CHAPTER 3

Fisherian theory

3.1 Introduction

There are alternatives to the Neyman–Pearson formulation of the problem of testing statistical hypotheses. Although it is important that we recognize and understand the differences between the various formulations, there is no standard terminology to help us. Many authors have distinguished between what we are calling Neyman–Pearson tests and tests that have a different form and purpose, often calling the latter significance tests and usually citing R.A. Fisher as a particularly influential developer or proponent. Although Fisher was not the originator of significance tests, we call them 'Fisherian' because of his consistent emphasis on the distinction between the problems addressed by the Neyman–Pearson theory of hypothesis testing and problems of evidential interpretation of scientific data, for which significance tests are intended. We will draw a further distinction, describing two varieties of significance test, both of which seem to have been advocated by Fisher. The first we will call ***p*-value procedures**, and will consider in sections 3.2–3.4. These are prominent in the statistical analyses used in science. The second variety, also influential in scientific applications, we call **rejection trials**. These are particularly interesting because they link statistical hypothesis testing directly to formal logic and to the philosophy of science; they will be discussed in section 3.5.

Later in this chapter we describe how the use of significance tests to measure evidence leads to a popular evidential interpretation of confidence intervals. We also consider briefly the general issue of alternative hypotheses in science.

3.2 A method for measuring statistical evidence: the test of significance

Statistical hypothesis tests, as they are most commonly used in analyzing and reporting the results of scientific studies, do not proceed as envisioned in the Neyman–Pearson theory, with a choice

between two specified hypotheses being made according to whether or not the observations fall into a pre-selected critical region. A more common procedure is described by Cox and Hinkley (1974, p. 66):

> Let $t = t(x)$ be a function of the observations and let $T = t(X)$ be the corresponding random variable. We call T a test statistic for testing H_0 if the following conditions are satisfied:
>
> (a) the distribution of t when H_0 is true is known at least approximately...
> (b) the larger the value of t the stronger the evidence of departure from H_0 of the type it is required to test...
>
> For given observations x we calculate $t_{obs} = t(x)$, say, and the *level of significance* p_{obs} by
>
> $$p_{obs} = \text{pr}(T \geq t_{obs}; H_0).$$

The result of this procedure is not a decision to choose one hypothesis or another, but a number, p_{obs}, called the level of significance, or ***p*-value**; the procedure is called a **significance test**.

For example, to test the hypothesis H_0 that the probability of success is one-half on each of 20 independent trials we might use as a test statistic T the total number of successes. When H_0 is true this statistic has a known probability distribution (binomial), and large values are evidence supporting hypotheses that specify a greater success probability over H_0. If we observe 14 successes then the p-value is $\Pr(T \geq 14) = 0.06$.

An essential component of significance tests is a concept that did not appear in the Neyman–Pearson theory of hypothesis testing, the concept of strength of evidence. A p-value is supposed to indicate 'the strength of the evidence against the hypothesis' (Fisher, 1958, p. 80), with conventional interpretations as described by Burdette and Gehan (1970, p. 9):

> Reasonable interpretations of the results of significance tests are as follows:

Significance Level of Data	*Interpretation*
Less than 1 per cent	Very strong evidence against the null hypothesis
1 per cent to 5 per cent	Moderate evidence against the null hypothesis
More than 5 per cent and less than 10 per cent	Suggestive evidence against the null hypothesis
10 per cent or more	Little or no real evidence against the null hypothesis.

eing made according to whether or
re-selected critical region. A more
/ Cox and Hinkley (1974, p. 66):

)servations and let $T = t(X)$ be the
e call T a test statistic for testing H_0
)fied:

H_0 is true is known at least

nger the evidence of departure from
test...

ate $t_{obs} = t(x)$, say, and the *level of*

$\geq t_{obs}; H_0)$.

: a decision to choose one hypoth-
, called the level of significance, or
significance test.

thesis H_0 that the probability of
independent trials we might use
ber of successes. When H_0 is true
)ility distribution (binomial), and
)rting hypotheses that specify a
H_0. If we observe 14 successes
0.06.

lificance tests is a concept that did
)n theory of hypothesis testing, the
A p-value is supposed to indicate
inst the hypothesis' (Fisher, 1958,
etations as described by Burdette

results of significance tests are as

Interpretation
Very strong evidence against the null hypothesis
Moderate evidence against the null hypothesis
Suggestive evidence against the null hypothesis
Little or no real evidence against the null hypothesis.

Another difference between hypothesis testing, in the sense of Neyman and Pearson, and significance testing is the role of alternative hypotheses: Neyman–Pearson tests are for choosing between two hypotheses, whereas significance tests are for measuring the evidence against one, the null hypothesis. Alternatives to the null hypothesis are often acknowledged to play a part in significance tests, as in Cox and Hinkley's (1974, p. 66) implicit reference to an alternative in their condition that 'the larger the value of t the stronger the evidence of *departure from H_0 of the type it is required to test*' (emphasis added). But alternative hypotheses do not have an essential explicit role analogous to the one they play in Neyman–Pearson theory. In fact, many authorities maintain that significance tests' freedom from dependence on explicit alternative hypotheses is essential in some important applications (such as 'goodness of fit' tests):

> Let us try the simple single hypothesis first. If the data do not fit that, then it is worth while going ahead [and constructing alternative hypotheses]. If it is consistent with the data let us not waste our time. (Barnard, in Savage, 1962, p. 85)

Box (1980) has defended this position more recently.

Here is a summary of some of the differences between these two approaches to testing hypotheses about the distribution of a random variable X:

Neyman–Pearson hypothesis tests	*Significance tests (p-value procedures)*
Purpose: To choose one of two specified hypotheses, H_1 and H_2, on the basis of an observation $X = x$.	For a single hypothesis H, to measure the evidence against H represented by an observation $X = x$.
Elements:	
1. Two hypotheses (families of probability distributions) H_1 and H_2.	1. One hypothesis H, called the 'null' hypothesis.
2. A test function $\delta(x)$ that specifies which hypothesis to choose when $X = x$ is observed: if $\delta(x) = 1$ we choose H_1, if $\delta(x) = 2$ we choose H_2.	2. A real-valued function $t(x)$ that gives an ordering of sample points as evidence against H: $t(x_1) > t(x_2)$ means that x_1 is stronger than x_2 as evidence against H.

3. Result is a decision or action, 'Choose H_1' or 'Choose H_2'.	3. Result is a number, the significance level, or p-value, interpreted as a measure of the evidence against H; the smaller the p-value the stronger the evidence.

The distinction between Neyman–Pearson tests and significance tests is not made consistently clear in modern statistical writing and teaching. Mathematical statistical textbooks tend to present Neyman–Pearson theory, while statistical methods textbooks tend to lean more towards significance tests. The terminology is not standard, and the same terms and symbols are often used in both contexts, blurring the differences between them. For example, descriptions of Neyman–Pearson theory often refer to the size, or Type I error probability, as the 'significance level'.

