SCIENTIFIC REASONING
THE BAYESIAN APPROACH

Colin Howson and Peter Urbach

THIRD EDITION

. . . if this [probability] calculus be condemned, then the whole of the sciences must also be condemned.
—Henri Poincaré

Our assent ought to be regulated by the grounds of probability.
—John Locke

OPEN COURT
Chicago and La Salle, Illinois
2006

urally to aspects of Bayesian logic, whereas non-Bayesian accounts fail more or less completely. So far, we have concentrated chiefly on deterministic theories. We shall see in the next and following chapters that the Bayesian approach applies equally well to statistical reasoning.

CHAPTER 5
Classical Inference: Significance Tests and Estimation

In the last chapter, we showed how leading aspects of scientific reasoning are illuminated by reference to Bayes's theorem, confining our attention, however, mainly to deterministic theories. We now consider theories that are not deterministic but probabilistic, or statistical. From the Bayesian viewpoint the division is artificial and unnecessary, the two cases differing only in regard to the probability of the evidence relative to the theory, that is, $P(e \mid h)$, which figures in the central theorem: when h is deterministic, this probability is either 1 or 0, depending on whether h entails e or is refuted by it; when h is statistical, $P(e \mid h)$ typically takes an intermediate value. The uniform treatment that this affords is unavailable in non-Bayesian methodologies, whose advocates have instead developed a specific system, known as *Classical Statistical Inference* or sometimes as *Frequentism*, to deal with statistical theories.

This system, with its 'significance tests', 'confidence intervals', and the rest, swept the board for most of the twentieth century, and its influence is still considerable. The challenge to Bayesian methodology posed by Frequentism requires an answer, and this we shall give in the present and succeeding chapters.

5.a | Falsificationism in Statistics

The simple and objective mechanism by which a hypothesis may, under certain circumstances, be logically refuted by observational evidence could never work with statistical hypotheses, for these ascribe probabilities to possible events and do not say of any that they will or will not actually occur. The fact that statistical theories have a respected place in science and are regularly tested and

evaluated through experiment is therefore an embarrassment to the methodology of falsificationism. In consequence, defenders of that methodology have tried to take account of statistical theories by modifying its central dogma.

The modified idea acknowledges that a statistical hypothesis is not strictly falsifiable, and what it proposes is that when an event occurs to which the hypothesis attaches a sufficiently small probability, it should be *deemed* false; scientists, Popper said, should make "a methodological decision to regard highly improbable events as ruled out—as prohibited" and he talked of hypotheses then being "practically falsified" (1959a, p. 191). The mathematician and economist Cournot (1843, p. 155) expressed the same idea when he said that events of sufficient improbability "are rightly regarded as physically impossible".

But is it right? After all, a distinctive feature of statistical hypotheses is that they do not rule out events that they class as improbable. For example, the Kinetic Theory attaches a tiny probability to the event of ice spontaneously forming in a hot tub of water, but does not rule it out; indeed the fact that the theory reveals so strange an event as a possibility, contrary to previous opinion, is one of its especially interesting features. And even though this particular unlikely event may never materialize, immensely improbable events, which no one would regard as refuting the Kinetic Theory, do occur all the time, for instance, the spatial distribution at a particular moment of the molecules in this jug of water.

Or take the simple statistical theory that we shall frequently use for the purpose of illustration, which claims of some particular coin that it has a physical probability, constant from throw to throw, of $\frac{1}{2}$ of landing heads and the same probability of landing tails (the coin is said then to be 'fair'). The probability of any particular sequence of heads and tails in, say, 10,000 tosses of the coin is $2^{-10,000}$, a minuscule value, yet it is the probability of every possible outcome of the experiment, one of which will definitely occur. The implication of the Cournot-Popper view that this definite occurrence should be regarded as physically impossible is clearly untenable.

5.b | Fisherian Significance Tests

Fisher was inspired by both the falsificationist outlook and the ideal of objectivity when, building on the work of Karl Pearson and W.S. Gossett (the latter, writing under the pen name 'Student'), he developed his system of *significance tests* for testing statistical theories. Fisher did not postulate a minimal probability to represent physical impossibility, and so avoided the problem that destroys the Cournot-Popper approach. His proposal, roughly speaking, was that a statistical hypothesis should be rejected by experimental evidence when it is, on the assumption of that hypothesis, contained in a certain set of outcomes that are *relatively* unlikely, relative, that is, to other possible outcomes of the experiment.

Before assessing how well they are suited to their task, let us set out more precisely the nature of Fisher's significance tests, which we shall illustrate using, as the hypothesis under test (what Fisher called the *null hypothesis*), the fair-coin hypothesis mentioned above. To perform the test, an experiment must be devised: in our example, it will involve flipping the coin a predetermined number of times, say 20, and noting the result; this result is then analysed in the following four stages.

1 First, specify the *outcome space*, that is, all the results that the experiment could have produced. In our example, this would normally be taken to comprise the 2^{20} possible sequences of 20 heads or tails. (We examine the assumptions underlying the specification of any outcome space in the next section when we discuss 'stopping rules'.) The result of a coin-tossing experiment would not normally be reported as a point in the outcome space just described but would be summarized in some numerical form, and for the purpose of our example, we shall select r, the number of heads in the outcome. Such a numerical summary when used in a significance test is known as a *test-statistic*; it is formally a random variable, as defined in Chapter 2. (We shall presently discuss the basis upon which test-statistics are chosen.)

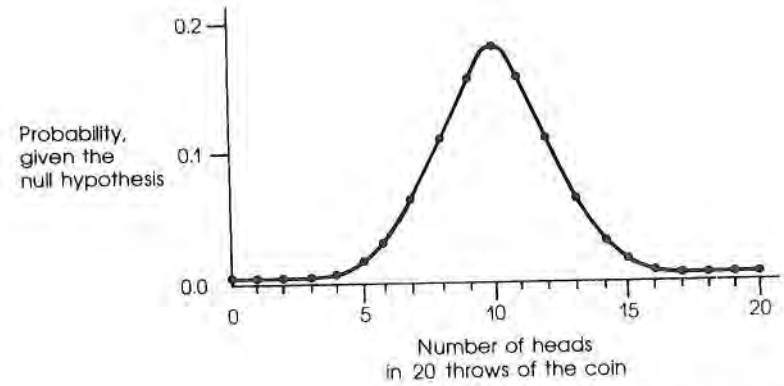
2 Next, calculate the probability, relative to the null hypothesis, of each possible value of the test-statistic—its *sampling distribution*. In general, if the probability of getting a head in a

coin-tossing experiment is p, and of getting a tail is q, then r heads will appear in n tosses of the coin with probability ${}^nC_r p^r q^{n-r}$.[1] In the present case, $p = q = \frac{1}{2}$ and $n = 20$. The required probabilities can now be directly calculated; they are shown in Table 5.1 and also displayed graphically below.

TABLE 5.1

The Probabilities of Obtaining *r* Heads in a Trial consisting of 20 Tosses of a Fair Coin

Number of Heads (r)	*Probability*	*Number of Heads* (r)	*Probability*
0	9×10^{-7}	11	0.1602
1	1.9×10^{-5}	12	0.1201
2	2×10^{-4}	13	0.0739
3	0.0011	14	0.0370
4	0.0046	15	0.0148
5	0.0148	16	0.0046
6	0.0370	17	0.0011
7	0.0739	18	2×10^{-4}
8	0.1201	19	1.9×10^{-5}
9	0.1602	20	9×10^{-7}
10	0.1762		



[1] This familiar fact is demonstrated in standard statistics textbooks. ${}^nC_r$ is equal to $\frac{(n!)}{(n-r)!\,r!)}$.

3 The third stage of Fisher's analysis requires us to look at all the results which *could have* occurred and which, relative to the null hypothesis, are, as Fisher put it, "more extreme than" the result that did occur. In practice, this vague expression is interpreted probabilistically, the requirement then being that we examine possible outcomes of the trial which, relative to the null hypothesis, have a probability less than or equal to the probability of the actual outcome. We should then calculate the probability (p^*) that the experimental result will fall within this group. (p^* is often called the *p-value* of the result.)

To illustrate, suppose our experiment produced 4 heads and 16 tails, which we see from the table occurs, if the null hypothesis is true, with probability 0.0046. The results with less or equal probability to this are $r = 4, 3, 2, 1, 0$ and $r = 16, 17, 18, 19, 20$ and the probability of any one of them occurring is the sum of their separate probabilities, *viz*:

$$p^* = 2 \times (0.0046 + 0.0011 + 2 \times 10^{-4} + 1.9 \times 10^{-5} + 9 \times 10^{-7}) = 0.012.$$

4 A convention has grown up, following Fisher, to *reject* the null hypothesis just in case $p^* \leq 0.05$. However, some statisticians recommend 0.01 or even 0.001 as the critical probability. The critical probability that is adopted is called the *significance level* of the test and is usually labelled α. If an experimental result is such that $p^* \leq \alpha$, it is said to be *significant at the α significance level*, and the null hypothesis is said to be *rejected at the α (or 100α percent) level*.

In our example, the coin produced 4 heads when flipped 20 times, corresponding to $p^* = 0.012$; since this is below 0.05, the null hypothesis should be rejected at the 0.05 or 5 percent level. But a result of 6 heads and 14 tails, with $p^* = 0.115$, would not be significant, and so the null hypothesis should then not be rejected at that level.

This simple example illustrates the bare bones of Fisher's approach. It is, however, not always so easy to apply in practice. Take the task often treated in statistics textbooks of testing whether two populations have the same means, for instance,

whether two groups of children have the same mean IQ. It may not be feasible to take measurements from every child, in which case, the recommended procedure is to select children at random from each of the groups and compare their IQs. But to perform a significance test on the results of this sampling one needs a test-statistic with a determinate and known distribution and these are often difficult to find. A solution was found in the present case by 'Student', who showed that provided the experimental samples were sufficiently large to ensure approximate normality, the so-called *t*-statistic[2] has the appropriate properties for use in a significance test.

Which Test-Statistic?

Fisher's theory as so far expounded is apparently logically inconsistent. This is because different random variables may be defined on any given outcome space, not all of them leading to the same conclusion when used as the test-statistic in a significance test; one test-statistic may instruct you to reject some hypothesis when another tells you not to.

We can illustrate this very simply in relation to our coin-tossing experiment. We there chose the number of heads in the outcome as the test-statistic, which, with 20 throws of the coin, takes values from 0 to 20. Now define a new statistic, r', with values from 0 to 18, derived from the earlier statistic by grouping the results as indicated in Table 5.2. In this slight modification, the outcome *5 heads* and the outcome *10 heads* are counted as a single result whose probability is that of obtaining either one of these; similarly, for the results *14* and *15 heads*. This new statistic is artificial, having no natural meaning or appeal, but according to the definition, it is a perfectly proper test-statistic.

It will be recalled that previously, with the number of heads as the test-statistic, the result *6 heads, 14 tails* was *not* significant at the 0.05 level. It is easy to see that using the modified statistic, this result now *is* significant at that level ($p^* = 0.049$). Hence Fisher's principles as so far described tell us both to reject and not to reject the null hypothesis, which is surely impossible. Clearly

[2] See Section 6.c for more details of the *t*-statistic.

