DEBORAH G. MAYO

AN OBJECTIVE THEORY OF STATISTICAL TESTING*

ABSTRACT. Theories of statistical testing ᵐay be seen as attempts to provide systematic means for evaluating scientific conjectures on the basis of incomplete or inaccurate observational data. The Neyman-Pearson Theory of Testing (NPT) has purported to provide an objective means for testing statistical hypotheses corresponding to scientific claims. Despite their widespread use in science, methods of NPT have themselves been accused of failing to be objective; and the purported objectivity of scientific claims based upon NPT has been called into question. The purpose of this paper is first to clarify this question by examining the conceptions of (I) the function served by NPT in science, and (II) the requirements of an objective theory of statistics upon which attacks on NPT's objectivity are based. Our grounds for rejecting these conceptions suggest altered conceptions of (I) and (II) that might avoid such attacks. Second, we propose a reformulation of NPT, denoted by NPT*, based on these altered conceptions, and argue that it provides an objective theory of statistics. The crux of our argument is that by being able to objectively control error frequencies NPT* is able to objectively evaluate what has or has not been learned from the result of a statistical test.

1. INTRODUCTION AND SUMMARY

In order to live up to the ideal of epistemological objectivity – whether it be in science or elsewhere – the essential requirement seems to be this: All claims are to be open to being tested and criticized by impartial criteria independent of the whims, desires, and prejudices of individuals. In its attempt to live up to this ideal of objectivity, empirical science, more than any other activity, self-consciously carries out observations and experiments to systematically test scientific claims. However, the incompleteness of observations, inaccuracies of measurements, and the general effects of environmental perturbations and "noise" cause observations to deviate from testable predictions, even when the scientific claims from which they are derived are approximately true descriptions of some aspect of a phenomenon. In order to distinguish observed deviations due to these extraneous sources from those due to the falsehood of scientific claims, observational tests in science are typically based on statistical considerations, and a theory of statistical testing attempts to make these considerations

Synthese **57** (1983) 297–340. 0039–7857/83/0573–0297 $04.40

precise. The procedures of the Neyman-Pearson Theory of Testing (which we abbreviate as NPT) have been widely used in science in the last fifty years to carry out such statistical tests in a purportedly objective manner. However, the procedures of NPT have themselves been accused of failing to be objective; for they appear permeated with the very subjective, personal, arbitrary factors that the ideal of objectivity demands one exclude. As such, the purported objectivity of scientific claims based upon NPT has been called into question; and it is this question around which our discussion revolves. That is, we want to ask the following: Can NPT provide objective scientific claims? Or: Is NPT an objective theory of statistical testing?

Despite the serious ramifications of negative answers to this question, the increasingly common arguments upon which such negative answers are based have gone largely unanswered by followers of NPT. We attempt to rectify this situation in this paper by clarifying the question of the objectivity of NPT, setting out a reconstruction of NPT, and defending its scientific objectivity.

We begin by giving a general framework for relating statistical tests to scientific inquiries, and for spelling out the formal components of NPT and comparing them to those of subjective Bayesian tests. The question of the objectivity of NPT then emerges as a "metastatistical" question concerning the relationships between the formal statistical inquiry and the empirical scientific one. One's answer to this question is seen to depend on one's conception of (I) the function served by NPT in science, and (II) the requirements of an objective theory of statistics. Firstly, the most serious attacks on the objectivity of NPT by philosophers and statisticians are seen to be based on the view that NPT functions to provide rules for using data to decide how to act with respect to a scientific claim. Secondly, they are seen to view only that which is given by purely formal logical principles as truly objective. Since these two conceptions are found in the typical formulations of NPT, such attacks are not without some justification. Nevertheless, we reject these conceptions, and our grounds for doing so are used to suggest the the sort of altered conceptions of (I) and (II) that avoid such attacks. To defend NPT against such objections we propose a reformulation, denoted by NPT*, which is based on these altered conceptions.

According to NPT*, statistical tests function in science as a means of *learning* about variable phenomena on the basis of limited data. This function is accomplished by detecting discrepancies between the (approximately) correct models of a phenomenon and hypothesized ones. As such, we hold that objective learning from a statistical test result can be accomplished if one can critically evaluate what it does and does not detect or indicate about the scientific phenomenon. Moreover, by clearly distinguishing the formal statistical result from its (metastatistical) critical evaluation, we can deny that any arbitrariness in specifying the test necessarily prevents this evaluation from being objective. Most importantly, we suggest how NPT* is able to objectively interpret the scientific import of its statistical test results by "subtracting out" the influences of arbitrary factors. NPT* permits these influences to be understood (and hence "subtracted out" or controlled) by making use of frequency relations between test results and underlying discrepancies (i.e., by making use of error frequencies).

Focusing on a common type of statistical test, we introduce two functions that allow this appeal to frequency relations to be made explicit. We show how, by using these functions, NPT* allows for an objective assessment of what has been learned from a specific test result. What is more, the ability to make objective (i.e., testable) assertions about these probabilistic relations (i.e., error probabilities or rates) is what fundamentally distinguishes NPT from other statistical inference theories. So, if we are correct, it is this feature that enables NPT* (and prevents other approaches) from providing an objective theory of statistical testing. While our focus is limited to the problem of providing a scientifically objective theory of statistical testing, this problem (and hopefully our suggested strategy for its resolution) is relevant for the more general problems that have increasingly led philosophers to the pessimistic conclusion that science is simply a matter of subjective, psychological preference (or "conversion") governed by little more than the whims, desires, and chauvinistic fantasies of individuals.

2. A THEORY OF STATISTICAL TESTING

Theories of statistical testing may be seen as attempts to provide systematic means for dealing with a very common problem in scientific inquiries. The problem is how to generate and analyze observational data to test a scientific claim or hypothesis when it is known that such data need not precisely agree with the claim, even if the claim correctly describes the phenomenon of interest. One reason for these deviations is that the hypothetical claim of interest may involve theoretical concepts that do not exactly match up with things that can be directly observed. Another is that observational data are necessarily finite and discrete, while scientific hypotheses may refer to an infinite number of cases and involve continuous quantities, such as weight and temperature. As such, the accuracy, precision and reliability of such data is limited by numerous distortions and "noise" introduced by various intermediary processes of observation and measurement. A third reason is that the scientific hypothesis may itself concern probabilistic phenomena, as in genetics and quantum mechanics. However, whether it is due to the incompleteness and inaccuracies of observation and measurement, or to the genuine probabilistic nature of the scientific hypothesis, the deviations, errors, and fluctuations found in observational data often follow very similar patterns; and statistical models provide standard paradigms for representing such variable patterns. By modelling the scientific inquiry statistically, it is possible to formally model (A) the scientific claims $\mathscr{C}$ in terms of *statistical hypotheses* about a population; (B) the observational analysis of the statistical hypotheses in terms of *experimental testing rules*; and (C) the actual observation $\mathcal{O}$ in terms of a *statistical sample* from the population. It will be helpful to spell out the formal components of these three models or structures of a statistically modelled inquiry.

2.1. *Models of Statistical Testing*

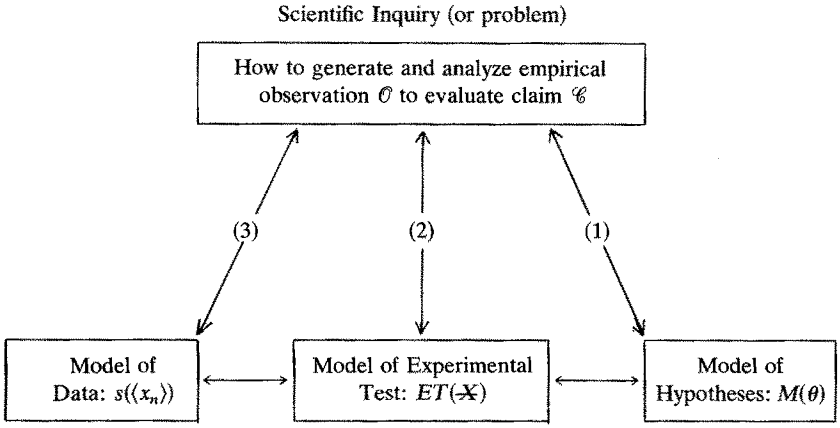
(A) *Statistical Models of Hypotheses*: The scientific claim is 'translated' into a claim about a certain *population* of objects, where this population is known to vary with respect to a quantity of interest, represented by X. For example, the population may be all houseflies in existence and the quantity X, their winglength. The variability of X in the population is modelled by viewing X as a random variable which takes on different values with given probabilities; that is, X is viewed as having a probability distribution P. The distribution P is governed by one or more properties of the population called *parameters*; and hypotheses about the values of such parameters are *statistical hypotheses*. An example of a parameter is the *average* wing length of houseflies, and for simplicity we will focus on cases where the statistical hypotheses of interest concern a single such parameter, represented by θ, where θ is the average (i.e., the mean) of a quantity X. The set of possible values of θ to be considered, called the *parameter space*, Ω, is specified and each statistical hypothesis is associated with some subset of Ω. The class of these statistical hypotheses can be represented by means of a probability model $M(\theta)$ which describes X as belonging to the class of probability distributions $P(X|\theta)$ where θ ranges over Ω. That is, $M(\theta) = [P(X|\theta), \theta \in \Omega]$.

(B) *Statistical Models of Experimental Tests*: The actual observational inquiry is modelled as observing n independent random variables $X_1, X_2, \ldots, X_n$, where each X_i is distributed according to $M(\theta)$, the population distribution. Each experimental outcome can be represented as an n-tuple of values of these random variables, i.e., $x_1, x_2, \ldots, x_n$ which may be abbreviated as $\langle x_n \rangle$. The set of possible experimental outcomes, the *sample space* $\mathcal{X}$, consists of the possible n-tuples of values. On the basis of the population distribution $M(\theta)$ it is possible to derive, by means of probability theory, the probability of the possible experimental outcomes, i.e., $P(\langle x_n \rangle | \theta)$, where $\langle x_n \rangle \in \mathcal{X}$, $\theta \in \Omega$. We also include in an experimental testing model the specification of a testing rule T that maps outcomes in $\mathcal{X}$ to various claims about the models of hypotheses $M(\theta)$; i.e., $T: \mathcal{X} \rightarrow M(\theta)$. That is, an experimental testing model $ET(\mathcal{X}) = [\mathcal{X}, P(\langle x_n \rangle | \theta), T]$.

(C) *Statistical Models of Data* (from the experiment): An empirical observation $\mathcal{O}$ is modelled as a particular element of $\mathcal{X}$; further modelling of $\langle x_n \rangle$ corresponds to various sample properties of interest, such as the sample average. Such a data model may be represented by $[\langle x_n \rangle, s(\langle x_n \rangle)]$, where in the simplest case $s(\langle x_n \rangle) = \langle x_n \rangle$.

We may consider the indicated components of these three models to be the basic formal elements for setting out theories of testing; many additional elements, built upon these, are required for different testing theories. Figure 2.1 sketches the approximate relationships between the scientific inquiry and the three models of a statistical testing inquiry.

The nature of the (metastatistical) relationships represented by rays (1)–(3), and their relevance to our problem, will become clarified throughout our discussion.


Scientific Inquiry (or problem)
How to generate and analyze empirical observation $\mathcal{O}$ to evaluate claim $\mathcal{C}$
(3)
(2)
(1)
Model of Data: $s(\langle x_n \rangle)$
Model of Experimental Test: $ET(X)$
Model of Hypotheses: $M(\theta)$

Statistical Inquiry: Testing Hypotheses

Fig. 2.1.

2.2. *Subjective Bayesian Tests vs. 'Objective' Neyman-Pearson Tests*

Using the models of data, experimental tests, and hypotheses, different theories of statistical testing can be compared in terms of the different testing rules each recommends for mapping the observed statistical data to various assertions about statistical hypotheses in $M(\theta)$. The subjective Bayesian theory, for example, begins by specifying not only a complete set of statistical hypotheses $H_0, \ldots, H_k$ about values of θ in $M(\theta)$, but also a *prior probability* assignment $P(H_i)$ to each hypothesis H_i, where probability is construed as a measure of one's degree of belief in the hypotheses. A subjective Bayesian test consists of mapping these prior probabilities, together with the sample data s, into a final probability assignment in a hypothesis of interest H_i, called a *posterior probability*, by way of *Bayes' Theorem*. The theorem states:

$$P(H_i \mid s) = \frac{P(s \mid H_i)P(H_i)}{\sum_{j=0}^{k} P(s \mid H_j)P(H_j)} \qquad (i = 0, \ldots, k)$$

The experimental test may use the measure of posterior probability $P(H_i \mid s)$ to reject H_i if the value is too small or to calculate a more elaborate test incorporating losses.

