CHAPTER NINE

Hunting and Snooping: Understanding the Neyman-Pearson Predesignationist Stance

If in sampling any class, say the *M*'s, we first decide what the character *P* is for which we propose to sample that class, and also how many instances we propose to draw, our inference is really made before these latter are drawn, that the proportion of *P*'s in the whole class is probably about the same as among the instances that are to be drawn. . . . But suppose we were to draw our inferences without the predesignation of the character *P*; then we might in every case find some recondite character in which those instances would all agree. That, by the exercise of sufficient ingenuity, we should be sure to be able to do this, even if not a single other object of the class *M* possessed that character, is a matter of demonstration. For in geometry a curve may be drawn through any given series of points.

—C. S. Peirce, *Collected Papers,* vol. 2, par. 737

To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the [chance] hypothesis is true.

—E. S. Pearson, *The Selected Papers of E. S. Pearson,* p. 127

9.1 Introduction and Overview

The debate about the novelty requirement in the arena of philosophy of science parallels an ongoing methodological debate in actual scientific practice, and the preceding results have direct ramifications for that dispute. That this dispute is alive and well is brought home by the program put forward in Glymour, Scheines, Spirtes, and Kelly 1987 and its subsequent extensions. The dispute concerns a principle often adhered to in statistical testing based on Neyman-Pearson (NP) methods. It is commonly supposed that the NP account, from which my error-statistical account derives, prohibits all after-trial constructions of hypotheses. Indeed, it is typically thought to mandate an even stricter requirement. It is supposed that a key feature of the NP model of tests is that all aspects of the tests, the hypotheses, the sample size, the significance level, and so on must be laid out in advance of running the experiment. All must be *predesignated.* If the predesignation of tests is always required, then in particular the temporal novelty of hypotheses is required. It is never OK to snoop at the data before formulating a hypothesis—at least not if the same data are to be used in testing that hypothesis. So if NP statistics requires predesignation, then it has all the problems of temporal novel accounts. Known data fail to provide good tests of hypotheses—NP theory has what is called an "old evidence" problem.

But is predesignation part and parcel of the NP methodology? Predesignation is part of what might be called "folk NP statistics." In practice it is often good advice. But violating predesignation does not necessarily conflict with NP principles. In fact, many of its own methods violate this predesignationist stance.

Examples of NP procedures that violate predesignation—by violating use-novelty—are those involved in checking the assumptions of an experimental test. The same data may lead to constructing a hypothesis—say, that the trials are not independent—and at the same time may be used to test that hypothesis. The rationale is analogous to the posttrial scrutiny of the eclipse data discussed in the last chapter. In checking if a particular set of data satisfies assumptions, such a double use of data is likely to offer a better test than looking to the data of some new experiment.

Roger Rosenkrantz puts his finger on this apparent dilemma for NP or "orthodox" statistics:

> It is difficult to live within the confines of a predesignationist methodology. Actual orthodox practice fully bears this out; indeed, standard orthodox texts are all replete with post-designated tests. . . . In analysis of variance, for example, upon rejecting the hypothesis that all means are equal, orthodox texts show you how to go on to test other more particular hypotheses about the means suggested by the data. . . . Those same texts—and their number is legion—also show how to test the underlying assumptions of the usual analysis of variance model . . . again using the same data. Similarly, they show how to test underlying assumptions of randomness, independence and stationarity, where none of these was the predesignated object of the test (the "tested hypothesis"). And yet, astoundingly in the face of all

> this, orthodox statisticians are one in their condemnation of "shopping for significance," picking out significant correlations in data *post hoc,* or "hunting for trends in a table of random digits." . . . It is little wonder that Orthodox texts tend to be highly ambivalent on the matter of predesignation. (Rosenkrantz 1977, 204–5)

Is the NP statistician being inconsistent in banning postdesignation in the form of shopping or hunting for significance, while condoning it in certain other tests? Are NP statisticians justified in insisting on predesignation with respect to certain kinds of tests and estimation procedures? And what exactly is the NP statistician condemning in condemning postdesignation? Answering these questions, which have been at the center of considerable controversy, is the goal of this chapter.

In chapter 8 I argued that the real aim of novelty is severity and that the novelty requirement was justified only to the extent that violating novelty precluded severity. Can an analogous move disentangle the predesignationist puzzle? The natural suggestion would be to propose that condoning violations of predesignation in one set of cases while condemning such violations in others may be perfectly justified if it turns out that severity is only problematic in the latter set of cases. Will this natural suggestion hold up?

I claim that it will. One must be careful, however, in understanding what it means for "severity to be problematic." Certainly severity would be problematic if violating predesignation led to a procedure of passing hypotheses where the severity was low. For then there would be a high probability of passing hypotheses erroneously, violating the NP low error probability requirement. But is predesignation necessary for severity? We already have our answer to that question as well. If predesignation is necessary for a good test, so is use-novelty. Hence the contexts in which one can have a severe test while violating UN are also contexts that yield a severe test despite violating predesignation. So having shown that UN is not necessary, we have shown predesignation is not.