TABLE 5.2

The Probability Distribution of the r'-Statistic

Value of Statistic (r')	*Probability*	*Value of Statistic (r')*	*Probability*
0 (0 heads)	9×10^{-7}	10 (11 heads)	0.1602
1 (1 heads)	1.9×10^{-5}	11 (12 heads)	0.1201
2 (2 heads)	2×10^{-4}	12 (13 heads)	0.0739
3 (3 heads)	0.0011	13 (14 or 15 heads)	0.0518
4 (4 heads)	0.0046	14 (16 heads)	0.0046
5 (6 heads)	0.0370	15 (17 heads)	0.0011
6 (7 heads)	0.0739	16 (18 heads)	2×10^{-4}
7 (8 heads)	0.1201	17 (19 heads)	1.9×10^{-5}
8 (9 heads)	0.1602	18 (20 heads)	9×10^{-7}
9 (5 or 10 heads)	0.1910		

test-statistics need some restriction that will ensure that all permissible ones lead to similar conclusions. And any such restriction must be recommended by more than the consistency it brings; it must produce the right consistent result, if there is one, for the right reasons, if there are any.

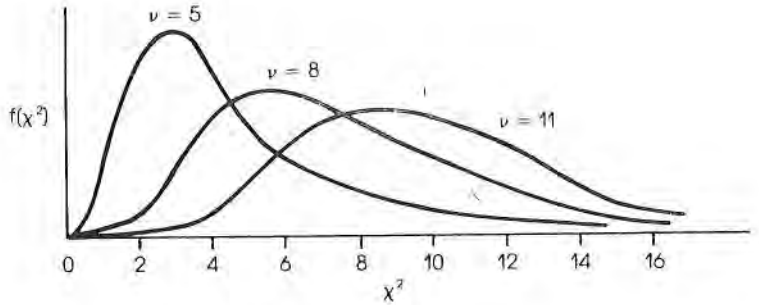
The Chi-Square Test

A striking illustration of the difficulties posed by the multiplicity of possible test-statistics is the chi-square (or χ^2) goodness-of-fit test, which bulks large in the literature and is widely used to test hypotheses that ascribe probabilities to several different types of event, for example, to the outcomes of rolling a particular die. Suppose the die were rolled n times and landed with a six, five, etc. showing uppermost with frequencies $O_6, O_5, \ldots, O_1$. If p_i is the probability that the null hypothesis ascribes to the outcome i, then np_i is the *expected frequency* (E_i) of that outcome. The null hypothesis is tested by the following so-called *chi-square statistic*:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

the sum being taken over all the possible outcomes of the trial.

Karl Pearson discovered the remarkable fact, very helpful for its application to significance tests, that the probability distribution of this statistic is practically independent of the unknown probabilities and of n and is dependent just on the number, ν, of the test's so-called *degrees of freedom*, where $\nu = J - 1$, and J is the number of separate cells into which the outcome was divided when calculating χ^2. The probability density distributions of χ^2, for various values of ν, are roughly as follows:



We may illustrate the χ^2-test with a simple example. Let the null hypothesis assert that a particular die is 'true', that is, has equal probabilities, of $\frac{1}{6}$, constant from throw to throw, of falling with each of its sides uppermost. Now consider an experiment involving 600 rolls of the die, giving the results, say: *six* (90), *five* (91), *four* (125), *three* (85), *two* (116), *one* (93).

To perform a chi-square test, we must calculate χ^2 for these data, as follows ($E_i = 600 \times \frac{1}{6} = 100$, for each i):

$$\chi^2 = \frac{1}{100}\,[(100-90)^2 + (100-91)^2 + (100-125)^2 + (100-85)^2 + (100-116)^2 + (100-93)^2] = 13.36.$$

Since the outcome involves six cells, the number of degrees of freedom is five and, as can be roughly gauged from the above sketches and more precisely established by consulting the appropriate tables, the probability of obtaining a value of χ^2 as large or larger than 13.36 is less than 0.05, so the result is significant, and the null hypothesis must therefore be rejected at the corresponding significance level.

Chi-square tests are also used to test theories asserting that some population has a particular, continuous probability distribution, such as the normal distribution. To test such a theory, the range of possible results of some sampling trial would be divided into several intervals and the numbers of subjects falling into each would be compared with the 'expected' number, proceeding then as with the example of the die.

Although the test has been much further developed and with great technical ingenuity, it is, we believe, vitiated by the absence of any principled rule for partitioning the outcomes into separate intervals or cells, for not all partitions lead to the same inferences when the significance test is applied. For instance, in our die-rolling example, if we had based χ^2 on just three cells, formed, say, by combining the pairs of outcomes [*six*, *five*], [*four*, *three*], and [*two*, *one*], the result would not now be significant at the 5 percent level.

This problem is rarely taken up in expositions of the chi-square test. When it is, it is resolved by considerations of convenience, not epistemology. For instance, Kendall and Stuart (1979, p. 457) argued that the class boundaries should be drawn so that each cell has the same probability (relative to the null hypothesis) of containing the experimental outcome, and they defended this rule on the epistemically irrelevant grounds that it is "perfectly definite and unique". But it is not even true that the equal-probability rule leads to a unique result, as we can see from our last example. We there considered partitioning the outcomes of the die-rolling experiment into three pairs: there are in fact fifteen distinct ways of doing this, all satisfying the equal-probability rule, and only two of them render the results significant at the 5 percent level.

The complacency of statisticians in the face of this difficulty is remarkable. Although Hays and Winkler (1970, p. 195) warn readers repeatedly and emphatically that "*the arrangement into population class intervals is arbitrary*", their exposition proceeds without recognizing that this renders the conclusions

of a chi-square test *equally* arbitrary. Cochran (1952, p. 335) claimed that the problem is a "minor" one, which merely calls for "more standardization in the application of the test". But standardization would only institute the universal application of arbitrary principles and would not address the central problem of the chi-square test, which is how to set it on a firm epistemic basis. No such basis appears to exist, and in view of this, the test should be abandoned.

It might be argued that, despite its epistemic difficulties, there are strong indications that the chi-square test is sound, because the conclusions that are in practice drawn from it generally fit so well with intuition. In many cases, intuition is indeed satisfied but the test can also produce quite counter-intuitive inferences, and once one such inference has been seen, it is easy to generate more. Suppose, for instance, that the above trial with the die had given the results: *six* (123), *five* (100), *four* (100), *three* (100), *two* (100), *one* (77). In a test of the hypothesis that the die was true, χ^2 takes the value 10.58, which is not significant at the 5 percent level, and any statistician who adopted this as the critical value would not be obliged to reject the null hypothesis, even though the results of the trial tell us pretty clearly that it is quite wrong.[3]

In the examples we have so far considered, only large values of χ^2 were taken as grounds for rejecting a hypothesis. But for Fisher both extremities of any test-statistic's distribution were critical. In his view, a null hypothesis is "as definitely disproved" when the observed and the expected frequencies are very similar, leading to a very small χ^2, as it is when the frequencies are sharply discrepant and χ^2 is large (Fisher 1970, Section 20). This forms the basis of Fisher's famous criticism of Mendel's experimental results, which we discussed above in 4.e. Those results, he said, were "too good to be true", that is to say, although they seemed to be in close accord with Mendelian theory, and were usually taken to be so, they corresponded to χ^2 values that were sufficiently small to imply its rejection in a significance test. For Fisher the chi-square test had to override intuitions in this case. But this is not the universal opinion amongst classical statisti-

[3] Good 1981, p.161, makes this point.

cians. For example, Stuart (1954) maintained that a small χ^2 is critical only if all "irregular" alternatives to the null hypothesis have been ruled out, where the irregularity might involve "variations due to the observer himself", such as "all voluntary and involuntary forms of falsification". Indeed, the Fisherian idea that a null hypothesis can be tested in isolation, without considering rival hypotheses, is not now widely shared and the predominant form of the significance test, that of Neyman and Pearson, which we discuss shortly, requires hypotheses to be tested against, or in the context of, alternative hypotheses.

Sufficient Statistics

It is sometimes claimed that consistency may be satisfactorily restored to Fisher's significance tests by restricting test-statistics to so-called *minimal-sufficient statistics*, because of their standard interpretation as containing all the information that is relevant to the null hypothesis and none that is irrelevant. We shall argue, however, that this interpretation is unavailable to Fisher, that there are no grounds for excluding irrelevant information from a test, and that the difficulty confronting Fisherian principles is unconnected with the amount of information in the test-statistic, but lies elsewhere.

Let us first examine the concept of a sufficient statistic. Some statistics clearly abstract more information from the outcomes than others. For instance, tossing a coin four times will result in one of the sixteen sequences of heads and tails *(HHHH), (THHH), . . . , (TTTT)*, and a statistic that assigns distinct numbers to each element of this outcome space preserves all the information produced by the experiment. But a statistic that records only the number of heads thereby discards information, so if you knew only that it took the value 3, say, you could not determine from which of the four different outcomes containing 3 heads it was derived. Whether some of the discarded information is relevant to an inference is a question addressed by the theory of sufficiency.

A sample statistic, t, is said to be *sufficient*, relative to a parameter of interest, θ, if the probability of any particular member of the outcome space, given t, is independent of θ. In our

example, the statistic representing the number of heads in the outcome is in fact sufficient for θ, the physical probability of the coin to land heads, as can be simply shown. The outcome space of the coin-tossing experiment consists of sequences $x = x_1, \ldots, x_n$, where each x_i denotes the outcome either *heads* or *tails*, and $P(x \mid t)$ is given as follows, remembering that, since the value of t is logically implied by x, $P(x \,\&\, t) = P(x)$:

$$P(x \mid t) = \frac{P(x \,\&\, t)}{P(t)} = \frac{P(x)}{P(t)} = \frac{\theta^r (1-\theta)^{n-r}}{{}^nC_r \theta^r (1-\theta)^{n-r}} = \frac{1}{{}^nC_r}$$

Since the binomial term, ${}^nC_r$, is independent of θ, so is $P(x \mid t)$; hence, t is sufficient for θ.

It seems natural to say that if $P(x \mid t)$ is the same whatever the parameter value, then x "can give us no information about θ that the sufficient statistic has not already given us" (Mood and Graybill 1963, p.168). Certainly Fisher (1922, p. 316) understood sufficiency that way: "The Criterion of Sufficiency", he wrote, is the rule that "the statistic chosen should summarize the whole of the relevant information supplied by the sample". But natural as it seems, this interpretation is unavailable to Fisher, for a hypothesis subjected to one of his significance tests may be rejected by one sufficient statistic and not by another. Our coin-tossing example illustrates this, for the statistic that summarizes the outcome as the number of heads in the sample, and the statistic that assigns separate numbers to each member of the outcome space are both sufficient, as is the artificial statistic r', described above, though these statistics do not generally yield the same conclusion when used in a Fisherian test of significance.

Since the sufficiency condition does not ensure a unique conclusion, the further restriction is sometimes argued for (for example by Seidenfeld 1979, p. 83) that the test-statistic should be *minimal-sufficient*; that is, it should be such that any further reduction in its content would destroy its sufficiency. A minimal-sufficient statistic is thought of as containing all the information supplied by the sample that is relevant, and none that is irrelevant. But this second restriction has received no adequate defence; indeed, it would be surprising if a case could be made for it, for if information is irrelevant, it should make no difference to a test, so there should be no need to exclude it. It is curious that, despite the

almost universal lip service paid to the sufficiency condition, the principal statistics that are in practice used in significance tests—the χ^2, t and F statistics—are none of them sufficient, let alone minimal-sufficient (Pratt 1965, pp. 169–170).

The idea of restricting admissible statistics according to their information content seems in any case misconceived as a way of saving Fisherian significance tests. For Neyman (1952, pp. 45–46) has shown that where the null hypothesis describes a continuous probability density distribution over the space, there may be pairs of statistics that are related by a 1-1 transformation, such that only one of them leads to the rejection (at a specified significance level) of the null hypothesis. Since these statistics necessarily carry the same information, there must be some other source of the trouble.

