look like objective. For instances, the probability of getting a '1' in throwing a particular die, the probability that a randomly selected Chinese is suffering from G6PD deficiency, and the probability that a particular radioactive element will disintegrate in 3 years all seem to be objective in nature. In other words, they are not a matter of opinion and are independent of how we believe them. Some subjectivists like De Finetti would reply that there are in fact no objective probabilities[286] and all probabilities including those apparently objective probabilities just described could only be interpreted as degrees of beliefs (De Finetti, 1937). This reply leaves two problems that they should address. First, if there were in fact no objective probabilities, they have to explain why most of us have an illusion that such probabilities exist. Second, frequency theory succeeds in many cases especially in figuring out the probabilities of the statistical events and in the games of chance. Could the subjective theory provide the same successful results in these cases? We will examine the subjectivists' answers to these questions by taking a simple example. Consider a particular die which is not known to be biased or not. Our intuition tells us there is a true though unknown probability of getting a '1' in a particular throw of this particular die (the event E). Its exact value may never be known to us but it could still be estimated by observing the number of occurrences of '1' (m) in throwing the die for a large number of times (n). Subjectivists like De Finetti, as we have already said, will claim that there is no such an absolute probability of E. Each person is free to have

[286] Not all subjectivists would agree with De Finetti on this point. For example, Ramsey would accept objective probability and advocate a kind of calibration between degrees of belief and frequencies (Williamson and Corfield, 2001, p.1).

whatever degrees of belief in E. We are certainly able to throw the die again and again in addition to the particular throw of the die (E), but each throw is an individual event and should not be regarded as a part of a set of repetitive events. In other words, a person may have a betting rate of 1/6 on E but he is still perfectly all right to have a betting rate of 1/2 on the next throw of the same die.

Now, suppose the die has been thrown for n times (where n is a big number, such as 5000) before E and we observe that the number of occurrences of '1' is m, experience tells us that any person who makes a betting rate on E which is much different from m/n will almost certainly lose in the long run. Let's see how the subjectivists could explain this phenomenon. Suppose we use a_i to denote that we get '1' in the ith throw, b_i to denote that we do not get '1' in the ith throw, and e_n the complete specification of the results of the first n throws. According to these definitions, $E = a_{n+1}$ and e_n is a particular n-tuple in which there are m a's. What the subjectivists are going to figure out is $P(a_{n+1} \mid e_n)$. If a person has coherent beliefs, then the betting rates will conform to the axioms of probability and we could thus obtain:

$$P(a_{n+1} \mid e_n) = \frac{p(a_{n+1} \wedge e_n)}{P(e_n)}.$$

Rather than using the notion of independent events, the subjectivists introduce the condition of exchangeability[287] which is given by:

> For any n, when there is more than one possible order in which a certain number of a appear in the n-tuple e_n we have to assign the same betting rate to any one of the particular n-tuple e_n.

[287] See De Finetti (1975, 211-224) or Jeffrey (2004, 78-81) for details of a formal treatment.

For example, when $n = 3$ there are three possible triples in which exactly one a appears:

$$a_1b_2b_3\ ,\ b_1a_2b_3\ ,\ b_1b_2a_3\ .$$

If we assign a betting rate r_1^3 to the event that there are exactly one '1' appears in the first 3 throws, then by the condition of exchangeability, we have to assign the same betting rates (r_1^3 / 3) to each of these three events (or triples).

Since e_n is a particular outcome of the first n throws in which there are m '1's, there will be C_m^n such possible outcomes. Hence, we have

$$P(e_n) = \frac{r_m^n}{C_m^n},$$

where r_m^n is a betting rate that a person assigns to the event that there are exactly m '1's appearing in the first n throws.

Similarly, $\quad P(a_{n+1} \wedge e_n) = \dfrac{r_{m+1}^{n+1}}{C_{m+1}^{n+1}}.$

Hence, we have

$$\begin{aligned} P(a_{n+1} \mid e_n) &= \frac{p(a_{n+1} \wedge e_n)}{P(e_n)} \\ &= \frac{r_{m+1}^{n+1} \times C_m^n}{C_{m+1}^{n+1} \times r_m^n} \\ &= \frac{(m+1)r_{m+1}^{n+1}}{(n+1)r_m^n} \end{aligned}$$

If we assume further that $\frac{r_{m+1}^{n+1}}{r_m^n}$ tends to 1 when n approaches infinity, then $P(a_{n+1} \mid e)$ will tend to $(m+1)/(n+1)$ which is approximately equal to m/n (the observed relative frequency) for large n. This explains why the observed relative frequency provides a good measure of $P(E)$. In other words, even though different persons may have different degrees of belief in the propositions that we will get m '1's in the first n throws of the die (i.e. assigning different values to r_m^n) initially, they will eventually change their degrees of belief in response to more observations and finally come to agreement as evidence mounts up to a certain extent. It is this very result that the subjectivists use it to explain why people have an illusion that there is an objective probability. Apparently the subjective theory is able to provide the same successful results in statistical events or games of chance as the objective interpretation of probability.

Let us now consider if this subjective interpretation is tenable or not. In the above derivation, the subjectivists have made two assumptions in addition to the axioms of probability (or the assumption of coherent betting rates). One is the assumption that the limit of the ratio r_{m+1}^{n+1}/r_m^n is equal to 1 and another is the condition of exchangeability. Although the limit assumption seems to be reasonable, it is sill legitimate for us to ask why we have to assume that it is true. As the subjective theory asserts, a person has his own right to assign any numbers to the betting rates r_m^n and r_{m+1}^{n+1}, provided the set of betting rates is coherent. We see no reason why for a subjectivist it is illegitimate to assign the better rates in a such a way that the ratio r_{m+1}^{n+1}/r_m^n does not converge to 1. In other words, this assumption is not justified within framework of the subjective theory.

Exchangeability is, in a certain sense, a substitute for the notion of independence. As Gillies (2000) notes, 'when an objectivist assumes independence and formulates corresponding mathematical equations, a subjectivist can simply reinterpret these equations as being about subjective probabilities and exchangeability' (p.77). But independence is a very crucial notion when we apply probabilities in enormous cases, we will show here that such an elimination of the notion of independence in favour of exchangeability is imprudent. Unlike independence which could be explicitly defined in mathematical terms, exchangeability is a loose concept and its use is a little bit arbitrary. For instance, in the above mentioned example, the die has been thrown for n times before E and we have used r_m^n to denote the betting rate that a subjectivist assigns to the event that there are exactly m '1's appearing in the first n throws. Since there are no compulsory rule in addition to the requirement of coherence, subjectivists could make any choice on the betting rates. Hence, though this is not mandatory, subjectivists are liable to apply exchangeability in this way: the number of '1's in the n throws could only be 0, 1, 2, ..., or n, we assign the same betting rates to each of these states, i.e., $r_0^n = r_1^n = ...r_n^n$. Since they sum up to 1, we have

$$r_0^n = r_1^n = ...r_n^n = \frac{1}{n+1}.$$

From previous result,

$$P(a_{n+1} \mid e_n) = \frac{(m+1)r_{m+1}^{n+1}}{(n+1)r_m^n}$$

Hence, $$P(a_{n+1} \mid e_n) = \frac{(m+1)\cdot\frac{1}{n+2}}{(n+1)\cdot\frac{1}{n+1}}$$

$$= \frac{m+1}{n+2}.$$

Now, without making an additional assumption that the limit of the ratio r_{m+1}^{n+1} / r_m^n equals 1, we could get this very famous result – Laplace's law of succession. This law is so famous because before the dawn of modern theory of probability Laplace has tried to use this law to justify induction. Given that the sun has risen daily for 5000 years or 1826213 days, to what extent we could be sure that the sun will rise tomorrow? Laplace uses this law to figure out the probability that the sun will rise tomorrow[288]:

$$\because m = n = 1826213$$

$$\therefore \text{the probability} = \frac{1826213+1}{1826213+2} = 0.999\ 999\ 452\ 419\ldots.$$

It seems to be able to justify why many people believe that the sun will rise tomorrow given that it has risen daily for so many years. Nevertheless, this law is plainly false. According to what we know about from astrophysics, our sun will not last forever but according to this law the probability for the sun to rise will only increase as time. We may also use a daily-life example to illustrate how this law would give absurd result. Suppose we have an old radio which has been programmed in such a way that it will automatically turn on at 6 am every morning. For the past 10 years or 3652 days, it works perfectly well every morning. But in this morning the radio doesn't turn on automatically. We try to turn it on manually but fail. Normally we will expect that this old radio will not work again unless we get someone to fix it. However, according to

[288] Feller, 1968, pp. 124 – 125.

Laplace's law of succession, the probability that it will work tomorrow only drops to a very small extent: from 3653/3654 to 3653/3655! The law of succession fails in this case because it doesn't take the dependence into proper account. Whether the radio will function properly tomorrow is not independent of how it functions today. The assumption of exchangeability only looked plausible in the first place because the case we consider – throwing of a die – is a sequence of independent events. It will lead to mistaken conclusion if it is applied to the sun rising or radio case in which the events are not independent. As a result, we cannot assume exchangeability in a priori way. There is, however, no posteriori guideline for us to determine when we should assume exchangeability or its consequence (law of succession), especially when the notion of independence is eliminated in the framework of subjective theory.