However, the real NP argument for predesignation differs from the argument given by the UN proponents we have considered. Those proponents thought UN necessary because its violation was thought to lead to zero- or low-severity tests. This is not what is going on in the contexts where NP statisticians insist upon predesignation—although it is often thought to be. The real argument is that in certain testing contexts violating predesignation alters the test procedure in such a way as to require that it be taken account of in assessing its severity. What is really being condemned, or so I shall argue, *is treating both predesignated and postdesignated tests alike.*

My focus in this chapter is on an important class of cases in which violating predesignation is condemned or deemed inadmissible by the error-probability statistician. Disagreement about whether predesignation should be required in this class of cases often masks a disagreement about whether error probability requirements matter altogether. The Bayesian (or other holder of the likelihood principle) says no, while the NP theorist says yes. But a debate also persists among users of the standard significance tests whose position on this matter is much less clear. The individual I have in mind does not deny the importance of error probabilities outright but views the predesignationist stance as unnecessarily hamstringing the researcher in actual practice.

This is how I view the position in the work of Glymour, Scheines, Spirtes, and Kelly (1987). Their project is a rare example of a joint effort by philosophers to articulate methods intended for scientific practice—in particular, for discovering causal hypotheses. But my focus is not on their computer algorithm TETRAD; it is on their discussion of what I take to be the main NP stance against postdesignation. This form of postdesignation falls under the class of cases *hunting for statistically significant correlations.* Glymour, Scheines, Spirtes, and Kelly maintain that despite the general adherence to the inadmissibility of postdesignation procedures, the arguments purporting to show their inadmissibility will not hold up. Admittedly, the real argument against postdesignation (to my knowledge) has never been articulated clearly. With the machinery I have been developing it will be seen that the NP statistician *does* have a legitimate objection to postdesignation, at least with respect to the class of cases of interest.

The general outline of the NP objection emerges naturally from the underlying aim of NP statistics. Despite the different interpretations NP procedures are open to, the following error probability principle stands:

> (EPP): An NP procedure of inference is inadmissible if its error probability characteristics are inconsistently reported or if it prevents the determination of valid error probabilities (even approximately).

In the contexts I will be considering, a test's error probabilities are altered when hypotheses are not predesignated. At the same time there are other tests of hypotheses and models that despite being postdesignated do not violate any NP principles. Several tests are even intended for use in cases that would ordinarily be thought to violate predesignation—Rosenkrantz is right. If my analysis is correct, however, then there is nothing inconsistent in this apparent schizophrenia as regards predesignation.

It is important to get clear on the real NP argument regarding postdesignation procedures, because supposing that NP methods rule them out altogether has opened those methods to unjust criticism. Furthermore, dismissing the NP argument as unsound has only made it easier to ignore the very special constraints that have to be satisfied for validly applying NP tests in such cases.

My strategy in appraising the rule of predesignation follows the general one that I advocate for a philosophy of experiment: the value of a methodological rule is determined by an analysis of how its application allows one to avoid particular types of experimental mistakes. But methodological rules are made to be broken.[1] By understanding the function that a rule of procedure has, we can identify the conditions under which it may be violated—even in testing contexts where it is normally of concern. In the opening epigraph of this chapter, Peirce explains that by prespecifying a test, "our inference is really made before" the data are collected. The "inference" that the NP statistician really wants to make before the trial is about the test's error probabilities. What is really of concern, we shall see, is the validity of the "before-trial" probabilistic guarantees.

The NP argument against violating predesignation says simply that if your worry is ensuring tests with high severity, then you must recognize that the manner of data generation can influence the probability of erroneously passing a hypothesis (the severity of the test). We must look at the entire experimental testing context to correctly assess severity. From this point of view, the NP admonishment against violating predesignation may be regarded as a kind of warning to the error statistician that additional arguments—possibly outside the simple significance test—may be required to rule out the error at hand. The tests themselves cannot be expected to provide the usual guarantees. This, I suggest, should be seen as an invitation for the error statistician to articulate *other* formal and informal considerations to arrive at a reliable experimental argument.

On my treatment of this issue it may be objected that I am assuming that it is unproblematic to determine what a test's error probabilities are, thereby discounting the ambiguities in describing a given test procedure. While I do not wish to minimize the problem of how to describe a given experiment—some might identify it with the *reference class* problem—I regard that as a problem of choosing an appropriate

[1] I am not just playing on a cliché. While obviously methodological rules are made to be followed, an equally important service they perform is to call attention to an assumption or goal that should be met, if not by following the rule, then (at least approximately) by an alternative route.

test procedure for a given inquiry, a problem distinct from the current issue. The basis for choosing a test will depend upon experimental background knowledge acquired from comparable cases, knowledge that would comprise a kind of repertoire of canonical inquiries into errors. Once one has specified the kind of argument from error one wishes to sustain, this background knowledge directs one to appropriate choices of test statistics and corresponding error characteristics. (I return to this issue in chapter 11.)