5.c | Neyman-Pearson Significance Tests

Fisher's significance tests were designed to provide for the statistical case something akin to the falsification available in the deterministic case; hence his insistence that the tests should operate on isolated hypotheses. But as we indicated earlier, statistical hypotheses cannot be refuted and, as we show later (Section 5.d), Fisher's own analysis of and arguments for a quasi-refutation are quite unsatisfactory. For this reason, Neyman felt that a different epistemic basis was required for statistical tests, in particular, one that introduces rival hypotheses into the testing process. The version of significance tests that he and Pearson developed resembled Fisher's however, in according no role to prior or posterior probabilities of theories, for they were similarly opposed to Bayesian methodology.

In setting out the Neyman-Pearson method, we shall first consider the simplest cases, where only two hypotheses, h_1 and h_2, are in competition. Neyman-Pearson tests permit two kinds of inference: either a hypothesis is rejected or it is accepted. And such inferences are subject to two sorts of error: you could regard h_1 as false when in fact it is true, or accept h_1 (and, hence, reject h_2) when it is false. When these errors can be distinguished by their gravity, the more serious is called a *type I* error

and the less serious a *type II* error. The seriousness of the two types of error is judged by the practical consequences of acting on the assumption that the rejected hypothesis is false and the accepted one true. For example, suppose two alternative hypotheses concerning a food additive were admitted, one that the substance is safe, the other that it is highly toxic. Under a variety of circumstances, it would be less dangerous to assume that a safe additive was toxic than that a toxic one was safe. Neyman and Pearson, adapting Fisher's terminology, called the hypothesis whose mistaken rejection is the more serious error the *null hypothesis*, and where the errors seem equally serious, either hypothesis may be so designated.

The possibilities for error are summed up in Table 5.3.

TABLE 5.3

Decision	*True Hypothesis*	
	h_1	h_2
Reject h_1	Error	/
Accept h_1	/	Error

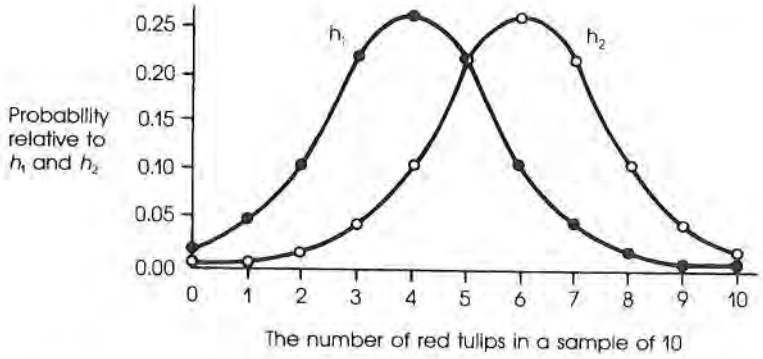
The Neyman-Peason approach aims to minimize the chance of committing both types of error. We will examine the Neyman-Pearson approach through an example borrowed from Kyburg (1974, pp. 26–35). The label on a particular consignment of tulip bulbs has been lost and the purchaser cannot remember whether it was the one that contained 40 percent of the red- and 60 percent of the yellow-flowering sort, or 40 percent of the yellow and 60 percent of the red. We shall designate these possibilities h_1 and h_2, respectively, and treat the former as the null hypothesis. An experiment to test these hypotheses might involve planting a predetermined number of bulbs, say 10, that have been randomly selected from the consignment, and observing which grow red and which yellow. The testing procedure is similar to Fisher's and involves the following steps.

First, specify the outcome space, which in the present case may be considered to comprise 2^{10} sequences, each sequence indicating the flower-colour of the tulip bulb that might be selected first, second, and so on, down to the tenth. Next, a test-statistic that summarizes the outcome in numerical form needs to be stipulated, and in this example we shall take the number of reds appearing in the sample for this purpose. (We discuss the basis for these arbitrary-seeming stipulations below.) Thirdly, we must compute the probabilities of each possible value of the test-statistic, relative to each of the two rival hypotheses. If, as we shall assume, the consignment of tulip bulbs is large, the probability of selecting r red-flowering bulbs in a random sample of n is approximated by the familiar binomial function ${}^nC_r p^r q^{n-r}$. We are assuming here that the probability, p, of selecting a red-flowering bulb is constant, an assumption that is more approximately true, the larger the population of bulbs. In the present case, h_1 corresponds to $p = 0.40$, and h_2 to $p = 0.60$. The sampling distributions for the imagined trial, relative to the two hypotheses, are given in Table 5.4 and displayed graphically, below.

TABLE 5.4

The Probabilities of Selecting *r* Red- and (10 – *r*) Yellow-flowering Tulips

Outcome (Red, Yellow)	h_1 *(p = 0.40)*	h_2 *(p = 0.60)*
0, 10	0.0060	0.0001
1, 9	0.0403	0.0016
2, 8	0.1209	0.0106
3, 7	0.2150	0.0425
4, 6	0.2508	0.1115
5, 5	0.2006	0.2006
6,4	0.1115	0.2508
7, 3	0.0425	0.2150
8, 2	0.0106	0.1209
9, 1	0.0016	0.0403
10, 0	0.0001	0.0060



Finally, the Neyman-Pearson method calls for a rule that will determine when to reject the null hypothesis. Consider the possibility of rejecting the hypothesis just in case 6 or more red-flowering plants appear in the sample. Then, if h_1 is true, the probability of a rejection may be seen from the table to be: 0.1115 + 0.0425 + 0.0106 + 0.0016 + 0.0001 = 0.1663, and this is therefore the probability of a type I error associated with the postulated rejection rule. This probability is called, as before, the significance level of the test, or its *size*. The probability of a type II error is that of accepting h_1 when it is false; on our assumption that one of the two hypotheses is true, this is identical to the probability of rejecting h_2 when it is true, which may be calculated from the table as 0.3664.

The *power* of a test is defined as 1 – *P*(type II error) and is regarded by advocates of this approach as a measure of how far the test 'discriminates' between the two hypotheses. It is also the probability of rejecting the null hypothesis when it is false, and in the present case has the value 0.6336.

In selecting the size and power of any test, a natural ideal might seem to be to try to minimize the former and maximize the latter, in order to reduce as far as possible the chances of both types of error. We shall, in due course, consider whether an ideal couched in terms of type I and type II errors is suited to inductive reasoning. We have to note straightaway, though, that the ideal is incoherent as it stands, for its twin aims are incompatible: in most cases, a diminution in size brings with it a contraction in power, and vice versa. Thus, in our example, if the rejection rule were

changed and h_1 rejected in the event of at least 7 red tulips in the sample, the size of the test would be reduced from 0.1663 to 0.0548, but its power would also be lower – 0.3823, compared with 0.6336. We see then that while the revised test has a smaller size, this advantage (as it is judged) is offset by its smaller power. For this reason, Neyman and Pearson proposed instead that one first fix the size of a test at an appropriate level and, thus constrained, then maximize its power.

Randomized Tests

It is generally held amongst classical statisticians that the size of a significance test should not exceed 0.05 and, for a reason we shall describe later, practitioners are often exhorted always to employ roughly the same significance levels. But with the methods introduced so far the size of a test cannot always be chosen at will. For this purpose, randomized tests have been devised, which Kyburg has lucidly explained in the context of the example we have been discussing. Suppose a test of size 0.10 were desired. Let the two tests considered above be labelled 1 and 2. As we showed, they have the following characteristics:

TABLE 5.5

	Probability of a type I error	*Probability of a type II error*	*Power*
Test 1	0.1663	0.3664	0.6336
Test 2	0.0548	0.6177	0.3823

Imagine, now, a third test which is carried out in the following manner: a pack of 200 cards, of which 119 are red and 81 black, is well shuffled, and one of these cards is then randomly selected. If the selected card is black, test 1 is applied, and if red, test 2. This *mixed* or *randomized test* has the required size of 0.10, given by

$$\frac{81}{200} \times 0.1663 + \frac{119}{200} \times 0.0548 = 0.100.$$

The corresponding probability of a type II error is similarly calculated to be 0.5159; so the power of the mixed test is 0.4841.

Readers might be surprised by the implication that inspecting a piece of coloured card, whose causal connexion to the tulip consignment is nil, can nevertheless provide an insight into its composition. Randomized tests are rarely if ever used, but they form a proper part of the Neyman-Pearson theory, so any criticism that they merit can quite correctly be re-directed to the Neyman-Pearson theory in general.

The Choice of Critical Region

An advantage that Neyman-Pearson significance tests enjoy over Fisher's is that they incorporate in a quite natural way a feature that Fisher seems to have adopted arbitrarily and in deference merely to apparent scientific practice, namely, to concentrate the critical region in (one or both of) the tails of the sampling distribution of outcomes. Fisher's reasoning seems to have been that evidence capable of rejecting a hypothesis must be very improbable and should lie in a region of very low probability (see Section 5.d). But Neyman pointed out that by this reasoning, Fisher could equally well have chosen for the rejection region a narrow band in the centre of a bell-shaped distribution as a broader band in its tails.

By contrast, in the Neyman-Pearson approach, the critical region is uniquely determined, according to a theorem known as the *Fundamental Lemma*. This states that the critical region of maximum power in a test of a null hypothesis, h_1, against a rival, h_2, is the set of points in the outcome space that satisfies the inequality:

$$\frac{P(x \mid h_1)}{P(x \mid h_2)} \leq k,$$

where k is a constant that depends on the hypotheses and on the significance level.[4] The probabilities may also be densities.

[4] Strictly speaking, the likelihoods $P(x \mid h_1)$, $P(x \mid h_2)$ should not be expressed here as conditional probabilities, for these presuppose that the hypotheses themselves have a probability, something that classical statisticians strenuously deny. Hence, they are sometimes written $P(x\ ;\ h)$ or $L(x \mid h)$. Bayesians, of course, need have no such qualms.

The lemma embraces randomized tests, the critical region then comprising those of the component non-randomized tests, which are selected at random, as already described.

Neyman-Pearson Tests and Sufficient Statistics

Neyman-Pearson tests have another fortunate consequence, namely, that for them sufficient statistics do contain all the relevant information. For if h_1 and h_2 ascribe different values to a parameter θ, and if t is a sufficient statistic relative to the outcomes $x = x_1, \ldots, x_n$, then, by definition, $P(x \mid t)$ is independent of θ, and it follows almost directly that

$$\frac{P(x \mid h_1)}{P(x \mid h_2)} = \frac{P(t \mid h_1)}{P(t \mid h_2)}.$$

The above lemma tells us that the left-hand ratio does not exceed some number k; hence, the same holds also for the right-hand ratio. So if the outcome were summarized in terms of t, rather than x, the region of maximum power would comprise the same outcomes, and consequently, none of the information in x that is omitted from t is relevant to the significance test inference.

5.d | Significance and Inductive Significance

'The null hypothesis was rejected at such-and-such a significance level' is a technical expression that simply records that an experimental result fell in a certain designated 'rejection region' of the outcome space. But what does it mean as an inductive conclusion about the hypothesis? There are three principal views on this amongst advocates of the significance test. None, we shall argue, is in the least satisfactory.

Fisher's View

Fisher took the process of logical refutation as the model for his significance tests. This is apparent in his frequently voiced claim that such tests could "disprove" a theory (for example, 1947, p. 16), and that "when used accurately, [they] are capable of rejecting or invalidating hypotheses, in so far as these are *contradicted by the data*" (1935; our italics).