In other words, the subjective Bayesian model translates the scientific problem into one requiring a means for quantifying one's subjective opinion in various hypotheses about θ, and using observed data to coherently change one's degree of belief in them. One can then decide to accept or reject hypotheses according to how strongly one believes in them and according to various personal losses one associates with incorrectly doing so. In contrast, the leaders of the so-called 'objective' tradition in statistics (i.e., Fisher, Neyman, and Pearson) deny that a scientist would want to multiply his initial subjective degrees of belief in hypotheses (assuming he had them) with the 'objective' probabilities provided by the experimental data to reach a final evaluation of the hypotheses. For the result of doing so is that one's final evaluation is greatly colored by the initial subjective degree of belief; and such initial degrees of belief may simply reflect one's ignorance or prejudice rather than what is true. (In an extreme case where one's prior probability is zero, one can never get a nonzero posterior probability regardless of the amount of data gathered.) Hence, the idea of testing a scientific claim by ascertaining so and so's degree of belief in it runs counter to our idea of scientific objectivity. While the subjective Bayesian test may be appropriate for making personal decisions, when one is only interested in being informed of what one *believes* to be the case, it does not seem appropriate for impartially testing whether beliefs accord with what really *is* the case.

In contrast, the Neyman-Pearson Theory of Testing, abbreviated NPT, sought to provide an objective means for testing. As Neyman (1971, p. 277) put it, "I find it desirable to use procedures, the validity of which does not depend upon the uncertain properties of the population." And since the hypotheses under test are uncertain, NPT rejects the idea of basing the result of the test on one's prior conceptions about them. Rather, the only piece of information to be formally mapped by the testing rules of NPT is the sample data. Since our focus will be on NPT, it will be helpful to spell out its components in some detail.

(A) NPT begins by specifying, in the *hypotheses model* $M(\theta)$, the region of the parameter space Ω to be associated with the *null hypothesis* H, i.e., Ω_H and the region to be associated with an *alternative hypothesis* J, i.e., Ω_J. The two possible results of a test of H are: reject H (and accept J) or accept H (and reject J).

(B) The *experimental testing model* $ET(\mathcal{X})$ specifies, before the sample is observed, which of the outcomes in $\mathcal{X}$ should be taken to reject H. This set of outcomes forms the *rejection or critical region*, CR, and it is specified so that, regardless of the true value of θ, the test has appropriately small probabilities of leading to erroneous conclusions. The two errors that might arise are rejecting H when it is true, the *type I error*; and accepting H when it is false, i.e., when J is true, the *type II error*. The probabilities of making type I and type II errors, called *error probabilities*, are denoted by α and β respectively. NPT is able to control the error probabilities by specifying (i) a *test statistic* S, which is a function of $\mathcal{X}$, and whose distribution is derivable from a given hypothesized value for θ and (ii) a *testing rule* T indicating which of the possible values of S fall in region CR. Since, from (i), it is possible to calculate the probability that $S \in CR$ under various hypothesized values of θ, it is possible to specify a testing rule T so that $P(S \in CR \mid \theta \in \Omega_H)$ is no greater than α and $1 - P(S \in CR \mid \theta \in \Omega_J)$ is no greater than β. Identifying the events $\{T \text{ rejects } H\}$ and $\{S \in CR\}$, it follows that

$$\begin{aligned} P(T \text{ Rejects } H \mid H \text{ is true}) &= \quad P(S \in CR \mid \theta \in \Omega_H) \leq \alpha \\ P(T \text{ Accepts } H \mid J \text{ is true}) &= 1 - P(S \in CR \mid \theta \in \Omega_J) \leq \beta. \end{aligned}$$

Since these two error probabilities cannot be simultaneously minimized, NPT instructs one to first select α, called the *size* or the *significance level* of the test, and then choose the test with an appropriately small β. However, if alternative J consists of a set of points, i.e., if it is *composite*, the value of β will vary with different values in Ω_J. The "best" NPT test of a given size α (if it exists) is the one which at the same time minimizes the value of β (i.e., the probability of type II errors) for all possible values of θ under the alternative J.

Error probabilities are deemed objective in that they refer, not to subjective degrees of belief, but to frequencies of experimental outcomes in sequences of (similar or very different) applications of a given experimental test. Since within the context of NPT parameter θ is viewed as fixed, hypotheses about it are viewed as either true or false. Thus, it makes no sense to assign them any probabilities other than 0 or 1; that is, a hypothesis is true either 100% of the time or 0% of the time, according to whether it is true or false. But the task of statistical tests arises in precisely those cases where the truth or falsity of a hypothesis is unknown. The sense in which NPT nevertheless accomplishes this task 'objectively' is that it controls the error probabilities of tests regardless of what the true, but unknown, value of θ is.

The concern with objective error probabilities is what fundamentally distinguishes NPT from other theories of statistical testing. The Bayesian, for example, is concerned only with the particular experimental outcome and the particular application of its testing rule (i.e., Bayes' Theorem); while error probabilities concern properties of T in a series of applications. As such, the Bayesian is prevented from ensuring that his tests are appropriately reliable. In order to clarify the relationship between the choice of α and the specification of the critical region of NPT, we will consider a simple, but very widespread, class of statistical tests.

2.3. *An Illustration: A NPT "Best" One Sided Test T^+*

Suppose the model $M(\theta)$ is the class of normal distributions $N(\theta, \sigma)$ with mean θ and standard deviation σ, where σ is assumed to be known, and θ may take any positive value. It is supposed that some quantity X varies according to this distribution, and it is desired to test whether the value of θ equals some value θ_0 or some greater value. Such a test involves setting out in $M(\theta)$ the following null and alternative hypotheses:

2.3(a): *Null Hypothesis H*: $\theta = \theta_0$ in $N(\theta, \sigma)$
Alternative Hypothesis J: $\theta > \theta_0$.

The null hypothesis is *simple*, in that it specifies a single value of θ, while the alternative is *composite*, as it consists of the set of θ values exceeding θ_0. The *experimental test statistic* S is the average of the n random variables $X_1, X_2 \ldots X_n$, where each X_i is $N(\theta, \sigma)$; i.e.,

2.3(b): *Test Statistic* $S = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$.

The test begins by imagining that H is true, and that X_i is $N(\theta_0, \sigma)$. But if H is true, it follows that statistic $\bar{X}$ is also normally distributed with mean θ_0; but $\sigma_{\bar{x}}$, the standard deviation, is $\sigma/n^{1/2}$. That is,

2.3(c): *Experimental Distribution* of $\bar{X}$ (under $H: \theta = \theta_0$): $N(\theta_0, \sigma/n^{1/2})$.

Since our test is interested in rejecting H just in case θ exceeds θ_0, i.e., since our test is *one sided* (in the positive direction from θ_0), it seems plausible to reject H on the basis of sample averages ($\bar{X}$ values) that sufficiently exceed θ_0; and this is precisely what NPT recommends. That is, our *test rule*, which we may represent by T^+, maps $\bar{X}$ into the critical region (i.e., T^+ rejects H) just in case $\bar{X}$ is "significantly far" (in the positive direction) from hypothesized average θ_0, where distance is measured in standard deviation units (i.e., in $\sigma_{\bar{x}}$'s). That is,

2.3(d): *Test Rule* T^+: Reject $H: \theta = \theta_0$ iff $\bar{X} \geq \theta_0 + d_\alpha \sigma_{\bar{x}}$

for some specified value of d_α where α is the *size* (significance level) of the test. The question of what the test is to count "significantly far" becomes a question of how to specify the value of d_α; the answer is provided once α is specified.

For, if the test is to erroneously reject H no more than $\alpha(100\%)$ of the time (i.e., if its *size* is α), it follows from 2.3(d) that we must have

2.3(e): $P(\bar{X} \geq \theta_0 + d_\alpha \sigma_{\bar{x}} | H) \leq \alpha$.

To ensure 2.3(e), d_α must equal that number of standard deviation units in excess of θ_0 exceeded by $\bar{X}$ no more than $\alpha(100\%)$ of the time, when H is true. And by virtue of the fact that we know the experimental distribution of $\bar{X}$ under H (which in this case is $N(\theta_0, \sigma/n^{1/2})$), we can derive, for a given α, the corresponding value for d_α. Conventional values for α are 0.01, 0.02, and 0.05, where these correspond to $d_{0.01} = 2.3$, $d_{0.02} = 2$, $d_{0.05} = 1.6$, respectively. For the distribution of $\bar{X}$ (under H) tells us that $\bar{X}$ exceeds its mean θ_0 by as much as $2.3\sigma_{\bar{x}}$ only 1% of the time; by as much as $2\sigma_{\bar{x}}$ only 2% of the time, and by as much as $1.6\sigma_{\bar{x}}$ only 5% of the time. In contrast, $\bar{X}$ exceeds θ_0 by as much as $0.25\sigma_{\bar{x}}$ as often as 40% of the time; so if H were rejected whenever $\bar{X}$ exceeded θ_0 by $0.25\sigma_{\bar{x}}$ one would erroneously do so as much as 40% of the time; i.e., $d_{0.4} = 0.25$. The relationship between α and distances of $\bar{X}$ from θ_0 is best seen by observing areas under the normal curve given by the distribution of $\bar{X}$ (under H) (see Figure 2.3).

It will be helpful to have a standard example of an application of test T^+ to which to refer our discussion.

Example ET$^+$: A good example of such an application may be provided by adapting an actual inquiry in biology.[1] We are interested in testing whether or not treating the larval food of houseflies with a

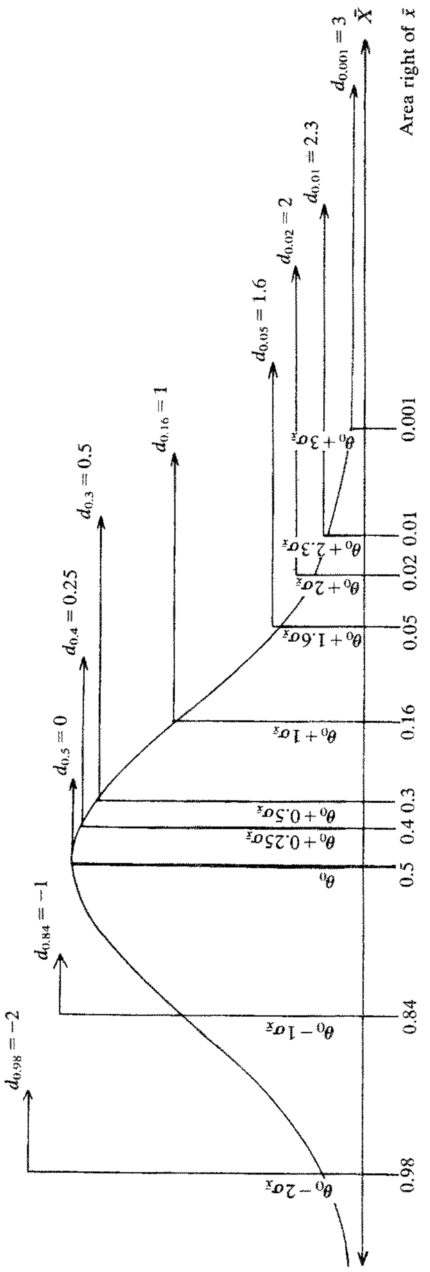

$d_{0.98} = -2$
$d_{0.84} = -1$
$d_{0.5} = 0$
$d_{0.4} = 0.25$
$d_{0.3} = 0.5$
$d_{0.16} = 1$
$d_{0.05} = 1.6$
$d_{0.02} = 2$
$d_{0.01} = 2.3$
$d_{0.001} = 3$
$\bar{X}$
Area right of $\bar{x}$
$\theta_0 - 2\sigma_{\bar{x}}$
$\theta_0 - 1\sigma_{\bar{x}}$
θ_0
$\theta_0 + 0.25\sigma_{\bar{x}}$
$\theta_0 + 0.5\sigma_{\bar{x}}$
$\theta_0 + 1\sigma_{\bar{x}}$
$\theta_0 + 1.6\sigma_{\bar{x}}$
$\theta_0 + 2\sigma_{\bar{x}}$
$\theta_0 + 2.3\sigma_{\bar{x}}$
$\theta_0 + 3\sigma_{\bar{x}}$
0.98
0.84
0.5
0.4 0.3
0.16
0.05
0.02 0.01
0.001

Fig. 2.3.

chemical (DDT) increases their size, as measured by their winglength X. Earlier studies show that the average winglength θ in the population of normal houseflies whose food is *not* so treated is 45.5 mm and the question is whether DDT increases θ. That is, the inquiry concerns the following scientific claim (or some version of it):

2.3(f): *Scientific Claim* $\mathscr{C}^+$: DDT increases the average winglength of houseflies (i.e., in DDT treated houseflies, θ is (importantly) greater than 45.5).