9.2 HUNTING FOR STATISTICALLY SIGNIFICANT DIFFERENCES

As is so often the case with contemporary statistical disputes, a history of this debate already exists from early applications of significance tests. A good source for the debates that took place nearly forty years ago is Morrison and Henkel's classic volume *The Significance Test Controversy.* Much of what I would wish to say emerged in these early discussions. I will focus on two contributors to the volume, both social scientists.

Leslie Kish's Example

Leslie Kish (1970, 138) refers to a study regarding infant training. The study, done in the late 1940s, sought to investigate a variety of infant training experiences regarding nursing, weaning, and toilet training that according to Freudian psychological theories of the day were thought advantageous for personality adjustment. In records on children spanning several years, a number of high (statistically significant) correlations were found between children exposed to a certain kind of infant training *T* and various personality traits. Among those subjected to training *T* "gradual weaning," for example, there was a significantly higher proportion of children with "high social standards" than among those abruptly weaned. Among the other high correlations found were between "no punishment for toilet accidents" and "good school relations," and between "late bladder training" and "little nail biting." The question is whether the evidence on correlations provides good evidence for the existence of a real effect in the population of children or whether the observed correlations are spurious or "due to chance."

The Key Question. The fact that the researchers searched among many factors for large correlations adds a new twist to this question. The question, which I pose to the reader, is this: Is it relevant to the assessment of the evidential import of the observed correlations that they are the outcome of a procedure of searching for effects large enough to be statistically significant? The Neyman-Pearson error statistician says yes. The question is why?

Let us use Kish's example to describe a procedure sometimes called "hunting with a shotgun." Here the sample size is fixed ahead of time, and even the cutoff for rejecting a test hypothesis may be preset, say at the .05 level. What varies is the hypothesis chosen for testing and reporting. (The hypotheses themselves may or may not have been thought up in advance.) Each such hypothesis asserts that some property or "treatment" T is genuinely correlated with some factor F in some population. Each factor is dichotomous, each subject has it or not. Before getting to the "hunting" aspect, let us recall how a test of such a hypothesis might go in nonhunting cases.

The situation shares several features with example 5.1 on birth control pills. The "null" or test hypothesis H is that the observed correlation is merely due to chance—that in fact the incidence of the effect (cancer, high feeling of belonging) is no different among those treated (with the pill, with gradual weaning in infancy) and those not so treated.[2] (In Kish's example the difference is sought in either a positive or a negative direction.) The null hypothesis is rejected when the observed correlation is sufficiently statistically significant (e.g., at the .05 level). The hypothesis that passes—let us abbreviate it as H^*—asserts that there is a genuine correlation in the population between the factors of interest (in either or in one direction).

As in example 5.1, a simple difference measure can be used for a statistic measuring observed correlations, and to avoid complexities I will stick to that case in this chapter. We observe the proportion of *T*s that are *F*s and the proportion of not-*T*s that are *F*s and record the difference D. We can compute the statistical significance of D by considering the number of standard deviations by which it differs from 0 (0 being the expected difference, given the truth of the null hypothesis). The 2-standard-deviation cutoff corresponds to a .05 level of statistical significance (see note 2).

The same mathematical procedure for calculating statistical significance is available even if the particular factor F is specified after hunting for large correlations. At issue is whether it is misleading or

[2] In example 5.1, only the existence of a positive difference was being tested. It was a *one-sided* test. Kish's example is a *two-sided* test, because it looks for differences in either direction. A 2-standard-deviation difference in one direction corresponds to a significance level of approximately .025 (I elsewhere round it to .03), so in two directions it corresponds to a significance level of about .05. That is why the two-sided significance tests discussed here have .05 significance levels.

fallacious to report the statistical significance of a difference *in the same way* as in the case where the hypothesis to be tested is prespecified. The "hunter," as Kish calls him, thinks not, and Kish alludes to the infant training example, as discussed by William Sewell (1952), to illustrate how the hunter would get into trouble.

The researchers in the infant training study conducted 460 statistical significance tests! Out of these they found that 18 were statistically significant at the .05 level (or beyond), 11 of these were in the direction expected by the popular psychological account. Sewell (1952) denies that we should be just as impressed with the 11 statistically significant results as we would be if they were the only 11 hypotheses to be tested. Kish agrees and explains why.

Note that the hunting procedure is an example of what I called a "use-constructing" test procedure in chapter 8. Introducing some abbreviations for this simple example will make it easy to characterize the more general argument later. For each factor F_j, calculate the difference statistic D_j, between the proportions with F_j among those given infant training T and those not given T. Finding a factor, F_j, on which the experimental subjects show a statistically significant difference would lead to testing the postdesignated null hypothesis:

> *Null hypothesis H_j:* In the population of children, treatment T is not correlated with factor F_j.