There is one more objection to the subjective theory of probability. One of the claimed merits of the subjective theory is that it allows us to apply Bayes' theorem to consider the probability of a hypothesis. Suppose in a research problem there are some competing hypotheses to be considered. According to the objectivists, a hypothesis H is either true or false. If it is true, then $P(H) = 1$; if it is false, $P(H) = 0$. Talking about any probabilities other than 0 and 1 seem to be absurd. The subjectivists assume that if the set of hypotheses is expressed as $\{H_\theta \mid \theta \in I\}$ there is an initial subjective distribution (also called prior distribution) $\mu(\theta)$ for the parameter θ. In response to new evidence e, the prior distribution $\mu(\theta)$ is revised by conditioning on e with the use of Bayes' theorem, which yields the posterior distribution $\mu(\theta \mid e)$. The whole revision process, which relies on the use of Bayes' theorem, is called 'Bayesian conditionalization' (Howson, 1995, p.8;

Gillies, 2000, p.36). The posterior distribution $\mu(\theta \mid e)$ is also over the set of hypotheses, the subjectivists then set $P(H_\theta) = \mu(\theta)$ and $P(H_\theta \mid e) = \mu(\theta \mid e)$. As a result, to the subjectivists, sentences like ' $P(H \mid e) = 0.98$' are perfectly meaningful. Bayesians believe that the process of Bayesian conditionalization will provide rational persons (such as scientists or researchers) a learning strategy so that whatever prior probabilities they adopt their posterior probabilities will converge towards the same value. This is the very point that we are going to argue against.

To illustrate the argument, we would like to use an example called the game of red or blue probably first described by Feller (1968, p.78 - 84).[289] A fair coin is first tossed to determine the initial mark of a player A. If it is head, his initial mark is 0; if it is tail, the initial mark is – 1. Then the coin will be tossed for a number of times. For each toss of the coin, if it is head, A gets one more mark; if it is tail, 1 mark will be deducted from A's current total mark. After each toss, the total mark A gets is either a non-negative point in which we will call the event 'Blue' or a negative number in which we call 'Red'. The game consists in making bets on the occurrence of one of these events 'Blue' or 'Red'. Since the coin is fair and the initial mark is either 0 or – 1, whose probability of occurrence is supposed to be equal, many of us would expect that in the long run the relative frequency of the occurrences of 'Blue' will be almost the same as that of 'Red'. The remarkable thing is that the result does not like what we expect! Here we perform a simulation with the use of Maple 8 of the game: the coin is tossed for 3000 times and we count the occurrences of 'Blue' and 'Red'. This is one trial and we will

[289] This example has also been used by Popper (1983) and Gillies (2000). It is Popper who first calls it the game red or blue (1983, p.303).

figure out the relative frequency of 'Blue'. The trial will be repeated for 1000 times. That means we would have 1000 data about the relative frequency of 'Blue' in each trial.

```
> restart: with(plots, display):
libname:="C:/mylib/stat",libname: with(stat):
> N:=1000:T:=3000:
> for j from 1 to N do:
> S[0]:=0:
> for i from 1 to T do:
> X:=Die(2,1):
> if X=[1] then RA:=-1 else RA:=1 end if:
> S[i]:=S[i-1]+RA:
> if S[i]>=0 then B[i]:=1: R[i]:=0 else B[i]:=0: R[i]:=1 end
if;
> od:
>
BB:=sum(B[m],m=1..T):RR:=sum(R[m],m=1..T):Rate_B[j]:=evalf(B
B/T,5);Rate_R[j]:=evalf(RR/T,5):
> od:
> for k from 1 to N do:
> Rate_B[k];Rate_R[k];
> od:
> L:=[seq(Rate_B[w],w=1..N)]:
> Histogram(L,0..1,5);
```

Its histogram is plotted as shown in Figure A11.1:

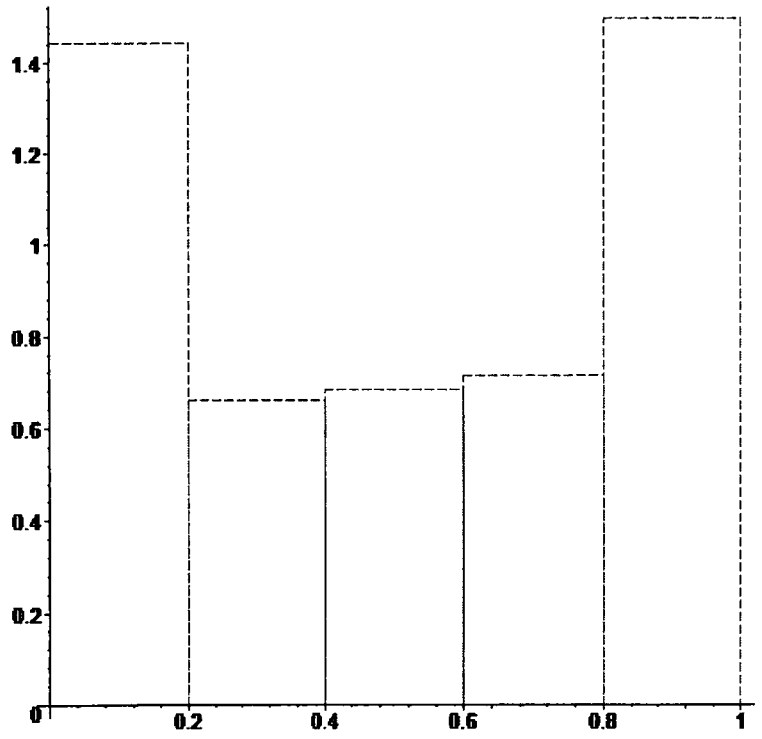


Figure A11.1 (N:=1000:T:=3000)

It is clearly from the graph that there is high probability of one of the colours appearing much more than the other. One may wonder if it is only an accidental result. Let's increase the number of tosses to 8000 in each trial and repeat the trials for 5000 times. We get the following similar result again:

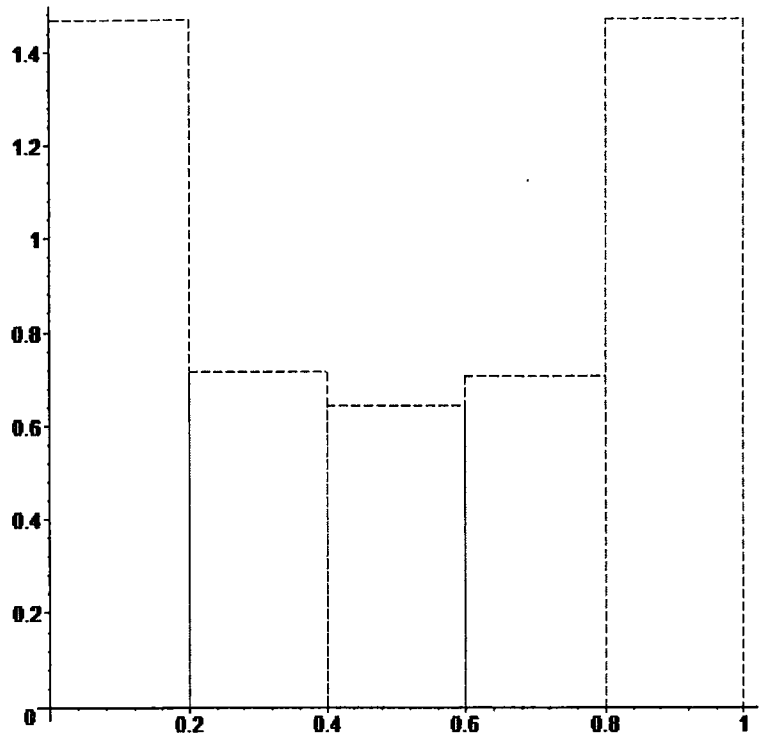


Figure A11.2 (N:=5000:T:=8000)

If a subjectivist S is asked to analysis a sequence of numbers (either 1 or 0) which are, unknown to S, generated by the game of red or blue (say, if the event 'blue' occurs, otherwise, the number is 0), it is quite natural for S to give a prior uniform distribution to r_m^n (the event that there are exactly m '1's appearing in the first n numbers of the sequence) but no matter how A chooses his betting rates the striking results would still appear. For example, if A finally gets the Laplace's law of succession by assuming that:

$$r_0^n = r_1^n = ...r_n^n = \frac{1}{n+1}.$$

In this case, if it is given that the first 100 numbers are 1 and the 101st and 102nd numbers are both 0, A will find that the probability of getting 1 for the 103rd number is quite high: 101/104 = 0.971 (3 sig. fig.). However, from the rule of game of red and blue,

we know that the probability should be 0. In other words, it is indeed impossible to get 1 for the 103rd number.[290] Moreover, from Figure A11.2, if the number of tosses is 8000, we find that about 73.2% of the 5000 trials will give the result that the relative frequency of one colour is more than 0.7. Hence we would expect that S is natural to estimate the probability of getting one of the two numbers in the number sequence is 0.7 or more. But in the real underlying game, the two colours are indeed symmetrical. In this example, S's calculations using exchangeability or Bayesian conditionalization will produce a sequences of probabilities at complete variance with reality. Since the events are indeed dependent whose possibility has been already ruled out in the use of exchangeability, no Bayesian conditionalization will bring S close to grasping what the real situation is. It thus shows that exchangeability or Bayesian conditionalization is hardly an effective learning strategy. One may argue that S fails because the events are independent and he should have considered not just exchangeability but also various form of Markov exchangeability and a broad and comprehensive class of hypotheses. But the point is: it is quite impossible to consider all the forms of exchangeability and all the possibilities which might arise at the very beginning of the study.

Besides, Max Albert (1999, 2001) has argued forcefully that no matter how a Bayesian chooses his learning strategy (i.e. how he figure out the probability of the nth number given the first $n-1$ numbers in the sequence) there must exist a prior probability distribution μ over the set of what he called the modified chaotic hypothesis such that S's

[290] The reason is simple. When the 100th number is 1 and the 101st number is 0, the 100th event and the 101st event must be 'Blue' and 'Red'. In this case, the 100th total mark must be 0, otherwise it is impossible for it to change to negative in the next event. In this connection, the total mark for the 101st and 102nd event must -1 and -2 respectively. Therefore, it is impossible for the total mark to become 0 or positive in the next event.

probabilities could be produced by Bayesian conditioning on μ. In other words, if S is going to consider a broader set of hypotheses, the chaos theory should be on the list and thus anything he does will, according to Albert's proof, will become Bayesian, rendering the whole approach empty.