Fisher seems to be saying here that statistical theories may actually be falsified, though, of course, he knew full well that this was impossible, and in his more careful accounts he took a different line.[5] The force of a test of significance, he said (1956, p. 39), "is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [i.e., the null hypothesis] is not true". But in thus avoiding an unreasonably strong interpretation, Fisher fell back on one that is unhelpfully weak, for the significant or critical results in a test of significance are by definition improbable, relative to the null hypothesis. Inevitably, therefore, a significant result is either a "rare chance" (an improbable event) or the null hypothesis is false, or both. And Fisher's claim amounts to no more than this empty truism.[6]

Significant Results and Decisions to Act

Neyman and Pearson proposed what is now the most widely adopted view, namely, that on 'accepting' a hypothesis after a significance test, one should act as if one believed it to be true, and if one 'rejects' it, one's actions should be guided by the assumption that it is false, "without", as Lindgren (1976, p. 306) put it, "necessarily being convinced one way or the other". Neyman and Pearson (1933, p. 142) defended their rule by saying that although

[5] The careless way that Fisher sometimes described the inductive meaning of a significant result is often encountered in statistics texts. For example, Bland (1987, p. 158) concludes from one such result "that the data are not consistent with the null hypothesis"; he then wanders to the further conclusion that the alternative hypothesis is "more likely".

[6] Hacking 1965, p.81, pointed this out.

a significance test "tells us *nothing* as to whether in a particular case h is true"[7], nevertheless

> it may often be proved that if we behave according to . . . [the rule], then in the long run we shall reject h when it is true not more, say, than once in a hundred times [when the significance level is 0.01], and in addition we may have evidence that we shall reject h sufficiently often when it is false [i.e., when the test's power is sufficiently large].

This is a surprising argument to encounter in this context. After all, the significance test idea was born out of the recognition that events with probability p cannot be proved to occur with any particular frequency, let alone with frequency p; indeed, they may never occur at all. This is acknowledged tacitly in the above argument, through the proviso "in the long run", a prevarication that suggests some largish but practically accessible number, yet at the same time also hints at the indefinite and infinite. The former suggestion is, as we have said, unsustainable; the latter would turn the argument into the unhelpful truism that with a significance level of 0.01, we would reject a true null hypothesis with probability 0.01. Either way, the argument does not uphold the Neyman-Pearson rejection rule.

There are also objections to the idea of acting as if a hypothesis were definitely true or definitely false when one is not convinced one way or the other. If, to go back to our earlier example, one were to reject the hypothesis that the tulip consignment contained 40 percent of the red variety and then act as if it were definitely false, there would be no incentive to repeat the experiment and every incentive to stake all one's worldly goods, and whatever other goods one might possess, on a wager offered at any odds on the hypothesis being true. The idea is clearly absurd.

Neyman (1941, p. 380), on the other hand, argued that there are in fact occasions when it is reasonable to behave as if what one believed to be false were actually true, and *vice versa*, citing the

[7] These are our italics. Neyman's view that no inductive inferences are licensed by sampling information is discussed in Section 5.f.2 below.

purchase of holiday insurance as a case in point. In making such a purchase, he said, "we surely act against our firm belief that there will be no accident; otherwise, we would probably stay at home". This, however, seems a perverse analysis of the typical decision to take out insurance. We surely do not firmly believe that there will be no accident when we go away, but regard the eventuality as more or less unlikely, depending on the nature of the holiday, its location, and so forth; and the degree of risk perceived is reflected in, for example, the sum we are prepared to lay out on the insurance premium.

Another example that is sometimes used to defend the idea of acting as if some uncertain hypothesis were true is industrial quality control.[8]

The argument is this. Suppose an industrialist would lose money by marketing a product-run that included more than a certain percentage of defective items. And suppose product-runs were successively sampled, with a view to testing whether they were of the loss-making type. In such cases, there could be no graduated response, it is claimed, since the product-run can either be marketed or not; but, the argument goes, the industrialist could be comforted by the thought that "in the long run" of repeatedly applying the same significance test and the same decision rule, only about, say, 5 percent of the batches marketed will be defective, and that may be a financially sustainable failure rate.

But this argument does not succeed, for the fact that only two actions are possible does not imply that only two beliefs can be entertained about the success of those actions. The industrialist might attach probabilities to the various hypotheses and then decide whether or not to market the batch by balancing those probabilities against the utilities of the possible consequences of the actions, in the manner described by a branch of learning known as Decision Theory. Indeed, this is surely the more plausible account.

[8] Even some vigorous opponents of the Neyman-Pearson method, such as, A.W.F. Edwards (1972, p. 176) accept this defence.

Significance Levels and Inductive Support

The fact that theories are not generally assessed in the black-and-white terms of acceptance and rejection is acknowledged by many classical statisticians, as we see from attempts that have been made to find in significance tests some graduated measure of evidential support. For example, Cramér (1946, p. 421–23) wrote of results being "almost significant", "significant" and "highly significant", depending on the value of the test-statistic. Although he cautiously added that such terminology is "purely conventional", it is clear that he intended to suggest an inverse relationship between the strength of evidence against a null hypothesis and the significance level that would just lead to its rejection.[9]

Indeed, he implies that when this significance level exceeds some (unspecified) value, the evidence ceases to have a negative impact on the null hypothesis and starts to support it; thus, when the χ^2 value arising from some of Mendel's experiments on pea plants was a good way from rejecting Mendel's theory at the 5 percent level, Cramér concluded that "the agreement must be regarded as good", and, in another example, when the hypothesis would only be rejected if the significance level were as high as 0.9, Cramér said that "the agreement is very good".

Classical statisticians commonly try to superimpose this sort of notion of strength of evidence or inductive support on their analyses. For instance, Weinberg and Goldberg (1990, p. 291): "The test result was significant, indicating that H_1 . . . was *a more plausible* statement about the true value of the population mean . . . than [the null hypothesis] H_0". The words we have italicised would, of course, be expected in a Bayesian analysis, but they have no legitimacy or meaning within classical philosophy. And the gloss which the authors then add is not any clearer or better founded: "all we have shown is that *there is reason to believe* that [H_1 is true]". And, of the same result, which was very improbable according to the null hypothesis and significant at the 0.0070

[9] The significance level that, for a given result, would just lead to the rejection of a null hypothesis is also called the p-value of that result, as we stated earlier. "The lower the p-value, the less plausible this null hypothesis . . . and the more plausible are the alternatives" (Wood 2003, p. 134).

level, they say that it is "*quite inconsistent* with the null hypothesis" (*ibid.*, p. 282).

But the result is not "inconsistent" with the null hypothesis, in the logical sense of the term. And as no useful alternative sense seems to exist—certainly none has been suggested—the term in this context is quite misleading. And the project of linking significance levels with strength of evidence has no prospect of success. To prove such a link, you would need to start with an appropriate concept of evidential or inductive support; in fact, no such concept has been formulated in significance test terms, nor is one likely to be. This, for two compelling reasons. First, the conclusions of significance tests often flatly contradict those that an impartial scientist or ordinary observer would draw. Secondly, significance tests depend on factors that it is reasonable to regard as extraneous to judgments of evidential support. We deal with these objections in the next three subsections.

A Well-Supported Hypothesis Rejected in a Significance Test

The first objection was developed in considerable generality by Lindley, 1957, and is sometimes referred to as Lindley's Paradox. We illustrate it with our tulip example. Table 5.6 lists the numbers of red tulips in random samples of size n that would just be sufficient to reject the null hypothesis at the 0.05 level.

It will be noticed that as n increases, the critical proportion of red tulips in the sample that would reject h_1 at the 0.05 level approaches more closely to 40 percent, that is, to the proportion hypothesized in h_1. Bearing in mind that the only alternative to h_1 that the example allows is that the consignment contains red tulips in the proportion of 60 percent, an unprejudiced consideration would clearly lead to the conclusion that as n increases, the supposedly critical values *support* h_1 more and more.

The table also includes information about the power of each test, and shows that the classical thesis that a null hypothesis may be rejected with greater confidence, the greater the power of the test is not borne out; indeed, the reverse trend is signalled.

Freeman (1993, pp. 1446–48) is one of the few to have proposed a way out of these difficulties, without abandoning the

TABLE 5.6

The sample size, n	*The number of red tulips (expressed as a proportion of* n*) that would just reject* h_1 *at the 5% level.*	*The power of the test against* h_2
10	0.70	0.37
20	0.60	0.50
50	0.50	0.93
100	0.480	0.99
1,000	0.426	1.0
10,000	0.4080	1.0
100,000	0.4026	1.0

basic idea of the significance test. He argued that Neyman and Pearson should not have formulated their tests as they did, by first fixing a significance level and then selecting the rejection region that maximizes power. It is this that renders them vulnerable to the Lindley Paradox, because it means that the inductive import of a rejection at a given significance level is the same whatever the size of the sample. Instead, Freeman proposes that the primary role should go to the *likelihood ratio*—that is, the ratio of the probabilities of the data relative to the null and an alternative hypothesis. And he argued that in a significance test, the rule should be to reject the null hypothesis if the likelihood ratio is less than some fixed value, on the grounds that this ensures that the probabilities of *both* the type I *and* the type II errors diminish as the sample size increases.

Freeman's rule is a version of the so-called *Likelihood Principle*, according to which the inductive force of evidence is contained entirely in the likelihood ratios of the hypotheses under consideration. This principle, in fact, follows directly from Bayes's theorem (see Section 4.c) and is unavoidable in Bayesian inductive inference. Freeman (1993, p. 1444) too regards this principle as essential—"the one secure foundation for all of statistics"—but

neither he nor any other non-Bayesian has proved it. And this is not surprising, for they strenuously deny that hypotheses have probabilities, and it is precisely upon this idea that the Bayesian proof depends. The likelihood principle therefore cannot save significance tests from the impact of Lindley's Paradox, which, it seems to us, shows unanswerably and decisively that inferences drawn from significance tests have no inductive significance whatever.

We now consider a couple more aspects of significance tests which reinforce this same point.

The Choice of Null Hypothesis

In a Neyman-Pearson test you need to choose which of the competing hypotheses to treat as the null hypothesis, and the result of that choice has a bearing on which is finally accepted and which rejected. Take the tulip example again: if an experiment showed 50 red-flowering plants in a random sample of 100, then h_1 (40 percent red) would be rejected at the 0.05 level if it were the null hypothesis, and h_2 (60 percent red) would be accepted. But with h_2 as null hypothesis, the opposite judgment would be delivered! It will be recalled that the role of null hypothesis was filled by considering the desirability, according to a personal scale of values, of certain practical consequences of rejecting a true hypothesis; and where the hypotheses were indistinguishable by this practical yardstick, the null hypothesis could be designated arbitrarily. But pragmatic and arbitrary decisions such as these have no epistemic meaning and cannot form the basis of inductive support.

Another sort of influence on significance tests that is also at odds with their putative role in inductive reasoning arises through the stopping rule.

The Stopping Rule

Significance tests are performed by comparing the probability of the outcome obtained with the probabilities of other possible outcomes, in the ways we have described. Now the space of possible outcomes is created, in part, by what is called the *stopping rule*; this is the rule that fixes in advance the circumstances under

which the experiment should stop. Our trial to test the fair-coin hypothesis, for example, was designed to stop after the coin had been flipped 20 times. Another stopping rule for that experiment might have instructed the experimenter to end it as soon as 6 heads appeared, which would exclude many of the outcomes that were previously possible and introduce an infinity of new ones. Expressed as the number of heads and tails in the outcome, the possibilities for the two stopping rules are: (*20,0*), (*19,1*), . . . , (*0,20*), in the first case, and (*6,0*), (*6,1*), (*6,2*), . . . , and so on, in the second. The two stopping rules have surprisingly and profoundly different effects.