A further source of confusion is that within the Neyman–Pearson framework it is sometimes recommended that the experimenter should report, not the result of testing H_1 versus H_2 at a pre-selected Type I error level α, but the smallest value of α that would have led to rejection of H_1. This enables the reader who prefers a different Type I error level, say α', to perform his own test, rejecting H_1 (choosing H_2) if the reported α is smaller than his α'. Such a reported α is mathematically equivalent to a p-value (and is sometimes called by that name). But this does not make the procedure into a significance test, which is defined, not simply by what number is calculated, but by what that number is supposed to mean. As we saw in sections 2.3 and 2.4, Neyman was quite right in his insistence on a narrow behavioral, or decision-making, interpretation of his theory – evidential interpretations are generally invalid.

The key difference between Neyman–Pearson tests and significance tests is in their purpose. Neyman–Pearson tests are rules for choosing between alternative actions, while significance tests purport to measure evidence. That is, Neyman–Pearson tests address the second of the physician's three questions in Chapter 1, 'What should I *do*?', while significance tests address the third, 'How should I interpret these observations as *evidence*?'. In his section on 'The simple test of significance', Fisher (1956, p. 42) complained that the Neyman–Pearson view 'that the purpose of the test is to discriminate or "decide" between two or more hypotheses' had 'greatly obscured' the understanding of tests.

3. Result is a number, the significance level, or p-value, interpreted as a measure of the evidence against H; the smaller the p-value the stronger the evidence.

an–Pearson tests and significance
lear in modern statistical writing
istical textbooks tend to present
tatistical methods textbooks tend
ce tests. The terminology is not
d symbols are often used in both
es between them. For example,
theory often refer to the size, or
significance level'.
s that within the Neyman–Pearson
mmended that the experimenter
testing H_1 versus H_2 at a pre-
out the smallest value of α that
H_1. This enables the reader who
level, say α', to perform his own
if the reported α is smaller than
ematically equivalent to a p-value
t name). But this does not make
test, which is defined, not simply
t by what that number is supposed
3 and 2.4, Neyman was quite right
avioral, or decision-making, inter-
ntial interpretations are generally

Neyman–Pearson tests and signi-
Neyman–Pearson tests are rules
e actions, while significance tests
That is, Neyman–Pearson tests
ian's three questions in Chapter 1,
nificance tests address the third,
observations as *evidence*?'. In his
significance', Fisher (1956, p. 42)
earson view 'that the purpose of
"decide" between two or more
ired' the understanding of tests.

He then offered

> a clear view of the nature of a test of significance applied to a single hypothesis by a unique body of observations.
>
> Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is in fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders.

3.3 The rationale for significance tests

Why should a small p-value be interpreted as signifying strong evidence against the hypothesis? Barnard (1967, p. 32) explains:

> The meaning of 'H is rejected at significance level α' is 'Either an event of probability α has occurred, or H is false,' and our disposition to disbelieve H arises from our disposition to disbelieve in events of small probability.

This echoes Fisher's (1959, p. 39) explanation – after calculating, under a random distribution hypothesis, that the probability of the event observed, or a more extreme event, was about 1/33 000, he proposed that this probability

> is amply low enough to exclude at a high level of significance any theory involving a random distribution.
>
> The force with which such a conclusion is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true.

According to the Fisher–Barnard explanation, significance tests rest on some principle like the following:

> *Law of improbability*: If hypothesis A implies that the probability that a random variable X takes on the value x is quite small, say $p_A(x)$, then the observation $X = x$ is evidence against A, and the smaller $p_A(x)$, the stronger that evidence.

More recently Cox (1977, p. 53) has cited this law ('the smaller is the probability under H_0, the stronger is the evidence against H_0') as the basis for significance tests in some circumstances. But the law of improbability has attracted criticism as well as support. Some have observed that it appears to be unacceptably hard on null hypotheses. Suppose, for example, that we have a computer program intended to

generate standard normal deviates. Consider the null hypothesis that the program is operating correctly. Now consider the evidence in a single observed output, $X = x$. Because the hypothesis implies that the probability of any single point is zero, the law of improbability would imply that whatever value, x, is produced, it is overwhelming evidence that the program is not working properly. The problem is not restricted to continuous distributions – if X is intended to have a $Bin(n, \frac{1}{2})$ distribution then the maximum probability on any outcome is roughly $(2/n\pi)^{1/2}$, so that if n is large then no matter what value x is observed, it will be judged to be strong evidence against the (true) binomial distribution hypothesis. This is the point that Hacking (1965, p. 82) made in discussing Fisher's argument quoted above: 'if Fisher's disjunction had any force, we should always have to exclude any hypothesis like that of random distribution, whatever happened. So it has no force'.

The binomial distribution assigns greater probability to values of x near $n/2$. Although the absolute probabilities are all small when n is large, the relative probabilities are not, and the ratio of the maximum probability to the minimum, which occurs at both $x = 0$ and at $x = n$, is quite large, roughly $2^n(2/n\pi)^{1/2}$. Thus although all possible outcomes have low probability under the hypothesis, some have much lower probabilities than others. To accommodate this phenomenon, we might try a modified version of the law stating that it is low probability *relative to other outcomes* that makes a given outcome evidence against a hypothesis.

> *Law of improbability II*: If hypothesis A implies that the probability that a random variable X takes on the value x is small compared to the probability of another value x', $p_A(x) \ll p_A(x')$, then the observation $X = x$ is evidence against A, and the smaller the ratio $p_A(x)/p_A(x')$, the stronger the evidence.

Law II is unsatisfactory on various counts, one of which is that it leaves some important hypotheses exempt from unfavorable evidence. Suppose X represents a series of n Bernoulli (success or failure, coded 1 or 0) trials, and consider the hypothesis that the trials are independent with common probability of success equal to one-half. Under this hypothesis every possible outcome is a series of n zeroes and ones, and they all have the same probability of occurrence, $(\frac{1}{2})^n$. Thus for every pair of possible outcomes, x and x', $p_A(x)/p_A(x') = 1$, indicating evidence of no strength at all; no outcome is less probable than any other, so none is evidence against the hypothesis.

tes. Consider the null hypothesis
rectly. Now consider the evidence
x. Because the hypothesis implies
point is zero, the law of improb-
r value, x, is produced, it is over-
ram is not working properly. The
ntinuous distributions – if X is
ibution then the maximum prob-
ily $(2/n\pi)^{1/2}$, so that if n is large
observed, it will be judged to be
binomial distribution hypothesis.
(1965, p. 82) made in discussing
: 'if Fisher's disjunction had any
exclude any hypothesis like that
happened. So it has no force'.
gns greater probability to values of
e probabilities are all small when n
es are not, and the ratio of the
ninimum, which occurs at both
large, roughly $2^n(2/n\pi)^{1/2}$. Thus
have low probability under the
wer probabilities than others. To
we might try a modified version
obability *relative to other outcomes*
ence against a hypothesis.

hesis A implies that the probability
n the value x is small compared to
', $p_A(x) \ll p_A(x')$, then the observa-
t A, and the smaller the ratio
dence.