Appendix 12 Maple 8 Worksheets for Chapter 5

Maple Worksheet for Figure 5.1

```
> restart:with(stats):with(plots):
Warning, the name changecoords has been redefined

> alpha:=0.05:power:=0.9:
> normalf:=(x,miu,sig)->exp(-(x-
miu)^2/(2*sig^2))/sqrt(2*Pi*sig^2):
> Z:=fsolve(statevalf[cdf,normald](x)=1-alpha/2,x);
```

$$Z := 1.959963985$$

```
>
difference:=solve(statevalf[cdf,normald](x)=power,x)+fsolve(
statevalf[cdf,normald](x)=1-alpha/2,x);
```

$$difference := 3.241515551$$

```
> zz:=(aa)->solve(statevalf[cdf,normald](x)=1-aa/2,x);
```

$$zz := aa \rightarrow \mathrm{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{1}{2}\,aa, x \right)$$

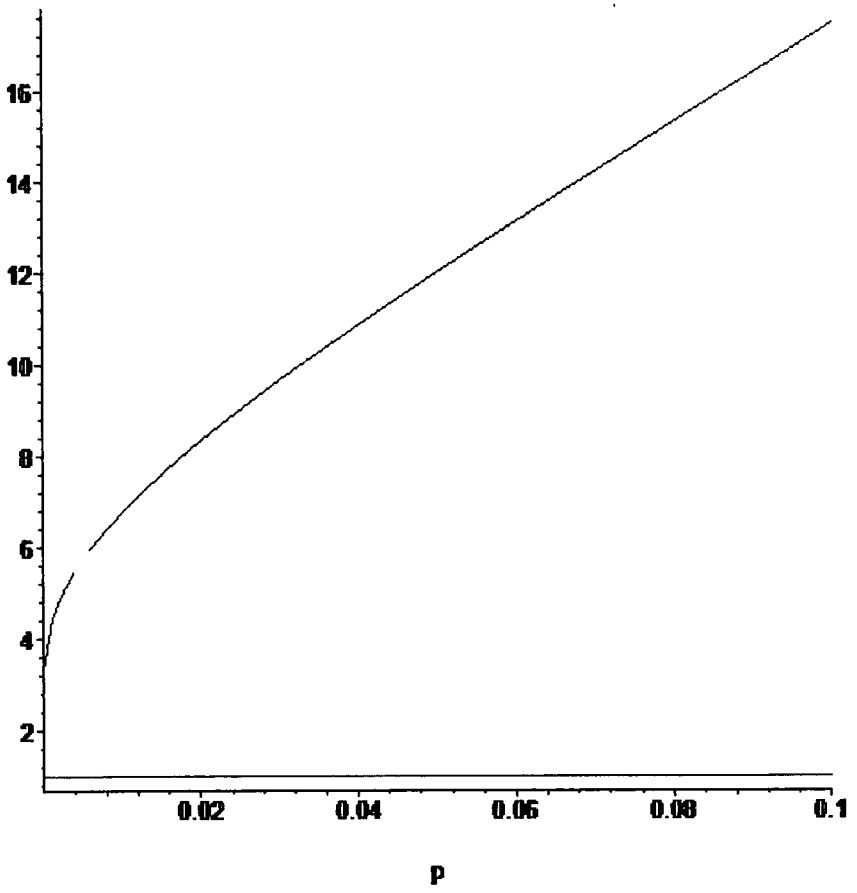
```
> ratio:=(p)->(normalf(zz(p),0,1)/normalf(difference -
zz(p),0,1)) /
((1-
statevalf[cdf,normald](zz(p)))/statevalf[cdf,normald](differ
ence - zz(p)));
```

$$ratio := p \rightarrow \frac{\mathrm{normalf}(\mathrm{zz}(p), 0, 1)\, statevalf_{cdf,\,normald}(difference - \mathrm{zz}(p))}{\mathrm{normalf}(difference - \mathrm{zz}(p), 0, 1)\,(1 - statevalf_{cdf,\,normald}(\mathrm{zz}(p)))}$$

```
>
a:=0.0001:b:=0.1:ratio(a);ratio(0.02);plots[display]({plot(r
atio(p), p=a..b, colour=blue)},{plot(1,a..b,colour=red)});
>
```

3.292177782

8.326763046



Maple Worksheet for Figure 5.2

```
> restart:with(stats):with(plots):
Warning, the name changecoords has been redefined

> alpha:=alpha:power:=0.8:
> normalf:=(x,miu,sig)->exp(-(x-
miu)^2/(2*sig^2))/sqrt(2*Pi*sig^2):
> Z:=(alpha)->solve(statevalf[cdf,normald](x)=1-alpha/2,x);
```

$$Z := \alpha \rightarrow \mathrm{solve}\left(statevalf_{cdf,\ normald}(x) = 1 - \frac{1}{2}\alpha, x \right)$$

```
> difference:=(alpha)-
>solve(statevalf[cdf,normald](x)=power,x)+solve(statevalf[cd
f,normald](x)=1-alpha/2,x);
```

$$difference := \alpha \rightarrow \text{solve}(statevalf_{cdf,\,normald}(x) = power, x) + \text{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{1}{2}\alpha, x\right)$$

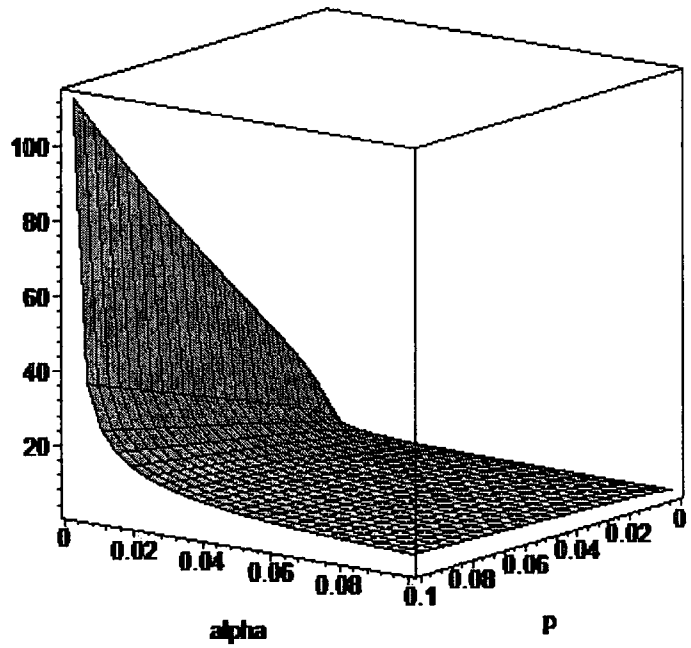
```
> zz:=(aa)->solve(statevalf[cdf,normald](x)=1-aa/2,x);
```

$$zz := aa \rightarrow \text{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{1}{2}aa, x\right)$$

```
> ratio:=(p,alpha)-
>(normalf(zz(p),0,1)/normalf(difference(alpha) - zz(p),0,1))
/                                                    ((1-
statevalf[cdf,normald](zz(p)))/statevalf[cdf,normald](differ
ence(alpha) - zz(p)));
```

$$ratio := (p, \alpha) \rightarrow \frac{\text{normalf}(\text{zz}(p), 0, 1)\, statevalf_{cdf,\,normald}(\text{difference}(\alpha) - \text{zz}(p))}{\text{normalf}(\text{difference}(\alpha) - \text{zz}(p), 0, 1)\,(1 - statevalf_{cdf,\,normald}(\text{zz}(p)))}$$

```
>
plot3d({ratio(p,alpha)},p=0.001..0.1,alpha=0.001..0.1,axes=b
ox);
```



Maple Worksheet for Figure 5.3

```
> restart:with(stats):with(plots):
```

```
Warning, the name changecoords has been redefined

> alpha:=0.05:power:=power:
> normalf:=(x,miu,sig)->exp(-(x-
miu)^2/(2*sig^2))/sqrt(2*Pi*sig^2):
> Z:=solve(statevalf[cdf,normald](x)=1-alpha/2,x);
```

$$Z := 1.959963985$$

```
> difference:=(power)-
>solve(statevalf[cdf,normald](x)=power,x)+solve(statevalf[cd
f,normald](x)=1-alpha/2,x);
```

$$difference := power \rightarrow \text{solve}(statevalf_{cdf,\,normald}(x) = power, x) + \text{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{\alpha}{2}, x\right)$$