The problem is that a sample from this population, whose food has been treated with DDT, may be observed to have an average winglength $\bar{X}$ in excess of the normal average 45.5 even though the chemical does *not* possess the positive effect intended. As such, the need for statistical considerations arises.

(A) *Statistical Hypotheses*: It is supposed that the variability of X, the winglength of houseflies in a given population, can be modelled by a normal distribution with known standard deviation σ equal to 4, and unknown mean θ, i.e., X is $N(\theta, 4)$. Hence, if the chemical does not increase winglengths as intended, the distribution of X among flies treated with it would not differ from what scientists have found in the typical housefly population; that is, θ would be 45.5 in $N(\theta, 4)$. On the other hand if $\mathscr{C}^+$ is true, and DDT does have its intended effect, θ will exceed the usual 45.5. These correspond to the following statistical hypotheses:

2.3(a′): *Null Hypothesis* H: $\theta = 45.5$ in $N(\theta, 4)$
Alternative Hypothesis J: $\theta > 45.5$ in $N(\theta, 4)$.

(B) *Experimental Test*: ($\mathrm{ET}^+ - 1$) Since we are within the context of the one-sided test above, each of the general components there apply to the present example. In particular, (b′) experimental statistic $S = \bar{X}$, and (c′) its distribution under H is $N(45.5, 4/n^{1/2})$; and the "best" testing rule, according to NPT, is (d′) T^+. Suppose, in addition, the following test specifications are made:

(i) Number of observations in sample, $n = 100$
(ii) Size of test $\alpha = 0.02$.

It follows from (i) that $\sigma_{\bar{x}} = 0.4$ and from (ii) that $d_\alpha = d_{0.02} = 2$. So $T^+ - 1$ is

2.3(d′): $T^+ - 1$: Reject H iff $\bar{X} \geq 45.5 + d_\alpha \sigma_{\bar{x}} = 46.3$.

(C) *Sample Data*: The average winglength of the 100 observed flies (treated with DDT), $\bar{x}$, is 46.5, a value in excess of the hypothesized average 45.5 by $2.5\sigma_{\bar{x}}$. Since this value exceeds 46.3, $T^{+}-1$ would deem such an $\bar{x}$ "significantly far" from H to reject it. Hence $T^{+}-1$ rejects H.

In addition to ensuring the maximum frequency of erroneously rejecting H is 0.02, this test ensures that the frequency of erroneously accepting H will be minimized for all possible θ values under the alternative J, (i.e., for all $\theta > \theta_0$). Such a test is a "best" NPT test of size 0.02, in the context of ET^{+}.

3. THE OBJECTIVITY OF NPT UNDER ATTACK

While NPT may provide tests that are good, or even "best" according to the criteria of low error probabilities, the question that concerns us is whether such tests are also good from the point of view of the aims of scientific objectivity. Answering this question takes us into the problem of how to relate an empirical scientific inquiry to the statistical models of NPT. But this problem is outside the domain of mathematical statistics itself, and may be seen as a problem belonging to *metastatistics*. For, its resolution requires stepping outside the statistical formalism and making statistics itself an object of study.

We can organize the problem of the objectivity of NPT around three major metastatistical tasks in linking a scientific inquiry with each of the three statistical models of NPT:

(1) relating the statistical hypotheses in $M(\theta)$ and the results of testing them to scientific claims;
(2) specifying the components of experimental test $\mathrm{ET}(\mathscr{X})$; and,
(3) ascertaining whether the assumptions of a model for the data of the experimental test are met by the empirical observations (i.e., relating $\langle x_n \rangle$ to $\mathscr{O}$).

These three concerns correspond to the three rays diagrammed in Figure 2.1 numbered (1)–(3), and what is asked in questioning the objectivity of NPT is whether NPT can accomplish the above tasks (1)–(3) objectively. As will be seen, these three tasks are intimately related, and hence, the problem of objectively accomplishing them are not really separate. However, the most serious attacks on the objectivity of NPT focus on the interrelated problems of (1) relating statistical conclusions to scientific claims and (2) specifying statistical tests in an objective manner; as such, our primary aim will be to clarify and resolve these problems. In Section 5, we will briefly discuss the problem of (3) validating model assumptions objectively and how it would be handled within the present approach.

3.1 *NPT as an 'Objective' Theory of Inductive Behavior*

Although once the various formal experimental specifications are made, NPT provides an objective means for testing a hypothesis, it is not clear that these specifications are themselves objective. For these specifications involve considerations that go beyond the formalism of statistical testing. As Neyman and Pearson (1933, p. 146) note:

> From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

Acknowledging that the tasks of (1) interpreting and (2) specifying tests go beyond the objective formalism of NPT, Neyman and Pearson consider a situation where there would be some clear basis for accomplishing them. Noting that the sorts of informal considerations NPT requires are similar to those involved in certain decision theoretic contexts, they are led to suggest the behavioristic interpretation of NPT. Although the behavioristic construal of NPT was largely advocated by Neyman and was not wholly embraced by Pearson (see Pearson 1955), NPT has generally been viewed as purporting to offer an *objective theory of inductive behavior.*

In an attempt to develop testing as an objective theory of behavior, tests are formulated as mechanical rules or "recipes" for reaching one of two possible decisions: accept hypothesis H or reject H, where "accept H" and "reject H" are interpreted as deciding to "act as if H were true" and to "act as if H were false", respectively. Neyman and Pearson's (1933, p. 142) own description of "rules of behavior" is a clear statement of such a conception:

> Here, for example, would be such a 'rule of behavior': to decide whether a hypothesis H, of a given type, be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H; if $x \leq x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that

> if we behave according to such a rule. . . we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.

On this view, the rationale for using such a rule of behavior is the desire for a rule with appropriately small probabilities α, β for leading to erroneous decisions in a given long run sequence of applications of the rule. Although the actual values of α, β are considered beyond the NPT formalism, it is suggested that the scientist first specify α as the maximum frequency with which he feels he can afford to erroneously reject H. He then selects the test that at the same time minimizes the value of β, i.e., of erroneously accepting H.

This rationale corresponds to the NPT view of the purpose of tests; namely, "as a result of the tests a decision must be reached determining which of two alternative courses is to be followed" (Neyman and Pearson, 1936, p. 204). For example, rejecting H in our housefly experiment may be associated with a decision to publish results claiming that DDT increases winglengths of flies, a decision to ban the use of DDT, a decision to do further research, and so on; and for each decision there are certain losses and costs associated with acting on it when in fact H is true. By considering such consequences the scientist is, presumably, able to specify the risks he can "afford". However, such considerations of consequences are deemed subjective by NPT's founders, (e.g., "this subjective element lies outside of the theory of statistics" (Neyman, 1950, p. 263)); and the desire to keep NPT objective leads to the intentional omission of any official incorporation of pragmatic consequences within NPT itself. Ironically, this attempt to secure NPT's objectivity is precisely what leads both to attacks on its objectivity and to strengthening the position of subjective theories of testing.

3.2 *The Objectivity of the Behavioristic Interpretation of NPT Under Attack*

The most serious problems concerning the objectivity of NPT center around its ability to objectively accomplish the tasks of (1) relating results of NPT to scientific claims and (2) specifying the components of experimental tests. The major criticism is this: Since the scientific import of a statistical test conclusion is dependent upon the specifications of the test, the objectivity of test conclusions is jeopardized if the test specifications fail to be objective. But NPT leaves the problem of test specifications up to the individual researcher who is to specify them according to his personal considerations of the risks (of erroneous decisions) he finds satisfactory. But if the aim of a scientist is to be seen as objectively finding out what is the case as opposed to finding out how it is most prudent for him to act, then he does not seem to be in the position of objectively making pragmatic judgments about how often he can afford to be wrong in some long run of cases. And if he is left to make these pragmatic judgments subjectively, it is possible for him to influence the test in such a way that it is overwhelmingly likely to produce the result he happens to favor. In short, it appears that the 'objectivity' of NPT is somewhat of a sham; its results seem to be permeated with precisely those subjective, personal factors that are antithetical to the aim of scientific objectivity.

Among the first to criticize the objectivity of NPT on these grounds was R. A. Fisher. While it was Fisher's ideas that formed the basis of NPT,[2] by couching them in a behavioristic framework he felt Neyman and Pearson had given up the ideal of objectivity in science. In his grand polemic style, Fisher (1955, p. 70) declared that followers of the behavioristic approach are like

> 3.2(a) Russians (who) are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation.

Not to suggest a lack of parity between Russia and the U.S., he continues:

> In the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money.

One finds analogous attacks on the objectivity of NPT voiced by a great many contemporary statisticians and philosophers. They typically begin by denying the objectivity of NPT specifications:

> 3.2(b) In no case can the appropriate significance level be determined in an objective manner. (Rubin, 1971, p. 373)

The basis of this denial is that fixing their levels require judgments – judgments which, it is claimed, (to use a phrase of I. J. Good) NPT tends to "sweep under the carpet" (i.e., SUTC). Good (1976, p. 143) claims

> 3.2(c) Now the *hidebound* objectivist tends to hide that fact; he will not volunteer the information that he uses judgment at all. . . .

At least the Bayesian theorist, it is argued, admits the need for pragmatic judgments and explicitly incorporates them into his testing theory. Rosenkrantz (1977, pp. 205–206) sums it up well:

> 3.2(d) In Bayesian decision theory, an optimal experiment can be determined, taking due account of sampling costs. But no guidelines exist for the analogous problem of choosing a triple (α, β, n) in Neyman–Pearson theory, . . . These judgments are left to the individual experimenter, and they depend in turn on his personal utilities. . . . But if we are 'interested in what the data have to tell us', why should one experimenter's personal values and evaluations enter in at all? How, in short, can we provide for *objective scientific reporting* within this framework? (emphasis added)

Implicit in these criticisms are various assumptions of what *would* be required for objective scientific reporting. A clear example of this is found in Fetzer's attack on the objectivity of NPT. Fetzer (1981, p. 244) argues that while the subjective judgments required to specify NPT "are not 'defects'" in decision-making contexts, in general, nevertheless,

> 3.2(e) to the extent to which principles of inference are intended to secure the desideratum of 'epistemic objectivity' by supplying standards whose systematic application . . . would warrant assigning all and only the same measures of evidential support to all and only the same hypotheses, they [NPT] appear inadequate.

On this view – a view which underlies the most serious attacks on the objectivity of NPT – a theory of statistics should provide means for measuring the degree of evidential strength (support, probability, belief, etc.) that data affords hypotheses. Correspondingly, a theory of statistics is thought to be objective only if it (to use Fetzer's words) "would warrant assigning all and only the same measures of evidential support to all and only the same hypotheses" (given the same data).

On this view, NPT, which only seeks to ensure the scientist that he will make erroneous decisions no more than a small percent of the time in a series of test applications, fails to be objective. For NPT cannot tell him whether a particular conclusion is one of the erroneous ones or not; nor can it provide him with a measure of how probable or of how well supported a particular conclusion is. For example if H is rejected with a test with size 0.02, it is not correct to assign alternative J 98% probability, support, or confidence – although such misinterpretations are common. The only thing 0.02 tells him about a specific rejection of H is that it was the result of a general testing *procedure* which erroneously rejects H only 2% of the time. Similarly, thu NPT rationale may permit null hypothesis H to be accepted without guaranteeing that H is highly supported or highly probable; it may simply mean a given test was unable to reject it. As Edwards (1971, p. 18) puts it:

> 3.2(f) Repeated non-rejection of the null hypothesis is too easily interpreted as indicating its acceptence, so that on the basis of no prior information coupled with little observational data, the null hypothesis is accepted.... Far from being an exercise in scientific objectivity, such a procedure is open to the gravest misgivings. What used to be called prejudice is now called null hypothesis....

