CHAPTER 2

Neyman–Pearson theory

2.1 Introduction

Much of contemporary statistical practice consists of using the methods of hypothesis testing, estimation, and confidence intervals in order to represent and interpret the evidence in a given set of observations. These same methods are used for other purposes as well, but here we are concerned only with their role in interpreting observed data as evidence, as typified by their conventional use in research reports in scientific journals. In particular, we are concerned with the rationale behind such applications. The most widely taught statistical theory, which is based on a paradigm of Neyman and Pearson (1933), explicitly views these statistical methods as solutions to problems of a different kind, so that these evidential applications fall outside the scope of that theory. In this chapter we describe the Neyman–Pearson theory and look at problems that arise when its results are used for interpreting data as evidence.

2.2 Neyman–Pearson statistical theory

At the heart of Neyman–Pearson theory is the problem of testing two simple hypotheses, which was considered briefly in section 1.8. In Chapter 1 we examined a rule for interpreting observations $X = x$ as evidence for one hypothesis *vis-à-vis* another. Neyman–Pearson theory is not concerned with such interpretations; instead, its focus is on using the observations to make a choice between the two hypotheses. In the words of Neyman (1950, p. 258):

> The problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B...

He goes on to explain that he is considering situations when the desirability of actions A and B depend on the unknown probability distribution of a random variable X, and our action is to be

determined by the observed value of X. Action A is preferred if the distribution belongs to one set of possible distributions for X and B is preferred if it belongs to another set:

> *any rule R prescribing that we take action A when the sample point ... falls within a specified category of points, and that we take action B in all other cases, is a test of a statistical hypothesis.*
>
> (Neyman, 1950, p. 258)

He then lets H denote the set of distributions where action A is preferred, and $\bar{H}$ the set where B is preferred.

> The choice between the two actions A and B is interpreted as the *adoption* or the *acceptance* of one of the hypotheses H or $\bar{H}$ and the *rejection* of the other. Thus, if the application of an adopted rule... leads to action A, we say that the *hypothesis H is accepted* (and, therefore $\bar{H}$ is rejected). On the other hand, if the application of the rule leads to action B, we say that the *hypothesis H is rejected* (and, therefore, the hypothesis $\bar{H}$ is accepted). Frequently it is convenient to concentrate our attention on a particular one of the two hypotheses H and $\bar{H}$. To do so, one of them is called the *hypothesis tested.* The outcome of the test is then reduced to either accepting or rejecting the hypothesis tested. Plainly it is immaterial which of the two alternatives H and $\bar{H}$ is labelled the hypothesis tested. (Neyman, 1950, p. 259)

Neyman then warns against interpreting the result of a test to mean anything except a decision to choose one action or the other:

> The terms 'accepting' and 'rejecting' a statistical hypothesis are very convenient and are well established. It is important, however, to keep their exact meaning in mind and to discard various additional implications which may be suggested by intuition. Thus, to accept a hypothesis H means only to decide to take action A rather than action B. This does not mean that we necessarily believe that the hypothesis H is true. Also if the application... 'rejects' H, this means only that the rule prescribes action B and does not imply that we believe that H is false. (Neyman, 1950, p. 259)

Here Neyman places his theory squarely in the domain of the second of the physician's three questions of Chapter 1, 'What should I *do*?'. He is careful to deny explicitly that it is intended to answer the first question, 'What do I believe?', while ignoring altogether the third question, the one that we are concerned with, 'How should I interpret this observation as evidence?'.

How are these decision rules to be evaluated? What criteria determine whether one test is better than another? The view of the Neyman–Pearson school is that a statistical test procedure should be evaluated in terms of its error probabilities, i.e. the probability

f X. Action A is preferred if the
)ssible distributions for X and B
set:

ction A when the sample point ...
ints, and that we take action B in
l hypothesis.
(Neyman, 1950, p. 258)

distributions where action A is
preferred.

, and B is interpreted as the *adop-*
potheses H or $\bar{H}$ and the *rejection*
ı of an adopted rule... leads to
H is accepted (and, therefore $\bar{H}$
e application of the rule leads to
H is rejected (and, therefore, the
ly it is convenient to concentrate
f the two hypotheses H and $\bar{H}$.
ypothesis tested. The outcome of
:pting or rejecting the hypothesis
)f the two alternatives H and $\bar{H}$ is
ıan, 1950, p. 259)

eting the result of a test to mean
:e one action or the other:

a statistical hypothesis are very
l. It is important, however, to
nd to discard various additional
l by intuition. Thus, to accept a
e to take action A rather than
we necessarily believe that the
application... 'rejects' H, this
ıction B and does not imply that
1950, p. 259)

squarely in the domain of the
uestions of Chapter 1, 'What
ıy explicitly that it is intended
t do I believe?', while ignoring
)ne that we are concerned with,
ation as evidence?'.
e evaluated? What criteria deter-
han another? The view of the
statistical test procedure should
)robabilities, i.e. the probability

of rejecting H when it is true, and the probability of accepting H when $\bar{H}$ is true. A good test is one with small error probabilities. In the simple-versus-simple case these are just the Type I and Type II error probabilities, α and β. If two tests have the same α, then the one with the smaller β is the better test. The fundamental lemma of Neyman and Pearson (1933) shows how, for any fixed value of α, to find the best test, the one with smallest β. If we use such a test then we still risk making a Type I error, but we have controlled that risk at α. And we can be sure that any test with a smaller Type II risk than ours carries a larger Type I risk.

The Neyman–Pearson theory of hypothesis testing, with its attractive pragmatic focus on minimizing the probabilities of making errors, provides a powerful paradigm that dominates contemporary statistical theory. Wald (1939; 1950) made basic generalizations showing that much of the rest of statistics could be modelled after the optimal decision-making approach of the Neyman–Pearson theory of hypothesis testing. For this reason the general theory is sometimes called the Neyman–Pearson–Wald theory (e.g. Basu, 1975; Carnap, 1950; Efron, 1986), and it views the basic subject matter of statistics as a collection of decision-making problems that are analogous to the hypothesis-testing problem in that they are formulated in terms of choosing between alternative actions. In the hypothesis-testing problem there are only two actions, corresponding to the two hypotheses. In estimation, the actions correspond to values of the parameter being estimated; the goal is to choose a value that is close to the true parameter. And in the confidence interval problem the actions correspond to sets of parameter values, the goal being to choose a set that contains the true value.

So according to the Neyman–Pearson–Wald formulation, statistics is primarily concerned with using observations to choose from a specified set of actions, the desirability of the actions being dependent on which probability distribution is generating the observations. Neyman's expression for this process is **inductive behavior**: 'If a rule R unambiguously prescribes the selection of action for each possible outcome..., then it is a rule of inductive behavior' (Neyman, 1950, p. 10). In his view, the generalized Neyman–Pearson theory encompasses the whole of statistics:

> **Scope of Mathematical Statistics.** *Mathematical statistics is a branch of the theory of probability. It deals with problems relating to performance characteristics of rules of inductive behavior based on random experiments.* (Neyman, 1950, p. 11)

This extravagant view of the scope of Neyman–Pearson theory has been widely accepted:

> In recent years, Statistics has been formulated as the science of decision making under uncertainty. This formulation represents the culmination of many years of development and, for the first time, furnishes a simple and straightforward method of exhibiting the fundamental aspects of a statistical problem. (Chernoff and Moses, 1959, p. vii)

And it remains fundamental – a recent course announcement for a basic statistical theory course at my own university (Johns Hopkins) explained that 'Statistics is the science of using data to make decisions'.

Neyman–Pearson theory formulates a statistical problem in terms of choosing from among a specified set of actions. A solution is a *procedure for choosing* an action (a 'rule of inductive behavior'), a protocol that specifies for every possible value of the random variable X whose probability distribution is under study, what action is to be taken if that value is observed. A solution to a testing problem may take the form 'Choose H_2 if $\bar{X} \geq 7$; otherwise choose H_1'. A solution to an estimation problem might be 'Estimate θ by $\bar{X}$' or '... by $\sum(X_i - \bar{X})^2/n$'.