Consider, for example, the result (6,14), which could have arisen with either stopping rule. When the rule was to stop after 6 heads, the null hypothesis would be rejected at the 0.05 level. This is shown as follows: the assumed stopping rule produces the result (*6, i*) whenever (*5, i*), appearing in any order, is then succeeded by a head. Thus, relative to the fair-coin hypothesis, the probability of the result (*6, i*) is given by ${}^{i+5}C_5(\frac{1}{2})^5(\frac{1}{2})^i \times \frac{1}{2}$. Table 5.7 shows the sampling distribution.

TABLE 5.7

The Probabilities of Obtaining *i* Tails with a Fair Coin in a Trial of Designed to Stop after 6 Heads Appear.

Outcome (H,T)	*Probability*	*Outcome (H,T)*	*Probability*
6,0	0.0156	6,11	0.0333
6,1	0.0469	6,12	0.0236
6,2	0.0820	6,13	0.0163
6,3	0.1094	6,14	0.0111
6,4	0.1230	6,15	0.0074
6,5	0.1230	6,16	0.0048
6,6	0.1128	6,17	0.0031
6,7	0.0967	6,18	0.0020
6,8	0.0786	6,19	0.0013
6,9	0.0611	16,20	0.0008
6,10	0.0458	6,21	0.0005
		etc.	etc.

We see from the table that the results which are at least as improbable as the actual one are (*6,14*), (*6,15*), . . . , and so on, whose combined probability is 0.0319. Since this is below the critical value of 0.05, the result (*6,14*) is significant at this level and the null hypothesis should therefore be rejected. It will be recalled that when the stopping rule predetermined a sample size of 20, the very same result was not significant.[10] So in calculating the significance of the outcome of any trial, it is necessary to know the stopping rule that informed it.

We have considered just two stopping rules that could have produced some particular result, but any number of others have that same property. And not all of these other possibilities rest the decision to stop on the outcomes themselves, which some statisticians regard as not quite legitimate. For instance, suppose that after each toss of the coin, you drew a playing card at random from an ordinary pack, with the idea of calling the trial off as soon as the Queen of Spades has been drawn. This stopping rule introduces a new outcome space, which will lead to different conclusions in certain cases. Or suppose the experimenter intends to continue the trial until lunch is ready: in this case, the sampling distribution could only be worked out with complex additional information about the chance, at each stage of the trial, that preparations for the meal are complete.

The following example brings out clearly how inappropriate it is to involve the stopping rule in the inductive process: two scientists collaborate in a trial, but are privately intent on different stopping rules; by chance, no conflict arises, as the result satisfies both. What then are the outcome space and the sampling distribution for the trial? To know these you would need to discover how each of the scientists would have reacted in the event of a disagreement. Would they have conceded or insisted, and if they had put up a fight, which of them would have prevailed? *We suggest that such information about experimenters' subjective intentions, their physical strengths and their personal qualities has no inductive relevance whatever in this context, and that in practice it is never sought or even contemplated. The fact that significance*

[10] This illustration of the stopping-rule effect is adapted from Lindley and Phillips 1976.

tests and, indeed, all classical inference models require it is a decisive objection to the whole approach.

Whitehead (1993) is one of the few to have defended the stopping rule as an essential component of the inductive process. He denies that the subjective intention underlying the stopping rule is irrelevant, illustrating his point with a football match, of all things, in which the captain of one side is allowed to decide when the game should finish, and in fact blows the whistle when his team is 1–0 ahead. Whitehead remarks that learning the stopping rule here would reduce his high opinion of the winning side. To revert to a case where classical statistics can more obviously be applied, this is analogous to an experimenter, who is predisposed in favour of one of the hypotheses, deciding to stop sampling as soon as more red than yellow tulips have flowered. If the final count were, say, 1 red and 0 yellow, we would indeed not be much swayed in favour of the experimenter's preferred hypothesis, but not because of the known bias, or the stopping rule, rather, we suggest, because of the smallness of the sample. To believe otherwise runs into the objection we raised earlier, namely, that if the biased experimenter were working with an impartial, or differently biased colleague, who was actuated by a different stopping rule, you would have to delve into the personal qualities of the experimenters in order to discover the outcome space of the experiment, and hence the inductive significance of the result.

Experimenters' prejudices can only have inductive significance for us if we believe them to have clairvoyant knowledge about future samples; but this is just what a random sampling experiment effectively precludes. On the other hand, the captain in charge of the stopping rule in the hypothetical football match does have information about the likely course of the game, since he may know the teams' recent form and can observe how well each side is presently playing. But a football game is not a random sampling experiment, and is therefore an unsuitable example in this context.

Gillies (1990, p. 94) also argued that the stopping rule is an essential part of a scientific inference. He claimed that "to those who adopt falsificationism (or a testing methodology)" it "seems natural and only to be expected" that the stopping rule should in general affect a theory's empirical support, because "wherever

possible the experimental method should be applied, and this consists in designing and carrying out a *repeatable* experiment . . . whose result might refute h [the null hypothesis]". This, he claims, means that the stopping rule is evidentially relevant.

In response, we certainly concede that it can do no harm and might do good to repeat an experiment. But why should it be repeat*able*? Many useful and informative tests cannot be repeated: for example, pre-election opinion polls and certain astronomical observations. Would our confidence in the age of the Turin Shroud be any different if the entire cloth had been consumed in the testing process, so precluding further tests? Surely not.

Moreover, in an important sense, no experiment is repeatable, for none could ever be done in exactly the same way again. Indefinitely many factors alter between one performance of an experiment and another. Of course, not all such changes matter. For instance, the person who tossed the coin might have worn yellow shoes or sported a middle parting; but these are irrelevant, and if you called for the experiment to be repeated, you would issue no instructions as to footwear or hairstyle. On the other hand, whether or not the coin had a piece of chewing gum attached to one side, or a strong breeze was blowing when it was tossed should be taken into account. The question then is whether the stopping rule falls into the first category of irrelevant factors or into the second of relevant ones. Gillies (*ibid.*, p. 94) simply presumes the latter, arguing, with reference to the coin trial, that the "test of h in this case consists of the whole carefully designed experimental procedure", and suggesting thereby that this procedure must include reference to the stopping rule. But Gillies neither states this explicitly nor provides any reason why it should be so—unavoidably, in our view.

We show in Chapter 8 that in the Bayesian scheme the posterior probabilities in each case are unaffected by the subjective intentions implicit in the stopping rules and depend on the result alone. Thus, if the experimental result is, for instance, *6 heads, 14 tails*, it does not matter whether the experimenter had intended to stop the trial after 20 tosses of the coin, or after *6 heads*, or after lunch, or after the Queen of Spades has made her entrance, or whatever.

5.e | Testing Composite Hypotheses

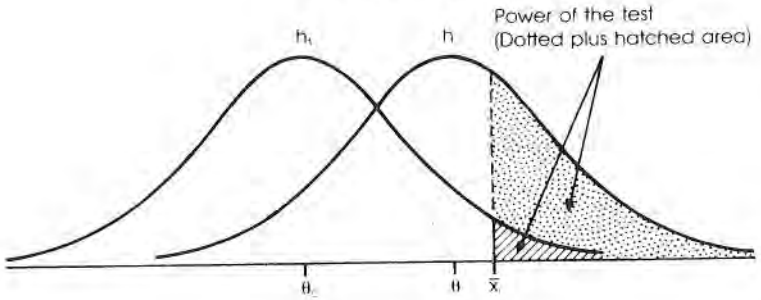
We have, so far, restricted our account of the Neyman-Pearson method to cases where just two, specific hypotheses are assumed to exhaust the possibilities. But such cases are atypical in practice, and we need to look at how Neyman and Pearson extended and modified their approach to deal with a wider range of alternative hypotheses. They considered, for instance, how to test a hypothesis, h_1, that some population parameter, θ, has a specific value, say θ_1, against the unspecific, composite hypothesis, h_2, that $\theta > \theta_1$.

The principle of maximizing the power of a test for a given significance level cannot be applied where these are the competing hypotheses. For, although the situation allows one to determine a critical region corresponding to any designated significance level, as before, the probability of a type II error is indeterminate. Neyman and Pearson responded by varying their central principle.

They first of all proposed that in such cases one should choose for the critical region one that has maximum power for each component of h_2. Tests satisfying this new criterion are called *Uniformly Most Powerful* (UMP). But they ran into the problem that, relative to some elements of h_2, the power of UMP tests might be very low, indeed, lower than the significance level, in which case there would be a greater chance of rejecting the null hypothesis when it is true than when it is false. This possibility was unacceptable to Neyman and Pearson and to avoid it, they imposed the further restriction that the test should be *unbiased*, that is, its power relative to each element of h_2 should be at least as great as its significance level.

We can illustrate the idea of tests that are both UMP and unbiased (UMPU) with a simple example. The test will be based on the mean, $\bar{x}$, of a random sample drawn from a normal population with known standard deviation and unknown mean, θ. The diagram shows the sampling distributions relative to the null hypothesis h_1: $\theta = \theta_1$, and relative to an arbitrary element, h_i, of the composite alternative hypothesis h_2: $\theta > \theta_1$.

AN UMPU TEST



Consider a critical region for rejecting h_1 consisting of points to the right of the critical value, $\bar{x}_c$. The area of the hatched portion is proportional to the significance level. The area under the h_1-curve to the right of $\bar{x}_c$ represents the probability of rejecting h_1 if h_i is true (and hence, h_1 false). Clearly, the closer are the means specified by h_1 and h_i, the smaller this probability will be. But it could never be less than the significance level. In other words, the test is UMPU.

Suppose now that the range of alternatives to the null hypothesis is greater still and that h_1: $\theta = \theta_1$ is to be tested against h_2: $\theta \neq \theta_1$, the standard deviation again being known. A critical region located in one tail of the sampling distribution associated with h_1 would not now constitute an unbiased test. But an UMPU test can be constructed by dividing the critical region equally between the two tails. This can be appreciated diagrammatically and also rigorously shown (see for instance Lehmann 1986). An UMPU test thus provides some basis for the two-tailed tests implied by Fisher's theory, for which he offered no rationale.

But UMPU tests are in fact rather academic, since they exist in very few situations. And they depart somewhat from the Neyman-Pearson ideal of maximum power for a given significance level, in that the power of such a test can never be determined; so in particular cases, it may, for all we know, be only infinitesimally different from the significance level. More seriously, the modifications introduced to meet the challenge of composite hypotheses are equally afflicted by the various difficulties we have shown to discredit even the most uncomplicated form of the significance test.

5.f | Classical Estimation Theory

Scientists often estimate a physical quantity and come thereby to regard a certain number, or range of numbers, as a more or less good approximation to the true value. Significance tests do not in general deliver such estimates and the need for them has prompted classical statisticians to develop a distinct body of doctrine known as Estimation Theory. The theory is classical, in that it purports to provide objective, non-probabilistic conclusions. It has two aspects, namely, point estimation and interval estimation, both of which we regard as fallacious, as we explain in the following review.