Let us focus on just one type of infant training T—gradual weaning. Suppose, for example, that a statistically significant difference is observed between gradual weaning and factor F_6, say, "a strong feeling of belonging." In that case, the procedure directs one to test the corresponding null hypothesis, H_6:

> H_6: Gradual weaning in babyhood is not correlated with a strong feeling of belonging in older children.

Next, the null hypothesis H_j (in this case, H_6) is then rejected if the observed difference is statistically significant at the .05 level. The hypothesis that passes when H_j is rejected is the non-null hypothesis H_j^*:

> *Non-null hypothesis H_j^*:* There is a genuine correlation between gradual weaning and factor (or personality trait) F_j.

But the only time H_j is tested according to this procedure is when the observed difference is statistically significant! So on this procedure, whenever H_j is tested, it is rejected.

What is wrong with this? In the actual study, out of 460 attempts to hunt for statistically significant correlations, 18 were found significant at the .05 level or beyond (11 in the expected direction "on the basis of psychoanalytic writings"). Kish (1970) remarks:

> Note that by chance alone one would expect 23 "significant" differences at the 5 percent level. A "hunter" would report either the 11 or the 18 and not the hundreds of "misses." . . . After finding a result improbable under the null hypothesis the researcher must not accept blindly the hypothesis of "significance" due to a presumed cause. Among the several alternative hypotheses is that of having discovered an improbable random event through sheer diligence. (P. 138)

Keep in mind Kish's statement about what the "hunter" would report. It is not just that the hunter postspecifies (and tests) hypotheses to fit samples, but that the hunter, or one who endorses being a hunter, is saying that *doing so calls for no difference in interpretation.* And because of that, the hunter reports the statistically significant cases *just as if* the successful cases had been predesignated.

However, if one's answer to the "key question" I posed at the start of this section is yes, then one does *not* think that the postdesignated cases should be reported just as if they had been predesignated. One thinks they should be distinguished. The NP error statistician distinguishes between the two on the very grounds Kish cites. But why, one might ask, should the import of the evidence depend upon whether the hypothesis is set out in advance? If the hypothesis, say H_6, had been set out in advance, and a 2-standard-deviation difference observed, we would have computed the statistical significance level in the usual way (.05). Why this change because it was one of the successfully hunted ones?

For the NP or error statistician, the altered interpretation is called for because the test procedure in the postdesignated case is very different from the case in which H_6 is preset as the hypothesis to test. What is allowed to vary—and hence the set of possible outcomes—is very different. With H_6 predesignated, the possible results are the possible differences in F_6 rates between the two differently trained groups—that is, the different values of statistic D_6. In the postdesignated case the possible results are the possible statistically significant F factors that one might hunt down. This difference is reflected in the difference in error probabilities. Let us turn to a second contributor to Morrison and Henkel (1970), Hanan Selvin.

Hanan Selvin's Example

Selvin, in an article first published in 1958, gives a very useful capsule statement of the problem:

> When the hypotheses are tested on the same data that suggested them and when tests of significance are based on such data, then a spurious impression of validity may result. The computed level of significance may have almost no relation to the true level. . . . Suppose that twenty sets of differences have been examined, that one difference seems large enough to test and that this difference turns out to be "significant at the 5 percent level." Does this mean that differences as large as the one tested would occur by chance only 5 percent of the time when the true difference is zero? The answer is *no,* because the difference tested has been *selected* from the twenty differences that were examined. The actual level of significance is not 5 percent, but 64 percent! (Selvin 1970, 104)[3]

So more than half the time one will be designating an observed difference (or correlation) unlikely to have been the result of mere chance error when in fact it is a result that easily (commonly) results from chance.

Selvin's distinction between "the computed" and "the true or actual" significance levels is a useful way of making out the NP argument, and it merits some additional clarification.

Computed versus Actual or True Significance Levels. The *computed level* of significance of a difference is the usual one: the improbability of observing such a large difference in the proportion with trait F_j given that in fact there is no real correlation, that is, the null hypothesis H_j is true. The computed level would be .05 if the observed difference were at least 2 standard deviations (the chance or null hypothesis being that the difference is 0). In the case of a prespecified test of null hypothesis H_j, the computed level equals the actual error probability of the procedure—the actual significance level. But the actual significance level differs if H_j arose from a procedure of searching through 20 factors on which the groups might be correlated. In this case, the *actual significance level* would be the probability of observing at least one such 2-standard-

[3] As Selvin notes, this can be calculated approximately by considering the probability of finding at least one statistically significant difference at the .05 level when 20 independent samples are drawn from populations having true differences of zero, $1 - P$ (no such difference). This is $1 - (.95)^{20} = 1 - .36$. In general, the probability of obtaining at least one statistically significant outcome (in either direction) with N independent tests and a 2α (computed) significance level is $1 - (1 - 2\alpha)^N$. This would give the *actual significance level,* that is, the actual probability of erroneously affirming a genuine correlation. The assumption of independent samples is made here for simplicity. With real data on a single population, Selvin remarks, this independence assumption does not hold "and the computation of the true level of significance would be extremely difficult" (ibid.).

deviation difference, given that there is no genuine correlation (in the population) on any of the 20 factors. Using various assumptions, Selvin calculates this probability to be .64.[4]

More generally, the probability of error in the postdesignated case is the probability of finding some such α-significant correlation *or other,* given that no real correlation exists. In both cases, one minus the probability of error is the severity of the test (for passing the hypothesized correlation). Thus, while $1 - \alpha$ is the severity (for passing H_j) when H_j is prespecified, severity is no longer $1 - \alpha$ in the postspecified case. In the postspecified case the actual significance level, the actual probability of erroneously finding some such α-significant correlation, is not generally equal to α. (Recall the altered severity criterion for hypotheses constructed from the data in chapter 6, section 6.6.)