The basic tenet of Neyman–Pearson theory is that solutions to statistical problems, that is, statistical procedures, should be evaluated in terms of their probabilistic properties ('performance characteristics', in Neyman's words). These properties measure the expected, or long-run average, performance of the procedures – a procedure with good probabilistic properties will, if used repeatedly, give good performance, on average. In the simple-versus-simple hypothesis-testing problem, procedures are evaluated in terms of their Type I and Type II error probabilities. An estimation procedure, or estimator, associates with every possible observation x an estimate $t(x)$ of the unknown parameter. If θ denotes this parameter and $X = x$ is observed, then $t(x)$ is used as an estimate of θ. The probabilistic properties that are most popular for evaluating estimators are the expected error, or bias, $\mathrm{E}[t(X) - \theta]$, the variance, $\mathrm{var}[t(X)]$, and the expected squared error, $\mathrm{E}[t(X) - \theta]^2$. A confidence interval procedure associates with every x an interval, $(\ell(x), u(x))$ of parameter values, and two key properties are the probability that the interval will contain the true value of the parameter, $\Pr(\ell(X) < \theta < u(X))$, and the expected width of the interval, $\mathrm{E}[u(X) - \ell(X)]$.

To illustrate the estimation theory we can again consider repeated independent draws from an urn. If X is the number of white balls in

of Neyman–Pearson theory has

nulated as the science of decision
ulation represents the culmina-
nd, for the first time, furnishes a
of exhibiting the fundamental
rnoff and Moses, 1959, p. vii)

ent course announcement for a
own university (Johns Hopkins)
of using data to make decisions'.
tes a statistical problem in terms
l set of actions. A solution is a
'rule of inductive behavior'), a
sible value of the random vari-
on is under study, what action
erved. A solution to a testing
oose H_2 if $\bar{X} \geq 7$; otherwise
ion problem might be 'Estimate

son theory is that solutions to
istical procedures, should be
ilistic properties ('performance
). These properties measure the
ormance of the procedures – a
operties will, if used repeatedly,
e. In the simple-versus-simple
ures are evaluated in terms of
abilities. An estimation proce-
very possible observation x an
eter. If θ denotes this parameter
used as an estimate of θ. The
t popular for evaluating estima-
ias, $E[t(X) - \theta]$, the variance,
d error, $E[t(X) - \theta]^2$. A confi-
es with every x an interval,
nd two key properties are the
ntain the true value of the para-
e expected width of the interval,

we can again consider repeated
X is the number of white balls in

ten draws, then to estimate the proportion of white balls in the urn, θ, we might first consider the estimator $t_1(X) = X/10$, which estimates θ by the proportion of draws that produce white balls. This estimator is unbiased, $E[t_1(X) - \theta] = 0$, and its variance and expected squared error are both equal to $\theta(1 - \theta)/10$. An alternative estimator is $t_2(X) = \frac{1}{2}$, which simply ignores X and estimates θ to equal $\frac{1}{2}$ regardless of the value of X. This estimator has a bias, of course, $E[t_2(X) - \theta] = \frac{1}{2} - \theta$. But its variance is small, $\text{var}[t_2(X)] = 0$, and its expected squared error is $(\frac{1}{2} - \theta)^2$. A third competitor is $t_3(X) = (1 - w)t_1(X) + \frac{1}{2}w$, where $w = 1/(1 + 10^{1/2})$ or about 0.24. This estimator represents a compromise between t_1 and t_2. Its bias is $E[t_3(X) - \theta] = w(\frac{1}{2} - \theta)$, its variance is $\theta(1 - \theta)w^2$, and its expected squared error is simply $(\frac{1}{2}w)^2$.

In terms of bias, t_1 is the best of the three estimators; in terms of the variance t_2 is best; and in terms of the expected squared error, Figure 2.1 shows that t_1 is best if θ is less than 0.17 or greater than 0.83, t_2 is best if $0.38 < \theta < 0.62$ (but much the worst if θ is close to zero or one), while t_3 is best for the remaining values of θ, $0.17 < \theta < 0.38$ and $0.62 < \theta < 0.83$.

This situation is typical – there is no best procedure. One is best with respect to one performance measure, but for a different criterion another procedure is best, while for a third criterion, one procedure is better for some values of the unknown parameter and another is better for other values. Here $t_1(X)$ happens to be the best,

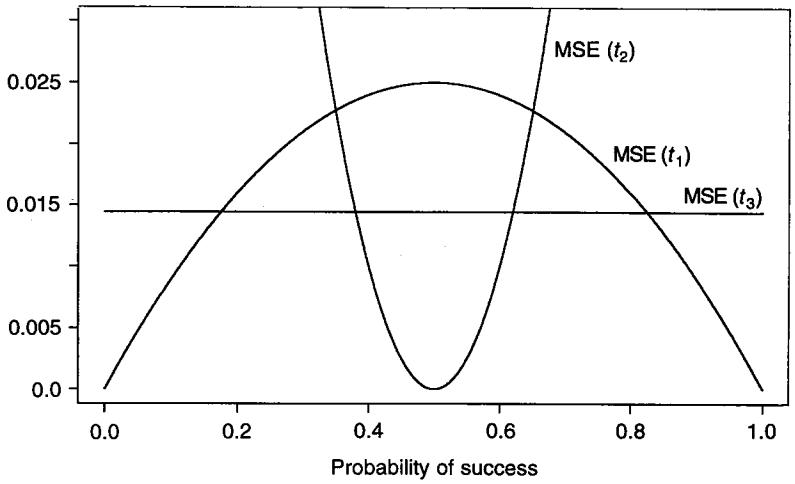

Figure 2.1 *Mean square errors for three estimators of a binomial probability.*

in one sense, among all of the estimators that are unbiased – no other unbiased estimator has smaller variance at any value of θ. How good $t_1(X)$ is, as measured by its variance or its expected squared error, $\theta(1-\theta)/10$, depends on the unknown value of θ. By contrast, the expected squared error of $t_3(X)$ is $(\frac{1}{2}w)^2$, the same for all θ (Figure 2.1). Now $t_3(X)$ is the best of all possible estimators with respect to an important property, the maximum possible value of the expected squared error: the use of $t_3(X)$ ensures that the expected squared error does not exceed $(\frac{1}{2}w)^2$, regardless of the true value of θ. Every other estimator has some values of θ at which the expected squared error exceeds this bound. The estimator $t_2(X) = \frac{1}{2}$ is the best possible estimator if it happens that $\theta = \frac{1}{2}$, and it is much better (smaller expected squared error) than either of the other two if θ is very close to $\frac{1}{2}$.

The user must decide whether the properties of $t_1(X)$, $t_2(X)$, or $t_3(X)$ are more important for his or her particular problem. Although a procedure may be disfavored because it lacks a property that is judged to be important (such as unbiasedness or minimizing the maximum possible mean squared error), neither procedure can be described as 'incorrect' or 'invalid'. For estimating the probability of 'heads' for a coin in my pocket, on the basis of ten tosses, I would use $t_2(X) = \frac{1}{2}$ in preference to t_1 or t_3, because I know that the actual probability is very close to one-half, certainly between 0.4 and 0.6, and in this range t_2 is much the best estimator of the three. On the other hand, for estimating the proportion of white balls in an urn about which I have no prior knowledge, t_1 and t_3 might both be more attractive than t_2.

A subtle distinction proves to be critical: the probabilistic properties that Neyman–Pearson theory uses to evaluate a decision procedure are properties of the procedure, not of its results. For example, a test procedure might have a Type I error probability of 0.05. This means that if H_1 is true then the probability that this test procedure will reject H_1 is only 0.05. It does not mean that if the procedure has rejected H_1 then the probability that a Type I error has been committed is 0.05. The probability, 0.05, refers to the test procedure, not to an outcome of the procedure. Having rejected H_1, we know that either H_1 is true and we have committed a Type I error, or H_1 is false and we have made the correct decision; but we do not know which.