rious counts, one of which is that
theses exempt from unfavorable
a series of n Bernoulli (success or
l consider the hypothesis that the
mon probability of success equal
esis every possible outcome is a
they all have the same probability
very pair of possible outcomes, x
ting evidence of no strength at all;
n any other, so none is evidence

Maybe we need to bring the 'more extreme' outcomes into the analysis. Since the occurrence of an event whose probability under H is small is interpreted as evidence against H, with the strength of evidence growing as the probability shrinks, the outcomes that are 'as extreme or more so' are apparently just those outcomes whose probabilities under H are as small or smaller. Suppose we try to state the law in terms of the probabilities of outcomes 'as extreme or more so' than the one observed:

> *Law of improbability III*: If hypothesis A implies that the probability that a random variable X takes on the value x is $p_A(x)$ and if the sum $S(x)$ of the probabilities of all values whose probabilities are less than or equal to $p_A(x)$ is small, then the observation $X = x$ is evidence against A, and the smaller the sum $S(x)$, the stronger the evidence.

But law III also fails in the simple case of a sequence of independent Bernoulli trials with success probability one-half. Since all possible outcomes have the same probability, $p_A(x) = (\frac{1}{2})^n$, for every one $S(x) = 1$, again indicating evidence of no strength at all. According to law III only outcomes that are *impossible* under this null hypothesis are evidence against it.

We will not continue to fiddle with the law of improbability, trying to adjust our statement of it until we get it right. It cannot be made right, as we already learned in section 1.4: it is not low probability under A that makes an event evidence against A – it is low probability under A relative to the probability under another hypothesis B that makes it evidence supporting B over A. And then it is not evidence against A, but evidence against A, *vis-à-vis* B.

Suppose I send my valet to bring my urn containing 100 balls, of which only two are white. I draw one ball and find that it is white. Is this evidence against the hypothesis that he has brought the correct urn? And is $p = 0.02$ a proper measure of the strength of this evidence? Suppose that I keep in my urn vault two urns, one with two white balls and another, identical in appearance, that contains no white balls. Now is my observation of a white ball evidence that he has not brought the right urn? Fisher's disjunction still applies – either a rare event has occurred or the null hypothesis (correct urn) is false. But although the observation of a white ball is rare under the null hypothesis, it is even rarer under the alternative (wrong urn). In this case, the observation is actually strong evidence *in favor* of the null hypothesis. Of course, we might consider other hypotheses as well. For example, if my valet likes to play tricks,

we might consider the hypothesis that he has added some more white balls to the urn. The evidence favors that hypothesis over the 'correct urn' null hypothesis by a factor that depends on how many white balls he might have added.

The point again is that evidence is relative (as we saw in section 1.4) – whether it counts for or against one hypothesis can only be determined with reference to an alternative (see Exercise 3.1). This point has been made well and often for decades. Before the birth of the Neyman–Pearson theory the inventor of the *t*-test, W.S. Gosset, explained to Neyman's coauthor, Egon Pearson, that an observed discrepancy between a sample mean and a hypothesized population mean

> doesn't in itself necessarily prove that the sample was not drawn randomly from the population even if the chance is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 . . . you will be very much more inclined to consider that the original hypothesis is not true.
>
> (Gosset [1926], quoted in Pearson, 1938)

So the Fisher–Barnard rationale for significance tests, as expressed in the law of improbability, is wrong. There is, in fact, no sound rationale for these tests. This is because they are incompatible with the law of likelihood. Specifically, significance tests depend critically on how the probability distribution is spread over unobserved points in the sample space (through their definition in terms of outcomes 'as extreme or more so' than the one observed) and are therefore incompatible with the law of likelihood's implication of the 'irrelevance of the sample space' (section 1.11). This point is pursued in the next section, where a conspicuous problem with the interpretation of significance tests is also described. The existence of such problems supports the above claim that the reason why a plausible rationale for significance tests has not yet been found is that none exists.

3.4 Troubles with *p*-values

Let us look at the role of outcomes 'as extreme or more so' in significance tests. In problems where there is a well-defined alternative hypothesis, we can certainly identify such outcomes: if f_1 and f_2 are densities corresponding to the null and alternative distributions respectively, and if x_0 is the observation, then the set

that he has added some more white
favors that hypothesis over the
a factor that depends on how
added.
e is relative (as we saw in section
gainst one hypothesis can only be
alternative (see Exercise 3.1). This
ften for decades. Before the birth
the inventor of the *t*-test, W.S.
coauthor, Egon Pearson, that an
sample mean and a hypothesized

that the sample was not drawn
en if the chance is very small, say
if there is any alternative hypothesis
f the sample with a more reasonable
very much more inclined to consider
true.
set [1926], quoted in Pearson, 1938)

onale for significance tests, as
bility, is wrong. There is, in fact,
s. This is because they are incom-
od. Specifically, significance tests
robability distribution is spread
iple space (through their definition
or more so' than the one observed)
vith the law of likelihood's impli-
sample space' (section 1.11). This
on, where a conspicuous problem
cance tests is also described. The
pports the above claim that the
for significance tests has not yet

omes 'as extreme or more so' in
where there is a well-defined
rtainly identify such outcomes: if
nding to the null and alternative
x_0 is the observation, then the set

$\{x; f_2(x)/f_1(x) \geq f_2(x_0)/f_1(x_0)\} \equiv S(x_0)$ consists of all the outcomes that are 'as extreme or more so' compared to x_0. These are the outcomes that would give a likelihood ratio supporting H_2 over H_1 as great as or greater than the ratio associated with x_0.

The *p*-value, $\Pr_1(S(x_0))$, consists not only of the probability of what was observed (x_0), but also of the probabilities of all the more extreme outcomes that did not occur. But a proper measure of strength of evidence should not depend on the probabilities of unobserved values. To see this, recall the example in section 1.10, where 20 tosses were made with a coin whose probability of heads (success), θ, is unknown. The result is reported in a code that is known to you; I, on the other hand, know only the code word for '6'. If the number of heads observed is six, then you and I obtain precisely the same evidence about θ. Thus if we both consider $H_1: \theta = 0.5$ and an alternative asserting that the proportion is somewhat lower, say $H_2: \theta = 0.3$, then your prior probability ratio, $\Pr(H_2)/\Pr(H_1)$, and mine will both be increased by the same factor, 5.18. But our *p*-values for testing H_1 versus H_2 do not agree. Yours is $p_1(X = 6) + p_1(X = 5) + \ldots + p_1(X = 0) = 0.06$. On the other hand, since I can observe only '6' or 'not-6', the observed outcome is the most extreme possible one for me, and my *p*-value is just its probability, $p_1(X = 6) = 0.04$. The *p*-values assert (incorrectly) that the outcome (six heads in 20 tosses) is stronger evidence against H_1 (in favor of H_2) for me than it is for you.