It may be objected that in calling this the "actual" significance level I am taking sides in favor of one description of the "actual" test procedure—one that takes into account the fact that searching has occurred. I am, but maintain that this aspect of the procedure cannot be ignored given the aim of the statistical significance test chosen. Remember, I am distinguishing the appropriateness of the test chosen (for a given inquiry) from the error probabilities, *given* that that test is chosen. One chooses a type of test corresponding to the type of argument from error that one wishes to sustain. In the present case, the interest is in arguing from error to infer a genuine correlation. "Hunting" raises a problem because it may invalidate the desired argument.

Ronald Giere's Example

Ronald Giere (1969) generalizes this kind of argument against hunting for a corresponding procedure for estimating a population proportion. Here it is imagined that we hunt through random samples for a property shared by all of the n members in the sample. Finding such a property, we construct a confidence interval estimate with some high confidence level $1 - \alpha$, say, .95 (Giere uses q). As we know, a standard .95 confidence interval estimation procedure includes the true population proportion 95 percent of the time—whatever its value might be. It is possible, as Giere shows, to describe a series of applications of the estimation procedure such that the probability (or the ex-

[4] Selvin's calculation, discussed in note 3, is just an application of the Binomial model that we have already considered. As always, each outcome is either a "success" or not. Here, however, a successful outcome is a test result statistically significant at level .05. The probability of getting no successes in 20 independent trials is the probability of not getting a significant difference in one trial, namely, .95 raised to the twentieth power, giving .36. One minus this probability is .64.

pected relative frequency) of successful estimates is not 95 percent but 0! As Giere remarks,

> This will be sufficient to prove [the inadmissibility of this method] because Neyman's theory asserts that the average ratio of success is independent of the constitutions of the population examined. (Giere 1969, 375)

Giere shows how to construct populations that, in effect, illustrate Peirce's point at the outset. Take a population of *A*s and to each set of *n* members from this population assign some shared property. The full population has *U* members where $U > 2n$ members. Then arbitrarily assign this same property to exactly $U/2 - n$ additional members.

> Given a sufficient store of logically independent properties, this can be done for all possible combinations of *n* *A*'s. The result is a population so constructed that while every possible *n*-membered sample contains at least one apparent regularity, every independent property has an actual ratio of exactly one-half in the total population. (Ibid., 376)

More generally, for any postdesignated selection of the property to be estimated whose population frequency is statistically dependent on its frequency in the observed sample used to arrive at the estimate, "one can always construct a possible series of populations leading to an expected ratio of successful estimates differing from [the predesignated confidence level $1 - \alpha$]" (p. 376). (In Mayo 1980, I give an analogous argument for tests.) This difference corresponds, in the case of postdesignated tests, to the difference between the actual significance level and the computed (or predesignated) level.

Summary

To sum up this section, the dependency of the correlation to be tested, or the proportion estimated, on the correlation or proportion that is observed results in changing the experiment—at least, so far as the NP error statistician is concerned. In the standard statistical significance test (on difference in proportions) where the factors whose correlation will be tested are predesignated, the possible outcomes are the possible different degrees of significance that might be observed with respect to that single predesignated correlation of interest. In the case of hunting for a statistically significant difference, in contrast, what is fixed is the particular level of statistical significance for which one is going to hunt. What varies now are the possible factors or possible correlations that might turn out to be statistically significant at that level.

This formal difference in error probabilities corresponds to an informal difference in the way the procedure can err. The hunting procedure has more and different ways of erring than a procedure of testing a predesignated hypothesis. The ability of a test to protect against the errors of one kind of procedure may have no relation to its ability to protect against errors in some other procedure that has many more ways of going wrong. This is what the NP stance against violating predesignation amounts to. Why, then, would anyone who answered yes to the question at the start of section 9.2 object to the NP stance?