Similarly, a 95% confidence interval procedure has probability 0.95 of generating an interval that will contain the true value of the target parameter, say θ. For example, if X_i, $i = 1, 2, \ldots, n$, are

mators that are unbiased – no
ler variance at any value of θ.
by its variance or its expected
s on the unknown value of θ.
rror of $t_3(X)$ is $(\frac{1}{2}w)^2$, the same
ıe best of all possible estimators
ty, the maximum possible value
use of $t_3(X)$ ensures that the
xceed $(\frac{1}{2}w)^2$, regardless of the
ıator has some values of θ at
ceeds this bound. The estimator
or if it happens that $\theta = \frac{1}{2}$, and it
uared error) than either of the

e properties of $t_1(X)$, $t_2(X)$, or
is or her particular problem.
ored because it lacks a property
as unbiasedness or minimizing
d error), neither procedure can
ılid'. For estimating the prob-
y pocket, on the basis of ten
eference to t_1 or t_3, because I
very close to one-half, certainly
ge t_2 is much the best estimator
ır estimating the proportion of
I have no prior knowledge, t_1
e than t_2.
ritical: the probabilistic proper-
ses to evaluate a decision proce-
, not of its results. For example,
I error probability of 0.05. This
bability that this test procedure
t mean that if the procedure has
ıt a Type I error has been com-
5, refers to the test procedure,
. Having rejected H_1, we know
committed a Type I error, or
correct decision; but we do not

rval procedure has probability
will contain the true value of
ample, if X_i, $i = 1, 2, \ldots, n$, are

independent and identically distributed (i.i.d.) $N(\theta, 1)$ then $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$ defines a 95% confidence interval procedure. The interval is random, depending on the random variable $\bar{X}$; it will contain θ if and only if $\bar{X}$ falls in the interval $(\theta - 1.96/\sqrt{n}, \theta + 1.96/\sqrt{n})$, and since $\bar{X}$ has a $N(\theta, 1/n)$ distribution the probability that this will occur is 0.95. But after $\bar{X}$ has been observed to equal $\bar{x}$ and we have the specific interval, $(\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n})$, the probability statement no longer applies – either θ is in this interval or it is not, and we do not know which. Both the interval and the parameter θ are fixed, not random. We can say that we used a procedure that generates an interval containing θ 95% of the time, and that in this instance it generated the interval $(\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n})$. But we cannot say that the probability that θ is in *this* interval is 0.95. That this is not semantic hair-splitting is illustrated by the confidence interval example in the next section, while Exercise 2.3 gives an example of a 95% confidence interval that contains all the possible values of θ!

2.3 Evidential interpretation of the results of Neyman–Pearson decision procedures

Contrary to the views quoted in the preceding section, many statistical applications in research are not well represented by the Neyman–Pearson model of choosing between alternative actions. And in addressing these applications many statisticians explicitly reject that formulation, instead describing the problems in terms of 'inductive reasoning' (Fisher, 1959, p. 109), representing 'what the data say' (Cox, 1958, p. 359), finding an 'index to or measure of weight of evidence' (Cornfield, 1966, p. 18), 'summarization of evidence' (Cox and Hinkley, 1974, p. 56), etc. Nevertheless, one approach to such applications employs many of the same statistical tools, methods, and even concepts as Neyman–Pearson theory. The same general problem areas are identified – hypothesis testing, estimation, and confidence intervals – and many of the same tests, estimators, and confidence interval procedures are used. Moreover, the procedures are evaluated, just as in Neyman–Pearson theory, in terms of their probabilistic properties – size, power, bias, variance, etc.

In these applications the probabilistic properties not only determine which procedure will be used, as envisioned by the Neyman–Pearson theory; after a procedure has been used its properties are reported alongside the result. Thus in addition to the fact that our test, when applied to the data observed in this experiment, leads

us to choose H_2, our published report also describes the test's size and, sometimes, its power. In addition to the value of the estimate, we report that the estimator (i.e. the estimation procedure used to derive this estimate) is unbiased, if it is, and we state the (estimated) value of its standard error. In addition to describing the confidence interval that our observations produced, we report the confidence coefficient, the coverage probability of the procedure that generated this particular interval.

The reason for reporting not only the result but also the probabilistic properties of the procedure leading to that result is that the procedure is actually being used, not to choose an action, but to indicate 'what the data say', that is, to interpret the data as evidence. The probabilistic properties are used to refine and quantify this interpretation. In these applications, when a test procedure leads to the rejection of H_1 it is not really taken to mean that we or anyone else actually decides to act as if H_1 were false. Rather 'reject H_1' is interpreted as a figure of speech meaning that the data in question are evidence against H_1; and the error probability α, or the two probabilities α and β, are interpreted as somehow measuring the strength of that evidence.

Likewise, although an estimated treatment effect in a published report of a clinical trial might represent a decision to act as if the treatment really has that precise effect, it more commonly is interpreted to mean that the observations are evidence that the effect has approximately the size of the estimate, and the standard error of the estimation procedure is supposed to show how strong the evidence is, a large standard error indicating that it is weak.

A confidence interval is also commonly given an evidential interpretation. In fact it is sometimes recommended that not one but a system of confidence intervals be reported, one for each of a series of confidence coefficients, such as 80%, 90%, 95%, and 99%. This system is not interpreted as some (very complicated) action, in the Neyman–Pearson sense. Rather its interpretation is evidential – it 'summarizes what the data tell us about θ, given the model' (Cox and Hinkley, 1974, p. 227).

Such applications and interpretations, although formally outside the scope of Neyman–Pearson theory, are not entirely unauthorized; one of the leading expositors of the Neyman–Pearson school acknowledges the use of confidence intervals for 'indicating what information is available concerning the unknown parameter' (Lehmann, 1959, p. 4). And Neyman (1976, p. 749) himself wrote that when the two possible results of a hypothesis test are described

ɔort also describes the test's size
ɪion to the value of the estimate,
he estimation procedure used to
it is, and we state the (estimated)
tion to describing the confidence
duced, we report the confidence
/ of the procedure that generated

ɪly the result but also the prob-
re leading to that result is that
ɔd, not to choose an action, but
hat is, to interpret the data as
es are used to refine and quantify
ions, when a test procedure leads
taken to mean that we or anyone
were false. Rather 'reject H_1' is
eaning that the data in question
error probability α, or the two
ɛted as somehow measuring the

l treatment effect in a published
resent a decision to act as if the
ffect, it more commonly is inter-
ons are evidence that the effect
ɛstimate, and the standard error
ɔposed to show how strong the
indicating that it is weak.
commonly given an evidential
mes recommended that not one
ɪls be reported, one for each of
such as 80%, 90%, 95%, and
ted as some (very complicated)
ɛnse. Rather its interpretation is
ɪe data tell us about θ, given the
227).
ɪtions, although formally outside
eory, are not entirely unauthor-
of the Neyman–Pearson school
ce intervals for 'indicating what
ning the unknown parameter'
ɪan (1976, p. 749) himself wrote
of a hypothesis test are described

with reference to a single hypothesis H, namely (1) reject H and (2) do not reject H, 'My own preferred substitute for "do not reject H" is "no evidence against H is found" '.

For these evidential applications the distinction between a statistical procedure and a result of using that procedure is critical. Column 1 in Table 2.1 lists some of the basic procedures pertaining to an unknown parameter θ. They are defined in terms of a generic random variable X and are themselves random. Column 2 lists some of their important probabilistic properties. Column 3 represents the time when a realized value x of the random variable X is observed, and column 4 shows the result that is generated by the statistical procedure when X is observed to take the value x. Finally, column 5 contains examples of statements in which the properties of the procedures are used for interpreting the observation x as evidence about θ.

The statements in the second column concern well-defined probabilistic properties of random variables and are subject to rigorous mathematical verification. The meanings of the statements in the fifth column are less clear. Yet these statements, and others like them, are an integral part of today's dominant statistical methodology, which uses tests, estimates, and confidence intervals for interpreting and representing statistical observations as evidence. Here is how Pratt (1961) described one evidential interpretation of confidence intervals:

> What has made the confidence interval popular is 'indicating what information is available'. Decision problems seem beside this point; a confidence interval probably contains the parameter, and the confidence coefficient measures how probably. But does it? By the formal definition, it no longer does, once we insert numerical values for the endpoints. Then no probability (except 0 or 1) can be attached to the event that the interval contains the parameter: either it does or it doesn't. Unfortunately we don't know which. We think, and would like to say, it 'probably' does; we can invent something else to say, but nothing else to think. We can say to an experimenter, 'A method yielding true statements with probability .95, when applied to your experiment, yields the statement that your treatment effect is between 17 and 29, but no conclusion is possible about how probable it is that your treatment effect is between 17 and 29'. The experimenter, who is interested not in the method, but in the treatment and this particular confidence interval, would get cold comfort from that if he believed it.