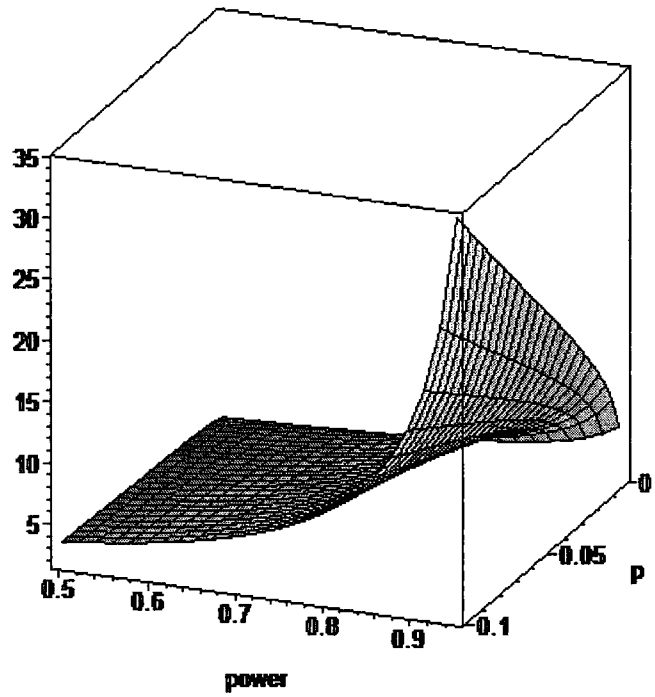
```
> zz:=(aa)->solve(statevalf[cdf,normald](x)=1-aa/2,x);
```

$$zz := aa \rightarrow \text{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{1}{2}aa, x\right)$$

```
> ratio:=(p,power)-
>(normalf(zz(p),0,1)/normalf(difference(power) - zz(p),0,1))
/                                                    ((1-
statevalf[cdf,normald](zz(p)))/statevalf[cdf,normald](differ
ence(power) - zz(p)));
```

$$ratio := (p, power) \rightarrow \frac{\text{normalf}(\text{zz}(p), 0, 1)\, statevalf_{cdf,\,normald}(\text{difference}(power) - \text{zz}(p))}{\text{normalf}(\text{difference}(power) - \text{zz}(p), 0, 1)\,(1 - statevalf_{cdf,\,normald}(\text{zz}(p)))}$$

```
>
plot3d({ratio(p,power)},p=0.001..0.1,power=0.5..0.95,axes=bo
x);
```



Maple Worksheet for Figure 5.4

```
> restart:with(stats):with(plots):
Warning, the name changecoords has been redefined
```

```
> z:=p->solve(statevalf[cdf,normald](x)=1-p/2,x);
```

$$z := p \to \text{solve}\left(statevalf_{cdf,\,normald}(x) = 1 - \frac{1}{2}p, x\right)$$

```
> normalf:=(x,miu,sig)->exp(-(x-
miu)^2/(2*sig^2))/sqrt(2*Pi*sig^2):
> LD:=a->normalf(z(p),0,1)/normalf(0,0,1);u:=p->LD(p)/p:
```

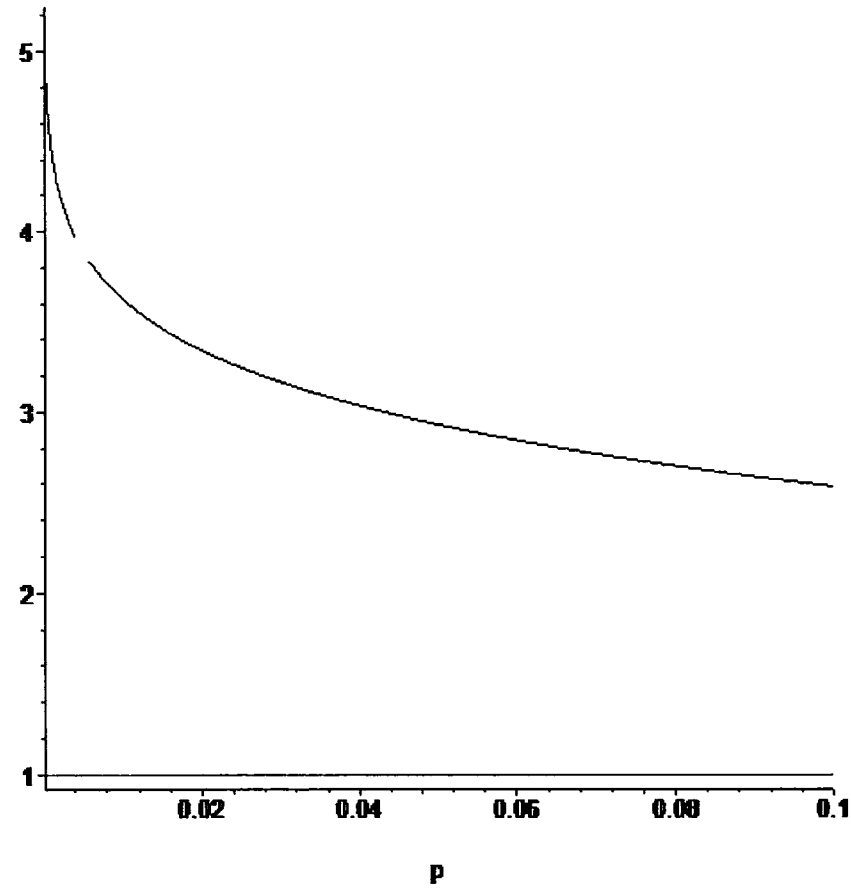
$$LD := a \to \frac{\text{normalf}(\text{z}(p), 0, 1)}{\text{normalf}(0, 0, 1)}$$

```
> LD(0.001);u(0.05);
```

$$e^{\left(-1/2\ \text{RootOf}(2\ statevalf_{cdf,\,normald}(\_Z) - 2 + p)^2\right)}$$

$$20.00000000\ e^{\left(-1/2\ \text{RootOf}(2\ statevalf_{cdf,\,normald}(\_Z) - 2 + p)^2\right)}$$

```
> a:=0.0001:b:=0.1:plots[display]({plot(u(p), p=a..b,
colour=black)},{plot(1,a..b,colour=red)});
```


5
4
3
2
1
0.02
0.04
0.06
0.08
0.1
p

Bibliography

Abelson,, R.P. (1997a). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8* (1), 12-15.

Abelson, R.P. (1997b). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) (1997), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Lawrence Erlbaum.

Adam, W.J. (1974). *The life and times of the central limit theorem.* New York: Kaedmon.

Adams, E.W. (1998). *A primer of probability logic.* Stanford, California: CSLI Publications.

Albert, M. (1999). Bayesian learning when chaos looms large. *Economics Letters*, 65, 1-7.

Albert, M. (2001). Bayesian learning and expectations formation: Anything goes. In J. Williamson & D. Corfield (2001), *Foundations of Bayesianism* (pp.341 – 362), Dordrecht: Kluwer. Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society, 23*, 186-190.

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician, 27*, 17-21.

APA (1994). *Publication manual of the American Psychological Association (4th Ed.).* Washington, DC: American Psychological Association.

APA (2001). *Publication manual of the American Psychological Association (5th Ed.).* Washington, DC: American Psychological Association.

Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions (1683-1775), 27*, 186-190.

Aron, A., Aron, E.N., & Coups, E.J. (2005). *Statistics for the behavioral and social sciences: A brief course (3rd ed.)*. Upper Saddle River, NJ: Pearson.

Atkinson, D. (1998). The light of quantum mechanics. *Dialectica, 52*, 103-126.

Ayer, A.J. (1973). *The concept of a person: And other essays*. London: Macmillan.

Azar, B. (1997). APA task force urges a harder look at data. *The APA Monitor, 28* (3), 26.

Baird, D.W. (1981). *Significance tests: Their logic and early history*. (Doctoral Dissertation, Stanford University, 1981)

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Baril, G.L., & Cannon, J.T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist, 50*, 1098-9.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*, 75-97.

Bayes, T. (1763/1958). An essay towards solving a problem in the doctrine of chances (with commentary by G.A. Barnard). *Biometrika, 45*(3/4), 293-315.

Bellhouse, D.R. (1989). A manuscript on chance written by John Arbuthnot. *International Statistical Review, 57*, 249-259.

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18*, 1-32.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berkson, J. (1941). Comments on Dr. Madow's "Note on tests of departure from normality" with some remarks concerning tests of significance. *Journal of the American Statistical Association, 36*, 539-543.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

Berkson, J. (1943). Experience with tests of significance: A reply to Professor R.A. Fisher. *Journal of the American Statistical Association, 38*, 242-246.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70*, 107-115.

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics*, *4*, 24-26.

Blackburn, S. (1994). *The Oxford dictionary of philosophy*. Oxford: Oxford University Press.

Blaikie, N. (2003). *Analyzing quantitative data: From description to explanation*. London: Sage.

Blume, J.D., & Royall, R.M. (2002). Illustrating the law of large numbers (and confidence intervals). *The American Statistician*, *57*, 51-57.

Boring, E.G. (1920). The logic of the normal law of error in mental measurement. *The American Journal of Psychology*, *31* (1), 1-33.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online*, *4*. Retrieved September 1, 2005, from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue7/art2/brandstaetter.pdf

Braselton, R.A. (2003). *Statistics with Maple*. San Diego, Ca: Academic Press.

Brewer, J.K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics*, *10*, 252-268.

Brody, T. (1993). *The philosophy behind physics*. New York: Springer-Verlag.

Brown, M. (2005). Extending the evidence base on the teaching and learning of primary mathematics. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 10-24.

Cai, J. (2005). Critical features of effective mathematics instruction: Issues and practices. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 51-70.

Campbell, R.B. (2001). John Graunt, John Arbuthnott, and the human sex ratio. *Human Biology*, *73*, 605-610.

Capiński, M., & Kopp, E. (1999). *Measure, integral and probability*. New York: Springer-Verlag.

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research*, *5* (4), 513-532.

Carnap, R. (1962). *Logical foundations of probability (2nd ed.)*. Chicago: The University of Chicago Press.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48* (3), 378-399.

Carver, R.P. (1993). The case against statistical significance testing., revisited. *Journal of Experimental Education, 61*, 287-292.

Chaitin, G.J. (2001). *Exploring randomness*. London: Springer.

Chistensen-Szalanski, J.J.J., & Bushyhead,L.R.(1981). Physicians' use of probabilistic information in real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 928-935.

Chow, S.L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

Chow, S.L. (1998a). Precis of Statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21* (2), 169-193.

Chow, S.L. (1998b). The null-hypothesis significance-test procedure is still warrant. *Behavioral and Brain Sciences, 21* (2), 228-235.

Clark, C.A. (1963). Hypothesis testing in relation to statistical methodology. *Review of Educational Research, 33*, 455-473.