The significance-test approach to measuring the evidence is wrong because its dependence on the sample space leads to different answers in situations where the evidence is the same. That is, it violates the principle of the 'irrelevance of the sample space' (section 1.11). This becomes even clearer if we provide some more details about this experiment. It turns out that you have memorized only the code-word for '6'. If any other result had occurred, you would have had to consult your code-book to find how many heads had been observed. Now, long after the experiment has been completed and the *p*-values have been published, we are storing some bent coins in your vault and we happen to notice that your code-book is missing.

So your situation was actually the same as mine – if $X = 4$ had occurred you could have recognized it only as 'not-6'. Therefore your sample space was the same as mine, {6, not-6}, and your calculated *p*-value, 0.06, is wrong. You conscientiously draft a letter to the journal where your result was published, apologizing for your error and reporting the corrected *p*-value, 0.04. But then your secretary,

when he sees the letter, sheepishly confesses that he threw away the code-book while tidying up the vault. Now the plot thickens. If he threw the book away *before* the outcome 'six successes in 20 tosses' was reported, then the appropriate p-value is the corrected one; but if the clean-up took place *later*, so that the code-book was still available when it might have been needed (but was not), then your sample space was $\{0, 1, \ldots, 20\}$ after all, and so your original p-value is still valid.

This is clearly silly – for the data actually observed, for the evidence actually obtained, the code-book was not needed. The evidence about θ, unlike the p-value, does not depend on when the book disappeared – that is, it does not depend on which sample space, $\{0, 1, \ldots, 20\}$ or $\{6, \text{not-}6\}$, you were sampling from when $X = 6$ was observed. (This example is a descendant of one constructed by Pratt (1961) that has figured prominently in modern discussions of the foundations of statistical inference.)

There is a significant piece of indirect evidence that something is seriously wrong with significance tests. According to the widely used 'Reasonable interpretations of the results of significance tests' described by Burdette and Gehan, and quoted earlier, a given p-value has a more or less fixed meaning. For example, a p-value between 1% and 5% is supposed to indicate 'moderate evidence against the null hypothesis'; a value less than 1% indicates 'very strong evidence'. This concept, that equal p-values represent equal amounts of evidence, at least approximately, was named the 'α-postulate' by Cornfield (1966). Fisher (1934, p. 182) states it as follows:

> It is not true... that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability [the p-value], we thereby make full allowance for the size of the sample, and should be influenced in our judgement only by the value of the probability indicated.

Berkson's (1942) statement was only slightly less forceful: 'the evidence provided by a small p correctly evaluated is broadly independent of the number in the sample'. The central role of significance tests in many research areas rests on the α-postulate – results with a p-value between 0.01 and 0.05 are flagged with an asterisk and declared to be 'statistically significant', while those with a p-value smaller than 0.01 are given two asterisks and declared 'highly significant'. The acceptability of a research report for publication often depends on whether key results are 'significant' or not.

y confesses that he threw away the
vault. Now the plot thickens. If he
he outcome 'six successes in 20
opropriate *p*-value is the corrected
lace *later*, so that the code-book
t have been needed (but was not),
, 1, . . . , 20} after all, and so your

e data actually observed, for the
code-book was not needed. The
alue, does not depend on when the
loes not depend on which sample
}, you were sampling from when
mple is a descendant of one con-
as figured prominently in modern
f statistical inference.)
indirect evidence that something is
ce tests. According to the widely
ns of the results of significance
id Gehan, and quoted earlier, a
fixed meaning. For example, a *p*-
supposed to indicate 'moderate
esis'; a value less than 1% indicates
cept, that equal *p*-values represent
ast approximately, was named the
). Fisher (1934, p. 182) states it as

asions cannot be drawn from small
used in calculating the probability
full allowance for the size of the
in our judgement only by the value

s only slightly less forceful: 'the
p correctly evaluated is broadly
e sample'. The central role of signi-
as rests on the α-postulate – results
).05 are flagged with an asterisk and
ificant', while those with a *p*-value
wo asterisks and declared 'highly
f a research report for publication
esults are 'significant' or not.

But the α-postulate is wrong. In their preface to the *New Cambridge Elementary Statistical Tables*, Lindley and Scott (1984, p. 3) explain:

> the interpretation to be placed on the phrase 'significant at 5%' depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one.

Thus a given *p*-value does not have a fixed meaning. If two experiments that are identical except for their sample sizes produce results with the same *p*-value, these results do not represent equally strong evidence against the null hypothesis – the evidence is stronger in the smaller experiment.

Ten of the world's most influential applied statisticians co-authored a paper in which they, too, explained that the α-postulate is false: 'A given *p*-value in a large trial is usually stronger evidence that the treatments really differ than the same *p*-value in a small trial of the same treatments would be' (Peto *et al.*, 1976, p. 593). Their interpretation is opposite that of Lindley and Scott.

Does a significance level of $p = 0.04$ indicate 'moderately strong' evidence against the null hypothesis, regardless of sample size, as the α-postulate and common practice imply? Or does it indicate stronger evidence in a small sample than in a large one, as Lindley and Scott state? Or does it indicate stronger evidence in a large sample as Peto *et al.* assert?

We should not be surprised to find that a statistical procedure that purports to measure evidence, but in a way incompatible with the law of likelihood, is mired in paradox and controversy (Royall, 1986; see also Morrison and Henkel, 1970).

3.5 Rejection trials

We have contrasted two ways to formulate statistical hypothesis-testing problems. The one developed by Neyman and Pearson addresses problems of choosing between two hypotheses, avoiding our central question of how to interpret statistical data as evidence. The other aims directly at out target – it seeks to measure the strength of evidence – but misses the mark. Neither of these formulations seems to capture the spirit of the definition given in the textbook from which many of today's practicing statisticians learned the basics:

> A procedure which details how a sample is to be inspected so that we may conclude that it either agrees reasonably with the hypothesis or

> does not agree with the hypothesis will be called a test of the hypothesis. (Dixon and Massey, 1969, p. 76)

Here a hypothesis test is seen as a decision procedure, *à la* Neyman–Pearson theory, but with some important differences. Although there are two possible conclusions or 'actions', only one hypothesis is mentioned. And the phrase 'either agrees reasonably with the hypothesis or does not' suggests that the two conclusions correspond to definite evidential interpretations of the sample. In many scientific applications of statistical tests a similar view is adopted. While the objective is a rule or procedure for choosing between two alternatives, as in the Neyman–Pearson paradigm, the two alternatives are now stated in terms of a single hypothesis – one is favorable to the hypothesis and the other unfavorable. And an essential part of the reasoning is that choosing the unfavorable conclusion is justified only when the sample represents sufficiently strong evidence against the hypothesis (as in the Dixon and Massey scenario when the sample does not agree 'reasonably' with the hypothesis).