Thus, although the Neyman–Pearson theory of confidence intervals stops at the fourth column of Table 2.1, typical applications

Table 2.1 *Properties of statistical procedures applied to results of the procedures for evidential interpretation of observations*

Procedure (depends on random variable X)	Property of procedure (determined by probability distribution of X)	Observation (realized value of random variable X)	Result of procedure (fixed by observation)	Evidential interpretation (property used to interpret observation)
Confidence interval $(\ell(X), u(X))$	We can be 0.95 confident that the random interval $(\ell(X), u(X))$ will contain θ. (Confidence coefficient: $\Pr(\ell(X) < \theta < u(X)) = 0.95$)	$X = x$	An interval $(\ell(x), u(x))$	The observation x is evidence that θ is in $(\ell(x), u(x))$. Large confidence coefficient means strong evidence.
Estimator $t(X)$	The expected value of $t(X)$ is θ and its standard error is σ. ($\mathrm{E}[t(X)] = \theta$; $\mathrm{var}[t(X)] = \sigma^2$)	$X = x$	An estimate, $t(x)$	The observation x is evidence that θ is near $t(x)$. The smaller σ, the stronger the evidence.
Hypothesis test $H_{\delta(x)}$	Type I and Type II error probabilities are α and β	$X = x$	A hypothesis, $H_{\delta(x)}$ (H_1 if $\delta(x) = 1$; H_0 if $\delta(x) = 0$)	The observation x is evidence in favor of the hypothesis $H_{\delta(x)}$. Small α and β mean strong evidence.

in scientific investigation and reporting, where the objective is to represent the evidence in a given set of observations, proceed to the fifth column.

Is it valid to use Neyman–Pearson theory in this way, interpreting the procedures of hypothesis testing, estimation, and confidence intervals as techniques for representing 'what the data say'? For instance, if a good procedure for testing H_1 versus H_2 leads to acceptance of H_2, does this mean that the data are evidence supporting H_2 over H_1? If a good estimation procedure leads to a specific estimate, does it mean that the data are evidence that the parameter lies near that value? If a good confidence interval procedure leads to the interval (a, b), does it mean that the data are evidence that the parameter lies between a and b? And do the probabilistic properties of the procedures, the error probabilities, standard errors, confidence coefficients, etc., measure the strength of the evidence? All of these questions have the same simple answer – no.

We have already seen in section 1.8 that the above evidential interpretation of Neyman–Pearson test results is not valid – a good test, one whose error probabilities are both very small, can call for choosing H_1 when the evidence favors H_2 and vice versa.

Randomized tests furnish another example of the problems that appear when we try to interpret Neyman–Pearson test procedures as showing 'what the data say'. Suppose we make five draws from an urn in which either (H_1) half of the balls are white or (H_2) three-fourths are white, replacing the ball after each draw. Let X represent the number of white balls that we observe. If we reject H_1 whenever $X = 5$ we will have the best (most powerful) test of size $\alpha = p_1(X = 5) = 1/32$, and if we reject whenever $X = 4$ or $X = 5$ we will have the best test of size $\alpha = 6/32$. If we want the best test having size $\alpha = 0.05$, we must use a randomized test; it calls for rejecting H_1 whenever $X = 5$ as well as rejecting sometimes, but not always, when $X = 4$. Specifically, when $X = 4$ it rejects 12% of the time. We might carry out such a test as follows: if $X = 5$, reject H_1; if $X = 4$, choose a random number U between 0 and 1, and, if $U \leq 0.12$, reject H_1; otherwise accept H_1. Our test has the desired size $\alpha = p_1(X = 5) + 0.12p_1(X = 4) = 0.05$, and the fundamental lemma of Neyman and Pearson assures us that there is not a better one – among all tests that have size 0.05 or less, ours has the smallest possible Type II error probability. In particular, ours has smaller Type II error probability than any non-randomized test with $\alpha \leq 0.05$.

Hypothesis test $H_{\delta(x)}$	Type I and Type II error probabilities are α and β	$X = x$	A hypothesis, $H_{\delta(x)}$ (H_1 if $\delta(x) = 1$; H_0 if $\delta(x) = 0$)	The observation x is evidence in favor of the hypothesis $H_{\delta(x)}$. Small α and β mean strong evidence.

and its standard error is σ, ($\mathrm{E}[t(X)] = \theta$; $\mathrm{var}[t(X)] = \sigma^2$) evidence that θ is near $t(x)$. The smaller σ, the stronger the evidence.

But despite its optimality, most statisticians would be reluctant to use this test in applications where the purpose of the statistical analysis is to indicate what the observations say about H_1 *vis-à-vis* H_2. They properly sense that whatever $X = 4$ means as evidence, it is not affected by whether $U \leq 0.12$ or not, and to let their assessment of the evidence depend on this clearly irrelevant event is inappropriate. The evidence about the unknown proportion of the white balls in the urn is the same when $X = 4$ and the test calls for rejecting H_1 as it is when $X = 4$ and the test calls for accepting.

Straightforward evidential interpretation of Neyman–Pearson confidence intervals is also invalid. This is illustrated in the following example, which is derived from a famous one of Cox (1958, p. 360). It concerns an experiment that is conducted in two stages. At the first stage we simply toss a coin; the outcome of the toss determines what happens at the second stage. If the coin falls heads then we observe a random variable X with a $N(\theta, \sigma^2)$ probability distribution. But if the coin falls tails we observe k independent random variables $X_1, X_2, \ldots, X_k$, all having the same $N(\theta, \sigma^2)$ distribution. That is, at the second stage we make either one or k observations, depending on the result of the coin toss. The value of σ^2 is known and we want a 95% confidence interval for θ with short expected length.

If instead of using a coin toss we choose the sample size, say n, deliberately, then it is well known that $\bar{X} \pm 1.96\sigma/n^{1/2}$ represents the best (shortest expected length) 95% confidence interval procedure. Thus in our two-stage experiment we might consider procedure A: if the coin falls heads and $X = x$ is observed, use the interval $x \pm 1.96\sigma$; if it falls tails and $X_1 = x_1, \ldots, X_k = x_k$ are the observations, use $\bar{x} \pm 1.96\sigma/k^{1/2}$. That is, regardless of which sample size our coin toss tells us to use, we employ the best 95% confidence interval procedure for that sample size. This is certainly a reasonable procedure; but we can do better.

When the sample size is determined by a coin toss the best (shortest expected length) 95% confidence interval depends on the value of k, the number of observations we make if the coin falls tails. To make the example concrete we let $k = 100$. In that case the best 95% confidence interval procedure (B) uses $x \pm 1.68\sigma$ if the coin falls heads and $\bar{x} \pm 2.72\sigma/10$ if it falls tails.

Both A and B are valid 95% confidence interval procedures: if the coin falls heads, the coverage probability of A is 0.95 while that of B is 0.91; if it falls tails then A again covers θ with probability 0.95 while B's probability is 0.99. Thus A's overall coverage probability

atisticians would be reluctant to
ıe purpose of the statistical ana-
tions say about H_1 *vis-à-vis* H_2.
$= 4$ means as evidence, it is not
ıt, and to let their assessment of
rrelevant event is inappropriate.
oportion of the white balls in the
test calls for rejecting H_1 as it is
ccepting.
pretation of Neyman–Pearson
This is illustrated in the following
mous one of Cox (1958, p. 360).
onducted in two stages. At the
outcome of the toss determines
e. If the coin falls heads then
with a $N(\theta, \sigma^2)$ probability dis-
ails we observe k independent
having the same $N(\theta, \sigma^2)$ distri-
we make either one or k observa-
he coin toss. The value of σ^2 is
dence interval for θ with short

e choose the sample size, say n,
that $\bar{X} \pm 1.96\sigma/n^{1/2}$ represents
) 95% confidence interval pro-
experiment we might consider
and $X = x$ is observed, use the
ıd $X_1 = x_1, \ldots, X_k = x_k$ are the
That is, regardless of which
ıse, we employ the best 95% con-
sample size. This is certainly a
lo better.
ed by a coin toss the best (short-
interval depends on the value of
make if the coin falls tails. To
$k = 100$. In that case the best
(*B*) uses $x \pm 1.68\sigma$ if the coin
ılls tails.
ìdence interval procedures: if the
ıbility of *A* is 0.95 while that of *B*
ı covers θ with probability 0.95
A's overall coverage probability

is $\frac{1}{2} \times 0.95 + \frac{1}{2} \times 0.95 = 0.95$ and *B*'s is the same: $\frac{1}{2} \times 0.91 + \frac{1}{2} \times 0.99 = 0.95$. But the expected width of an interval produced by *A* is $1.96\sigma + 1.96\sigma/10 = 2.16\sigma$, while that of an interval produced by *B* is only $1.68\sigma + 2.72\sigma/10 = 1.95\sigma$; *B* generates intervals that are, on the average, about 10% shorter. Note that even if we increase the value of k in procedure *A*, its expected interval width cannot be made as small as that produced by *B* with $k = 100$.