5.f.1. Point Estimation

Point estimation differs from interval estimation, which we deal with below, in offering a single number as the so-called 'best estimate' of a parameter. Suppose a population parameter, such as its mean, is in question. The technique for estimating this is to draw a random sample of predetermined size, n, from the population, and to measure each element drawn. Then, letting $x = x_1, \ldots, x_n$ denote the measurements thus derived, the next step is to pick an estimating statistic, t, this taking the form of a calculable function $t = f(x)$. Finally, the best estimate of the unknown parameter is inferred to be t_0, the value of t yielded by the experiment. But not every statistic is accepted as an estimator. The authors of this approach to estimation have specified certain conditions that any estimator must meet; the most frequently mentioned being the conditions of *sufficiency*, *unbiasedness*, *consistency* and *efficiency*. We discuss these in turn.

Sufficient Estimators

It will be recalled from Section 5.b, that a statistic t is sufficient for θ when $P(x \mid t)$ is independent of θ. The present requirement is that any estimating statistic should be sufficient in this sense. The mean of a random sample satisfies the requirement when it is

used to estimate a population mean, but the sample range, for example, is not. Nor is the sample median.[11]

Sufficiency is a Bayesian requirement too. Expressed in Bayesian terms a statistic, t, is sufficient for θ, just in case $P(x \mid t) = P(x \mid t \ \& \ \theta)$, for all θ. It follows straightforwardly from Bayes's theorem that t is sufficient for θ if, and only if, $P(\theta \mid t) = P(\theta \mid x)$. Hence, when a sample statistic is sufficient, it makes no difference whether you calculate the posterior distribution using it or using the full experimental information in x; the results will be the same. In other words, a sufficient statistic contains all the information that in Bayesian terms is relevant to θ.

A compelling intuition tells us that, in evaluating a parameter, we should not neglect any relevant information. There is a satisfactory rationale for this. Suppose you have two bits of information, a and b. There are then three posterior distributions to consider: $P(\theta \mid a)$, $P(\theta \mid b)$ and $P(\theta \mid a \ \& \ b)$. If these differ, which should describe your current belief state? The Bayesian has no choice, for $P(\theta \mid a)$ is your distribution of beliefs were you to learn a and nothing else. But you have in fact learned $a \ \& \ b$ and nothing else. Therefore, your current belief state must be described by $P(\theta \mid a \ \& \ b)$, rather than by the other mathematical possibilities.

The injunction to use all the relevant evidence in an inductive inference, which Carnap (1947) called the *Total Evidence Requirement*, is often considered to be an independent postulate. This is true, at any rate, within classical estimation theory, which also has to rely on the intuition, which it cannot prove either, that a sufficient statistic captures all the relevant evidence. The intuitions are well founded, but their source, in our opinion, is Bayes's theorem, applied unconsciously.

Unbiased Estimators

These are defined in terms of the expectation, or expected value, of a random variable, which is given by $E(x) = \Sigma x_i P(x_i)$, the sum

[11] The sample *range* is the difference between the highest and lowest measurements; if the sample measurements are arranged in increasing order, and if n is odd, their *median* is the middle element of the series; if n is even, the median is the higher of the middle two elements.

or in the continuous case, the integral, being taken over all possible values of x_i. We mentioned in Chapter 2 that the expectation of a random variable is also called the *mean* of its probability or density distribution; when the distribution is symmetrical, its mean is also its geometric centre. A statistic is defined as an *unbiased estimator* of θ just in case its expectation equals the parameter's true value. The idea is often glossed by saying that the value of an unbiased statistic, averaged over repeated samplings, will "in the long run", be equal to the parameter being estimated.[12]

Many intuitively satisfactory estimators are unbiased, for instance the proportion of red counters in a random sample is unbiased for the corresponding proportion in the urn from which it was drawn, and the mean of a random sample is an unbiased estimator of the population mean. However, sample variance is not an unbiased estimator of population variance and is generally "corrected" by the factor $\frac{n}{n-1}$.

But unbiasedness is neither a necessary nor a sufficient condition for a satisfactory estimation. We may see this through an example. Suppose you draw a sample, of predetermined size, from a population and note the proportion of individuals in the sample with a certain trait, and at the same time, you toss a standard coin. We now posit an estimating statistic which is calculated as the sample proportion plus k (> 0), if the coin lands heads, and plus k', if it lands tails. Then, if $k = -k'$, the resulting estimator is unbiased no less than the sample proportion itself, but its estimates are very different and are clearly no good. If, on the other hand, $k' = 0$, the estimator is biased, yet, on the occasions when the coin lands tails, the estimates it gives seem perfectly fine.

Not surprisingly, then, one finds the criterion defended, if at all, in terms which have nothing to do with epistemology. The usual defence is concerned rather with pragmatics. For example, Barnett claimed that, "within the classical approach unbiasedness is often introduced as a *practical* requirement to limit the class of estimators" (1973, p. 120; our italics). Even Kendall and Stuart, who wrote so confidently of the need to correct biased

[12] See for example Hays 1963, p. 196.

estimators, conceded that they had no epistemic basis for this censorious attitude:

> There is *nothing except convenience* to exalt the arithmetic mean above other measures of location as a criterion of bias. We might *equally well* have chosen the median of the distribution of t or its mode as determining the "unbiased" estimator. The mean value is used, as always, *for its mathematical convenience*. (1979, p. 4; our italics)

These authors went on to warn their readers that "the term 'unbiased' should not be allowed to convey overtones of a non-technical nature". But the tendentious nature of the terminology makes such misleading overtones hard to avoid. The next criterion is also named in a way that promises more than can be delivered.

Consistent Estimators

An estimator is defined to be *consistent* when, as the sample size increases, its probability distribution shows a diminishing scatter about the parameter's true value. More precisely, a statistic derived from a random sample of size n is a consistent estimator for θ if, for any positive number, ε, $P(|t - \theta| \leq \varepsilon)$ tends to 1, as n tends to infinity. This is sometimes described as t tending probabilistically to θ.

There is a problem with the consistency criterion as described, because it admits estimators that are clearly inadmissible. For example, if T_n is a consistent estimator, so is the estimator, T_n', defined as equal to zero for $n \leq 10^{10}$ and equal to T_n for $n > 10^{10}$. Fisher therefore added the further restriction that an admissible estimator should, in Rao's words, be "an explicit function of the observed proportions only". So, if the task is to estimate a population proportion, θ, the estimator should be a consistent function just of the corresponding sample proportion, and it should be such that when the observed and the population proportions happen to coincide, the estimator gives a true estimate (Rao 1965, p. 283). This adjustment appears to eliminate the anomalous estimators.

Fisher believed that consistency was the "fundamental criterion of estimation" (1956, p. 141) and that non-consistent estimators "should be regarded as outside the pale of decent usage" (1970, p. 11). In this, Neyman (1952, p 188) agreed "perfectly" with Fisher, and added his opinion that "it is definitely not profitable to use an inconsistent estimate."[13] Fisher defended his emphatic view in the following way:

> as the samples are made larger without limit, the statistic will usually tend to some fixed value characteristic of the population, and, therefore, expressible in terms of the parameters of the population. If, therefore, such a statistic is to be used to estimate these parameters, there is only one parametric function to which it can properly be equated. If it be equated to some other parametric function, we shall be using a statistic which even from an infinite sample does not give a correct value. . . . (1970, p. 11)

Fisher's claim here is that because a consistent estimator converges to some parameter value, it "can properly be equated" to that value, and it should be equated to no other value, because, if the sample were infinite, it would then certainly give the wrong result. This is more assertion than argument and in fact is rather implausible in its claims. Firstly, one should not, without qualification, equate an unknown parameter with the value taken by a statistic in a particular experiment; for, as is agreed on all sides, such estimates may, almost certainly will be in error, a consideration that motivates interval estimation, which we discuss below. Secondly, the idea that a consistent estimator becomes more accurate as the sample increases, and perfectly so in the limit, implies nothing at all about its accuracy on any particular occasion. Arguing for an estimator with this idea in mind would be like defending the use of a dirty measuring instrument on the grounds that if it were cleaner it would be better; in assessing a result, we need to know how good the instrument was in the experiment at hand, not how it might have performed under different conditions.

And (rebutting Fisher's last point) just as estimates made by consistent estimators may be quite inaccurate and clearly wrong, those from non-consistent ones might be very accurate and

[13] Presumably this should read: 'estimator'.

clearly right. For instance, suppose $\bar{x} + (n - 100)\bar{x}$ were chosen to estimate a population mean. This odd statistic is non-consistent, for, as the sample size grows, it diverges ever more sharply from the population mean. Yet for the special case where $n = 100$, the statistic is equivalent to the familiar sample mean, and gives an intuitively satisfactory estimate.

Efficient Estimators

The above criteria are clearly incomplete, because they do not incorporate the obvious desideratum that an estimate should improve as the sample becomes larger. So, for instance, a sample mean that is based on a sample of 2 would be 'sufficient', 'unbiased' and 'consistent', yet estimates of the population mean derived from it would not inspire confidence, certainly not as much as when there are 100, say, in the sample. This consideration is addressed by classical statistics through the efficiency criterion: the smaller an estimator's variance about the parameter value, the more *efficient* it is said to be, and the better it is regarded. And since the variance of a sample statistic is generally inversely dependent on the size of the sample, the efficiency criterion reflects the preference for estimates made with larger samples.

But it is not easy to establish, in classical terms, why efficiency should be a measure of quality in an estimator. Fisher (1970, p. 12) stated confidently that the less efficient of two statistics is "definitely inferior . . . in its accuracy"; but since he would have strayed from classical principles had he asserted that particular estimates were certainly or probably correct, even within a margin of error, this claim has no straightforward meaning. Kendall and Stuart's interpretation is the one that is widely approved. A more efficient statistic, they argued, will "deviate less, on the average, from the true value" and therefore, "we may reasonably regard it as better" (1979, p. 7). Now it is true that if e^i_1 and e^i_2 are the estimates delivered by separate estimators on the ith trial and if θ is the true value of the parameter, then there is a calculable probability that $| e^i_1 - \theta | < | e^i_2 - \theta |$, which will be greater the more efficient the first estimator is than the second. Kendall and Stuart translate this probability into an average frequency in a long run of trials, which, as we already remarked, goes beyond

logic. But even if the translation were correct, the performance of an estimator over a hypothetical long run implies nothing about the closeness of a particular estimate to the true value. And since estimates are usually expensive and troublesome to obtain and often inform practical actions, what is wanted and needed are just such evaluations of particular estimates.

5.f.2. Interval Estimation

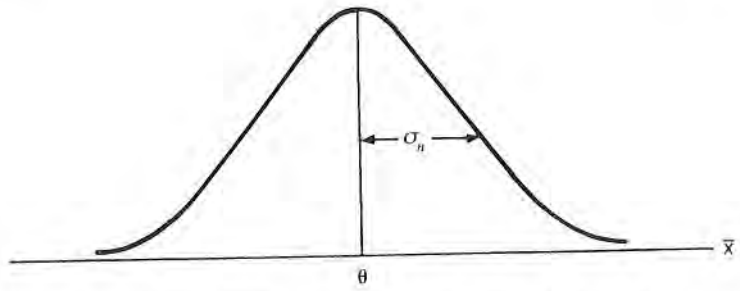
In practice, this demand is evidently met, for estimates are normally presented as a range of numbers, for example, in the form $\theta = a \pm b$, not as point values, which, as Neyman observed, "it is more or less hopeless to expect . . . will ever be equal to the true value" (1952, p. 159). Bayesians would qualify an interval estimate by the subjective probability that it contains the true value (see the discussion of 'credible intervals' in Section 8.a). Neyman's theory of confidence intervals, developed around 1930, and now dominant in the field, was intended to give a classical expression to this idea.