Edwards is referring to the fact that if one has a prejudice in favor of null hypothesis H one can specify a test so that it has little chance of rejecting H. Consider our housefly winglength example. In $ET^+ - 1$ the test size α was set at 0.02, but if a researcher were even more concerned to avoid erroneously rejecting H he might set α to an even smaller value. (Possibly the researcher manufactures DDT and rejecting H would lead to banning the chemical.) Consider $ET^+ - 2$, where α is now set at 0.001. The corresponding test $T^+ - 2$ would not reject H unless $\bar{X}$ exceeded θ_0 (45.5) by at least 3 standard deviation units, as compared to only 2 standard deviation units when α was set at 0.02. But with the size set at 0.001, the observed average of 46.5 in the sample would *not* be taken to reject H as it was in $ET^+ - 1$; rather H would be accepted. (The test T^+-2 is: reject H iff $\bar{X} \geq 46.7$.) In the extreme case, one can ensure that H is never rejected by setting $\alpha = 0$!

The fact that the same data leads to different conclusions depending on the specification of α is entirely appropriate when such specifications are intended to reflect the researcher's assessment of the consequences of erroneous conclusions. For, as Neyman and Pearson (1936, p. 204) assert, "clearly established consequences which appear satisfactory to one individual may not be so regarded by another." But if "same data, same test result" is taken as a requirement for an objective theory of testing, this feature of NPT will render it entirely inappropriate for objective scientific reporting. As Kyburg (1971, pp. 82–83) put it:

> 3.2(g) To talk about accepting or rejecting hypotheses, for example is *prima facie* to talk epistemologically; and yet in statistical literature to accept the hypothesis that the parameter θ is less than θ^* is often merely a fancy and roundabout way of saying that Mr Doe should offer no more than \$36.52 for a certain bag of bolts. . . .
>
> When it comes to general scientific hypotheses (e.g., that $f(x)$ represents the distribution of weights in certain species of fish. . .) then the purely pragmatic, decision theoretic approach has nothing to offer us.

If Kyburg is right, and NPT "has nothing to offer us" when testing hypotheses about the distribution of fish weights, it will not do much better when testing hypotheses about the distribution of housefly winglengths! And since our aim is to show that NPT does provide objective means for testing such hypotheses, we will have to be able to respond to the above criticisms. By making some brief remarks on these criticisms our grasp of the problem may be enriched and the present approach elucidated.

3.3 *Some Presuppositions About NPT and Objectivity: Remarks on the Attacks*

How one answers the question of whether NPT lives up to the aim of scientific objectivity depends upon how one answers two additional questions: (I) What functions are served by NPT?; and (II) What is required for a theory of statistical testing to satisfy the aim of scientific objectivity? If the attacks on the objectivity of NPT turn out to be based upon faulty answers to either (I) or (II), we need not accept their conclusion denying NPT's objectivity.

Generally, the attacks on NPT are based on the view that NPT functions to provide objective rules of behavior. But there are many interpretations one can place on a mathematical theory, and the behavioristic one is not the only interpretation of which NPT admits. Moreover, it seems to reflect neither the actual nor the intended uses of NPT in science. In an attempt "to dispel the picture of the Russian technological bogey" seen in Fisher's attack (3.2(a)), Pearson (1955, p. 204) notes that the main ideas of NPT were formulated a good deal before they became couched in decision-theoretic terms. Pearson insists that both he and Neyman "shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'." Nevertheless, within NPT they failed to include an indication of how to use tests as learning tools – deeming such extrastatistical considerations subjective. As such, it appears that they shared the view of objectivity held by those who attack NPT as hopelessly subjective.

Underlying such flat out denials of the possibility of objectively specifying α as Rubin's (3.2(b)) is the supposition that only what is given by formal logical principles alone is truly objective. But why should a testing theory seek to be objective in this sense? It was such a conception of objectivity that led positivist philosophers to seek a logic of inductive inference that could be set out in a formal way (most notably, Carnap and his followers). But their results – while logical masterpieces – have tended to be purely formal measures having questionable bearing on the actual problems of statistical inference in science. The same mentality, of course, has fostered the uncritical use of NPT, leading to misuses and criticisms.

Similarly, we can question Good's suggestion (3.2(c)) that a truly objective theory must not include any judgments whatsoever. Attacking the objectivity of NPT because judgments are required in applying its tests is to use a straw man argument; that judgments are made in using NPT is what allows tests to avoid being sterile formal exercises. A ban on judgments renders all human enterprises subjective. Still, the typical Bayesian criticism of NPT follows the line of reasoning seen in Good and Rosenkrantz (3.2(d)) above. They reason that since NPT requires extrastatistical judgments and hence is not truly objective, it should make its judgments explicit (and stop sweeping them under the carpet) – which (for them) is to say, it should incorporate degrees of belief, support, etc. in the form of prior probabilities. However, the fact that both NPT and the methods of Bayesian tests require extrastatistical judgments does not mean that the judgments required are equally arbitrary or equally subjective in both – a point to be taken up later. Nor does it seem that the type of judgments needed in applying NPT are amenable to quantification in terms of prior probabilities of hypotheses. But their arguments are of value to us; they make it clear that our defense of the objectivity of NPT must abandon the "no extrastatistical judgments" view of objectivity. At the same time we need to show that it is possible to avoid the sort of criticisms that the need for such judgments is thought to give rise.

Edwards (3.2(f)), for example, attacks NPT on the grounds that the extrastatistical judgments needed to specify its tests may be made in such a way that the hypothesis one happens to favor is overwhelmingly likely to be accepted. What could run counter to the ideal of scientific objectivity more! – or so this frequently mounted attack supposes. But this attack, we maintain, is based on a misconception of the function of NPT *in science*; one which, unfortunately, has been encouraged by the way in which NPT is formulated. The misconception is that NPT functions to directly test scientific claims from which the statistical hypotheses are derived.[3] As a result it is thought that since a test with appropriate error probabiities warrants the resulting statistical conclusion (e.g., reject H), it automatically warrants the corresponding scientific one (e.g., $\mathscr{C}^+$). This leaves no room between the statistical and the scientific conclusions whereby one can critically interpret the former's bearing on the latter (i.e., ray (1) in Figure 2.1 is collapsed). As we will see, by clearly distinguishing the two such a criticism *is* possible; and the result of such a criticism is that Edwards' attack can be seen to point to a possible *misinterpretation* of the import of a statistical test, rather than to its lack of objectivity (see Section 4.4).

From Fetzer's (3.2(e)) argument we learn that our defense of NPT must abandon the supposition that objectivity requires NPT to satisfy the principle of "same data, same assignment of evidential support to hypotheses". That NPT fails to satisfy this principle is not surprising since its only quantitative measures are error probabilities of test procedures, and these are not intended to provide measures of evidential support.[4] Moreover, as we saw, NPT does allow one to have "Same (sample) data, *different* error probabilities." So, it is also not surprising that if error probabilities are misconstrued as evidential support measures – something which is not uncommon – the failure of NPT to satisfy the principle of "same data, same evidential support" will follow. Our task in the next section will be to show that without misconstruing error probabilities it is possible to use them to satisfy an altered conception of scientific objectivity.

4. DEFENSE OF THE OBJECTIVITY OF (A REFORMULATED) NPT: NPT*

If we accept the conception of (I) the function of NPT in science, and (II) the nature of scientific objectivity that the above attacks presuppose, then, admittedly, we must conclude that NPT fails to be objective. But since the correctness of these conceptions is open to the questions we have raised, we need not accept this conclusion. Of course, rejecting this conclusion does not constitute a positive argument showing that NPT is objective. However, our grounds for doing so point the way to the sort of altered conceptions of (I) and (II) that *would* permit such a positive defense to be provided; and it is to this task that we now turn.

Like Pearson (3.3), we view the function of statistical tests in a scientific inquiry as providing "a means of learning" from empirical data. However, unlike the existing formulation of NPT, we seek to explicitly incorporate its learning function within the model of statistical testing in science itself. That is, our model of testing will include some of the metastatistical elements linking the statistical to the scientific inquiry. These elements are represented by ray (1) in Figure 2.1. To contrast our reformulation of the function of NPT with the behavioristic model, we may refer to it as the *learning model*, or just NPT*. Having altered the function of tests we could substitute existing test conclusions, i.e., accept or reject H, by assertions expressing what has or has not been learned. However, since our aim is to show that existing statistical tests can provide a means for objective scientific learning, it seems preferable to introduce our reformulation in terms of a metastatistical interpretation of existing test results.

While our learning (re)interpretation of NPT (yielding NPT*) goes beyond what is found in the formal apparatus of NPT, we argue that it is entirely within its realm of legitimate considerations. It depends every step of the way on what is fundamental to NPT; namely, being able to use the distribution of test statistic S to objectively control error probabilities. Moreover, we will argue, it is the objective control of error probabilities that enables NPT* to provide objective scientific knowledge.

4.1 *NPT as an Objective Means for Learning by Detecting Discrepancies*: *NPT**

Rather than view statistical tests in science as either a means of deciding how to behave or a means of assigning measures of evidential support we view them as a means of learning about variable phenomena on the basis of limited empirical data. This function is accomplished by providing tools for *detecting* certain *discrepancies* between the (approximately) correct models of a phenomenon and the hypothesized ones; that is, between the pattern of observations that would be observed (if a given experiment were repeated) and the pattern hypothesized by the model being tested. In ET^+, for example, we are interested in learning if the actual value of θ is positively discrepant from the hypothesized value, θ_0. In the case of our housefly inquiry, this amounts to learning if DDT-fed flies would give rise to observations describable by $N(45.5, 4)$, or by $N(\theta, 4)$ where θ exceeds 45.5.

Our experiment, however, allows us to observe, not the actual value of this discrepancy, but only the *difference* between the sample statistic S (e.g., $\bar{X}$) and a hypothesized population parameter θ (e.g., 45.5). A statistical test provides a standard measure for classifying such observed differences as "significant" or not. Ideally, a test would classify an observed difterence significant, in the statistical sense, just in case one had actually detected a discrepancy of scientific importance. However, we know from the distribution of S that it is possible for an observed s to differ from a hypothesis θ_0, even by great amounts, when no discrepancy between θ and θ_0 exists. Similarly, a small or even a zero observed difference is possible when large underlying discrepancies exist. But, by a suitable choice of S, the size of observed differences can be made to vary, in a *known manner*, with the size of underlying discrepancies. In this way a test can be specified so that it will very infrequently classify an observed difference as significant (and hence reject H) when no discrepancy of scientific importance is detected, and very infrequently fail to do so (and so accept H) when θ is importantly discrepant from θ_0. As such, the rationale for small error probabilities reflects the desire to detect all and only those discrepancies about which one wishes to learn in a given scientific inquiry – as opposed to the desire to infrequently make erroneous decisions. That is, it reflects epistemological rather than decision-theoretic values.

This suggests the possibility of objectively (2) specifying tests by appealing to considerations of the type of discrepancies about which it would be important to learn in a given scientific inquiry. Giere (1976) attempts to do this by suggesting objective grounds upon which "professional judgments" about scientifically important discrepancies may be based. With respect to such judgments he remarks:

> It is unfortunate that even the principal advocates of objectivist methods, e.g., Neyman, have passed off such judgments as merely 'subjective'.... They are the kinds of judgments on which most experienced investigators should agree. More precisely, professional judgments should be approximately invariant under interchange of investigators. (p. 79)

Giere finds the source of such intersubjective judgments in scientific considerations of the uses to which a statistical result is to be put. For example, an accepted statistical hypothesis "is regarded as something that must be explained by any proposed general theory of the relevant kind of system... ."