9.3 "No Peeking!": Glymour, Scheines, Spirtes, and Kelly

Glymour, Scheines, Spirtes, and Kelly, henceforth abbreviated as GSSK (1987), develop a computerized procedure for using observed correlations in data to construct a (linear) model that fits the observed correlation or difference. I will use their term "model" interchangeably with my "hypothesis." Although their approach is largely intended as a (computerized) method for *finding* models, presumably to be tested on other evidence, the authors claim they "also believe that there is often nothing wrong with using one and the same body of data to discover a theory and to confirm it or test it" (p. 46). The discovery procedure they provide is essentially a computerized program for carrying out the postdesignated searching procedure described above. (The correlations looked for will typically involve many more variables, but this will not alter the points that they or I wish to make.) The bulk of social scientists and statisticians, steeped as they are in NP methodology, object to such a hunting procedure on the grounds that it entails using the same data both to find as well as to test statistically significant correlations. Because they comprise a key group for whom their method of causal modeling is intended, GSSK are led to examine the basis of such objections.

Under the apt subtitle "No Peeking," GSSK (1987) consider what they take to be the best arguments for prohibiting this kind of double use of data, and find them wanting (p. 45). My focus is on the argument that they regard as most promising. They dub it the "worst case argument" and model it on Giere's argument above. We can use it to elicit what I view to be "the real NP stance on predesignation." Toward this end it is sufficient to keep to the type of correlation hypothesis described above, leaving to one side the additional difficulties of warranting causal hypotheses (which we might imagine to be higher up in the hierarchy of models).

The Worst Case Argument

GSSK (1987) spell out what they view as the NP argument against hunting for statistically significant correlations in terms of an argument against their computerized search procedure. The NP argument would begin with a fact that they recognize, namely, that the procedure of hunting for statistically significant correlations "will produce *some* model, even for data that are in fact randomly generated from independent variables" (p. 55). That is, the hunting procedure would find some statistically significant correlation even if there were none. Moreover, they recognize, it would do so with high probability. Nevertheless, GSSK want to reject the NP objection to such a procedure by denying the soundness of what they regard as the NP argument. Let us quote directly from their gloss on the NP argument against computerized hunting procedures:

> 1. Computer-aided heuristic searches for statistical models must examine the data for statistical dependencies among the variables, search for the model or models that best explain [fit] those dependencies, subject the models thus obtained to statistical tests based on the data, and output those models that survive the tests.
> 2. No procedure for searching for hypotheses is acceptable if there are circumstances in which it is very probable that that procedure will yield a false conclusion.
> 3. For any procedure as in 1, a number r of independent random variables and a sample size n can be found such that it is very probable that a sample of size n will show k statistically significant correlations (or other statistic) among h of the r variables, for some number h and for some number k. . . .
> 4. In the circumstance described in 3, it is very probable that a procedure, such as described in 1, will output false hypotheses.
> 5. Therefore, by 4 and 2, a computer-aided heuristic search procedure is unacceptable. (GSSK 1987, 55–56)

Does this argument capture the NP objection? On a first reading it appears tò distort the NP argument, but this is mainly because it is not in the language of NP criteria for judging *inference* procedures. The conclusion, for example, might make it sound as if the NP statistician disallows searching for hypotheses not specified beforehand. In actuality, the only prohibition relates to a procedure of first using data to arrive at a correlation that is statistically significant at some level α, and then using that same data to test the corresponding null hypothesis *H*, that there is no real correlation. However, that is precisely what GSSK mean by a computer-aided heuristic search procedure. We can avoid misunderstandings by spelling out, in the conclusion of the argument (clause 5), the procedure that is being declared NP inadmissible. That change calls for corresponding adjustments to the premises as well. Once these alterations are made, this argument turns out to be a plausible rendering of the NP objection; but contrary to what GSSK suppose, the resulting argument is sound.

Let us turn to the adjustments. To begin with premise 1, we need to be clear about their phrase that the searches "output those models that survive the tests." First, the output of any NP procedure is never just a hypothesis or model or estimate by itself, it is always accompanied by a statement of the error characteristics of the test or estimation procedure. Here the error characteristic is the significance level of the test, α, such as .05. Second, it is rather curious to say that the outputs here are hypotheses that *survive* statistical tests. The outputs are those hypotheses of form H_j^*: treatment T is correlated with F_j. And this output occurs when the corresponding *null hypothesis* H_j does not survive, that is, when H_j is rejected by the statistical significance test. Accordingly we may rewrite premise 1 as follows:

> 1. Computer aided heuristic searches . . . search for statistical dependencies among the variables, reject those null hypotheses H_j if the data show a difference d_j that is statistically significant at some small level α (e.g., .05) and output the non-null hypothesis H_j^* (at level α).

(In more abbreviated form, the procedure is to search for variables for which observed differences d_j satisfy $P(D_j \geq d_j | H_j) \leq \alpha$, declare d_j statistically significant at level α, reject H_j and pass H_j^*.)

In premise 2, once again, "procedure for searching" has to be filled out as the procedure just described. Premise 2 also talks about yielding "a false conclusion," which could be ambiguous were it not for premises 3 and 4. Those premises make it clear that yielding a false conclusion means outputting a statistically significant correlation (at some small level α) even though the variables in question are *not* correlated (in the population) but are independent. (The false conclusion here is a type I error: rejecting the null hypothesis even though it is true.) Premise 2 becomes

> 2. No procedure for significance testing is acceptable if there are circumstances in which it is very probable that that procedure will reject the null hypothesis H_j at a low significance level α (and pass the non-null hypothesis H_j^*) even though the null hypothesis is actually true.