Although *B* is the better procedure, if we apply it to the observations from an experiment for the purpose of indicating what those observations say about θ, then it appears to be misleading in every instance. When we take only one observation, it seems wrong to present $x \pm 1.68\sigma$ as a 95% confidence interval – the confidence coefficient, 0.95, seems too large. Similarly, when the coin falls tails and we make k observations, it seems wrong to attach to the interval $\bar{x} \pm 2.72\sigma/k^{1/2}$ a confidence coefficient of only 0.95.

Another way to look at this example is to compare the cases when we have taken 100 observations deliberately and when we have made this choice of sample size by the toss of a coin. In the first case the best 95% confidence interval procedure uses $\bar{x} \pm 1.96\sigma/10$, while in the second the best procedure uses $\bar{x} \pm 2.72\sigma/10$. Many people agree that, whatever the evidence concerning θ in the observations $x_1, \ldots, x_{100}$, it is unaffected by whether the number of observations was fixed by considerations of costs, for example, or by a coin toss. That is, for interpreting a given set of observations as evidence about θ, it does not matter whether they arose in the first case or the second. If a particular interval is appropriate for showing what the data say in one case, then it is also appropriate in the other. The Neyman–Pearson theory leads to different results in two situations where the evidence is the same, and in applications where the purpose of the statistical analysis is to represent and interpret the data as evidence, this is unacceptable.

Problems appear also when Neyman–Pearson estimation theory is used in applications where the goal is evidential interpretation. We have seen a simple example of this in section 2.2 – for estimating the probability of heads on the basis of ten tosses of a coin the estimator $t_2(X) = \frac{1}{2}$ is a good one in the Neyman–Pearson sense if performance is measured in terms of expected squared error. Now our sample does constitute evidence concerning the probability of heads, but this estimator, which ignores the sample altogether, in no sense represents that evidence.

Here is a much less trivial example. Suppose that one colleague brings me his measurements on the length of butterfly wings in

Ecuador, another brings her observations on tensile strength of wire samples, and I have some data of my own showing the weight loss in laboratory rats on a special diet. If all of the measurements have normal probability distributions, James and Stein (1961) showed that the naive procedure which uses the three sample means to estimate their respective parameters can be improved. Specifically, the naive procedure is not as good as one devised by Stein that uses a statistic depending on all three means, the butterfly wings, the tensile strengths, and the rat weights, to estimate the mean length of butterfly wings, another statistic depending on all three to estimate the mean tensile strength, etc. The James–Stein estimation procedure is better in the sense that the average of the three expected squared errors is smaller with that procedure than with the naive one, no matter what the true values of the three parameters are. But for interpreting our observations as evidence about butterflies, etc., this estimation procedure makes no sense. Whatever evidence we have about butterfly wings is contained in the butterfly data. (The likelihood ratio measuring the relative support for two values of the wing parameter depends only on the butterfly data.) It is inappropriate to allow our assessment of that evidence to depend on irrelevant observations related to wires and rats. Just as in the Cox confidence interval example, we find that the procedure that is better in the Neyman–Pearson sense of expected or average performance is unsatisfactory as a tool for interpreting data as evidence, because it leads to different results in situations where the evidence is the same.

Of course, not all attempts to use Neyman–Pearson methodology for evidential interpretation of data produce results that are as obviously unsatisfactory as the examples above. If they did, the discipline of statistics would look very different than it does today. In countless applications every day, statistical evidence is interpreted, analyzed, and reported in terms of hypothesis tests, estimates, and confidence intervals. Many of these applications seem to be reasonable and helpful, both to experimenters and to readers of their research reports, for representing and communicating 'what the data say'. We will pursue this point in Chapters 3 and 4.

Although Neyman–Pearson test procedures do not have a valid evidential interpretation in general, there is one interesting exception. In that special case the interpretation is derived from the law of likelihood. Suppose it is reported that a test with small error probabilities, α and β, has led to the choice of H_2. In this situation it seems reasonable to claim that the report represents evidence favoring H_2 over H_1, and that the smaller α and β are, the stronger

ʼations on tensile strength of wire
ıy own showing the weight loss in
If all of the measurements have
James and Stein (1961) showed
ːs the three sample means to esti-
an be improved. Specifically, the
one devised by Stein that uses a
ns, the butterfly wings, the tensile
timate the mean length of butter-
ling on all three to estimate the
mes–Stein estimation procedure
ıge of the three expected squared
ure than with the naive one, no
ıree parameters are. But for inter-
ːe about butterflies, etc., this esti-
Whatever evidence we have about
e butterfly data. (The likelihood
ʼt for two values of the wing para-
data.) It is inappropriate to allow
lepend on irrelevant observations
ːhe Cox confidence interval exam-
is better in the Neyman–Pearson
rmance is unsatisfactory as a tool
ɔecause it leads to different results
the same.
ɜ Neyman–Pearson methodology
ata produce results that are as
xamples above. If they did, the
very different than it does today.
lay, statistical evidence is inter-
ı terms of hypothesis tests, esti-
Many of these applications seem
ı to experimenters and to readers
ɜpresenting and communicating
.e this point in Chapters 3 and 4.
t procedures do not have a valid
ıl, there is one interesting excep-
pretation is derived from the law
ɾted that a test with small error
ːhe choice of H_2. In this situation
ːt the report represents evidence
smaller α and β are, the stronger

the evidence is. The test procedure usually chooses the correct hypothesis (because α and β are small), and in this instance it has chosen H_2. Is this not evidence that H_2 is correct? Is it not right to presume that what usually happens (i.e. the procedure chooses correctly) *has* happened, in the absence of any evidence to the contrary? The law of likelihood confirms this judgement. The key here is that the evidence we are evaluating is not the observation, $X = x$, that led to the choice of H_2, but simply an indicator showing which hypothesis was chosen – instead of observing X itself, we see only $Z(X)$, where $Z(x) = 2$ if x is in the critical region (so H_2 is chosen) and $Z(x) = 1$ otherwise (and H_1 is chosen). Thus when we observe $Z = 2$ we have the likelihood ratio of $\Pr(Z = 2|H_2)/\Pr(Z = 2|H_1) = (1 - \beta)/\alpha$ in favor of H_2 over H_1. That is, according to the law of likelihood, the report 'A test of H_1 versus H_2 having size α and power $1 - \beta$ led to rejection of H_1 in favor of H_2' represents evidence favoring H_2 by the factor $(1 - \beta)/\alpha$ (Barnard, Jenkins, and Winsten, 1962, p. 331; see also Birnbaum, 1977). Note that precisely the same reasoning applies when we are told that the test has led to the choice of H_1, rather than H_2. The report 'A test of H_1 versus H_2 having size α and power $1 - \beta$ led to acceptance of H_1' represents evidence favoring H_1 by the factor $(1 - \alpha)/\beta$.

That $Z = 2$ is evidence for H_2 over H_1 does not mean that the observation, $X = x$, on which the test is based is evidence for H_2 over H_1. A data reduction has been made, and evidence has been discarded. That is, although we can give a valid evidential interpretation to the result of a Neyman–Pearson test procedure, that interpretation does not necessarily represent even crudely the evidence in the original observation $X = x$. In section 1.7 we observed X itself and found that some of the values that fell in the critical region (leading to rejection of H_1 for H_2) were in fact evidence favoring H_1 over H_2. Here we observe only whether X is in the critical region or not. Our conclusion that when H_2 is selected we have evidence in favor of H_2 over H_1 refers to the evidence in the limited information given to us, not to the evidence in the observation $X = x$, that caused H_2 to be selected. Thus, if we are not told the value of x, but only that it produced the test result $Z = 2$, say, then we can give a proper evidential interpretation of this very limited information (via the likelihood ratio $(1 - \beta)/\alpha$). But it is not a proper evidential interpretation of x.

The hypothesis-testing procedures that are most often used for interpreting and reporting scientific data are not of the

Neyman–Pearson variety. Before turning to the more commonly used test procedures, which are discussed in Chapter 3, we consider a place in science where Neyman–Pearson theory does play an important role.