This third formulation views statistical hypothesis testing as a process analogous to testing a proposition in formal logic via the argument known as *modus tollens*, or 'denying the consequent': if A implies B, then not-B implies not-A. We can test A by determining whether B is true. If B is false, then we conclude that A is false. But, on the other hand, if B is found to be true we cannot conclude that A is true. That is, A can be proven false by such a test, but it cannot be proven true – either we disprove A or we fail to disprove it. (This is the form of argument that is used in mathematics when a false proposition is disproved by a counter-example.) When B is found to be true, so that A survives the test, this result, although not proving A, does seem intuitively to be evidence supporting A. Whether this evidential interpretation is correct or not is the subject of Hempel's famous 'paradox of the ravens', which is discussed in the Appendix. This form of reasoning is at the heart of the philosophy of science, according to Popper (see Putnam, 1974). Its statistical manifestation is in this third formulation of hypothesis testing that we will call 'rejection trials'.

In applications of this third form of testing, a statistical hypothesis H_0, the 'null' hypothesis, plays a role analogous to that of the proposition A in that it can be disproved but not proved, rejected but not accepted (Noether, 1971, p. 64). Fisher (1966, section II.8)

hesis will be called a test of the
(Dixon and Massey, 1969, p. 76)

decision procedure, *à la* Neyman–
important differences. Although
is or 'actions', only one hypothesis
either agrees reasonably with the
gests that the two conclusions
interpretations of the sample. In
statistical tests a similar view is
a rule or procedure for choosing
the Neyman–Pearson paradigm,
ted in terms of a single hypothesis
thesis and the other unfavorable.
oning is that choosing the unfavor-
ly when the sample represents
ist the hypothesis (as in the Dixon
ample does not agree 'reasonably'

statistical hypothesis testing as a
roposition in formal logic via the
ens, or 'denying the consequent':
es not-A. We can test A by deter-
s false, then we conclude that A is
f B is found to be true we cannot
, A can be proven false by such a
ue – either we disprove A or we
form of argument that is used in
osition is disproved by a counter-
be true, so that A survives the
roving A, does seem intuitively to
ther this evidential interpretation
of Hempel's famous 'paradox of
in the Appendix. This form of
hilosophy of science, according to
statistical manifestation is in this
testing that we will call 'rejection

m of testing, a statistical hypothesis
a role analogous to that of the
disproved but not proved, rejected
., p. 64). Fisher (1966, section II.8)

explains:

> it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.... The notion of an error of the so-called 'second kind,' due to accepting the null hypothesis 'when it is false'... has no meaning with reference to simple tests of significance...

The experimenter identifies a rejection region R that has small probability under H_0, so that, if H_0 is true, then the event 'X is not in R' has high probability. This event, call it E, has a role analogous to that of the proposition B in the *modus tollens* argument. The analogy is imperfect because whereas A implied B (with perfect certainty), H_0 implies E with high probability; that is, if A is true then not-B is impossible, while if H_0 is true then not-E (X in R) is merely improbable. But the form of reasoning in the statistical version of the problem parallels that in deductive logic: if H_0 implies E (with high probability) then not-E justifies rejecting H_0.

The term 'test of significance' is often used, as it was in section 3.2, to refer to procedures that produce p-values for measuring the evidence against a hypothesis. The same term is also used, as it was in the above quote from Fisher's *Design of Experiments*, to refer to the (reject/do not reject) procedures just described. This latter usage seems more apt – here the hypothesis is subjected to a test. If it fails the test, it is rejected; if not, it survives, perhaps to face another test. We use the terms 'p-value procedures' and 'rejection trials' to distinguish between these two visions of 'tests of significance'.

Rejection trials are similar to Neyman–Pearson procedures in some respects. First, their objective is stated not as measuring the evidence against the hypothesis, but as choosing between two alternative actions, that is, choosing whether to reject the hypothesis or not. And like Neyman–Pearson tests, they require that an error rate α be selected and that a rejection region having probability no greater than α under H_0 be determined. Then H_0 is rejected if the observed value of the random variable falls in the region.

An important difference between rejection trials and Neyman–Pearson tests becomes clear when the observation does not fall in the rejection region: Neyman–Pearson tests require choosing

between two alternatives, H_0 and H_1, so that the complement of the region where H_0 is rejected (and H_1 accepted) is the region where H_1 is rejected (and H_0 accepted). Rejection trials, on the other hand, are viewed as challenges to the single null hypothesis H_0. If the observation is in the rejection region, then H_0 fails the challenge and is rejected; otherwise the result is 'Do not reject H_0'. That is, the symmetry described by Neyman (1950, p. 259), 'it is immaterial which of the two alternatives... is labelled the hypothesis tested', is clearly missing in the trials described by Fisher (1966, section II.8) 'in which the only available expectations are those which flow from the null hypothesis being true'. Thus Fisz (1963, p. 426) writes: 'In general, a significance test [rejection trial] allows us to make decisions only in one direction'. If the observation is in the rejection region 'then H_0 may be rejected', but if not 'then we can only state that the experiment does not contradict H_0'.

We are concerned in this monograph with how statistical data are interpreted as evidence. From this viewpoint the key difference between Neyman–Pearson tests and rejection trials is not in the existence, explicit or not, of an alternative statistical hypothesis, nor in the relationship between such an alternative and the null hypothesis. The key difference is that, unlike Neyman–Pearson tests, rejection trials entail evidential interpretation of the observations. In these trials the rejection of H_0 is justified when x falls in the rejection region, it is said, because such observations 'do not agree with' or 'do not fit' the hypothesis; they 'are inconsistent with', 'contradict', or even 'disprove' it. If under H_0 the probability of the rejection region is α, then the observations are said 'to provide sufficient evidence to cause rejection', or to be 'statistically significant' at level α. Whatever expression is used, the implication is that observations in the rejection region are evidence against the hypothesis; and observations in a rejection region with very small α are very strong evidence.

In section 2.3 we considered an example of Cox (1958) in which a coin toss is used to determine whether one or k i.i.d. $N(\theta, \sigma^2)$ observations will be made. There we looked at confidence intervals for θ. However, Cox's original example was stated in terms of hypothesis tests, and it dramatizes the difference between Neyman–Pearson tests and significance tests of the 'rejection-trial' variety. For simplicity let $\sigma^2 = 1$, and suppose the sample size when the coin falls tails is $k = 100$. The hypotheses are H_0: $\theta = 0$ and H_1: $\theta = 1$. Cox (1958) observed that if instead of using the

d H_1, so that the complement of
(and H_1 accepted) is the region
ccepted). Rejection trials, on the
nges to the single null hypothesis
:ejection region, then H_0 fails the
wise the result is 'Do not reject
:ribed by Neyman (1950, p. 259),
:wo alternatives... is labelled the
issing in the trials described by
:h the only available expectations
null hypothesis being true'. Thus
ıeral, a significance test [rejection
ns only in one direction'. If the
egion 'then H_0 may be rejected',
ate that the experiment does not

graph with how statistical data are
this viewpoint the key difference
and rejection trials is not in the
alternative statistical hypothesis,
such an alternative and the null
is that, unlike Neyman–Pearson
ıtial interpretation of the observa-
ı of H_0 is justified when x falls in
ıecause such observations 'do not
hypothesis; they 'are inconsistent
ove' it. If under H_0 the probability
ıe observations are said 'to provide
:tion', or to be 'statistically signi-
ression is used, the implication is
n region are evidence against the
a rejection region with very small