Clements, D.H. (1999). 'Concrete' manipulatives, concrete ideas. *Contemporary Issues in Early Childhood,* 1(1), 45-60.

Clements, D. H., & McMillen, S. (1996). Rethinking Concrete Manipulatives. *Teaching Children Mathematics*, 2(5), 270-279.

Cochran, W.G. (1976). Early development of techniques in comparative experimentation. In D.B. Owen (Ed.), *On the history of statistics and probability* (pp. 3-25), New York and Basel: Marcel Dekker.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Coover, J.E. (1917/1975). *Experiments in psychical research*. New York: Arno Press.

Copi, I.M., & Cohen, C. (1998). *Introduction to logic (10 Ed.)*. Upper Saddle River, NJ: Prentice-Hall.

Cortina, J.M., & Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Pyschological Methods, 2*, 161-172.

Cowles, M. (2001). *Statistics in psychology: An historical perspective (2nd ed.)*. Mahwah, NJ: Lawrence Erlabaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37, 553-558.

Cox, D.R., & Hinkley, D.V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Cronbach, L.J. (1975). Beyond two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and non-central distribution. *Educational and Psychological Measurement, 61*, 532-574.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.

Curriculum Development Council (1992). *Syllabus for Secondary Schools: Applied Mathematics (Advanced Level)*. Hong Kong: The Printing Department.

Curriculum Development Council (1998). *Syllabus for Secondary Schools: Applied Mathematics (Advanced Supplementary Level)*. Hong Kong: The Printing Department.

Dale, A.I. (1999). *A history of inverse probability: From Thomas Bayes to Karl Pearson (2nd ed.)*. New York: Springer-Verlag.

Daly, F., Hand, D.J., Jones, M.C., Lunn, A.D., & McConway, K.J. (1995). *Elements of statistics*. London: Pearson.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge: Cambridge University Press.

Dar, R., Serlin, R.C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

David, F.N. (1962/1998). *Games, gods, and gambling: A history of probability and statistical ideas*. Mineola, NY: Dover.

David, H.A. (2001). First (?) occurrence of common terms in statistics and probability. In H.A. David and A.W.F. Edwards (2001), *Annotated readings in the history of statistics* (pp.209-246), New York: Springer-Verlag.

David, H.A., & Edwards, A.W.F. (2001). *Annotated readings in the history of statistics*. New York: Springer-Verlag.

Davis, J.A. (1958). Some pitfalls of data analysis without a formal criterion. In D.E. Morrison, & R.E. Henkel (Eds.) (1970), *The significance test controversy – A reader* (pp.91-93). London: Butterworths.

Dawes, R.M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

De Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In H.E. Kyburg and H.E. Smokler (Eds.) (1964), *Studies in subjective probability* (pp.93-158), New York: Wiley.

De Finetti, B. (1975). *Theory of probability: Volume 2.* New York: Wiley.

DeGroot, M.H., & Schervish, M.J. (2002). *Probability and statistics (3rd ed.).* New York: Addison-Wesley.

de Queiroz, K. (2004). The measurement of test severity, significance tests for resolution, and a unified philosophy of phylogenetic inference. *Zoologica Scripta, 33,* 463-473.

de Queiroz, K., & Poe, S. (2001). Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology, 50,* 305-321.

de Queiroz, K., & Poe, S. (2003). Failed refutations: Further comments Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology, 50,* 305-321.

Diamond, G.A., & Forrester, J.S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine, 98,* 385-394.

Dracup, C. (1995). Hypothesis testing – what it really is. *The Psychologist, 8,* 359-362.

Drumm, D.A. (1995). Statistical error, power and clinical significance. BNI Quarterly, 11. Retrieved August 31, 2005 from http://www.emergemd.com/bniq/article.asp?article_ref_id=11-2-5

Dunn, S. (2001). The significance of "statistical significance". *CancerGuide: Statistics.* Retrieved March 1, 2005 from http://www.cancerguide.org/significance.html

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, P. Slovic, & A. Tversky (Eds.) (1982), *Judgment*

under uncertainty: Heuristics and biases (pp. 249-267). Cambridge: Cambridge University Press.

Edgell, S.E. (1995). Commentary on "Accepting the null hypothesis". *Memory & Cognition, 23*, 525.

Edgeworth, F.Y. (1919). Psychical research and statistical method. *Journal of the Royal Statistical Society*, 82(2), 222-228.

Edwards, A.W. (1992). *Likelihood (Expanded ed.)*. Baltimore and London: The Johns Hopkins University Press

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70, 193-242.

Egeworth, F.Y. (1885). Methods of statistics. *Journal of the Royal Statistical Society*, Jubilee meeting, 181-217.

Eisenhart, C., & Birnbaum, A. (1967). Anniversaries in 1966-1967 of interest to statisticians. *The American Statistician, 21*, 22-29.

Ellis, B. (1973). On the logic of subjective probability. *The British Journal for the Philosophy of Science, 24*, 125-152.

Erwin, E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences, 21* (2), 197-198.

Etes, W.K. (1997). Significance testing in psychological research: Some persisting issue. *Psychological Science, 8* (1), 18-19.

Eysenck, H.J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review, 67*, 269-271.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*, 83-96.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*(7), 798-799.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.

Feller, W. (1968). *An introduction to probability theory and its application, Volume 1*. New York: John Wiley & Sons.

Fetzer, J.H. (1981). *Scientific knowledge: Causation, explanation, and corroboration*. Dordrecht: D. Reidel.

Fidler, F., & Cumming, G. (2005). *Teaching confidence intervals: Problems and potential solutions*. Paper presented at International Association for Statistical Education 55th Session IPM 49. Retrieved September 1, 2005, from http://www.stat.auckland.ac.nz/~iase/publications/13/Fidler-Cumming.pdf

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15*, 119-126.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 503-513. Retrieved August 25, 2005, from http://www.library.adelaide.edu.au/digitised/fisher/48.pdf

Fisher, R.A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research, 39*, 189-192. Retrieved August 25, 2005, from http://www.library.adelaide.edu.au/digitised/fisher/79.pdf

Fisher, R.A. (1943). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association, 38*, 103-104.

Fisher, R.A. (1945). The logical inversion of the notion of the random variable. *Sankhyâ, 7*, 129-132. Retrieved August 25, 2005, from http://www.library.adelaide.edu.au/digitised/fisher/203.pdf

Fisher, R.A. (1960). Scientific thought and the refinement of human reasoning. *Journal of the Operations Research Society of Japan, 3*, 1-10. Retrieved August 25, 2005, from http://www.library.adelaide.edu.au/digitised/fisher/282.pdf

Fisher, R.A. (1971). *The design of experiment (9th ed.)*. New York : Hafner Press. (First edition published in 1935)

Fisher, R.A. (1973a). *Statistical methods for research workers (13 ed.)*. New York: Hafner Publishing. (First edition published in 1925)

Fisher, R.A. (1973b). *Statistical methods and scientific inference (3nd ed.)*. New York: Hafner Press. (First edition published in 1956)

Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics (3rd ed.)*. New York: W.W. Norton.

Freudenthal, H. (1970). Arbuthnot, John. In C. C. Gillispie (Ed.), *Dictionary of Scientific Biography Volume 1* (pp. 208-209). New York: Scribner.

Freund, J.E., & Perles, B.M. (1993). Observations on the definition of p-values. *Teaching Statistics, 15*, 8-9.

Frick, R.W. (1995a). Accepting the null hypothesis. *Memory & Cognition, 23*, 132-138.

Frick, R.W. (1995b). A reply to Edgell. *Memory & Cognition, 23*, 526.

Frick, R.W. (1995c). A problem with confidence interval. *American Psychologist, 50*, 1102-1103.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.

Galavotti, M.C. (2005). *Philosophical introduction to probability*. Stanford, Ca: CSLI.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44-63.

Gavarrett, J. (1840). *Principles Généraux de Statistique Médicale*. Paris.

Geary, R.C. (1947). Testing for normality. *Biometrika, 34*, 209-242.

Gelman, A., & Nolan, D. (2002). You can load a die, but you can't bias a coin. *The American Statistician, 56*(4), 308-311.

Giere, R.N. (1979). *Understanding scientific reasoning*. New York: Holt, Rinehart and Winston.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C.A. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences – Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). The empire of chance: How probability changed science and everyday life. Cambridge: Cambridge University Press.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52* (3), 647-674.

Gillies, D. (1971). A falsifying rule for probability statements. *The British Journal for the Philosophy of Science, 22*, 231-261.

Gillies, D. (1986). Discussion: In defense of the Popper-Miller argument. *Philosophy of Science, 53*, 110-113.

Gillies, D. (2000). *Philosophical theories of probability.* London and New York: Routledge.

Glaser, D.N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care, 8*, 291-296.

Gliner, J.A., Leech, N.L., & Morgan, G.A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*, 83-92.

Glymour, C. (1992). *Thinking things through: An introduction to philosophical issues and achievements.* Cambridge, Mass: A Bradford Book, The MIT Press.

Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist, 4*, 42-46. Reprinted in D.E. Morrison, & R.E. Henkel (Eds.) (1970), *The significance test controversy – A reader* (pp.172-181). London: Butterworths.

Goldfried, M.R. (1959). One tailed tests and "unexpected" results. *Psychological Review, 66*, 79-80.

Goldman, A. (1986). *Epistemology and cognition.* London: Harvard University Press.

Good, I.J. (1983). *Good thinking: The foundations of probability and its application.* Minneapolis, MN: University of Minnesota Press.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130*, 995-1004.

Gott, R., & Duggan, S. (2003). *Understanding and using scientific evidence: How to critically evaluate data.* London: Sage.

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61.

Grayling, A.C. (1996). *Russell.* Oxford: Oxford University Press.

Grayling, A.C. (1997). *An introduction to philosophical logic (3rd ed.).* Oxford: Blackwell.