2.4 Neyman–Pearson hypothesis testing in planning experiments: choosing the sample size

We have observed that Neyman–Pearson tests are not designed for interpreting statistical evidence, and that their use for that purpose can lead to serious errors in which observations that are evidence supporting H_1 over H_2 are given the opposite interpretation. Strict Neyman–Pearson test procedures are in fact rarely used for interpreting and reporting scientific data, but they are routinely used in another important phase of research. When a study or experiment is being planned, the researcher often uses Neyman–Pearson theory to determine how many observations will be made. He models the study as a procedure for choosing between two hypotheses, H_1 and H_2, and specifies the maximum tolerable error probabilities, α and β. Then two objectives, stated in terms of the Neyman–Pearson hypothesis-testing paradigm, determine the sample size: 'We want to be pretty sure (probability $1-\alpha$ or greater) that we will not reject H_1 when it is true, and also pretty sure (probability $1-\beta$ or greater) that we will reject H_1 when H_2 is true'.

For any sample size we can choose to test at any size we like, so we can always accomplish the first objective. But in order to accomplish the second we must make the sample size large enough.

This approach leads to standard formulas for the number of observations required. For example, consider the simple case of independent $N(\theta, \sigma^2)$ observations, where the variance σ^2 is known from pilot data or from results of previous studies, with hypotheses H_1: $\theta = \theta_1$ and H_2: $\theta = \theta_1 + \delta$. The usual calculation shows that the number of observations must be at least

$$n_{\mathrm{NP}} = [(z_{1-\alpha} + z_{1-\beta})\sigma/\delta]^2, \tag{2.1}$$

where $z_{1-\alpha}$ is the $100(1-\alpha)$th percentile of the standard normal distribution. Using a sample size $n \geq n_{\mathrm{NP}}$ ensures that a test with size α will have power of at least $1-\beta$: the Type I and Type II error probabilities will not exceed the specified values, α and β.

This approach to determining sample size is often used in studies whose actual purpose is more accurately described in terms of

e turning to the more commonly
iscussed in Chapter 3, we consider
an–Pearson theory does play an

testing in planning experiments:

Pearson tests are not designed for
nd that their use for that purpose
ch observations that are evidence
the opposite interpretation. Strict
s are in fact rarely used for inter-
ata, but they are routinely used in
rch. When a study or experiment
er often uses Neyman–Pearson
observations will be made. He
or choosing between two hypoth-
s the maximum tolerable error
o objectives, stated in terms of
-testing paradigm, determine the
retty sure (probability $1-\alpha$ or
$_1$ when it is true, and also pretty
) that we will reject H_1 when H_2

ose to test at any size we like, so we
jective. But in order to accomplish
ple size large enough.
ard formulas for the number of
ple, consider the simple case of
ns, where the variance σ^2 is
results of previous studies, with
$\theta = \theta_1 + \delta$. The usual calculation
ations must be at least

$$+ z_{1-\beta})\sigma/\delta]^2, \qquad (2.1)$$

ercentile of the standard normal
$n \geq n_{NP}$ ensures that a test with
st $1-\beta$: the Type I and Type II
the specified values, α and β.
ample size is often used in studies
accurately described in terms of

evidence than decisions. The objective is not really to choose between $\theta = \theta_1$ and $\theta = \theta_1 + \delta$, but to generate evidence about θ, with particular interest in how strongly that evidence supports one of these two values, θ_1 and $\theta_1 + \delta$, versus the other. When this is true, the Neyman–Pearson approach is unsatisfactory. We will show that in the simple important case of the normal probability distribution described above, the sample size n_{NP} is too small to ensure that the researcher has an adequate chance to meet his actual objectives, and that using the Neyman–Pearson hypothesis-testing procedure to interpret the data as evidence leads to misinterpretation with high frequency. That is, the misinterpretations that were shown to be possible in section 2.3 are not confined to extreme or pathological cases, but are common when the sample size is determined by the 'usual' formula. Specifically, we will see that when $\alpha = 0.05$ and $\beta = 0.20$ the evidence will be misinterpreted more than 30% of the time.

The objectives can be restated as 'We want to be pretty sure (probability $1-\alpha$ or greater) that we will not find strong evidence in favor of H_2 when H_1 is true, and also pretty sure (probability $1-\beta$ or greater) that we will find strong evidence in favor of H_2 when H_2 is true'. Suppose that these are our actual objectives, but that we use the Neyman–Pearson paradigm, determining our sample size by equation (2.1).

First, we ask 'How often will the study produce strong evidence?'. The results, $X = x$, will be strong evidence for H_2 over H_1 if (and only if) the likelihood ratio L_2/L_1 is at least k, where k is determined by the expression 'strong' evidence. Now most readers will agree that, in the canonical urn scheme of section 1.6, the evidence in favor of the 'all white' urn over the 'half white' one, when the number of consecutive white balls observed is two (a likelihood ratio of $2^2 = 4$) is only weak, but that six consecutive white balls (a likelihood ratio of $2^6 = 64$) are not just 'strong' but 'quite strong' evidence. Thus we will focus on values of $k = 8$, 16, 32, corresponding to 3, 4, or 5 white balls in the urn scheme.

The likelihood ratio for H_2: $\theta = \theta_1 + \delta$ versus H_1: $\theta = \theta_1$, where θ is the mean of the normal distribution, equals

$$\exp\{[\bar{x} - (\theta_1 + \delta/2)]n\delta/\sigma^2\},$$

so that we have strong evidence for H_2 when we have observations x for which this quantity exceeds k, that is, for which

$$n^{1/2}(\bar{x} - \theta_1)/\sigma > n^{1/2}\delta/2\sigma + \sigma \ln(k)/\delta n^{1/2}.$$

Table 2.2 *Probabilities of undesirable results when n is chosen so that* $\alpha = 0.05$ *for* $\beta = 0.20, 0.05$*: (a) finding strong evidence in favor of false hypothesis; (b) failing to find strong evidence in favor of true hypothesis*

	(a) $\Pr_1(L_2/L_1 \geq k)$		(b) $\Pr_2(L_2/L_1 < k)$	
k	$\beta = 0.20$	$\beta = 0.05$	$\beta = 0.20$	$\beta = 0.05$
8	0.019	0.011	0.342	0.156
16	0.009	0.006	0.449	0.212
32	0.004	0.003	0.560	0.277

If the sample size is the value n_{NP} given in expression (2.1), then the right-hand side of this inequality equals $c/2 + \ln(k)/c$, where $c = z_{1-\alpha} + z_{1-\beta}$, and the probability of finding strong evidence in favor of H_2 when H_1 is true is

$$\Pr_1(L_2/L_1 \geq k) = 1 - \Phi(c/2 + \ln(k)/c). \tag{2.2}$$

It is easily shown that this probability of misleading strong evidence is the same as the probability of misleading evidence in the other direction, $\Pr_2(L_1/L_2 \geqslant k)$. Similarly, the probability of finding strong evidence for H_2 when H_2 is true is the same as the probability of finding strong evidence for H_1 when H_1 is true, and that probability is

$$\Pr_2(L_2/L_1 \geq k) = 1 - \Phi(\ln(k)/c - c/2). \tag{2.3}$$

Table 2.2 gives the values of expressions (2.2) and (2.3) for selected values of k when $\alpha = 0.05$ and $\beta = 0.20$ and 0.05.

If the researcher actually wants to be pretty sure (probability at least 0.95) that the study will not produce strong evidence supporting H_2 when H_1 is true, then columns 1 and 2 show that for both $\beta = 0.20$ and $\beta = 0.05$ the sample size n_{NP} is adequate. But so are smaller ones. In fact it is easy to show (Exercise 1.5) that the probability of misleading evidence, $\Pr_1(L_2/L_1 \geq k)$, cannot exceed $1 - \Phi(\sqrt{2\ln(k)})$ for any choice of n, θ_1, and δ. For $k = 8$, 16, 32 this bound equals 0.021, 0.009, and 0.004, respectively. That is, the choice of a reasonable value of k ensures that the probability of generating misleading evidence is small, less than 0.021 when $k = 8$ and 0.01 when $k = 16$, regardless of n.