n example of Cox (1958) in which
: whether one or k i.i.d. $N(\theta, \sigma^2)$
: we looked at confidence intervals
example was stated in terms of
matizes the difference between
ificance tests of the 'rejection-trial'
: 1, and suppose the sample size
.00. The hypotheses are H_0: $\theta = 0$
erved that if instead of using the

coin toss we choose the sample size – say, n – deliberately, then the best (most powerful) test of size $\alpha = 0.05$ is to reject H_0 if and only if $\bar{x} > 1.645/\sqrt{n}$. So again we consider procedure A: if the coin falls heads, so $n = 1$, and $X = x$ is observed, reject H_0 if $x > 1.645$; if it falls tails and 100 observations are made, reject H_0 if $\bar{x} > 1.645/10$. Procedure A consists of using, for each sample size, 1 and 100, the best test of size 0.05. It has power $\frac{1}{2} \times 0.259 + \frac{1}{2} \times 1.000 = 0.63$. But again we can do better. The most powerful test of size 0.05 is given by procedure B: reject H_0 if the coin falls heads and $x > 1.282$ or if it falls tails and $\bar{x} > 5.078/10$. Procedure B's Type I error rate is $\frac{1}{2} \times 0.100 + \frac{1}{2} \times 0.000 = 0.05$, the same as A's, but its power is greater: $\frac{1}{2} \times 0.389 + \frac{1}{2} \times 1.000 = 0.69$.

For one whose problem is accurately represented by the Neyman–Pearson formulation, one who truly seeks to minimize the Type II error rate subject to the constraint that the Type I rate not exceed 0.05, it might come as a surprise that A is not the better procedure. But B's superiority, though surprising, is real, and there is no reason to prefer A. On the other hand, if the rejection-trial formulation is more apt, procedure B is not better – in fact, it is widely considered to be quite wrong. Cox (1958), calling procedure A the conditional test and B the unconditional one, wrote

> Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test... is in order. ... If, however, our objective is to say what we can learn from the data we have, the unconditional test is surely no good. Suppose that we know we have [only one] observation... The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution [i.e. we might make 100 observations instead of only one].

Procedure B is 'no good' because when only one observation is made it rejects at the 5% level whenever $X > 1.282$, and this is evidently too liberal – to properly claim 5% significance we should require, as A does, $X > 1.645$. Procedure B compensates 'on the average' by being overly conservative when $n = 100$, rejecting H_0 at the 5% significance level only on the basis of quite extreme outcomes, $\sqrt{n}\bar{X} > 5.078$. From the significance-testing viewpoint, Procedure B will not do because the objective is to characterize the evidence properly in each case; B allows the claim of 5% significance on the basis of evidence that is too

weak when $n = 1$, requiring evidence that is too strong when $n = 100$.

3.6 A sample of interpretations

The distinctions between the three views of hypothesis testing that we have considered are useful for understanding the rationale and interpretation of statistical tests. It is quite possible, however, that none of the three is a precise representation of what any one statistical author means by 'hypothesis testing'. The following quotations certainly do not represent a single viewpoint. Instead each author describes a slightly different vision, each drawing elements from all of the three formulations that we have tried to distinguish. But the point of view that we have called 'rejection trials' is influential in each description.

> In the testing process the null hypothesis either is rejected or is not rejected. If the null hypothesis is not rejected, we will say that the data on which the test is based do not provide sufficient evidence to cause rejection. (Daniel, 1991, p. 192)

> A nonsignificant result does not prove that the null hypothesis is correct – merely that it is tenable – our data do not give adequate grounds for rejecting it. (Snedecor and Cochran, 1980, p. 66)

> The verdict does not depend on how much more readily some other hypothesis would explain the data. We do not even start to take that question seriously until we have rejected the null hypothesis.
>
> ... The statistical significance level is a statement about *evidence* ... If it is small enough, say $p = 0.001$, we infer that the result is not readily explained as a chance outcome if the null hypothesis is true and we start to look for an alternative explanation with considerable assurance. (Murphy, 1985, p. 120)

> If [the p-value] is small, we have two explanations – a rare event has happened, or the assumed distribution is wrong. This is the essence of the *significance test* argument. Not to reject the null hypothesis ... means only that it is accepted for the moment on a provisional basis. (Watson, 1983)

> *Test of hypothesis.* A procedure whereby the truth or falseness of the tested hypothesis is investigated by examining a value of the test statistic computed from a sample and then deciding to reject or accept the tested hypothesis according to whether the value falls into the critical region or acceptance region, respectively. (Remington and Schork, 1970, p. 200)

ridence that is too strong when

ee views of hypothesis testing that
r understanding the rationale and
It is quite possible, however, that
representation of what any one
ypothesis testing'. The following
resent a single viewpoint. Instead
y different vision, each drawing
ormulations that we have tried to
ew that we have called 'rejection
ption.

pothesis either is rejected or is not
not rejected, we will say that the
o not provide sufficient evidence to
(Daniel, 1991, p. 192)

prove that the null hypothesis is
e – our data do not give adequate
Snedecor and Cochran, 1980, p. 66)

iow much more readily some other
i. We do not even start to take that
ejected the null hypothesis.
vel is a statement about *evidence*...
, we infer that the result is not readily
ie null hypothesis is true and we start
tion with considerable assurance.
(Murphy, 1985, p. 120)

two explanations – a rare event has
ution is wrong. This is the essence
Not to reject the null hypothesis...
the moment on a provisional basis.
(Watson, 1983)

whereby the truth or falseness of the
by examining a value of the test
ole and then deciding to reject or
rding to whether the value falls into
egion, respectively.
Lemington and Schork, 1970, p. 200)

> Although a 'significant' departure provides some degree of evidence against a null hypothesis, it is important to realize that a 'nonsignificant' departure does not provide positive evidence *in favour* of that hypothesis. The situation is rather that we have failed to find strong evidence against the null hypothesis. (Armitage and Berry, 1987, p. 96)

> If that value [of the test statistic] is in the region of rejection, the decision is to reject H_0; if that value is outside the region of rejection, the decision is that H_0 cannot be rejected at the chosen level of significance ... The reasoning behind this decision process is very simple. If the probability associated with the occurrence under the null hypothesis of a particular value in the sampling distribution is very small, we may explain the actual occurrence of that value in two ways; first we may explain it by deciding that the null hypothesis is false or, second, we may explain it by deciding that a rare and unlikely event has occurred. (Siegel and Castellan, 1988, Chapter 2)

3.7 The illogic of rejection trials

The above quotes suggest that the rejection trial is a method for determining when a given set of observations represents sufficiently strong evidence against a hypothesis to justify rejecting that hypothesis. But when it is given this interpretation the method defies the rules of logic.