Confidence Intervals

Consider the task of estimating the mean height, θ, of the people in a large population, whose standard deviation, σ, is known. A sample of some predetermined size, n, is randomly selected, and its mean, $\bar{x}$, is noted. This mean can take many possible values, some more probable than others; the distribution representing this situation is approximately normal (the larger the population, the closer the approximation) with a mean equal to that of the population and a standard deviation given by $\sigma_n = \sigma n^{-\frac{1}{2}}$.

The sampling distribution plots possible sample means against probability densities, not probabilities, and, as explained earlier, this signifies that the probability that $\bar{x}$ lies between any two points is proportional to the area enclosed by those points and the curve. Because the distribution is essentially normal, it follows that, with probability 0.95,

$$-1.96\,\sigma_n \leq \theta - \bar{x} \leq 1.96\,\sigma_n.$$



The sampling distribution of means from a population with mean θ and standard deviation σ_n

And this implies that, with probability 0.95,

$$\bar{x} - 1.96\,\sigma_n \leq \theta \leq \bar{x} + 1.96\,\sigma_n.$$

Let m be the value of $\bar{x}$ in a particular experimental sample; since θ and n are known, the terms $m - 1.96\,\sigma_n$ and $m + 1.96\,\sigma_n$ can be computed. The interval between these two values is called a *95 percent confidence interval* for θ. Clearly there are other confidence intervals relating to different regions of the sampling distribution, and others, too, associated with different probabilities. The probability associated with a particular confidence interval is called its *confidence coefficient*.

The probability statements given above are simply deductions from the assumptions made and are unquestionably correct; and what we have said about confidence intervals, being no more than a definition of that concept, is also uncontroversial. Controversy arises only when confidence intervals are assigned inductive meaning and interpreted as estimates of the unknown parameter. What we shall call the *categorical-assertion* interpretation and the *subjective-confidence* interpretation are the two main proposals for legitimizing such estimates. We deal with these in turn.

The Categorical-Assertion Interpretation

This interpretation was first proposed by Neyman and has been widely adopted. Neyman said (1937, p. 263) that the "practical statistician", when estimating a parameter, should calculate a confidence interval and then "state" that the true value lies between the two confidence bounds, in the knowledge that (when the confidence coefficient is 0.99) "in the long run he will be correct in about 99 percent of all cases". The statistician's statement should not signify a belief in its truth, however. Indeed, Neyman (1941, p. 379) rejected the very idea of reasoning inductively to a conclusion, because he believed that "the mental process leading to knowledge . . . can only be deductive". Induction, for Neyman, was rather a matter of behaviour, and in the case of interval estimates, the proper outcome was a decision "to behave as if we actually knew" that the parameter lies within the confidence bounds.

We have already discussed this interpretation (in Section 5.d) in the context of significance tests and argued that typically and more reasonably scientists evaluate theories by degree; they do not, and, moreover, should not act in the way that Neyman recommended. A further indication that the interpretation is wrong arises from the fact that confidence intervals are not unique, as we explain next.

Competing Intervals

It is obvious from the sampling distribution of means depicted above that indefinitely many regions of that normal distribution cover 95 percent of its area. So instead of the usual 95 percent confidence interval located at the centre of the distribution, one could consider asymmetrical confidence intervals or ones that extend further into the tails while omitting smaller or larger strips in the centre. Neyman's categorical-assertion interpretation requires one to "assert" and "behave as if one actually knew" that the parameter lies in each and every one of this multiplicity of possible 95 percent confidence intervals, which is clearly unsatisfactory, and indeed, paradoxical.

Defenders of the interpretation have reacted in two ways, both we believe unsatisfactory. The first discriminates between confidence intervals on the basis of their length, and claims that, for a given confidence coefficient, the shortest interval provides the best estimate. In the words of Hays (1969, p. 290), there is "naturally

. . . an advantage in pinning the population parameter within the narrowest possible range with a given probability". By this criterion, the centrally symmetrical interval $m \pm 1.96\sigma_{\bar{y}}$ is the preferred 95 percent confidence interval for the population mean in the example cited above. The preference is based on the idea that the width of a confidence interval is a measure of the 'precision' of the corresponding estimate, and that this is a desirable feature. Thus Mood (1950, p. 222), when comparing two 95 percent confidence intervals, stated that the longer one was inferior, "for it gives less precise information about the location" of the parameter.

But it is not true that the length of a confidence interval measures its precision. For, consider the interval $| a, b |$ as an estimate of θ, and the interval $| f(a), f(b) |$ as an estimate of $f(\theta)$. If f is a 1-1 function, the two estimates are equivalent and must be equally informative and therefore equally precise. But while the first may be the shortest 95 percent confidence interval for θ, the second might not be the shortest such interval for $f(\theta)$; this would be the case, for instance, when $f(a) = a^{-1}$.

Another difficulty is that different sample statistics may yield different minimum-length confidence intervals, a fact that has prompted the proposal to restrict interval estimates to those given by statistics with the smallest possible variance. It is argued that although this new criterion does not guarantee the shortest possible confidence interval in any particular case, it does at least ensure that such intervals "are the shortest on average in large samples" (Kendall and Stuart 1979, p. 126). We have already criticized both the long-run justification and the short-length criterion, and since two wrongs don't make a right, we shall leave the discussion there.

Neyman (1937, p. 282) suggested another way of discriminating between possible confidence intervals. He argued, in a manner familiar from his theory of testing, that a confidence interval should not only have a high probability of containing the correct value but should also be relatively unlikely to include wrong values. More precisely: a best confidence interval, I_o, should be such that for any other interval, I, corresponding to the same confidence coefficient, $P(\theta' \in I_o \mid \theta) \leq P(\theta' \in I \mid \theta)$; moreover, the inequality must hold whatever the true value of the parameter, and for every value θ' different from θ. But as Neyman himself showed, there are no 'best' intervals of this kind for most of the cases with which he was originally concerned.

The Subjective-Confidence Interpretation

Neyman's categorical-assertion interpretation, contrary to its main intention, does, in fact, contain an element that seems to imply a scale by which estimates may be evaluated and qualified, namely the confidence coefficient. For suppose some experimentally established range of numbers constituted a 90 percent confidence interval for θ, rather than the conventionally approved 95 or 99 percent, we would still be enjoined to assert that θ is in that range and to act as if we believed that to be true, though with a correspondingly modified justification that now referred to a 90 percent frequency of being correct "in the long run". But if the justification had any force (we have seen that it does not), it would surely be stronger the lower the frequency of error. So the categorical assertion that θ is in some interval must after all be qualified by an index running from 0 to 100 indicating how well founded it is, and this is hard to distinguish from the index of confidence that is explicit in the subjective-confidence interpretation that we deal with now.

In this widely approved position, a confidence coefficient is taken to be "a measure of our confidence" in the truth of the statement that the confidence interval contains the true value (for example, Mood 1950, p. 222). This has some surface plausibility. For consider again the task of estimating a population mean, θ. . . We know that in experiments of the type described, θ is included in any 95 percent confidence interval with an objective probability of 0.95; and this implies that if the experiment were performed repeatedly, θ would be included in such intervals with a relative frequency that tends, in the limit, to 0.95. It is tempting, and many professional statisticians find it irresistible, to infer from this limit property a probability of 0.95 that the particular interval obtained in a particular experiment does enclose θ. But drawing such an inference would commit a logical fallacy.[14]

[14] This fallacy is repeated in many statistics texts. For example, Chiang 2003, (pp. 138–39): "The probability that the interval contains μ is either zero or one;

The subjective-confidence interpretation seems to rely on a misapplication of a rule of inference known as the Principle of Direct Probability (see Chapter 3), which is used extensively in Bayesian statistics. The principle states that if the objective, physical probability of a random event (in the sense of its limiting relative frequency in an infinite sequence of trials) is known to be r, then, in the absence of any other relevant information, the appropriate subjective degree of belief that the event will occur on any particular trial is also r. Expressed formally, if the event in question is a, and $P^*(a)$ is its objective probability, and if a_t describes the occurrence of the event on a particular trial, the Principle of Direct Probability says that $P[a_t \mid P^*(a) = r] = r$, where P is a subjective probability function.

For example, the physical probability of getting a number of heads, K, greater than 5 in 20 throws of a fair coin is 0.86 (see Table 5.1 above), that is, $P^*(K > 5) = 0.86$. By the Principle of Direct Probability,

$$P[(K > 5)_t \mid P^*(K > 5) = 0.86] = 0.86.$$

That is to say, 0.86 is the confidence you should have that any particular trial of 20 throws of a fair coin will produce more than 5 heads. Suppose one such trial produced 2 heads. To infer that we should now be 86 percent confident that 2 is greater than 5 would, of course, be absurd; it would also be a misapplication of the principle. For one thing, if it were legitimate to substitute numbers for K, why would such substitution be restricted to its first occurrence in the principle? But in fact, no such substitution is allowed. For the above equation does not assert a general rule for each number K from 0 to 20; the K-term is not a number, but a function that

no intermediate values are possible. What then is the initial probability of 0.95? Suppose we take a large number of samples, each of size n. For each sample we make a statement that the interval observed from the sample contains μ. Some of our statements will be true, others will not be. According to [the equation we give in the text, above] . . . 95 percent of our statements will be true. In reality we take only one sample and make only one statement that the interval contains μ. Thus [*sic*] we do have confidence in our statement. The measure of our confidence is the initial probability 0.95."

takes different values depending on the outcome of the underlying experiment.

Mistaking this appears to be the fallacy implicit in the subjective-confidence interpretation. It is true that the objective probability of θ being enclosed by experimentally determined 95 percent confidence intervals is 0.95. If I_1 and I_2 are variables representing the boundaries of such confidence intervals, the Principle of Direct Probability implies that

$$P[(I_1 \leq \theta \leq I_2)_t \mid P^*(I_1 \leq \theta \leq I_2) = 0.95] = 0.95,$$

and this tells us that we should be 95 percent confident that any sampling experiment will produce an interval containing θ. Suppose now that an experiment that was actually performed yielded I'_1 and I'_2 as the confidence bounds; the subjective-confidence interpretation would tell us to be 95 percent confident that $I'_1 \leq \theta \leq I'_2$. But this would commit exactly the same fallacy as we exposed in the above counter-example. For I_1 and I_2, like K, are functions of possible experimental outcomes, not numbers, and so the desired substitution is blocked.[15]

In response, it might be said that the subjective-confidence interpretation does not depend on the Principle of Direct Probability (a Bayesian notion, anyway), that it is justified on some other basis. But we know of none, nor do we think any is possible, because, as we shall now argue, the interpretation is fundamentally flawed, since it implies that one's confidence in a proposition should depend on information that is manifestly irrelevant, namely, that concerning the stopping rule, and should be independent of prior information that is manifestly relevant. We address these two points next.

The Stopping Rule

Confidence intervals arise from probability distributions over spaces of possible outcomes. Although one of those outcomes

[15] Howson and Oddie 1979 pointed out this misapplication of the principle in another context. See also 3.f above.

will be actualized, the space as a whole is imaginary, its contents depending in part on the experimenters' intentions, embodied in the adopted stopping rule, as we explained earlier. Estimating statistics employed in point estimation are also stopping-rule dependent, because the stopping rule dictates whether or not those statistics satisfy the various conditions that are imposed on estimators. So, for instance, the sample mean is an unbiased estimator of a population mean if it is based on a fixed, predetermined sample size, but not necessarily otherwise.[16]

The criticism we levelled at this aspect of classical inference in the context of tests of significance applies here too, and we refer the reader back to that discussion. In brief, the objection is that having to know the stopping rule when drawing an inference from data means that information about the experimenters' private intentions and personal capacities, as well as other intuitively extraneous facts, is ascribed an inductive role that is highly inappropriate and counter-intuitive. This is, in a way, tacitly acknowledged by most classical statisticians, who in practice almost always ignore the stopping rule and standardly carry out any classical analysis *as if* the experiment had been designed to produce a sample of the size that it did, without any evidence that this was so, and even when it clearly was not.