But columns 3 and 4 show that the sample size n_{NP} is not adequate with respect to producing strong evidence in favor of H_2 when H_2 is true. When β is set at 0.20 and H_2 is true, a sample of size n_{NP} will fail

'e results when n is chosen so that ing strong evidence in favor of false vidence in favor of true hypothesis

	(b) $\Pr_2(L_2/L_1 < k)$	
	$\beta = 0.20$	$\beta = 0.05$
	0.342	0.156
	0.449	0.212
	0.560	0.277

given in expression (2.1), then the ity equals $c/2 + \ln(k)/c$, where lity of finding strong evidence in

$$- \Phi(c/2 + \ln(k)/c). \qquad (2.2)$$

lity of misleading strong evidence misleading evidence in the other arly, the probability of finding true is the same as the probability when H_1 is true, and that prob-

$$- \Phi(\ln(k)/c - c/2). \qquad (2.3)$$

ssions (2.2) and (2.3) for selected = 0.20 and 0.05.

to be pretty sure (probability at produce strong evidence support- umns 1 and 2 show that for both size n_{NP} is adequate. But so are how (Exercise 1.5) that the prob- $\Pr_1(L_2/L_1 \geq k)$, cannot exceed f n, θ_1, and δ. For $k = 8$, 16, 32 d 0.004, respectively. That is, the ensures that the probability of is small, less than 0.021 when rdless of n.

he sample size n_{NP} is not adequate vidence in favor of H_2 when H_2 is s true, a sample of size n_{NP} will fail

to produce strong evidence for H_2 more than one-third of the time (column 3). That is, the sample is not large enough that the researcher can be pretty sure that, if H_2 is true, the study will produce strong evidence in its favor. And column 4 shows that when $\alpha = \beta = 0.05$, the probability of failing to produce strong evidence in favor of H_2 when it is true is greater than 0.15, three times the value at which the researcher probably thought he was controlling this risk when he fixed the Type II error probability at $\beta = 0.05$.

If the study is structured as a Neyman–Pearson testing procedure, it always leads to a decision, to choose H_1 or to choose H_2. It does not always lead to strong evidence; in fact, Table 2.2 shows that at $\alpha = 0.05$, when either of the two hypotheses is true the study will fail to produce evidence strong enough to give a likelihood ratio as large as 8 in favor of one or the other about one-third of the time ($0.342 - 0.019 = 0.323$) when $\beta = 0.20$, and about 15% of the time when $\beta = 0.05$.

In this example, if either hypothesis is true the probability of producing strong evidence supporting that hypothesis over the other one is the same, $\Pr_2(L_2/L_1 \geq k) = \Pr_1(L_1/L_2 \geq k)$. To ensure that this probability is at least 0.95, we need at least $n_L = \{1.645 + [(1.645)^2 + 2\ln(k)]^{1/2}\}^2\sigma^2/\delta^2$ observations (Exercise 2.1). Table 2.3 shows these values for $k = 8$, 16, 32. They are larger than the value n_{NP} given by the Neyman–Pearson formula (2.1) with $\alpha = \beta = 0.05$, which equals 10.824 times σ^2/δ^2. For instance, to be pretty sure (probability at least 0.95) that we will obtain strong evidence in favor of the true hypothesis ($LR \geqslant 8$) we require about two-thirds again as many observations as are required to achieve Type I and Type II error probabilities of $\alpha = \beta = 0.05$: $n_L/n_{NP} = 18.191/10.824 = 1.68$. Even if we reduce α to 0.025, so that n_{NP} is the sample size required for a two-sided Neyman–Pearson test with Type I error rate $\alpha = 0.05$, formula (2.1) gives $n_{NP} = 12.996\sigma^2/\delta^2$, so that $n_L/n_{NP} = 1.40$. We actually need 40% more observations.

Perhaps we have chosen the value of k that identifies 'pretty strong evidence' badly. Perhaps $k = 8$ is more extreme than we realize, and a more enlightened choice of this critical value, say $k = 4$, would produce a sample size n_L that agrees with the Neyman–Pearson value n_{NP}. Is there some value of $k > 1$ for which $n_L = n_{NP}$? No. Exercise 2.1 shows that for any specified k we can make $\Pr_1(L_1/L_2 \geq k) = \Pr_2(L_2/L_1 > k) = 0.95$ by choosing n large enough, $n \geq n_L(k)$, where the required sample size increases as k increases. It also shows that for $k = 1$ the required sample size,

Table 2.3 *Sample size n_L required to give the stated probability of producing strong evidence for the true hypothesis, and (α, β) values for which formula (2.1) gives the required sample size*

	Probability = 0.95			Probability = 0.80		
		$\alpha = 0.05$	$\alpha = \beta$		$\alpha = 0.05$	$\alpha = \beta$
k	$n_L\delta^2/\sigma^2$	α β	α β	$n_L\delta^2/\sigma^2$	α β	α β
8	18.191	(0.05, 0.004)	(0.016, 0.016)	9.292	(0.05, 0.080)	(0.064, 0.064)
16	20.408	(0.05, 0.002)	(0.012, 0.012)	11.175	(0.05, 0.045)	(0.047, 0.047)
32	22.557	(0.05, 0.001)	(0.009, 0.009)	13.004	(0.05, 0.025)	(0.036, 0.036)

8	18.191	(0.05, 0.004)	(0.016, 0.016)	9.292	(0.05, 0.080)	(0.064, 0.064)
16	20.408	(0.05, 0.002)	(0.012, 0.012)	11.175	(0.05, 0.045)	(0.047, 0.047)
32	22.557	(0.05, 0.001)	(0.009, 0.009)	13.004	(0.05, 0.025)	(0.036, 0.036)

$n_L(1)$, equals the sample size n_{NP} given by the Neyman–Pearson formula with $\alpha = \beta = 0.05$. This sample size is not adequate for any $k > 1$. We are not finding that the required sample size n_L is greater than the Neyman–Pearson value n_{NP} because we are unwittingly setting our standard too high; n_{NP} is really too small.

If we are willing to settle for a probability as low as 0.80 that the study will produce strong evidence in favor of the true hypothesis, column 4 of Table 2.3 shows that we need only $9.292\sigma^2/\delta^2$ observations, about half the number required for a probability of 0.95. This is still 50% more than the sample size given by the Neyman–Pearson formula (2.1) with $\alpha = 0.05$, $\beta = 0.20$, which is $6.185\sigma^2/\delta^2$.

In order to obtain the required sample size, $n_L(k)$, from the Neyman–Pearson formula, that is, to make $n_{NP} = n_L$, we must choose smaller values of α and β than the conventional ones. Column 2 of Table 2.3 shows the β values needed if $\alpha = 0.05$, and column 3 shows the common value needed if we set $\alpha = \beta$.

Now suppose that we have done the study, taking the number of observations n_{NP} given by formula (2.1). Furthermore, suppose that we use the Neyman–Pearson test procedure to interpret the evidence in our sample. We perform the test of H_1 versus H_2 at level α; because of the way we chose n, we know the test has the specified power, $1 - \beta$. When the test calls for choosing H_2, we will interpret this to mean that the sample represents pretty strong evidence supporting H_2 over H_1 (and vice versa).

We saw in section 2.3 that this interpretation can be wrong, that it can lead to classifying observations as evidence supporting H_2 over H_1 (or vice versa) when the opposite is true. But maybe this is not often the case. Maybe the Neyman–Pearson procedure usually produces a correct interpretation of the evidence in situations like the present example, where the sample size is chosen to control α and β at conventional levels. Let us see.

We will interpret the observations as evidence for H_2 when the test rejects H_1, that is, when $n_1^{1/2}(\bar{x} - \theta)/\sigma > z_{1-\alpha}$, and as evidence for H_1 when this inequality is reversed. When H_1 is true, how often will this interpretation be incorrect? From formula (2.2) we find that

$$\begin{aligned} &\Pr_1(L_2/L_1 < k \mid n^{1/2}(\bar{X} - \theta_1)/\sigma > z_{1-\alpha}) \\ &\quad = 1 - [1 - \Phi(c/2 + \ln(k)/c)]/\alpha \end{aligned} \tag{2.4}$$

(assuming that $c/2 + \ln(k)/c$ is greater than $z_{1-\alpha}$). This is the probability, given that H_1 is rejected in favor of H_2, that the evidence

supporting H_2 over H_1 is not strong. Similarly, the probability, given that H_1 is accepted, that the evidence in favor of that hypothesis is not strong is given by

$$\Pr_1(L_1/L_2 < k | n^{1/2}(\bar{X} - \theta_1)/\sigma < z_{1-\alpha}) = 1 - [\Phi(c/2 - \ln(k)/c)/(1-\alpha)]. \quad (2.5)$$

These and the corresponding probabilities of misinterpretation when H_2 is true are given in Table 2.4 for $\alpha = 0.05$ when $\beta = 0.20$ and when $\beta = 0.05$. There we see that if a likelihood ratio of at least $k = 8$ defines 'strong' evidence, then almost two-thirds (0.625) of the 'Type I error' rate results from rejecting H_1 when the evidence in favor of H_2 is not strong.