Grayson, D.A. (1998). The frequentist façade and the flight from evidence inference. *British Journal of Psychology, 89*, 325-345.

Grevholm, B., Persson, L-E., & Wall, P. (2005). A dynamic model for education of doctoral students and guidance of supervisors in research groups. *Educational Studies in Mathematics, 60*, 173-197.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge: Cambridge University Press.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inferenc*e. Cambridge: Cambridge University Press.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University Press.

Hacking, I. (2001). *An introduction to probability and inductive logic.* Cambridge: Cambridge University Press.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52* (1), 15-24.

Hagen, R.L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53* (7), 801-3.

Hájek, A. (1997). "Mises redux" – Redux: Fifteen arguments against finite frequentism. *Erkenntnis, 45*, 209-227.

Hald, A. (1990). *A history of probability and statistics and their applications before 1750.* New York: John Wiley & Sons.

Hald, A. (1998). A *history of mathematical statistics from 1750 to 1930.* New York: John Wiley & Sons.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7.* Retrieved September 1, 2005, from http://www.dgps.de/fachgruppen/methoden /mpr-online/issue16/art1/haller.pdf

Harlow, L.L. (1997). Significance testing introduction and overview. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) (1997), *What if there were no significance tests?* (pp. 1-17). Mahwah, NJ: Lawrence Erlbaum.

Hamming, R.W. (1991). *The art of probability: For scientists and engineers.* Redwood City, CA: Addison-Wesley.

Harris, R.J. (1997). Significance tests have their place. *Psychological Science, 8* (1), 8-11.

Harshbarger, T.R. (1977). *Introductory statistics: A decision map (2nd ed.)*. New York: Macmillan.

Hays, W.L. (1981). *Statistics (3rd ed.)*. New York: Holt, Rinehart & Winston.

Hempel, C.G. (1945a). Studies in the logic of confirmation (I.). *Mind, 54*, 1-26.

Hempel, C.G. (1945b). Studies in the logic of confirmation (II.). *Mind, 54*, 97-121.

Hesse, M.B. (1974). *The structure of scientific inference*. London: Macmillan.

Hillage, J., et al (1998). *Excellence in research on schools*. London: DFEE.

Hogben, L.T. (1957). *Statistical theory: The relationship of probability, credibility, and error; an examination of the contemporary crisis in statistical theory from a behaviourist viewpoint*. London: Allen & Unwin.

Hogg, R.V., & Craig, A.T. (1995). *Introduction to mathematical statistics (5th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Hogg, R.V., & McKean, J.W., & Craig, A.T. (2005). *Introduction to mathematical statistics (6th ed.)*. Upper Saddle River, NJ: Pearson Prentice Hall.

Hopkins, K.D., Hopkins, B.R., & Glass, G.V. (1996). *Basic statistics for the behavioral sciences (3rd ed.)*. Boston: Allyn and Bacon.

Howson, C. (1995). Theories of probability. *British Journal for the Philosophy of Science, 46*, 1-32.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach (2nd ed.)*. Chicago and La Salle: Open Court.

Huang, R., & Mok, I.A.C. (2005). Repetition or Variation – "Practice" in the mathematics classrooms in Hong Kong and Shanghai. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 118-125. English version retrieved August 25, 2005, from http://www.math.ecnu.edu.cn/earcome3/TSG4/EARCOME3_HUANG%20and%20MOK_TSGf().doc

Hubbard, R. (2004). Alphabet soup: Blurring the distinction between p's and α's in psychological research. *Theory & Psychology, 14*, 295-327.

Hubbard, R, & Bayarri, M.J. (2003a). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing, *The American Statistician, 57*, 171–178.

Hubbard, R, & Bayarri, M.J. (2003b). Rejoinder, *The American Statistician, 57*, 181–182.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. *Educational and Psychological Measurement, 60*, 661-681.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333.

Hume, D. (2000). *A treatise of human nature*. Oxford: Oxford University Press. (Original work published 1739)

Humphreys, P. (1985). Why propensities cannot be probabilities. *The Philosophical Review, 94*, 557-570.

Hunter, G. (1971). *Metalogic: An introduction to the metatheory of standard first order logic*. Berkely, CA: University of California Press.

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science, 8* (1), 3-7.

Jeffrey, R. (1964). If. *Journal of Philosophy, 65*, 702-3.

Jeffrey, R. (1977). Mises redux. Reprinted in R. Jeffrey (1992), *Probability and the art of judgment* (pp.192-202). Cambridge: Cambridge University Press.

Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge: Cambridge University Press.

Johnson, D.H. (1995). Statistical sirens: the allure of nonparametrics. *Ecology, 76*, 1998-2000.

Johnstone, D.J. (1986). Tests of significance in theory and practice. *The Statistician, 35*, 491-498.

Johnstone, D.J. (1987). Tests of significance following R.A. Fisher. *The British Journal for the Philosophy of Science, 38*, 481-499.

Kaestle, C. (1993). The awful reputation of education research. *Educational Researcher, 22*(1): 23-31.

Karian, Z.A., & Tanis, E.A. (1999). *Probability and statistics: Explorations with Maple (2nd ed.)*. Upper Saddle River, NJ: Prentice Hall.

Kendall, M., & Plackett, R.L. (Eds.) (1977). *Studies in the history of statistics and probability: Volume II*. London: Charles Griffin & Co., Ltd.

Kendall, P. (1997). Editorial. *Journal of Consulting and Clinical Psychology, 65*, 3-5.

Kennedy, C.A. (2002). *The sampling distribution and the Central Limit Theorem: What they are and why they're important.* (ERIC Document Reproduction Service No. ED 463320).

Keuth, H. (2005). *The philosophy of Karl Popper.* Cambridge: Cambridge University Press.

Keynes, J.M. (1921). *A treatise on probability.* London: Macmillan.

Khrennikov, A. (1999). *Interpretations of probability.* Utrecht, the Netherlands : VSP

Kirk, R.E. (1996). Practical significance: A concept *whose* time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*, 213-218.

Kirwan, C. (1995). Identity. In T. Honderich (Ed.), *The Oxford Companion to Philosophy* (pp.390-391), Oxford: Oxford University Press.

Kline, R.B. (2005). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kluge, A.G. (1997). Testability and the refutation and corroboration of cladistic hypotheses, *Cladistics, 13*, 81-96.

Kluge, A.G. (2001). Philosophical conjectures and their refutation. *Systematic Biology, 50*, 322-330.

Kneale, W.C. (1949). *Probability and induction.* Oxford: Clarendon.

Kolmogorov, A.N. (1956). *Foundations of the theory of probability (2nd English ed.).* Prvidence, Rhode Island: AMS Chelsea Publishing.

Kruskal, W.H., & Stigler, S.M. (1997). Normative terminology: 'normal' in statistics and elsewhere. In B.D. Spencer (Ed.), *Statistics and public policy* (pp.85-111), Oxford: Clarendon Press.

Kuhn, T.S. (1977). *The essential tension: Selected studies in scientific tradition and change.* Chicago: University of Chicago Press.

Laplace, P.-S. (1994). *Philosophical essay on probabilities* (Translated from the fifth French edition of 1825 with notes by A.I. Dale). New York: Springer-Verlag.

Lehmann, E.L. (1986). *Testing statistical hypothesis (2nd ed.).* New York: John Wiley & Sons.

Lehmann, E.L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of *Amercian Statistical Association, 88* (424), 1242-1249.

Leung, F.K.S., & Park, K. (2005). Is there anything special about mathematics teaching in East Asia? Results from the TIMSS 1999 Video Study and the learners' perspective study. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 83.

Levin, J.R., & O'Donnell, A.M. (1999). What to do about educational research's credibility gaps. *Issues in Education*, 5(2), 177-229.

Levin, J.R., & Robinson, D.H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review, 11* (2), 143-155.

Limpert, E., Stahel, W.A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience, 51*, 341-352.

Lind, D.A., Marchal, W.G., & Mason, R.D. (2002). *Statistical techniques in business and economics (11th ed.)*. Boston, Mass: McGraw-Hill.

Liu, T., & Stone, C.C. (1999). A critique of one-tailed hypothesis test procedures in business and economics statistics textbooks. *Journal of Economic Education, 30* (1), 59-63.

Loftus, G.R. (1993). Editorial comment. *Memory & Cognition, 21*, 1-3.

Lubin, A. (1962). Statistics. *Annual Review of Psychology, 13*, 345-370.

Lunt, P.K., & Livingstone, S.M. (1989). Psychology and statistics: Testing the opposite of the idea you first thought of. *The Psychologist, 2*, 528-531.

Macdonald, R.R. (1997). On statistical testing in psychology. *British Journal of Psychology, 88*, 333-347.

Macdonald, R.R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Understanding Statistics, 1*, 167-189.

Macdonald, R.R. (2004). Statistical inference and Aristotle's *Rhetoric*. *British Journal of Mathematical and Statistical Psychology, 57*, 193-203.

Man, Y.K., Leung, C.K., & Ng, Y.K. (1997). *Introduction to the foundations of mathematics (Chinese)*. Hong Kong: Hong Kong Educational Publishing.

Martin-Löf, P. (1969). The literature on von Mises' Kollective revisited. *Theoria, 35*, 12-37.

Marton, F. ,Tsui, A.B.M., Chik, P., Ko, P.Y., Lo, M. L., & Mok, I. A. C. (2004). *Classroom discourse and the space of learning*. N.J.: Lawrence Erlbaum.

Matthews, R.A.J. (1998, September). *Facts versus factions: The use and abuse of subjectivity in scientific research – Part 2*. Working Paper for the European Science and Environment Forum. Retrieved August 1, 2003, from http://ourworld.compuserve.com/homepages/rajm/twooesef.htm

Matthews, R.A.J. (1999). *Statistical snake-oil: The use and abuse of significance tests in science*. Cambridge: ESEF.