Premise 3, while stated a bit confusingly, simply generalizes the situation we discussed in section 9.2. Since the population imagined is one where all the variables are independent, to assert a statistically significant correlation, that is, to reject the chance hypothesis, is to commit an error. Premise 3 describes a circumstance such that the searching procedure has a high probability of committing such a type I error. This is affirmed in premise 4. It then follows, in 5, that the significance test consisting of the procedure of searching described in premise 1 is unacceptable or inadmissible, on NP grounds.

So, fleshing out the argument given by GSSK turns out to give a reasonable rendition of how an NP argument might go. Will it hold up? GSSK think not.

Their Response and an NP Rejoinder

The main objection of GSSK to the argument concerns premise 2. Premise 2, they object, assumes that a procedure ought to be judged by the worst imaginable case, namely, the circumstances described in premise 3.[5] Why consider the worst possible case, they ask, where all of the variables searched are actually independent?

> In the majority of cases researchers are pretty confident that the statistical dependencies they find are due to some causal structure or other. . . . If the investigator were not strongly inclined to think that there is some explanation other than chance (or bad measurement design) for the patterns found in the data, a causal model would not be sought in the first place. Unless the researcher thinks there is a large probability that the dependencies in the data are spurious, there is no sufficient reason not to use the data to search for the best explanation of it. (GSSK 1987, 56–57)

They continue that,

> of course, some of the correlations found may be due to chance, and that is the more likely the smaller the sample size in proportion to the number of variables considered. The investigator should certainly

[5] GSSK (1987) also object that premise 2 "puts all of the weight in judging a procedure on the desirability of avoiding *false* theories" (p. 56). But NP criteria also consider the probability that a true hypothesis is accepted—by seeking a small type II error probability. It is just that the situation to which this NP argument is referring is one where it is given as part of the procedure that some hypothesis (or estimate) is going to be accepted. Thus the error of concern is the probability of erroneous acceptance. Of course nothing in the NP argument says there is anything wrong with searching for statistically significant results for the sake of getting some hypotheses to consider, and then testing them on other data.

> take account of that fact and, where appropriate, test a model on new samples. (P. 57)

I will return later to this continuation, which suggests a different tack than the earlier parts of the passage. What about the first part of their response? For starters, merely being ignorant of spuriousness is not sufficient for most researchers to be "pretty confident that the statistical dependencies they find are due to some causal structure." If, alternatively, this confidence is well founded from background knowledge, then why are the researchers running statistical significance tests? Perhaps what GSSK mean is that, most of the time, the researchers have a strong belief that any dependencies found to be "statistically significant" are real: they assume there's a causal story to be had; the function of the test is to tell them which factors are really connected.

For the test to tell them which are really connected, however, it has to be able to calculate the actual statistical significance—the actual error probability of the test procedure. That is, the researchers' reliance on an analysis of statistical significance to arrive at a reliable argument from error depends critically on that analysis being done correctly. Significance tests are useful only to the extent that they can be relied on to alert us to the lack of statistical significance when there is none. Tests *should* declare the statistical significance level high and *not* low, when the correlation observed is of the sort that would, very probably, disappear in subsequent trials. This ability to make reference to what would probably occur in additional trials is altogether central to NP principles.

The error statistician might put the rejoinder this way. This type of statistical significance test is designed for a case in which one is concerned with ruling out the error that an observed correlation is merely due to chance. It is designed for a case in which one wants to know what it would be like if this were a population in which it *would* be an error to declare correlations genuine. As we have seen in previous chapters, one can then use the sample to see whether this error can be ruled out.[6] If your case is not one in which there is a need of such information, then you are not in need of this particular NP test. But premise 2 does not rest on any claims about the *relevance* of significance tests for a given research situation. Claims about researchers' confi-

[6] What is wanted in such a situation is a standard signal or warning that an observed correlation is of the sort that can often occur by chance: that it is the sort of result frequently generated even in working with a population in which the factors are independent. The actual significance level is a measure of this frequency, so its being high indicates that such a correlation is highly probable by chance alone.

dence, warranted or not, are quite beside the point of the NP argument, which is really just an argument, a demonstrable one, about the *properties of these testing tools.* The argument shows that if you change the test procedure the error probabilities change, and if you report significance levels in the usual way—if you are a "hunter" in Kish's sense—then you are going to get your error probabilities wrong.

The upshot of premise 2, we can imagine the NP test saying, is this: "I cannot do my job if you are a hunter. My logic breaks down. Don't blame me if you declare correlations real far more often than the computed level arrived at in searching for significance." Or as Selvin (1970) put it, if you apply statistical significance tests to a hunted hypothesis, "a spurious impression of validity may result. The computed level of significance may have almost no relation to the true level" (p. 104).