We take up the discussion of the stopping rule again in Chapter 8, where we show why it plays no role in Bayesian induction.

Prior Knowledge

Estimates are usually made against a background of partial knowledge, not in a state of complete ignorance. Suppose, for example, you were interested in discovering the average height of students attending the London School of Economics. Without being able to point to results from carefully conducted studies, but on the basis of common sense and what you have learned informally about students and British universities' admission standards, you would feel pretty sure that this could not be below four feet, say, nor above six. Or you might already have made an

[16] See, for example, Lee 1989, p. 213.

exhaustive survey of the students' heights, lost the results and be able to recall with certitude only that the average was over five feet. Now if a random sample, by chance, produced a 95 percent confidence interval of 3' 10" ± 2", you would be required by classical principles to repose an equivalent level of confidence in the proposition that the students' average height really does lie in that interval. But with all you know, this clearly would not be a credible or acceptable conclusion.

A classical response to this difficulty might take one of two forms, neither adequate, we believe. The first would be to restrict classical estimation to cases where no relevant information is present. But this proposal is scarcely practicable, as such cases are rare; moreover, although a little knowledge is certainly a dangerous thing, it would be odd, to say the least, if it condemned its possessor to continue in this condition of ignorance in perpetuity. A second possibility would be to combine in some way informal prior information with the formal estimates based on random samples. The Bayesian method expresses such information through the prior distribution, which then contributes to the overall conclusion in a regulated way, but there is no comparable mechanism within the confines of classical methodology.

5.g | Sampling

Random Sampling

The classical methods of estimation and testing that we have been considering purport to be entirely objective, and it is for this reason that they call for the sampling distribution of the data also to be objective. To this end, classical statisticians require the sample that is used for the estimate to have been generated by an impartial, physical process that ensures for each element of the population an objectively equal chance of being selected. Here is a simple instance of such a process: a bag containing similar counters, each corresponding to a separate member of the population, and marked accordingly, is shaken thoroughly and a counter selected blindfold; this selection is repeated the prescribed number of times, and the population members picked out by the

selected counters then constitute a random sample. There are of course other, more sophisticated physical mechanisms for creating random samples.

What we call the *Principle of Random Sampling* asserts that satisfactory estimates can only be obtained from samples that are objectively random in the sense indicated.

Judgment Sampling

The Principle of Random Sampling may be contrasted with another approach, which is motivated by the wish to obtain a *representative sample*, one that resembles the population in all those respects that are correlated with the characteristic being measured. Suppose the aim were to measure the proportion of the population intending to vote Conservative in a forthcoming election. If, as is generally agreed, voting preference is related to age and socio-economic status, a representative sample should recapitulate the population in its age and social class structure; quite a number of other factors, such as gender and area of residence, would, no doubt, also be taken into account in constructing such a sample. A representative sample successfully put together in this way will have the same proportion of intending Conservative voters as the parent population. Samples selected with a view to representativeness are also known as *purposive*, or *judgment samples*; they are not random. A kind of judgment sampling that is frequently resorted to in market research and opinion polling is known as *quota sampling*, where interviewers are given target numbers of people to interview in various categories, such as particular social classes and geographical regions, and invited to exercise their own good sense in selecting representative groups from each specified category.

Some Objections to Judgment Sampling

Judgment sampling is held to be unsatisfactory by many statisticians, particularly those of a classical stripe, who adhere to the random sampling principle. Three related objections are encountered. The first is that judgment sampling introduces an undesirable subjectivity into the estimation process. It does have a subjective aspect, to be sure, for when drawing such a sample, a view needs to be taken on which individual characteristics are correlated with the population parameter whose value is being sought, and which are not: a judgment must be made as to whether a person's social class, age, gender, the condition of his front garden, the age of her cat, and so forth, are relevant factors for the sampling process. Without exhaustively surveying the population, you could not pronounce categorically on the relevance of the innumerable, possibly relevant factors; there is, therefore, considerable room for opinions to vary from one experimenter to another. This may be contrasted with random sampling, which requires no individual judgment and is quite impersonal and objective.

The second objection, which is, in truth, an aspect of the first, is that judgment samples are susceptible to bias, due to the experimenter's ignorance, or through the exercise of unconscious, or even conscious, personal prejudices. Yates (1981, pp. 11–16) illustrates this danger with a number of cases where the experimenter's careful efforts to select representative samples were frustrated by a failure to appreciate and take into account crucial variables. Such cases are often held up as a warning against the bias that can intrude into judgment sampling.

Sampling by means of a physical randomizing process, on the other hand, cannot be affected by a selector's partiality or lack of knowledge. On the other hand, it might, by chance, throw up samples that are as unrepresentative as any that could result from the most ill-informed judgment sampling. This seeming paradox[17] is typically turned into a principal advantage in the standard classical response, which says that when sampling is random, the probabilities of different possible samples can be accurately computed and then systematically incorporated into the inference process, using classical estimation methods. But judgment sampling—so the third objection goes—does not lend itself to objective methods of estimation.

There is another strand to the classical response, which invokes the idea of *stratified random sampling*. This involves partitioning

[17] See below, 8.d, for a discussion of Stuart's description of this situation as the "paradox of sampling".

the population into separate groups, or strata, and then sampling at random from each. The classically approved estimate of the population parameter is then the weighted average of the corresponding strata estimates, the weighting coefficients being proportional to the relative sizes of the population and the strata. Stratified random sampling seems clearly intended as a way of reducing the chance of obtaining seriously unrepresentative samples. But the orthodox classical rationale refers instead to the greater 'efficiency' (in the sense defined above) of estimates derived from stratified samples. For, provided the strata are more homogeneous than the population, and significantly different from one another in relation to the quantity being measured, estimation is more 'efficient' using stratified random sampling than ordinary random sampling, and so, by classical standards, it is better.

This rationale is however questionable; indeed, it seems quite wrong. The efficiency of an estimator, it will be recalled, is a measure of its variance. And the more efficient an estimator, the narrower any confidence interval based on it. So, for instance, a stratified random sample might deliver the 95 percent confidence interval 4' 8" ± 3" as an estimate of the average height of pupils in some school, while the corresponding interval derived from an unstratified random sample (which, by chance, is heavily biased towards younger children) might be, say, 3' 2" ± 6". Classical statisticians seem committed to saying that the first estimate is the better one because, being based on a more efficient estimating method, its interval width is narrower. But this surely misappraises the situation. The fact is that the first estimate is probably right and the second almost certainly wrong, but these are words that should not cross the lips of a classical statistician.

Some Advantages of Judgment Sampling

Judgment sampling has certain practical advantages. A pre-election opinion poll, for example, needs to be conducted quickly, and this is feasible with judgment sampling; on the other hand, drawing up a random sample of the population, finding the people who were selected and then persuading them to be interviewed is costly, time consuming, and sometimes impossible, and the election might well be over before the poll has begun. Practical considerations such as these have established the dominance of quota sampling in market research. "Probably 90 percent of all market research uses quota sampling and in most circumstances it is sufficiently reliable to provide consistent results" (Downham 1988, p. 13).

A second point in favour of judgment and quota samples is that they are evidently successful in practice. Opinion polls conducted by their means, insofar as they can be checked against the results of ensuing elections, are mostly more or less accurate, and market research firms thrive, their services valued by manufacturers, who have a commercial interest in accurately gauging consumers' tastes.

A third practical point is that inferences based on non-random samples are often confidently made and believed by others; indeed they seem inevitable when conclusions obtained in one sphere need to be applied to another, as commonly happens. For instance, in a study by Peto *et al.* (1988), a large group of physicians who had regularly taken aspirin and another group who had not showed similar frequencies of heart attacks over a longish period. Upon this basis, the authors of the study advised against adopting aspirin generally as a prophylactic, their implicit and plausible assumption being that the doctors taking part in the study typified the wider population in their cardiac responses to aspirin. Although the rest of the statistical procedures employed in the study were orthodox, this assumption was not checked by means of random samples taken from the population.[18]

We consider the question of sampling methods again when we discuss Bayesian inference in Chapter 8.

5.h | Conclusion

Classical estimation theory and significance tests, in their various forms, are still immensely influential; they are advocated in hundreds of books that are recommended texts in thousands of institutions of higher education, and required reading for hundreds of

[18] Smith 1983 makes the same point in relation to another study. On the arguments concerning sampling in this section, see also Urbach 1989.

thousands of students. And the classical jargon of 'statistical significance', 'confidence', and so on, litters the academic journals and has slipped easily into the educated vernacular. Yet, as we have shown, classical 'estimates' are not estimates in any normal or scientific sense, and, like judgments of 'significance' and 'non-significance', they carry no inductive meaning at all. Therefore, they cannot be used to arbitrate between rival theories or to determine practical policy.

A number of other objections that we have explored in this chapter show, moreover, that classical methods are set altogether on the wrong lines, and are based on ideas inimical to scientific method. Principal here is the objection that all classical methods involve an outcome space and hence a stopping rule, which we have argued brings to bear on scientific judgment considerations that are highly counter-intuitive and inappropriate in that context. And classical methods necessarily introduce arbitrary elements that are at variance not only with scientific practice and intuition, but also with the objectivist ideals that motivated them. The founders of the classical philosophy were seeking an alternative to the Bayesian philosophy, which they dismissed as unsuited to inductive method because it was tainted by subjectivity. It is therefore particularly curious and telling that classical methods cannot operate except with their own, hefty subjective input. This was frankly confessed, in retrospect, by one of the founders of the classical approach:

> Of necessity, as it seemed to us [him and Neyman], we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgement in such matters . . . as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities. (Pearson 1966, p. 277)

Classical distaste for the subjective element in Bayesian inference puts one in mind of those who were once accused of taking infinite trouble to strain out a gnat, while cheerfully swallowing a camel!

CHAPTER 6
Statistical Inference in Practice: Clinical Trials

We have thus far discussed various methodologies in terms sufficiently abstract to have perhaps created the impression that the question as to which of them is philosophically correct has little practical bearing. But this is far from being the case. We illustrate this point in the present chapter by looking at Classical and Bayesian approaches to the scientific investigation of causal connections, particularly in agricultural (or 'field') and medical (or 'clinical') trials. Large numbers of such trials are under way at any one time; they are immensely expensive; and their results may exert profound effects on farming and clinical practice. And a further practical effect, in the case of clinical trials, is the inconvenience to which the participants are put and the risks to which they may be exposed.

6.a | Clinical Trials: The Central Problem

A clinical trial is designed with a view to discovering whether and to what extent a particular drug or medical procedure alleviates certain symptoms, or causes adverse side effects. And a typical goal of an agricultural field trial would be to investigate whether a putative fertilizer increases the yield of a certain crop, or whether a new, genetically engineered potato has improved growth qualities.

Clinical trials typically involve two groups of subjects, all of whom are currently suffering from a particular medical condition; one of the groups, the *test group*, is administered the experimental therapy, while the other, the *control group*, is not; the progress of each group is then monitored over a period. An agricultural trial to compare a new variety of potato (A) with an established