May, K. (2003). A note on the use of confidence intervals. *Understanding Statistics, 2*, 133-135.

Mayo, D.G. (1996). *Error and the growth of experimental knowledge*. Chicago and London: The University of Chicago Press.

McCall, R.B. (2001). *Fundamental statistics for behavioral sciences (8th ed.)*. Belmont, CA: Wadsworth.

McClure, J., & Suen, H.K. (1994). Interpretation of statistical significance testing: A matter of perspective. *Topics in Early Childhood Special Education, 14* (1), 88 – 100.

McCurdy, C.S.I. (1996). Humphreys's paradox and the interpretation of inverse conditional propensities. *Synthese, 108*, 105-125.

McKnight, C., Magid, A., Murphy, T.J., & McKnight, M. (2000). *Mathematics education research: A guide for the research mathematician*. Providence, Rhode Island: American Mathematical Society.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15*, 295-300.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Mellor, D.H. (2005). *Probability: A philosophical introduction*. London & New York: Routledge.

Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5(1)*, 4-18.

Michalewicz, Z., & Fogel, D.B. (2000). *How to solve it: Modern heuristics*. New York: Springer.

Miller, D. (1980). Falsification versus inductivism. In J.L. Cohen, & M. Hesse (Eds.), *Applications of inductive logic* (pp.109-129). Oxford: Clarendon Press.

Miller, D. (1994). *Critical rationalism: A restatement and defence*. Chicago and La Salle, Illinois: Open Court.

Miller, D. (2002). Propensities may satisfy Bayes's theorem. In R. Swinburne (Ed.) (2002), *Bayes's theorem* (pp. 111-116). Oxford: Oxford University Press.

Miller, D. (2005). Do we reason when we think we reason, or do we think? *Learning for Democracy, 1*, 57-71.

Miller, D. (2006). Induction: A problem solved. In D. Miller, *Out of error: Further Essays on Critical Rationalism* (pp. 111-132). Aldershot, Hampshire: Ashgate. (Original work published 2002)

Miller, I., & Miller, M. (1999). *John E. Freund's mathematical statistics (6th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Mittage, K.C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*, 14-20.

Mok, I.A.C. (2005). How Chinese learn mathematics – Lessons from Shanghai. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 25-34.

Montori, V.M., Kleinbart, J., Newman, T.B., Keitz, S., et al. (2004). Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *Canadian Medical Association Journal, 171*, 611-615.

Moore, D.S. (1997). *Statistics: Concepts and controversies (4th ed.)*. New York: W.H. Freeman and Company.

Moore, D.S., & McCabe, G.P. (2006). *Introduction to the practice of statistics (5th ed.)*. New York: W.H. Freeman and Company.

Morgan, B. (1972). *Men and discoveries in mathematics*. London: John Murray.

Morrison, D.E., & Henkel, R.E. (Eds.) (1970). *The significance test controversy: A reader*. London: Butterworth.

Mosteller, F. (1965). *Fifty challenging problems in probability with solutions.* Reading, Mass: Addison-Wesley.

Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997). There is a time and a place for significance testing. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) (1997), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum.

Murphy, K.R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45*, 403-404.

Musgrave, A. (1999a). How to do without inductive logic. *Science & Education, 8*, 395-412.

Musgrave, A. (1999b). *Essays on realism and rationalism.* Amsterdam: Rodopi.

Musgrave, A. (2004). How Popper [might have] solved the problem of induction. *Philosophy, 79*, 19-31.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability (2nd ed., revised and enlarged).* Washington, D. C.: Graduate School, U.S. Dept. of Agriculture.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36*, 97-131.

Neyman, J., & Pearson, E.S. (1928a). On the use of interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175-240.

Neyman, J., & Pearson, E.S. (1928b). On the use of interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 263-294.

Neyman, J., & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.

Ng, Y.K., & Wu, K.F. (2003, August 29). *A study on the recent use of statistical procedures in mathematics education research.* Report presented at the Seminar of Department of Mathematics, The Hong Kong Institute of Education.

Ng, Y.K. (1999). Implications of artificial intelligence for secondary school mathematics. *Curriculum Forum, 8*, 28-43.

Nicholls, N. (2000). The insignificance of significance testing. *Bulletin of the American Meteorological Society, 81*, 981-986.

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5* (2), 3-14.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*, 641-650.

Nunnally, J. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester: Wiley.

Oakes, W.F. (1975). On the alleged falsity of the null hypothesis. *The Psychological Record, 25*, 265-272.

Onwuegbuzie, A. J., & Daniel L. G. (2003, February 19). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education* [On-line], *6*(2). Available: http://cie.ed.asu.edu/volume6/number2/

Ottenbacher, K.J. (1996). The power of replications and replications of power. *American Statistician, 50* (3), 271-275.

Pauker, S.P., & Pauker, S.G. (1979). The amniocentesis decision: An explicit guide for parents. In C.J. Epstein, C.J.R. Curry, S. Packman, S. Sherman, & B.D. Hall (Eds.), *Risk, communication, and decision making in genetic counseling (Birth defects Original article series, Vol. 15, Number 5C)* (pp. 289-324). New York: Liss.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood.* Oxford: Clarendon Press.

Pearce, S.C. (1992). Introduction to Fisher (1925): Statistical methods for research workers. In S. Kotz, & N.L. Johnson (Eds.), *Breakthroughs in statistics Vol. 2: Methodology and distribution* (pp.59-65). New York: Springer-Verlag.

Pearson, E., & Kendall, M.G. (Eds.) (1970). *Studies in the history of statistics and probability: Volume 1.* London: Griffin.

Pearson, E. S. (1990). *'Student': A statistical biography of William Sealy Gosset.* Oxford: Clarendon Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine (Series V)*, 50, 157-175. Reprinted in Karl Pearson (1948), *Karl Pearson's early statistical papers* (pp.339-357), Cambridge: Cambridge University Press.

Pearson, K. (1924). Historical note on the origin of the normal curve of errors. *Biometrika, 16* (3/4), 402-404.

Pearson, K. (1978). *The history of statistics in the 17th and 18 centuries against the changing background of intellectual, scientific and religious thought. Lectures by Karl Pearson given at University College London during the academic sessions 1921-1933 (Edited by E.S. Pearson).* London: Charles Griffin.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Phillips, D.C. (1999). How to play the game: A Popperian approach to the conduct of educational research. In G. Zecha (Ed.), *Critical rationalism and educational discourse* (pp. 170-190). Amsterdam: Rodopi.

Pollard, P., & Richardson, J.T.E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10*, 159-163.

Popper, K. (1957). The propensity interpretation of the calculus of probability, and the quantum theory. In S. Körner (Ed.) (1957), *Observational and interpretation in the philosophy of physics* (pp.65-70). London: Butterworth & Company. Reprinted in D. Miller (Ed.) (1985), *Popper selections* (pp. 199-206). Princeton, NJ: Princeton University Press.

Popper, K. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science, 10*, 25-42.

Popper, K. (1963). *Conjectures and refutations.* London: Routledge & Kegan Paul.

Popper, K. (1974). Replies to my critics. In P.A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp.961-1197). La Salle, Illinois: Open Court.

Popper, K. (1980). *The logic of scientific discovery.* London & New York: Routledge. (First edition published in 1957)

Popper, K. (1983). *Realism and the aim of science*. London: Routledge.

Popper, K. (1989). *Conjectures and refutations: The growth of scientific knowledge (5th ed, revised)*. London: Routledge. (First edition published in 1963)

Popper, K. (1990). *A world of propensities*. Bristol: Thoemmes Press.

Popper, K., & Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature, 302*, 687-688.

Popper, K., & Miller, D. (1984). Reply to Levi (1984), Jeffrey (1984) and Good (1984). *Nature, 310*, 434.

Popper, K., & Miller, D. (1987). Why probabilistic support is not inductive. *Philosophical transactions of the Royal Society of London. Series A*, 321, 569-591.

Pring, R. (2000). *Philosophy of educational research.* London and New York: Continuum.

Quine, W.V.O. (1980). *From a logical point of view (2nd ed., revised).* Cambridge, Mass: Harvard University Press.

Quine, W.V.O. (1982). *Methods of logic (4th ed.).* Cambridge, Mass: Harvard University Press.

Ramsey, F.P. (1931). *The foundations of mathematics and other logical essays.* London: Routledge & Kegan Paul.

Randall, L. (2005). *Warped passages: Unraveling the mysteries of the universe's hidden dimensions*. New York: HarperCollins.

Reeves, C.A., & Brewer, J.K. (1980). Hypothesis testing and proof by contradiction: An analogy. *Teaching statistics, 2*, 57-59.

Reichenbach, H. (1949). *The theory of probability: An inquiry into the logical and mathematical foundations of the calculus of probability* (English translation by E.H. Hutten and M. Reichenbach). Berkeley and Los Angeles: University of California Press.

Reichenbach, H. (1976). *Laws, modalities, and counterfactuals.* Berkeley and Los Angeles: University of California Press.

Rinskopf, D.M. (1997). Testing "small", not null, hypotheses: Classical and Bayesian approaches. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) (1997), *What if there were no significance tests?* (pp. 319-332). Mahwah, NJ: Lawrence Erlbaum.

Rosenthal, J.S. (2000). *A first look at rigorous probability theory.* Singapore: World Scientific.

Royall, R.M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416-428.

Russell, B. (1912/1980). *The problems of philosophy.* Oxford: Oxford University Press.

Sackrowitz, H., & Samuel-Cahn, E. (1999). P values as random variables – Expected p values. *The American Statistician,* 53 (4), 326-331.

Sainsbury, M. (1991). *Logical forms: An introduction to philosophical logic.* Oxford: Basil Blackwell.