However, critics (e.g., Rosenkrantz (1977, p. 211) and Fetzer (1981, p. 242)) may still deny that Giere's appeal to "professional judgments" about discrepancies renders test specifications any more objective than the appeal to pragmatic decision-theoretic values. In addition, the conclusion of a statistical test (accept or reject *H*) – even if it arose from a procedure which satisfied the (pre-trial) professional judgments of investigators – may still be accused of failing to express the objective import of the particular result that happened to be realized. While we deem the appeal to discrepancies of scientific importance, exemplified in Giere's strategy, to be of great value in resolving the problem of the objectivity of NPT, we attempt to use it for this end in a manner that avoids such criticisms.

As Giere notes, there do often seem to be good scientific grounds for specifying a test according to the discrepancies deemed scientifically important. However, we want to argue, even if objective grounds for (2) specifying tests are lacking, it need not preclude the possibility of objectively accomplishing the task of (1), relating the result of a statistical test to the scientific claim. For, regardless of how a test has been specified, the distribution of test statistic *S* allows one to determine the probabilistic relations between test results and underlying discrepancies. Moreover, we maintain, by making use of such probabilistic relations (in the form of error probabilities) it is possible to objectively understand the discrepancies that have or have not been detected on the basis of a given test result; and in this way NPT* accomplishes its learning function objectively. Since error probabilities do not provide measures of evidential support, our view of objective scientific learning must differ from the widely held supposition that objective statistical theories must provide such measures. Hence, before going on to the details of our reformulation of NPT, it will help to reconsider just what an objective theory of statistical testing seems to require.

4.2. *An Objective Theory of Statistical Testing Reconsidered*

If the function of a statistical test is to learn about variable phenomena by using stastically modelled data to detect discrepancies between statistically modelled conjectures about it, then an objective theory of statistical testing must be able to carry out this learning function objectively. But what is required for objective learning in science? We take the veiw that objectivity requires assertions to be checkable, testable, or open to criticism by means of fair standards or controls. Correspondingly, objective learning from the result of a statistical test requires being able to critically evaluate what it does and does not indicate about the scientific phenomenon of interest (i.e., what it does or does not detect). In NPT, a statistical result is an assertion to either reject or accept a statistical hypothesis H, according to whether or not the test classifies the observed data as significantly far from what would be expected under H. Hence, objective learning from the statistical result is a matter of being able to critically evaluate what such a statistically classified observation does or does not indicate about the variability of the scientific phenomenon. So, while the given statistical result is influenced by the choice of classification scheme provided by a given test specification, this need not preclude objective learning on the basis of an observation so classified. For, as Scheffler (1967, p. 38) notes, "having adopted a given category system, our hypotheses as to the actual distribution of items within the several categories are not prejudged." The reason is that the result of a statistical test "about the distribution of items within the several categories" is only partly determined by the specifications of the categories (significant and insignificant); it is also determined by the underlying scientific phenomenon. And what enables objective learning to take place is the possibility of devising ingenious means for "subtracting out" the influence of test specifications in order to detect the underlying phenomenon.

It is instructive to note the parallels between the problems of objective learning from statistical experiments and that of objective learning from observation in general. The problem in objectively interpreting observations is that observations are always relative to the particular instrument or observation scheme employed. But we often are aware not only of the fact that observation schemes influence what we observe, but also of *how* they influence observations and of how much "noise" they are likely to produce. In this way its influence may be subtracted out to some extent. Hence, objective learning from observation is not a matter of getting free of arbitrary choices of observation scheme, but a matter of *critically evaluating the extent of their influence* in order to get at the underlying phenomenon. And the function of NPT on our view (i.e., NPT*) may be seen as providing a systematic means for accomplishing such learning. To clarify this function we may again revert to an analogy between tests and observational instruments.

In particular, a statistical test functions in a manner analogous to the way an instrument such as an ultrasound probe allows one to learn about the condition of a patient's arteries indirectly. And the manner in which test specifications determine the classification of statistical data is analogous to the way the sensitivity of the ultrasound probe determines whether an image is classified as diseased or not. However, a doctor's ability to learn about a patient's condition is not precluded by the fact that different probes classify images differently. Given an image deemed diseased by a certain ultrasound probe, doctors are able to learn about a patient's condition by considering how frequently such a diseased image results from observing arteries diseased to various extents. In an analogous fashion, learning from the result of a given test is accomplished by making use of probabilistic relations between such results and underlying values of a parameter θ, i.e., via error probabilities.

Having roughly described our view of objective learning from statistical tests, it will help to imagine an instrument, which, while purely invented, enables a more precise description of how NPT* learns from an experimental test such as ET^+: This test may be visualized as an instrument for categorizing observed sample averages according to the width of the mesh of a netting on which they are "caught". Since the example on which our discussion focuses concerns the average winglength of houseflies, it may be instructive to regard ET^+ as using a type of flynet to catch flies larger than those arising from a population of flies correctly described by $H: \theta = \theta_0$. Suppose that a netting of size α catches an observed average winglength $\bar{x}$ just in case $\bar{x} \geq \theta_0 + d_\alpha \sigma_{\bar{x}}$. Then a test with size α categorizes a sample as "significant" just in case it is caught on a size α net. Suppose further that it is known that $\alpha(100\%)$ of the possible samples (in $\mathscr{X}$) from a fly population where $\theta = \theta_0$ would land on a netting of size α. Then specifying a test to have a small size α (i.e., a small probability of a type I error) is tantamount to rejecting $H: \theta = \theta_0$ just in case a given sample of flies is caught on a net on which only a small percentage of samples from fly population H would be caught.

For T^+ tests an $\bar{x}$ value may be "caught" on the net deemed significant by one test and not by another because the widths of their significant nets differ. But while "significantly large" is a notion relative to the 'arbitrarily' chosen width of the significant net, the notions "significant according to test T" (or, "caught by test T") may be understood by anyone who understands the probabilistic relation between being caught by a given test and various underlying populations. And since NPT provides such a probabilistic relation (by means of the distribution of the experimental test statistic) it is possible to understand what has or has not been learned about θ by means of a given test result. In contrast, a Bayesian views considerations about results that have not been observed irrelevant for reasoning from the particular result that happened to be observed. As Lindley (1976, p. 362) remarks.

> In Bayesian statistics . . . once the data is to hand, the distribution of X [or $\mathscr{X}$], the random variable that generated the data, is irrelevant.

Hence, one is unable to relate the result of a Bayesian test to other possible experimental tests – at least not by referring to the distribution of the test statistic (e.g., $\bar{X}$). So if our view of statistical testing in science is correct, it follows that the result of a Bayesian test cannot provide a general means for objectively assessing a scientific claim (e.g., $\mathscr{C}^+$). For such assessments implicitly refer to other possible experiments (whether actual or hypothetical) in which the present result could be repeated or checked. But according to Lindley (p. 359) "the Bayesian theory is about *coherence*, not about right or wrong," so the fact that it allows consistent incorrectness to go unchecked may not be deemed problematic for a subjective Bayesian.

It follows, then, that what fundamentally distinguishes the subjective Bayesian theory from NPT* is not the inclusion or exclusion of extrastatistical judgments, for they both require these. Rather, what fundamentally distinguishes them is that NPT*, unlike Bayesian tests, allows for objective (criticizable) scientific assertions; that is, for objectively accomplishing the task of (1) relating statistical results to scientific claims. Having clearly distinguished the statistical conclusion from the subsequent scientific one, we can deny that any arbitrariness on the level of formal statistics necessarily injects arbitrariness into assertions based on statistical results. Moreover, we have suggested the basic type of move whereby we claim NPT* is able to avoid arbitrariness in evaluating the scientific import of its tests. In the next two sections, we attempt to show explicitly how this may be carried out by objectively interpreting two statistical conclusions from ET^+.

Before doing so, however, two things should be noted. First, it is not our view that the result of a single statistical inquiry suffices to find out all that one wants to learn in typical scientific inquiries; rather, numerous statistical tests are usually required – some of which may only be interested in learning how to go on to learn more. As we will be able to consider only a single statistical inquiry, our concern will be limited to objectively evaluating what may or may not be learned from it alone.[5] Secondly, it is not our intention to legislate what information a given scientific inquiry requires, in the sense of telling a scientist the magnitude of discrepancy he should deem important. On the contrary, our aim is to show how an objective interpretation of statistical results (from ET^+) is possible *without* having to precisely specify the magnitude of the discrepancy that is of interest for the purposes of the scientific inquiry. This interpretation will consist of indicating the extent to which a given statistical result (from ET^+) allows one to learn about various θ values that are positively discrepant from θ_0.

4.3 *An Objective Interpretation of Rejecting a Hypothesis (with T^+)*

In order to objectively interpret the result of a test we said one must be able to critically evaluate what it does and does not indicate, and such a critical evaluation is possible if one is able to distinguish between correct interpretations and incorrect ones (i.e., misinterpretations). Hence, the focus of our metastatistical evaluation of the result of a statistical test will be various ways in which its scientific import may be *misconstrued*. A rejection of a statistical hypothesis H is misconstrued or misinterpreted if it is erroneously taken to indicate that a discrepancy of scientific importance has been detected. But which discrepancies are of "scientific importance"? In ET^+, for example, is any positive discrepancy, no matter how small, important? Although our null hypothesis H states that θ is precisely 45.5 mm, in fact, even if a population is that of normal untreated flies the assertion that $\theta = 45.5$ is not precisely correct to any number of decimal places. Since winglength measurements and $\mathscr{C}^+$ are in terms of 0.1 mm, no distinction is made between, say, 45.51 and 45.5. So, at the very least, the imprecision of the scientific claim prevents any and all discrepancies from being of scientific importance. However, we would also be mistaken in construing an observed difference as indicative of an importantly greater θ if in fact it was due to the various sources of genetic and environmental error that result in the normal, expected, variability of winglengths, all of which may be lumped under *experimental error*.

The task of distinguishing differences due to experimental error or "accidental effects" is accomplished by NPT by making use of knowledge as to how one may fail to do so. In particular, since a positive difference between $\bar{X}$ and its mean θ of less than 1 or 2 standard deviation units arises fairly frequently from experimental error, such small differences may often erroneously be confused with effects due to the treatment of interest (e.g., DDT). By making α small (e.g., 0.02), only observed differences as large as d_α (e.g., 2) $\sigma_{\bar{x}}$'s are taken to reject H, i.e., to deny the difference is due to experimental error alone. But even a small α does not protect one from misconstruing a rejection of H. For even if the test result rarely arises (i.e., no more than $\alpha(100\%)$ of the time) from experimental error alone, it may still be a poor indicator of a θ value importantly greater than 45.5; that is, it may be a misleading *signal* of $\mathscr{C}^+$. The reason for this is that, regardless of how small α is a test can be specified so that it will almost always give rise to a $\bar{x}$ that exceeds θ_0 (45.5) by the required $d_\alpha\sigma_{\bar{x}}$'s, even if the underlying θ exceeds θ_0 by as little as one likes. Such a sensitive, or *powerful*, test results by selecting an appropriately large sample size n. In this way $\sigma_{\bar{x}}$, and correspondingly, $\theta_0 + d_\alpha \sigma_{\bar{x}}$, can be made so small that even the smallest flywings are caught on the α-significant netting. That NPT thereby allows frequent misconstruals of the scientific import of a rejection of H is often taken to show its inadequacy for science. But such misconstruals occur only if a rejection of H is automatically taken to indicate $\mathscr{C}^+$.

How, on the metastatistical level, is such a misconstrual to be avoided? To answer this, consider how one interprets a failing test score on an (academic or physical) exam. If it is known that "such a score" frequently arises when only an unimportant deficiency is present, one would deny that a large important deficiency was indicated by the test score. Reverting to our ultrasound probe analogy, suppose a given image is classified as diseased by the probe, but that the doctors know that "such a diseased image" frequently arises (in using this probe) when no real disease has been detected (perhaps it is due to a slight shadow). Then, the doctors would deny that the image was a good indicator of the existence of a seriously diseased artery. (Remember, we are not concerned here with what action the doctors should take, but only with what they have learned.) Similarly, a rejection of H with a given $\bar{x}$ is not a good indicator of a scientifically important value of θ, if "such a rejection" (i.e., such a statistically significant result) frequently results from the given test when the amount by which θ exceeds θ_0 is scientifically unimportant. By "such a score" or "such a rejection", we mean one deemed as or even more significant than the one given by the test. Since the ability to criticize an interpretation of a rejection involves appealing to the frequency relation between such a result and underlying values of θ, it will be helpful to introduce a function that makes this relationship precise.