Consider the $Bin(n, \theta)$ model for X and the hypothesis H_0: $\theta = \frac{1}{2}$. When the observed value is x, we are justified in rejecting H_0 at level α if $\Pr_0(X \geq x) \leq \alpha/2$. If, on the other hand, we are testing the hypothesis H_0': $\theta \leq \frac{1}{2}$, our observation x is strong enough evidence to justify rejecting if $\Pr_0(X \geq x) \leq \alpha$. Thus a value x for which $\alpha/2 < \Pr_0(X \geq x) \leq \alpha$ represents strong enough evidence to justify rejecting the composite hypothesis that either $\theta = \frac{1}{2}$ or $\theta < \frac{1}{2}$, but it is not strong enough evidence to justify rejecting the simple hypothesis that $\theta = \frac{1}{2}$. We may conclude (at significance level α) that both $\theta = \frac{1}{2}$ and $\theta < \frac{1}{2}$ are false, but we may not conclude that $\theta = \frac{1}{2}$ alone is false. We may conclude 'neither A nor B' but we may not conclude 'not-A'. Odd.

This interpretation of rejection trials makes no more sense if it is expressed in terms of the alternatives to the hypotheses tested. If, when we reject $\theta = \frac{1}{2}$, we are concluding that either $\theta < \frac{1}{2}$ or $\theta > \frac{1}{2}$, then clearly this is justified by any evidence that justifies the stronger conclusion that $\theta > \frac{1}{2}$. That is, if the evidence justifies the conclusion that A is true, then surely it justifies the weaker conclusion that either A or B is true. Rejection trials do not conform to this logic.

3.8 Confidence sets from rejection trials

Rejection trials provide the basis for an evidential approach to defining and interpreting confidence sets. If we have for each possible value of a parameter θ a level-α test of significance (rejection trial) of the hypothesis that the parameter equals that value, then we can define a $100(1-\alpha)\%$ confidence set. This set consists simply of all the values of θ that would not be rejected by the corresponding test. That is, if the hypothesis H_0: $\theta = \theta_0$ is not rejected on the basis of the observation $X = x$, then θ_0 is in the set $S(x)$. That this procedure does indeed produce a $100(1-\alpha)\%$ confidence set follows directly from the fact that for every θ_0 the random set $S(X)$ includes θ_0 if and only if a value of X is observed which does not lead to the rejection of H_0: $\theta = \theta_0$, and the probability of this, when H_0 is true, is at least $1-\alpha$. Thus $\Pr_\theta(S(X)$ will include $\theta) \geq 1-\alpha$ for every θ, which is to say, $S(X)$ is a valid $100(1-\alpha)\%$ confidence-set procedure.

This approach gives an explicit evidential interpretation to the confidence set, which now consists of all the values of θ that are consistent with the observation $X = x$ in the sense that this observation would not justify their rejection at significance level α. Values excluded from the confidence set are those against which $X = x$ represents evidence strong enough to warrant rejection at level α.

This interpretation is sometimes invoked in order to 'make sense' of a confidence set that seems paradoxical when interpreted in terms of one's confidence that it contains the true parameter value. A popular example is the confidence set for a ratio of two normal means (Exercise 2.3). The 95% confidence set can turn out to be the whole real line. Since this set contains all possible values of the ratio, it seems ridiculous to assign to it a confidence coefficient of only 0.95 – we are actually 100% confident that it contains the true ratio of means. The rejection-trial interpretation is attractive: the confidence set excludes only those values against which we have sufficiently strong evidence to justify rejection of the corresponding hypothesis at the 5% level. Now in this example the samples that give the entire line as the confidence set are those in which the estimates of both numerator and denominator are very close to zero. Such samples tell us very little about the ratio; as Exercise 7.6 shows, they represent only weak evidence. They do not justify our rejecting any of the possible values of the ratio. All of the values are 'consistent with the observations at the 5%

trials

s for an evidential approach to
ence sets. If we have for each
vel-α test of significance (rejection
rameter equals that value, then we
nce set. This set consists simply of
be rejected by the corresponding
H_0: $\theta = \theta_0$ is not rejected on the
ıen θ_0 is in the set $S(x)$. That this
a $100(1-\alpha)\%$ confidence set
ıat for every θ_0 the random set
a value of X is observed which
H_0: $\theta = \theta_0$, and the probability of
east $1-\alpha$. Thus $\Pr_\theta(S(X)$ will
which is to say, $S(X)$ is a valid
edure.
it evidential interpretation to the
ts of all the values of θ that are
$= x$ in the sense that this obser-
rejection at significance level α.
ence set are those against which
g enough to warrant rejection at

s invoked in order to 'make sense'
adoxical when interpreted in terms
ains the true parameter value. A
ce set for a ratio of two normal
confidence set can turn out to be
contains all possible values of the
n to it a confidence coefficient of
% confident that it contains the
n-trial interpretation is attractive:
those values against which we
ıce to justify rejection of the
5% level. Now in this example
ine as the confidence set are those
numerator and denominator are
tell us very little about the ratio;
sent only weak evidence. They do
the possible values of the ratio.
with the observations at the 5%

level', and this is what the (very large) confidence region correctly shows.

The evidential interpretation of confidence sets that is provided by the significance-testing (rejection-trial) approach is attractive. But it is valid only if the evidential interpretation of rejection trials is valid. And this is not the case, because the rationale for rejection trials is the same as that for p-value procedures – it rests on Fisher's disjunction, as explained by Watson and by Siegel and Castellan in the quotations in section 3.6. Rejection trials fail, as tools for evidential interpretation of statistical data, for the same reasons that p-value procedures fail. Rejection trials lead to different answers in situations where the evidence is the same, just as p-value procedures were shown to do in section 3.4. In terms of the urn example discussed there, whether the coded report of six successes in 20 tosses of the bent coin is or is not 'significant at the 5% level' for testing H_0: $\theta = \frac{1}{2}$ depends on whether the code-book would have been available if a different number of successes had occurred. The immediate problem is the dependence of the significance test procedures, of both the p-value and the rejection-trial varieties, on the sample space. The underlying reason, explained in section 3.3, is that the law of improbability is not tenable.

3.9 Alternative hypotheses in science

As we discussed in section 3.3, the law of likelihood applies to pairs of hypotheses and suggests that a sound theory of evidence in relation to a single statistical hypothesis is impossible. Unfortunately, the use of significance-testing methodology has trained many scientists as well as statisticians to think in terms of evidence against single hypotheses, as illustrated in the quotations in section 3.6. Since the problem can be formulated in terms of one hypothesis and a test statistic (as in the description by Cox and Hinkley in section 3.2), with no explicit alternative required, it is easy to overlook the essential role played by alternative hypotheses.

Are there statistical 'null' hypotheses that are scientifically important? If so, they are rare. The reason is the familiar observation that our statistical models are only approximations to real-world phenomena and processes. The answer to the question 'Is the null hypothesis correct?' is always the same – no! Does the odds ratio equal 1? No. Does the regression coefficient equal zero? No. Are the two distributions identical? No. If the purpose of experiments

were to answer such questions, there would be no point in doing experiments, since we already know the answers.