Overall, if we think that we have strong evidence in favor of the hypothesis that is selected by the testing procedure with $\alpha = 0.05$ and $\beta = 0.20$, Table 2.4 shows that we will be wrong about one-third of the time: when H_1 is true we will accept H_2 5% of the time, and Table 2.4 shows that when that happens, the probability that we will actually have only weak evidence (likelihood ratio less than $k = 8$) in favor of H_2 is 0.625; similarly, we will accept H_1 95% of the time, but when that happens, we will actually have only weak evidence in favor of H_1 30.7% of the time. Thus when H_1 is true we will misinterpret the evidence a fraction $0.05 \times 0.625 + 0.95 \times 0.307 = 0.03 + 0.29 = 0.32$ (about one-third) of the time, usually in the direction of thinking that we have strong evidence for H_1 when we do not. When H_2 is true the same analysis shows that we will misinterpret the evidence about the same fraction of the time, $0.80 \times 0.177 + 0.20 \times 0.906 = 0.14 + 0.18 = 0.32$, incorrectly thinking that we have strong evidence in favor of H_2 (14% of the time) almost as often as we incorrectly think that we have strong evidence for H_1. If we change β so that $\beta = \alpha = 0.05$ then Table 2.4 shows that, regardless of which hypothesis is true, we will misinterpret the evidence 14.4% of the time: $0.05 \times 0.772 + 0.95 \times 0.111 = 0.144$.

At the beginning of this section, we noted that strict Neyman–Pearson procedures are rarely used for interpreting and reporting scientific data as evidence. Table 2.4 shows that this is appropriate – Neyman–Pearson procedures should not be used for that purpose. The more commonly used procedures are discussed in the next chapter. If they are valid, they must rest on a different theoretical basis than the one provided by Neyman and Pearson.

.g. Similarly, the probability, given
ence in favor of that hypothesis is

$$(\bar{X} - \theta_1)/\sigma < z_{1-\alpha})$$

$$- \ln(k)/c)/(1 - \alpha)]. \qquad (2.5)$$

robabilities of misinterpretation
le 2.4 for $\alpha = 0.05$ when $\beta = 0.20$
e that if a likelihood ratio of at
idence, then almost two-thirds
e results from rejecting H_1 when
strong.
ve strong evidence in favor of the
testing procedure with $\alpha = 0.05$
hat we will be wrong about one-
ue we will accept H_2 5% of the
hen that happens, the probability
eak evidence (likelihood ratio less
625; similarly, we will accept H_1
t happens, we will actually have
H_1 30.7% of the time. Thus when
rpret the evidence a fraction
$3 + 0.29 = 0.32$ (about one-third)
ection of thinking that we have
do not. When H_2 is true the same
interpret the evidence about the
e, $0.80 \times 0.177 + 0.20 \times 0.906 =$
inking that we have strong evi-
the time) almost as often as we
strong evidence for H_1. If we
en Table 2.4 shows that, regardless
ll misinterpret the evidence 14.4%
$\times 0.111 = 0.144$.
n, we noted that strict Neyman–
sed for interpreting and reporting
2.4 shows that this is appropriate
should not be used for that
used procedures are discussed
are valid, they must rest on a
he one provided by Neyman and

Table 2.4 *Probability that the evidential interpretation of the Neyman–Pearson test result will be incorrect when* $\alpha = 0.05$ *and* $\beta = 0.20$ *(*$\beta = 0.05$*)*

k	$\Pr_1(L_2/L_1 < k \mid \text{rej. } H_1)$	$\Pr_1(L_1/L_2 < k \mid \text{acc. } H_1)$	$\Pr_2(L_2/L_1 < k \mid \text{rej. } H_1)$	$\Pr_2(L_1/L_2 < k \mid \text{acc. } H_1)$
8	0.625 (0.772)	0.307 (0.111)	0.177 (0.111)	0.906 (0.772)
16	0.816 (0.871)	0.420 (0.170)	0.311 (0.170)	0.954 (0.871)
32	0.916 (0.930)	0.536 (0.239)	0.450 (0.239)	0.979 (0.930)

2.5 Summary

Neyman–Pearson statistical theory is aimed at finding good rules for choosing from a specified set of possible actions. It does not address the problem of representing and interpreting statistical evidence, and the decision rules derived from Neyman–Pearson theory are not appropriate tools for interpreting data as evidence.

Exercises

2.1 Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ with σ^2 known, and consider the two simple hypotheses H_1: $\theta = \theta_1$ and H_2: $\theta = \theta_1 + \delta$.
   (a) Derive a formula for the sample size n_{NP} required to make both the Type I and Type II error probabilities equal 0.05.
   (b) Derive a formula for the sample size $n_{\mathrm{L}}(k)$ required to make both the probabilities $\Pr_1(L_1/L_2 \geq k)$ and $\Pr_2(L_2/L_1 \geq k)$ equal to 0.95.
   (c) Show that $n_{\mathrm{L}}(k) > n_{\mathrm{NP}}$ for all $k > 1$, and that this remains true if the probabilities 0.05 in (a) and 0.95 in (b) are replaced by α and $1 - \alpha$, for any $0 < \alpha < 1$.

2.2 (Continuation of Exercise 2.1)
   (a) Show that if H_1 is true then the probability of misleading strong evidence in favor of H_2 $(L_2/L_1 \geq k)$ is greatest when $n = 2(\sigma/\delta)^2 \ln(k)$.
   (b) Find the maximum probability in (a).
   (c) If H_2 is true, what is the maximum probability of (misleading) strong evidence in favor of H_1?
   (d) For the sample size in (a), what is the power of the most powerful size-α Neyman–Pearson test of H_1 versus H_2?

2.3 For independent random variables $X \sim N(\mu, 1)$ and $Y \sim N(\eta, 1)$, consider the ratio $\theta = \mu/\eta$.
   (a) Use the fact that $(X - \theta Y)/(1 + \theta^2)^{1/2}$ has a standard normal distribution to derive a 95% confidence region for θ.
   (b) Show that with positive probability the 95% confidence region will consist of the entire real line (Fieller, 1954).

2.4 Consider a model in which both the sample space and the parameter space consist of M points, $x_1, \ldots, x_M$ and $\theta_1, \ldots, \theta_M$. $P(X = x_i; \theta_i) = \alpha$ for $i = 1, \ldots, M$, with the remainder of the probability spread uniformly over the remainder of the sample space. If α is less than $1/M$, what is the best $100(1 - \alpha)\%$

ory is aimed at finding good
ified set of possible actions. It
of representing and interpreting
ion rules derived from Neyman–
ate tools for interpreting data as

$N(\theta, \sigma^2)$ with σ^2 known, and con-
es H_1: $\theta = \theta_1$ and H_2: $\theta = \theta_1 + \delta$.
sample size n_{NP} required to make
II error probabilities equal 0.05.
ample size $n_L(k)$ required to make
$(L_1/L_2 \geq k)$ and $\Pr_2(L_2/L_1 \geq k)$

r all $k > 1$, and that this remains
0.05 in (a) and 0.95 in (b) are
for any $0 < \alpha < 1$.

)
ıen the probability of misleading
of H_2 $(L_2/L_1 \geq k)$ is greatest

bility in (a).
naximum probability of (mislead-
vor of H_1?
), what is the power of the most
Pearson test of H_1 versus H_2?

variables $X \sim N(\mu, 1)$ and
o $\theta = \mu/\eta$.
$\theta Y)/(1 + \theta^2)^{1/2}$ has a standard
ive a 95% confidence region for θ.
probability the 95% confidence
ntire real line (Fieller, 1954).

th the sample space and the para-
oints, $x_1, \ldots, x_M$ and $\theta_1, \ldots, \theta_M$.
$\ldots, M$, with the remainder of the
over the remainder of the sample
, what is the best $100(1 - \alpha)\%$

confidence region $R(X)$? Give a precise interpretation of an observation $X = x$ as evidence in relation to the hypotheses H_{in}: $\theta \in R(x)$ and H_{out}: $\theta \notin R(x)$. When $\alpha = 0.05$ and $M = 19$, is the observation fairly strong evidence supporting H_{in} over H_{out}?