Suppose $H: \theta = \theta_0$ is rejected by a test T^+ (in favor of alternative $J: \theta > \theta_0$) on the basis of an observed average $\bar{x}$. Using the distribution of $\bar{X}$ for $\theta \in \Omega$ define

4.3(a): $\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geqslant \bar{x} | \theta)$

That is, $\hat{\alpha}(\theta)$ is the area to the right of observed $\bar{x}$, under the Normal curve with mean θ (with known σ). In the special case where $\theta = \theta_0$, $\hat{\alpha}$ (sometimes called the observed significance level) equals the frequency of erroneously rejecting H with "such an $\bar{x}$" (i.e., the frequency of "such a Type I error.") Then, test T^+ rejects H with $\bar{x}$ just in case $\hat{\alpha}(\theta_0) \leqslant$ the (preset) size of the test.[6] To assert alternative J, that $\theta > \theta_0$, is to say that the sort of observations to which the population being observed (e.g., DDT-treated flies) gives rise are *not* those that typically arise from a population with θ as small as θ_0. That is, it is denied that the observations are describable as having arisen from a statistical distribution $M(\theta_0)$. Then, a T^+ rejection of $H: \theta = \theta_0$ is a good indicator that $\theta > \theta_0$ to the extent that such a rejection truly is not typical when θ is as small as θ_0 (i.e., to the extent that $\bar{x}$ goes beyond the bounds of typical experimental error from θ_0.) And this is just to say that $\hat{\alpha}(\theta_0)$ is small. Hence, rejecting H with a T^+ test with small size α indicates that $J: \theta > \theta_0$. It follows that if any and all positive discrepancies from θ_0 are deemed scientifically important, then a small size α ensures that construing such a rejection as indicating a scientifically important θ would rarely be erroneous. But so long as some θ values in excess of θ_0 are still not deemed scientifically important, even a small size α does not prevent a T^+ rejection of H from often being misconstrued when relating it to $\mathscr{C}^+$.

Let us define θ_{un} as follows:

4.3(b): θ_{un} = the largest scientifically unimportant θ value in excess of θ_0.

Then we can represent the task of relating the statistical to the scientific claim $\mathscr{C}^+$ (i.e., task (1)) by:

4.3(c): $$\frac{\text{Statistical Result:}}{\text{Reject } H: \theta = \theta_0 \text{ with } \bar{x}} \xrightarrow{(1)} \frac{\text{Scientific Claim } \mathscr{C}^+\text{:}}{\theta > \theta_{un}}$$

Even without knowing the value of θ_{un}, we can discriminate between legitimate and illegitimate construals of a statistical result by considering the values of $\hat{\alpha}(\theta')$ for $\theta' \in \Omega_J$. For such values provide an objective measure of the extent to which a T^+ rejection serves as an indicator that $\theta > \theta'$. For the same reasons we noted above, a T^+ rejection with $\bar{x}$ successfully indicates that $\theta > \theta'$ to the extent that $\hat{\alpha}(\theta')$ is small. For, if $\hat{\alpha}(\theta')$ is small, it is *correct* to assert that such a rejection infrequently arises if $\theta \leqslant \theta'$; it can infrequently be reproduced if $\theta \leqslant \theta'$.

In contrast, the larger $\hat{\alpha}(\theta')$ is, the poorer a T^+ rejection is as an indicator that $\theta > \theta'$ (i.e., that is *not* due to θ as small as θ'). For, if $\hat{\alpha}(\theta')$ is fairly large, then such a rejection *is* the sort of event that fairly frequently arises when $\theta \leqslant \theta'$ (by definition of $\hat{\alpha}$). Hence, if such a rejection is taken to signal that $\theta > \theta'$, it will be *mis*taken fairly frequently. So, if one is interested in learning only of the existence of θ

values larger than θ' (i.e., if $\theta' \leqslant \theta_{un}$), a result for which $\hat{\alpha}(\theta')$ is large fails to advance one's learning. To make this more concrete, consider the result of ET^+-1 in 2.3(c). The average winglength of the 100 observed (DDT-treated) houseflies, $\bar{x}$, was 46.5 mm. So, the statistical result is: T^+ rejects $H: \theta = 45.5$ with $\bar{x} = 46.5$. How, on our approach, is this result to be interpreted? To answer this, it will help to observe some of the values of $\alpha(46.5, \theta)$ (abbreviated $\hat{\alpha}(\theta)$). This will tell us how frequently such a T^+-1 rejection arises when various fly populations are being observed (see Figure 4.3.).

It can oe seen that this result is a good indication that one is observing a population where $\theta > 45.5$ ($\hat{\alpha}(45.5) = 0.01$). But, there is also a good indication that even more has been learned; that is, the result not only indicates that H is not precisely true, it also says something about how far from the truth it is. Since $\hat{\alpha}(45.7)$ is very small (0.02), there is a good indication that θ exceeds 45.7 as well. For, if one were catching (average) flywings with T^+-1 in a population of flies where θ was no greater than 45.7, only 2% of our catches would be this large. More generally, the fly populations for which $\hat{\alpha}$ is small are ones which have a correspondingly small chance of producing a 46.5-rejection with our test. So to easily reproduce the observed effect (i.e., the observed rejection) one must move to a fly population "to the right" of these. However, as soon as one moves far enough to the right to render the result fairly easy to reproduce, i.e., as soon as one reaches a θ (in excess of 45.5) for which $\hat{\alpha}$ is no longer small, there is *no indication* that one is in a population any further to the right.

Since $\hat{\alpha}(46.9) = 0.84$, our rejection does not indicate any further move to the right of 46.9. Suppose θ_{un} is 46.9, and $\mathscr{C}^+$ asserts that $\theta \geqslant 47$. Then if our 46.5-rejection is taken as a result that is *not* typical in a fly population where θ exceeds the normal 45.5 mm by an unimportant amount, i.e., if it is taken to indicate $\mathscr{C}^+$, then it will be *mis*taken. For such a rejection *is* typical of a fly population where $\theta = 46.9$ (it occurs 84% of the time).

By incorporating such metastatistical reasoning in NPT*, the $\hat{\alpha}(\theta)$ curve provides a nonsubjective tool for understanding the θ values about which one has or has not learned on the basis of a given T^+-rejection. Rather than report the entire curve, certain key $\hat{\alpha}(\theta)$ values succeed in conveying the sort of θ values that are being detected. Corresponding to a T^+-rejection with $\bar{x}$ define:

4.3(d): $\theta^{\hat{\alpha}}$ = the value of θ (in Ω) for which $\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = \hat{\alpha}$.

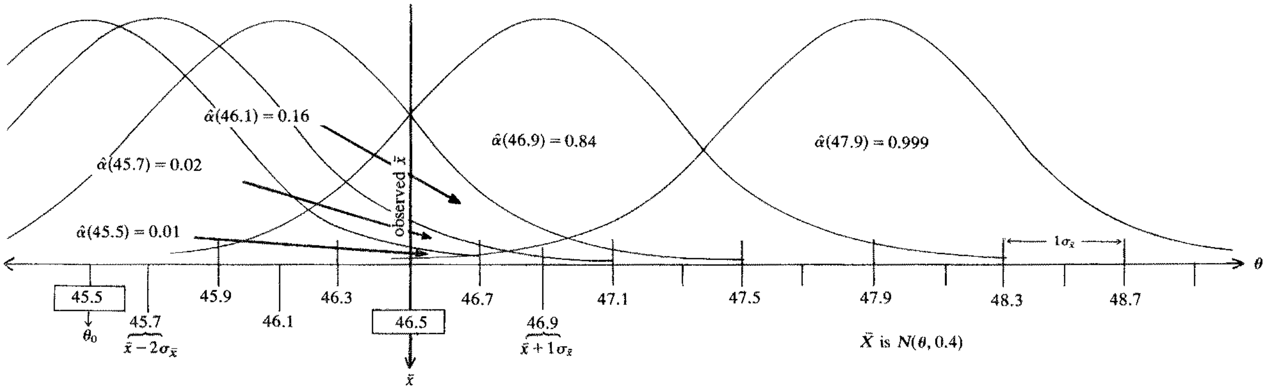


Fig. 4.3: $\hat{\alpha}(\theta) = \alpha(46.5, \theta)$, $= P(\bar{X} \geqslant 46.5 \mid \theta)$ in ET$^+ - 1$.

For example, $\theta^{0.02} = \bar{x} - 2\sigma_{\bar{x}}$, and $\theta^{0.84} = \bar{x} + 1\sigma_{\bar{x}}$.[7] Hence, a T^+ rejection is a good indication or signal of $\theta > \theta^{0.02}$; while it is a poor indication that $\theta > \theta^{0.84}$. So, if $\theta^{0.84} \leqslant \theta_{un}$ (or, more generally, if $\hat{\alpha}(\theta_{un})$ is large) then taking the statistical result as a signal of scientific claim $\mathscr{C}^+$ is illegitimate.

4.4 *An Objective Interpretation of Accepting a Hypothesis* (*With* T^+)

Just as rejecting H with too sensitive a test (i.e., too small a significance net) may indicate scientificially unimportant θ's have been found, accepting H with too insensitive a test (i.e., too coarse a significance net) may fail to indicate that *no* scientificially important θ's have been found. That is, a too sensitive test may detect unimportant discrepancies, while too insensitive a test may fail to detect important ones. As we saw in 3.2, tests are criticized because a given acceptance of H may be due, not to the non-existence of an importantly discrepant θ, but to deliberately specifying a test to be too coarse. The coarseness of a test (i.e., of its significance net) may be increased by decreasing its size α; for, as example $\text{ET}^+ - 2$ showed, this increases d_α and so increases $\theta_0 + d_\alpha \sigma_{\bar{x}}$. But this can also be accomplished by specifying a sufficiently small sample size n, while keeping the same α; for this increases $\sigma_{\bar{x}}$. However, once the influence of sample size is taken into account in interpreting an acceptance, one can avoid being misled.

Let experimental test $\text{ET}^+ - 3$ be the extreme case where n is only 1, while α, as in $\text{ET}^+ - 1$ is 0.02. Suppose a single fly (after having been DDT-treated) is observed to have a winglength of 46.5 mm – the same as the average winglength in the 100 flies in $\text{ET}^+ - 1$. With $n = 1$, the distribution of statistic $\bar{X}$ (which is just X) is the population distribution $N(\theta, 4)$. While the magnitude of the observed difference is the same as $\text{ET}^+ - 1$ (i.e., 1mm), it is now equivalent to only $0.25\sigma_{\bar{x}}$ (vs. $2.5\sigma_{\bar{x}}$). And since $0.25 = d_{0.4}$, our observation lands on a 0.4-net, but falls through the 0.02-significance net of $T^+ - 3$, as this net only catches $\bar{x}$'s that exceed 45.5 by 8 mm (i.e., $T^+ - 3$ reject H iff $\bar{x} \geqslant 53.5$.) Hence, $T^+ - 3$ accepts H with 46.5; 46.5 is deemed well-within the bounds of differences due to experimental error from $\theta = 45.5$ alone. However, "such an acceptance" (i.e., one with so insignificant an $\bar{x}$) does *not* indicate that θ is precisely 45.5 and that all θ's in Ω_J are ruled out. For, such an acceptance may not be infrequent even if one is observing populations to the right of $\theta = 45.5$. Define

4.4(a): $\beta(\bar{x}, \theta) = \hat{\beta}(\theta) = P(\bar{X} \leqslant \bar{x} | \theta)$.

That is, $\hat{\beta}(\theta)$ is the area to the left of $\bar{x}$, under the distribution of $\bar{X}$, which in $\mathrm{ET}^+ - 3$ is $N(\theta, 4)$. We can understand the import of a $T^+ - 3$ acceptance with $\bar{x} = 46.5$ by reporting various values of $\hat{\beta}(46.5)$, as seen in Figure 4.4.