Salmon, W.C. (2005). *Reality and rationality.* Oxford: Oxford University Press.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York: W.H. Freeman.

Salvage, I.R. (1957). Nonparametric statistics. *Journal of the American Statistical Association, 52,* 331-344.

Salvitz, D.A., Tolo, K., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology,* 1970-1900. *American Journal of Epidemiology,* 139, 1047-1052.

Schield, M. (1997). *Interpreting statistical confidence.* Paper presented at the Joint Statistical Meetings 1997 of American Statistical Association. Retrieved April 1, 2005, from http://web.augsburg.edu/~schield/MiloPapers/97ASA.pdf

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.) (1997), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum.

Schmidt, F.L., Hunter, J.E., & Urry, V.E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Schoenfeld, A.H. (2000). Purposes and methods of research in mathematics education. *Notices of the American Mathematical Society, 47*, 641-649.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R.A. Fisher*. Dordrecht: D. Reidel.

Selvin, H.C. (1957). A critique of tests of significance in survey research. *American Sociological Review, 22*, 519-527. Reprinted in Reprinted in D.E. Morrison, & R.E. Henkel (Eds.) (1970), *The significance test controversy – A reader* (pp.94-106). London: Butterworths.

Shafer, G. (1978). Non-addicitve probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19, 309-370.

Shea, C. (1996). Psychologists debate accuracy of 'significance test'. *The Chronicle of Higher Education, 42*, A12, A17.

Shenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*, 182-186.

Shoesmith, E. (1985). Nicholas Bernoulli and the argument for divine providence. *International Statistical Review, 53*(3), 255-259.

Shoesmith, E. (1987). The continental controversy over Arguthnot's argument for divine providence. *Historia Mathematica, 14*, 133-146.

Shrout, P.E. (1997). Should significance tests be banned? Introduction to a Special Section exploring the pros and cons. *Psychological Science, 8* (1), 1-2.

Shulman, L.S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-393.

Siddall, M.E., & Kluge, A.G. (1997). Probabilism and phylogenetic inference. *Cladistics, 13*, 313-336.

Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy, 79*, 186-195.

Simon, J.L. (1968). What does the normal curve "mean"? *The Journal of Educational Research*, 61 (10), 435-438.

Simon, J.L. (unpublished). *The Philosophy and Practice of Resampling Statistics.* Retrieved September 1, 2005, from http://www.resample.com/content/teaching/philosophy/index.shtml or http://www.juliansimon.com/writings/Resampling_Philosophy/

Skipper, J.K., Guenter, A.L., & Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social sciences: In D.E. Morrison & R.E. Henkel (Eds.), *The significance tests controversy: A reader* (pp.155-160). Chicago: Aldine.

Smith, P. (1998). *Explaining chaos.* Cambridge: Cambridge University Press.

Smithson, M. (2000). *Statistics with confidence.* London: Sage.

Smithson, M. (2003). *Confidence intervals.* London: Sage Publications.

Sober, E. (2002). Bayesianism – its scope and limts. In R. Swinburne (Ed.) (2002), *Bayes's theorem* (pp. 21-38). Oxford: Oxford University Press.

Sober, E. (2005). *What does a confidence interval mean?* Retrieved August 31, 2005, from http://philosophy.wisc.edu/sober/confidence%20interval%20ho%20 2005.pdf

Sowell, E.J. (1989). Effects of Manipulative Materials in Mathematics Instruction. *Journal for Research in Mathematics Education*, 20, 498-505.

Speiser, D. (Ed.) (1975). *Die werke von Jakob Bernoulli: Band 3.* Basel: Birkhäuser Verlag.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science, 41*, 211-226.

Stigler, S.M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, Mass.: The Belknap of Harvard University Press.

Stigler, S.M. (1999). *Statistics on the table: The history of statistical concepts and methods.* Cambridge, Mass.: Harvard University Press.

Streiner, D.L. (2003). Unicorns do exists: A tutorial on "proving" the null hypothesis. *The Canadian Journal of Psychiatry, 48*, 756-761.

Student (Gosset, W.S.) (1908). The probable error of a mean. *Biometrika*, 6, 1-25. [Reprinted in E.S. Pearson, & J. Wishart (Eds.) (1958), *"Student's" collected papers* (pp.11-34). Cambridge: Cambridge University Press.]

Sun, W. (2003). *Interpretations of probability.* Unpublished doctoral dissertation, University of Connecticut, Connecticut.

Suydam, M.N. (1986). Research Report: Manipulative Materials and Achievement. *Arithmetic Teacher*, 33, 10 & 32.

Swan, J. (2003). How science can contribute to the improvement of educational practice. *Oxford Review of Education, 29*, 253-268.

Swinburne, R. (2002). Introduction. In R. Swinburne (Ed.), *Bayes's theorem* (pp.1-20). Oxford: Oxford University Press.

Swinburne, R. (Ed.) (2002). *Bayes's theorem.* Oxford: Oxford University Press.

Tankard, J.W. (1984). *The statistical pioneers.* Cambridge, Mass.: Schenkman.

Thompson, A.G., Philipp, R.A., Thompson, P.W., & Boyd, B.A. (1994). Calculational and conceptual orientation in teaching mathematics. In D.B. Aichele & A.F. Coxford, *Professional development for teachers of mathematics* (pp.79-92). Reston, VA: National Council of Teachers of Mathematics.

Thompson, B. (1994). *The concept of statistical significance testing.* (ERIC Document Reproduction Service No. ED 366654).

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments.. *Educational Researcher, 26*, 29-32.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1999a). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practice.' *The Journal of Psychology, 133* (2), 133-140.

Thompson, B. (1999b). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9* (2), 165-181.

Thompson, B. (1999c). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65* (3), 329-337.

Thompson, B. (1999d). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* as to importance. *Educational Psychology Review, 11* (2), 157-169.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31.

Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the Journal of Experimental Education. *Journal of Experimental Education, 66* (1), 75-83.

Thompson, W.D. (1987). Statistical criteria in the interpretation of epidemiologic data. *American Journal of Public Health, 77*, 191-194.

Tooley, J., & Darby , D. (1998). *Educational research: a critique : a survey of published educational research : report presented to OFSTED.* London: Office for Standards in Education.

Torretti, R. (1999). *The philosophy of physics.* Cambridge: Cambridge University Press.

Truran, J.M. (1998). The development of the idea of the null hypothesis in research and teaching. In L. Periera-Mendoza, L.S. Kea, T.W. Kee, & W. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp.1067-1073). Voorburg, The Netherlands: ISI Permanent Office.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) (1982), *Judgment under uncertainty: Heuristics and biases* (pp. 153-160). Cambridge: Cambridge University Press.

Twaite, J.A., & Monroe, J.A. (1979). *Introductory statistics.* Glenview, IL: Scott, Foresman.

Tyler, R.W. (1931). What is statistical significance? *Educational Research Bulletin, 10*, 115-118, 142.

Tymoczko, T., & Henle, J. (1995). *Sweet reason: A field guide to modern logic.* New York: W.H. Freeman.

Uspensky, J.V. (1937). *Introduction to mathematical probability.* New York and London: McGraw-Hill.

Venn, J. (1889). Cambridge anthropometry. *Journal of the Anthropological Institute of Great Britain and Ireland*, 18, 140-154.

von, Mises, R. (1957). *Probability, statistics and truth (2nd rev English ed.)*. New York: Dover.

von Plato, J. (1994). *Creating modern probability: Its mathematics, physics and philosophy in historical perspective*. Cambridge: Cambridge University Press.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.

Warnock, G.J. (1960). New books: *The Logic of Scientific Discovery*. *Mind, 69*, 99-101.

Weatherford, R. (1982). *Philosophical foundations of probability theory*. London: Routledge & Kegan Paul.

Weigle, D.C. (1994, January 27). *Historical origins of contemporary statistical testing practices: How in the world did significance testing assume its current place in contemporary analytic practice?* Paper presented at the Annual Meeting of the Southwest Educational Research Association, San Antonio, Tx. (ERIC Document Reproduction Service No. ED367678).

Weitzman, R.A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports, 54*, 355-363.

Wells, D. (1986). *The Penguin Dictionary of Curious and Interesting Number*. London: Penguin Books.

Wilcox, R.R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.

Wild, C.J., & Seber, G.A.F. (2000). *Chance encounters: A first course in data analysis and inference*. New York: John Wiley & Sons.

Williams, A.M. (1999). Novice student's conceptual knowledge of statistical hypothesis testing. In J.M. Truran and K.M. Truran (Eds.), *Making the difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australia* (pp.540-560). Adelaide, South Australia: MERGA.

Williamson, J. & Corfield, D. (2001). Introduction: Bayesianism into the 21st Century. In J. Williamson & D. Corfield (2001), *Foundations of Bayesianism* (pp.1 – 16), Dordrecht: Kluwer.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologists*, *54*, 594-604.

Wolfram, S. (1989). *Philosophical logic: An introduction.* London and New York: Routledge.

Wong, N.Y. (2005). An investigation of the mathematics curricula in various regions around the world: What lessons can we learn from it. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 35-50.

Wood, T.B., & Stratton, F.J.M. (1910). The interpretation of experimental results. *Journal of Agricultural Science*, 3, 417-440.

Wu, K.F., Ng, Y.K., & Sze, C.L. (2003, August 29). *Base-rate fallacy.* Report presented at the Seminar of Department of Mathematics, The Hong Kong Institute of Education.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 19-34.

Zecha, G. (1999). Critical rationalism and educational discourse: Introductory overview. In G. Zecha (Ed.), *Critical rationalism and educational discourse* (pp. 7-16). Amsterdam: Rodopi.

Zhang, D. (2005). Math War in China and a personal view on the appearance of the nature of mathematics. *Proceedings of Conference on Mathematics Education 2005*, Hong Kong, 89-96.