Experiments like the following are sometimes cited as counter-examples to the above claim. To test whether a subject is capable of extrasensory perception (ESP), a random sequence of images is generated but concealed from the subject. The images may be the cards in a well-shuffled deck or a sequence of zeroes and ones generated by a process such as tossing a coin. The subject is asked to reproduce the sequence. Early experiments of this sort were plagued by the possibility that subjects were given inadvertent cues to the correct responses (via normal sensory channels) or were able to cheat (Hansel, 1966). Let us assume that we can eliminate these flaws in the experimental setup. If the subject has no ESP ability then the number of terms that he correctly matches has a simple probability distribution that becomes the null hypothesis. Any departure from that distribution would show ESP ability. For simplicity, suppose the images are generated by a sequence of independent Bernoulli trials with probability $\theta = \frac{1}{2}$. If the subject's success probability is anything different from $\frac{1}{2}$, this is taken to reflect ESP. If his probability is truly greater than $\frac{1}{2}$, this clearly means that he is receiving some extrasensory information. But a probability less than $\frac{1}{2}$ means the same (and that he is misinterpreting the information). Any departure from the null hypothesis that his number of successes in n trials has a $Bin(n, \frac{1}{2})$ probability distribution proves the existence of ESP. It seems that we really do want to answer the question 'Is the null hypothesis true?'. If it is not, then ESP exists.

The problem, of course, is that no one can generate a perfect sequence of i.i.d. $Bernoulli(\frac{1}{2})$ trials. Certainly it cannot be done by tossing a coin, for all coins are imperfect and the probability of heads is never exactly one-half. Likewise, the subject who has no ESP ability, but is simply guessing, cannot produce a perfect sequence of i.i.d. $Bernoulli(\frac{1}{2})$ guesses. Then there is always some probability of error in recording and transmitting the results. This means that the null hypothesis is always false, whether or not the subject has ESP ability. The $Bin(n, \frac{1}{2})$ probability distribution is only an imperfect model for the number of matches observed in n trials.

The key question then becomes 'Does the probability distribution differ from the $Bin(n, \frac{1}{2})$ by more than can be reasonably explained in terms of the inevitable imperfections in the mechanism for generating the sequence of images, checking for matches, and recording the results?'. This question refers not only to the null hypothesis but also

ıere would be no point in doing
w the answers.
are sometimes cited as counter-
test whether a subject is capable
a random sequence of images is
subject. The images may be the
a sequence of zeroes and ones
ssing a coin. The subject is asked
y experiments of this sort were
subjects were given inadvertent
ia normal sensory channels) or
6). Let us assume that we can
imental setup. If the subject has
f terms that he correctly matches
on that becomes the null hypoth-
ribution would show ESP ability.
s are generated by a sequence of
probability $\theta = \frac{1}{2}$. If the subject's
erent from $\frac{1}{2}$, this is taken to reflect
ıter than $\frac{1}{2}$, this clearly means that
nformation. But a probability less
ıe is misinterpreting the informa-
ıll hypothesis that his number of
$\frac{1}{2}$) probability distribution proves
ıat we really do want to answer
true?'. If it is not, then ESP exists.
ıt no one can generate a perfect
s. Certainly it cannot be done by
imperfect and the probability of
Likewise, the subject who has no
ısing, cannot produce a perfect
esses. Then there is always some
and transmitting the results. This
always false, whether or not the
$n(n, \frac{1}{2})$ probability distribution is
ıumber of matches observed in n

'Does the probability distribution
ıan can be reasonably explained in
ons in the mechanism for generat-
ng for matches, and recording the
nly to the null hypothesis but also

to alternatives. Results leading to rejection of the null hypothesis at a very small p-value do not necessarily represent evidence for ESP. If $n = 100$ million and $x = 50.02$ million successes are observed then $2\sqrt{n}(\bar{x} - 0.5) = 4.0$, giving a very small p-value, 0.000 03. These observations are quite strong evidence for a success probability of 0.5002 versus 0.5000 ($LR > 2900$). But a difference this small, an excess of two expected successes per 10 000 trials, might well be explained in terms of imperfections in the experiment, and at any rate would appear to represent the absence of an empirically meaningful ESP phenomenon.

The meaningful question, as explained by Gossett in the quote in section 3.3, is not 'Are the observations evidence against the null hypothesis?' but 'Are there scientifically meaningful alternative hypotheses that are better supported?'.

3.10 Summary

Today's statistical practice is directed by an informal blending of Neyman–Pearson theory with concepts and interpretations that are not a part of that theory. We call this approach Fisherian. Scientific applications of hypothesis testing, for example, are usually of a type so different from the procedures described by Neyman–Pearson theory that they are given a special name, tests of significance. There are actually two distinct types of significance test, namely p-value procedures and rejection trials. Both explicitly attempt to do what Neyman–Pearson theory does not – to quantify the strength of statistical evidence. Significance tests fail in this endeavor because they rest on the faulty foundation of the law of improbability. Fisherian methods in general, as tools for representing and interpreting statistical data as evidence, fail for the same reason – they rest on the law of improbability and violate the law of likelihood.

Exercises

3.1 (a) Suppose you observe a random variable X and are interested in the simple hypothesis H_0: $X \sim Bin(100, 0.5)$. Is the observation $X = 37$ strong evidence against H_0? How about $X = 50$? Explain. [Some numbers that you might want to consider are: $\Pr_0(X = 37) = 0.003$, $\Pr_0(X \leq 37) = 0.006$; $\Pr_0(X = 50) = 0.080$, $\Pr_0(X \leq 50) = 0.540$.]

(b) Now suppose you learn that X was produced by making 100 draws from an urn containing 100 balls, 50 black and 50 white, and counting the number of draws on which a black ball was seen. The hypothesis H_0 in (a) is true if the draws were made *with* replacement. Is the observation $X = 37$ strong evidence against H_0 *vis-à-vis* the alternative hypothesis H_1 stating that the draws were made *without* replacement? How about $X = 50$?

(c) Consider another alternative, H_2, stating that the draws were with replacement, but that only 25 of the 100 balls are black. Is the observation $X = 37$ strong evidence against H_0 *vis-à-vis* H_2? [$\Pr_2(X = 37) = 0.002$, $\Pr_2(X \leq 37) = 0.997$.]

3.2 Verify that for n observations i.i.d. $N(\theta, \sigma^2)$, with σ^2 known, the 1/8 likelihood interval for θ is $\bar{x} \pm 2.039\sigma/n^{1/2}$ and that this is a 95.9% confidence interval. Find the 1/32 likelihood interval and its confidence coefficient.

3.3 One form of reasoning that is sometimes used in efforts to give confidence intervals an evidential interpretation is as follows: The fact that a confidence interval procedure rarely results in the true value's being excluded implies that when a value is excluded, there is strong evidence that it is not the true one. Use the example in Exercise 2.4 to show that this reasoning is faulty.