In comparing Figure 4.4 to Figure 4.3, it can be seen that while $\mathrm{ET}^+ - 1$ distinguished H from θ values relatively close (e.g., only $\frac{1}{2}\sigma$ from 45.5), $\mathrm{ET}^+ - 3$ only distinguishes H from rather distant θ values (1 or 2σ from 45.5.) And an acceptance of H only indicates that one can rule out θ values it is capable of distinguishing. Since an acceptance is more informative the smaller the value of Ω_J it indicates can be ruled out, the more sensitive the test (the smaller the significance net) from which an acceptance arises, the more that is learned. More precisely, a T^+ acceptance of H does not rule out those alternative values of θ that fairly frequently give rise to such an acceptance (i.e., those for which $\hat{\beta}$ (θ) is *not* small.) And the less sensitive the test, the further to the right of H one must go before values of $\hat{\beta}(\theta)$ begin to get small. Corresponding to a T^+-acceptance define:

4.4(b): $\theta_{\hat{\beta}}$ = the value of θ (in Ω) for which $\beta(\bar{x}, \theta) = \hat{\beta}(\theta) = \hat{\beta}$.

Then a T^+ acceptance with $\bar{x}$ succeeds in indicating that $\theta \leqslant \theta'$ (i.e., that θ's in excess of θ' have not been found) to the extent that $\hat{\beta}(\theta')$ is small. Important examples of values for which $\hat{\beta}(\theta)$ is small are $\theta_{0.16} = \bar{x} + 1\sigma_{\bar{x}}$ and $\theta_{0.02} = \bar{x} + 2\sigma_{\bar{x}}$. In $\mathrm{ET}^+ - 3$, these correspond to $\theta = 50.5$ and $\theta = 54.5$, respectively. However, having learned that θ values larger than, say, $\theta_{0.16}$ have not been found, it cannot automatically be assumed that no scientifically important θ values have been found.

Rather than work solely with θ_{un}, let us define θ_{imp} as follows:

4.4(c): θ_{imp} = the smallest scientifically important θ value in excess of θ_0.

Then interpreting a T^+-acceptance as indicating that $\theta \leqslant \theta_{un}$ (i.e., as indicating one can *rule out* $\mathscr{C}^+ : \theta \geqslant \theta_{imp}$) when in fact $\theta \geqslant \theta_{imp}$ is to misinterpret its scientific import. This misinterpretation may be seen as a metastistical version of the Type II error; but now it consists of erroneously construing a T^+-acceptance of H as indicating the denial of $\mathscr{C}^+$. Correspondingly, $\hat{\beta}(\theta_{imp})$ gives the maximum frequency of such a

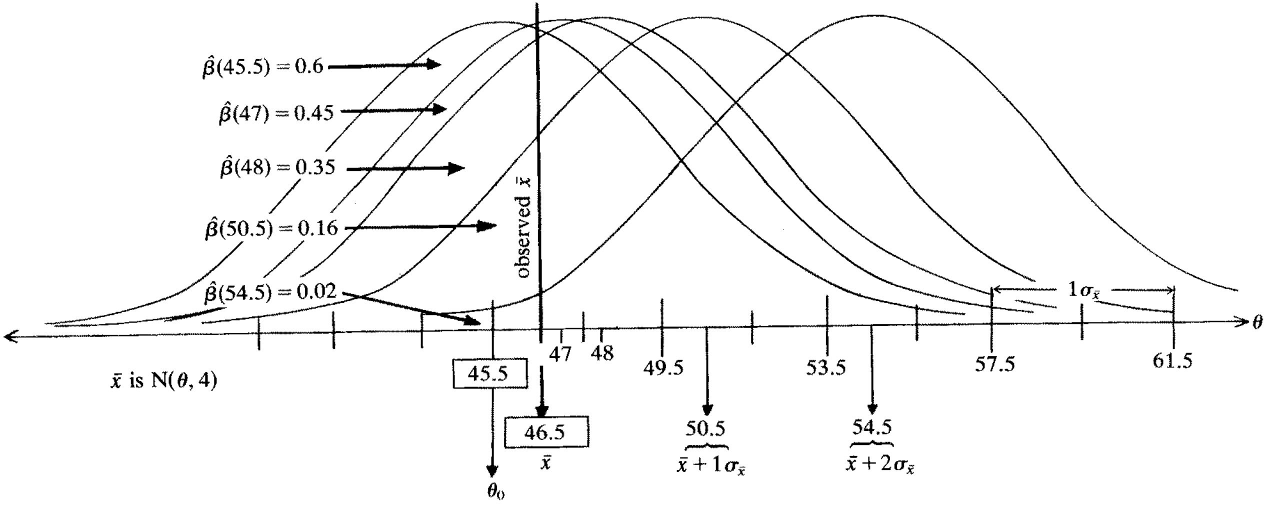


Fig. 4.4: $\hat{\beta}(\theta) = \beta(46.5, \theta), \quad = P(\bar{X} \leqslant 46.5 \mid \theta)$ in ET$^+ - 3$.

misconstrual. That is why if this value is fairly large NPT* deems $\bar{x}$ a poor indicator of $\theta < \theta_{imp}$. Reverting to our ultrasound probe analogy, suppose a given image is classified as non-diseased by the probe. If in fact such an image (or one deemed even healthier) frequently arises when a seriously diseased artery is being observed, it would be a poor indicator that no disease was present.

Suppose, as before, that scientists conducting the winglength study are interested in learning of θ values of 47 mm or more, i.e., $\theta_{imp} = 47$. But, since $\hat{\beta}(47) = 0.45$, it follows that an underlying θ of 47 would be concealed 45% of the time by reporting a $T^+ - 3$ acceptance with our observed $\bar{x}$ of 46.5. Imagine that the average winglength of a population of giant houseflies (possibly produced by having had their larval food contaminated with some other chemical) is 48 mm. Surely, in that case one would want to learn if DDT-treated flies may be described as having winglengths as far from the normal 45.5 as one finds in giant fly populations. Yet 48 is still not deemed clearly distinct from 45.5 by our statistical result; 35% of the time, the fact that such a giant fly population was being observed would be concealed by such an acceptance (i.e., $\theta_{0.35} = 48$). Unfortunately, this is a typical situation in a great many experiments on treatment effects, since large samples are rarely available. And since failure to reject the null hypothesis H is generally taken as a sign that no scientifically important treatment effect has been found, the existence of important effects is often concealed. In NPT*, however, such illegitimate interpretations of a statistical acceptance can be avoided, provided that one reports not only that H has been accepted, but also the experimental test used and the specific data observed. In the case of $\mathrm{ET}^+ - 3$, if it is reported that $T^+ - 3$ accepts $H: \theta = 45.5$ with $\bar{x} = 46.5$, it should only be taken as a signal that a θ in excess of $\theta_{\hat{\beta}}$ for small $\hat{\beta}$ (e.g., 0.16 or less) has not been found. Since $\theta_{0.16} = 50.5$, our result is not a good indication that $\theta < \theta_{imp}$ which was 47. It indicates only that $\theta < 50.5$ (or so).

While $\mathrm{ET}^+ - 1$ and $\mathrm{ET}^+ - 3$ both dealt with the same observation (46.5) the former rejects, while the latter accepts H. If, as we supposed, $\theta_{imp} = 47$, then the resulting $T^+ - 1$ rejection fails to signal that a scientifically important θ has been detected; and the $T^+ - 3$ acceptance fails to signal that *no* scientifically important θ has been found. But our assignment of θ_{imp} was simply to illustrate the metastatistical reasoning of NPT*. Without specifying θ_{imp} we have seen how such reasoning allows us to understand what sort of θ values would have to be of interest if statistical results from T^+ are to be informative. Moreover, regardless of the key values of $\hat{\alpha}$ and $\hat{\beta}$ selected as "fairly large", they may be used as standard indices for distinguishing what has or has not been learned. By using them in interpreting various T^+ results (perhaps distinguishing types of inquiries) certain values may turn out to be most insightful. In addition, after evaluating tests according to how well they perform their learning function, the means for specifying tests so that they are likely to perform this function may start to emerge. However, what is important from our point of view is that regardless of how they have been specified, it should be possible for the objective import of test results to be extracted; and we have seen how, in the case of ET$^+$, NPT* accomplishes this task.

5. A NOTE ON THE PROBLEM OF OBJECTIVELY EVALUATING TEST ASSUMPTIONS

Our ability to objectively interpret statistical results of T^+ rested on the ability to make use of probabilistic relationships between test results and underlying θ values given by $\hat{\alpha}$ and $\hat{\beta}$. However, these relationships refer to sample averages that are distributed according to the experimental test statistic $\bar{X}$; and the failure of empirical observation $\mathcal{O}$ to satisfy the assumptions of this distribution may render assertions about $\hat{\alpha}$ and $\hat{\beta}$ invalid. The task of evaluating whether $\mathcal{O}$ meets these assumptions was seen as part of task (3) in Section 3 (ray (3) in Figure 2.1). The problems that arise in accomplishing task (3) plague all theories of statistical testing, and perhaps for this reason they are less often the focus of attacks on NPT than are the problems of accomplishing (1) and (2). Nevertheless, this task is an important one, and while we can only briefly consider it here, an indication of how NPT* accomplishes task (3) objectively hopefully will be provided.

An (n-fold) sample observation $\mathcal{O}$ is statistically modelled as the result of n independent observations of a random variable distributed according to $M(\theta)$, i.e., as $\langle x_n \rangle$ (see section 2.1). But critics of NPT deny its ability to objectively verify that the data generation procedures give rise to a sample that satisfies the formal assumptions of the statistical data model. As Dempster (1970, p. 70) maintains:

> Although frequentists often welcome the label objective, every frequentist model rests on

> a necessarily subjective judgment that a particular set of elements were drawn from a larger collection of elements in a way which favored no elements over any other....

Once again the charge of subjectivity concerns the necessity of extrastatistical judgments; here in validating formal test assumptions. But the fact is, the formal assumptions need not be precisely met for the learning function of tests to be accomplished. It is required only that one be able to distinguish the major effects of the primary treatment of interest from extraneous factors.

For example, in the study on houseflies, the ability to learn about the effects of DDT on winglengths depends on being able to distinguish its effects from other factors influencing growth. However, if the sample is *biased* in various ways, such as by observing only male flies or including a growth hormone as well as DDT into the larval food, then the difference between the sample average and the normal average winglength (45.5) may be due – not to DDT – but to various extraneous factors. Ideally, the treated flies would differ from normal, untreated flies – at least with respect to factors affecting winglength – only in being given DDT. But the value of the theory of statistically designed experiments is in providing means for "subtracting out" or "designing out" factors for which one cannot physically control. Here, as elsewhere in our discussion, "subtracting out" something is a matter of taking its effect into account, and this is accomplished by making use of information as to its statistical effect. Just as a hypothesis test makes use of known patterns of variability from experimental error to distinguish it from genuine, systematic effects, knowledge of the influence from extraneous factors enables them to be distinguished from the primary effect of interest (e.g., DDT). And by generating the sample according to one of various *probabilistic schemes*,[8] the variability due to nonprimary sources may be approximated by a known statistical model. Hence, by probabilistic sampling schemes (e.g., randomization, stratification) NPT is able to ensure that claims about error frequencies are approximately equal to the actual error frequencies that would arise in other applications of its methods.

Admittedly, judgments and background information about relevant factors are required; but this information is used to intentionally avoid biasing or misreading the data. Moreover, NPT provides extensive methods for checking whether such bias is avoided by testing whether test assumptions are approximately met. For example, the assumption of independence may be checked by testing a null hypothesis of the form: $\langle x_n \rangle$ are independent samples from $M(\theta)$, and using various Goodness-of-Fit tests. The assumption in ET$^+$ that the standard deviation is unchanged may be tested using techniques of analysis of variance. Ideally the test assumptions can be checked prior to testing $H: \theta = \theta_0$; however numerous tests are designed to enable them to be checked after the data is already observed. Nevertheless, NPT has been criticized for failing to allow such an after-trial check of test assumptions – at least not without running an additional experiment. As Rosenkrantz (1977, p. 205) maintains:

> Imagine that we were testing an hypothesis about a binomial parameter or a normal mean, but afterwards, upon examination of the data, it appeared that the trials were not truly random or the population not truly normal. Again, these latter assumptions, not being themselves the object of the test, cannot be said to be counterindicated by the present data. Instead, a new experiment must be designed to test whether they are realistic assumptions about our current experiment!

But this ignores the ways in which NPT enables one to test whether the data from a current experiment obeys the test assumptions and still avoid the biases that can result from a double use of data. For the data from the current experiment (e.g., ET$^+$) may be *remodelled* in a variety of ways to serve as the data of various (*after-trial*) tests of its assumptions. Each such test corresponds to a different test statistic S. For instance, to check whether the assumption of normality is met, one might look, not at the average winglength (as one would in carrying out T^+), but at the number of flies whose winglengths were observed to be within certain ranges.

In fact, such tests often reveal that test assumptions are *not* precisely met! But the important thing is that the learning function of tests may be accomplished despite the violations of these assumptions – something that is not usually realized in foundational discussions. That is, NPT methods are, to a great extent, *robust* against such violations, and their robustness is made explicit in NPT*. A series of tests for checking assumptions may be developed within NPT*, as well as metastatistical principles for their interpretation – along the lines of those developed for T^+ (Section 4). However, while there our aim was to avoid confusing scientifically (un)important effects with statistically (in)significant ones; here our aim would be to avoid confusing the effects of our primary treatment (e.g., DDT) with the influences of extraneous factors – at least to the extent that we would be prevented from objectively interpreting the primary statistical result (i.e., of test T^+) via $\hat{\alpha}$ and $\hat{\beta}$. Moreover, these tests themselves (at least within the context of ET$^+$, but also for other 1-parameter cases) are either distribution-free or robust; or, if not, techniques exist for determining the extent to which possible violations hamper one's ability to learn about the scientific phenomenon of interest. And this is all that is required for objective scientific learning.

In contrast, a subjective Bayesian report of a posterior probability (in H) is not open to such a critical check – at least not by Bayesian principles alone. For it is not required that the influence of the prior be reported; and one may have no way of knowing if the prior was based on objective information, subjective prejudice, or on ignorance (of unequal probabilities). Admittedly, it is possible that in a given case a sample selected on the basis of subjective beliefs turns out to satisfy test assumptions better than one obtained from a NPT probabilistic scheme; the problem is that there is no objective means of knowing this – so long as the underlying population being studied is unknown.

In conclusion, we can agree with a slogan cited by Good (1981, p. 161) (which he attributes to Rubin) that "A good Bayesian does better than a non-Bayesian but a bad Bayesian gets clobbered". But one could likewise say that when a self proclaimed psychic like Jean Dixon has a good day, she predicts the future better than any scientific means. But the whole point of scientific objectivity is to systematically distinguish fortuitous guesswork from reliable methods – and one cannot do this in the case of Dixon and the subjective Bayesian. And this is why a follower of an objective theory of testing, such as NPT*, is – unlike the subjective Bayesian – able to clobber his hypotheses and assumptions, instead of getting clobbered himself!

NOTES

[†] I would like to thank Ronald Giere for very useful comments and numerous valuable conversations concerning this paper and his own work. This research was carried out during tenure of a National Endowment for the Humanities Fellowship; I gratefully acknowledge this support.

[1] Our example is adapted from data in the study in Sokal and Hunter (1955). This study, discussed in Sokol and Rohlf (1969), has been greatly simplified in our adaptation.

[2] Fisherian (Significance) Tests specify only a null hypothesis H, and use sample data to either reject or fail to reject H (where failing to reject is *not* taken as accepting H). H is

rejected just in case the probability of a result as or more deviant (from what would be expected under *H*) than the one observed is sufficiently small – for some specified value of "small". Since no alternative hypothesis is specified, however, there is a problem in ascertaining what is to count as "more deviant".

[3] Although Neyman (1950) notes that if the data fails to satisfy the assumptions of the experimental test, the test's error probabilities may hold for the statistical hypothesis but not for the corresponding scientific claim; as long as these assumptions are approximately satisfied, the statistical hypothesis is simply identified with the scientific claim – or so he suggests.

[4] A detailed discussion of the distinction between NPT and statistical theories that aim to provide measures of evidential strength that data afford hypotheses occurs in Mayo (1981a) and Mayo (1982). We show that the major criticisms of NPT rest on the assumption that for NPT to be adequate it should provide such evidential-strength measures. We reject these criticisms by arguing that (a) NPT does not intend to provide such measures, and (b) the critics fail to provide a non-question begging argument showing that NPT should seek to do so.

[5] Ultimately we envision NPT* as providing a systematic means for learning about a scientific phenomenon by imbedding individual statistical inquiries within a larger, more complex model of a scientific inquiry or *learning program*. To this end, a system of metastatistical principles, along the lines of those developed in 4.3 and 4.4 for T^+ (though more complex) may be developed for a variety of statistical tests. Then, by means of (meta-metastatistical?) principles spanning several different theories (e.g., theories of data, of observation, of the primary scientific phenomenon), individual tests may be both specified, interpreted, and evaluated by reference to other statistical tests (and their corresponding test models) within the larger model of the overall learning effort.

[6] For, from 4.3(a), $\hat{\alpha}(\theta_0)$ (shorthand for $\alpha(\bar{x}, \theta_0)$) is $\leqslant \alpha$ just in case $P(\bar{X} \geqslant \bar{x} | \theta_0) \leqslant \alpha$; and this occurs just in case $\bar{x}$ is in the critical region of NPT test T^+ with size α.

[7] Values of $\theta^{\hat{\alpha}}$ can be derived from our knowledge of the probability that $\bar{X}$ exceeds its mean θ by various amounts (i.e., by various $d_\alpha \sigma_{\bar{x}}$'s) as given in Figure 2.3. Consider, for example, $\theta^{0.02}$. If $\theta = \bar{x} - 2\sigma_{\bar{x}}$ then $\hat{\alpha}(\theta) = P(\bar{X} \geqslant \bar{x} | \theta = \bar{x} - 2\sigma_{\bar{x}})$ (from 4.3(a)). And this is just the probability that $\bar{X}$ exceeds its mean θ by $2\sigma_{\bar{x}}$ (where $\bar{X}$ is $N(\theta, \sigma_{\bar{x}})$), which we know is 0.02. That is, $2 = d_{0.02}$. It follows that $\hat{\alpha}(\bar{x} - 2\sigma_{\bar{x}}) = 0.02$, so $\theta^{0.02} = \bar{x} - 2\sigma_{\bar{x}}$ (for ET^+).

[8] In a *probabilistic* or *random sampling* scheme the elements of a population are selected or assigned to some "treatment" (e.g., DDT) in accordance with a specific probability. In *simple* random sampling each is given the same selection probability – but equiprobability is not required for NPT. In *stratified* random sampling, for example, the population is divided into groups and random samples are drawn from each. Ideally, these groups are known (from other studies) to share a property relevant to the population property of interest. For instance, in the housefly winglength data, flies had been grouped according to the culture jar in which they were incubated, so that samples would include flies from each. Since there is less variability within each group with respect to the property of interest, (e. g., size of adult fly), by suitably weighing average winglengths from each the final sample data (e.g., $\bar{x}$) will vary less from θ than in simple random sampling from the population (it will be more "representative").

But the primary value of probabilistic sampling for NPT* is that the pattern of variability of experimental data collected in this way may be approximated by means of standard statistical models. For, in this way one can validly make use of known probabilistic relations between statistical data and underlying population parameters, i.e., make use of the distribution of one or more experimental test statistics. And this provides an objective basis for assertions about error probabilities and therefore for assertions about $\hat{\alpha}$ and $\hat{\beta}$; and this, we have argued, is all that is required for NPT* to accomplish the task of objective learning.

REFERENCES

Birnbaum, A.: 1977, 'The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; With a Criticism of the Lindley-Savage Argument for Bayesian Theory', *Synthese* **36**, 19–50.

Carnap, R.: 1950, *Logical Foundations of Probability*, University of Chicago Press, Chicago.

Dempster, A. P.: 1971, 'Model Searching and Estimation in the Logic of Inference', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, 56–77.

Edwards, A. W. F.: 1971, 'Science, Statistics and Society', *Nature* **233**, 17–19.

Fetzer, J. H.: 1981, *Scientific Knowledge*, Reidel, Dordrecht.

Fisher, R. A.: 1955, 'Statistical Methods and Scientific Induction', *Journal of the Royal Statistical Society* (B) **17**, 69–78.

Giere, R. N.: 1976, 'Empirical Probability, Objective Statistical Methods and Scientific Inquiry', in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. II, Reidel, Dordrecht, 63–101.

Giere, R. N.: 1977, 'Testing vs. Information Models of Statistical Inference', in R. G. Colodny (ed.), *Logic Laws and Life*, University of Pittsburgh Press, Pittsburgh, 19–70.

Good, I. J.: 1976, 'The Bayesian Influence, or How to Sweep Subjectivism Under the Carpet', in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. II, Reidel, Dordrecht, 125–174.

Good, I. J.: 1981, 'Some Logic and History of Hypothesis Testing', in J. C. Pitt (ed.), *Philosophy in Economics*, Reidel, Dordrecht, 149–174.

Hacking, I.: 1965, *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hacking, I.: 1980, 'The Theory of Probable Inference: Neyman, Peirce and Braithwaite', in D. H. Mellor (ed.), *Science, Belief and Behavior: Essays in Honor of R. B. Braithwaite*, Cambridge University Press, Cambridge, 141–160.

Kalbfleisch, J. G.: 1979, *Probability and Statistical Inference*, Vol. II, Springer-Verlag, New York.

Kempthorne, O. and Folks, L.: 1971, '*Probability, Statistics, and Data Analysis*', Iowa State University Press, Ames.

Kyburg, H. E. Jr.: 1971, 'Probability and Informative Inference', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, 82–103.

Kyburg, H. E. Jr.: 1974, *The Logical Foundations of Statistical Inference*, Reidel, Dordrecht.

Lehmann, E. L.: 1959, *Testing Statistical Hypotheses*, John Wiley, New York.

Levi, I.: 1980, *The Enterprise of Knowledge*, MIT Press, Cambridge.

Lindley, D. V.: 1976, 'Bayesian Statistics', in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. II, Reidel, Dordrecht, 353–363.

Mayo, D.: 1981a, 'In Defense of the Neyman-Pearson Theory of Confidence Intervals', *Philosophy of Science* **48**, 269–280.

Mayo, D.: 1981b, 'Testing Statistical Testing', in J. C. Pitt (ed.), *Philosophy in Economics*, Reidel, Dordrecht, 175–203.

Mayo, D.: 1982, 'On After-Trial Criticisms of Neyman-Pearson Theory of Statistics', in P. Asquith (ed.), *PSA 1982*, Vol. 1, East Lansing Philosophy of Science Association, 145–158.

Neyman, J.: 1950, *First Course in Probability and Statistics*, Henry Holt, New York.

Neyman, J.: 1971, Comments on R. M. Royall, 'Linear Regression Models in Finite Population Sampling Theory', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, 276–278.

Neyman, J. and Pearson, E. S.: 1933, 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', in *Philosophical Transactions of the Royal Society* A, 231, 289–337. (As reprinted in *Joint Statistical Papers*, University of California Press, Berkeley, 1967, 276–283.)

Neyman, J. and Pearson, E. S.: 1936, 'Contributions to the Theory of Testing Statistical Hypotheses', *Statistical Research Memoirs* **1**, 1–37. (As reprinted in *Joint Statistical Papers*, University of California Press, Berkeley, 1967, 203–239.)

Pearson, E. S.: 1955, 'Statistical Concepts in Their Relation to Reality', *Journal of the Royal Statistical Society* B, **17**, 204–207.

Popper, K. R.: 1972, *Objective Knowledge*, Oxford University Press, Oxford.

Rosenkrantz, R. D.: 1977, *Inference, Method and Decision*, Reidel, Dordrecht.

Rubin, H.: 1971, 'Occam's Razor Needs New Blades', in V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, 372–374.

Savage, L.: 1954, *The Foundations of Statistics*, Wiley & Sons, New York.

Scheffler, I.: 1967, *Science and Subjectivity*, Bobbs-Merrill, New York.

Seidenfeld, T.: 1979, *Philosophical Problems of Statistical Inference*, Reidel, Dordrecht.

Sokal, R. R. and Hunter, P. E.: 1955, 'A Morphometric Analysis of DDT-Resistant and Non-Resistant Housefly Strains', *Annals of the Entomology Society of America* **48**, 499–507.

Sokal, R. R. and Rohlf, F. J.: 1969, *Biometry*, W. H. Freeman, San Francisco.

Spielman, S.: 1972, 'A Reflection on the Neyman-Pearson Theory of Testing', *British Journal for the Philosophy of Science* **24**, 201–222.

Dept. of Philosophy and Religion
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
U.S.A.