What GSSK call the "worst case" is precisely the case that the NP statistical test *must* consider here, because that is just what the type I error would be. It is often said that no one really thinks (point) null hypotheses are exactly true. But it is a mistake to regard this as a criticism of their use in tests. We use them in getting the probability of a type I error, or the significance level, because we seek an objective way of learning how far from true they are. Hunting, on the other hand, allows correlations to be described as improbably far from what would be expected due to chance, when in fact they are quite typical of what would be expected even if chance alone were responsible.

Using Computed Levels of Significance as Fit Measures. Granted, one may only be interested in giving a kind of summary measure of the observed correlation, and the computed significance level could be used for this. (The smaller the computed level, the larger the deviation from the corresponding null hypothesis H_j. This corresponds to a greater "fit" with the alternative hypothesis H_j^*.) However, the process of giving a summary measure of fit is no longer an NP test process. An NP test process always asks about the error probability of the observed correlation or fit measure—it would ask about the actual significance level. Without such an error probability, *there is at most a data summary and not a statistical inference* from the observed correlation to the population. To such a data summary the whole question of NP inadmissibility in premise 2 would not apply. Nevertheless, the soundness of this premise, which after all pertains to NP *tests*, still stands.

The Honest Hunter: Defeating the Worst Case

We have justified the argument of the NP inadmissibility of postdesignated or searching procedures of significance testing. The objection of GSSK exposed no unsoundness in the NP argument—once that argument is properly understood. However, the later part of GSSK's statement quoted earlier, as well as their subsequent research efforts, leads me to suspect that GSSK agree. For there they allude to the advisability of (*a*) ensuring a large enough sample size relative to the number of variables to be hunted or (*b*) testing on new samples. Both (*a*) and (*b*) presuppose taking seriously the threat posed by hunting for statistical significance. Avenue *b* is tantamount to setting out a new test and not violating predesignation altogether. Avenue *a,* however, may be seen as a way of defeating the worst case. I want to pursue this avenue a bit.

Let us use the designation "hunters" to refer to those who engage in "hunting with a shotgun" and allow postdesignated hypotheses about correlations to be reported *just as if* they were predesignated. One might agree that the NP argument above is a sound argument against being a hunter, but deny that this bars all postdesignated tests that double count data. It does not bar the practices of what we might call "honest hunters." Honest hunters report, as far as possible, the true or actual significance level, taking into account the way this is altered by the fact of hunting. In suggesting avenue *a,* however briefly, GSSK seem to be taking the line of the honest hunter.

NP statisticians can have no *in principle* objection to hunting and reporting the actual significance level[7]—although they may have a practical one, which I will return to in a moment. Quite the contrary, that is what the NP statistician would recommend in cases that violate predesignation. To state it more generally, avenue *a* advises that by appropriately specifying the test (number of variables searched, sample size, the significance level required), the actual significance level, even in postdesignated tests, may be sufficiently small.

One way that can happen is if many of the hypotheses tried turn out to be statistically significant. Let us return to Selvin. He explains how "curiously enough" the same argument against improperly using significance tests in postdesignation cases

> can be extended to show how the tests might legitimately be used on such hypotheses. Consider once more the twenty differences drawn from populations where the true differences are zero. We have seen that the probability of *at least one* difference "significant" at the 5 percent level is 0.64. By similar calculation it can be shown that the prob-

[7] That is so whether or not it is high or low. It cannot bar an entire procedure because it is possible for it to be applied in a case with large error probabilities, for that is true for all NP procedures. It may say it is not a particularly good test, but not that it is inadmissible.

> ability of at least *two* "significant" differences is 0.26, that the probability of at least *three* is 0.07, and that the probability of at least *four* is 0.01. In other words, if one examines twenty differences and finds four or more "significant" at the 5 percent level, then the *set* of differences is significant at the 1 percent level, since this combined result would have happened only one time in a hundred if the true differences were zero. (Selvin 1970, 104–5)

Several caveats remain. First, while one can reject the worst case here, one still cannot say any particular hypothesized correlation has passed a severe test. Second, as Selvin notes, sustaining this argument would require carefully considering correlated biases and the lack of independence of the 20 factors (p. 105). For ease of computation, Selvin does not take these into account. Third, here the number of differences looked at was fixed at 20. If the number of variables sought could be open-ended—in a diligent hunting expedition—then it would be far more difficult to get a low error probability. It may not even be clear how to determine what the error probability is in such an open-ended case. Nevertheless, the honest hunter could argue that particular contexts impose a limit on the possible variables and corresponding hypotheses.[8]

So the task for honest hunters might be put as that of finding ways of showing that the overall error probability is fairly low. I take this to be the idea pursued in subsequent work by GSSK. By means of Monte Carlo simulations, and considerations of background constraints (based on knowledge of which variables are or are not related), they have made progress in investigating the reliability of their search procedures. The impetus of these investigations, in my view, is to show that their use-construction rules are of the sort that I called R-π rules, with fairly high severity π in chapter 8, section 8.5. The NP philosophy

[8] One should not overlook a related and very serious problem that can arise. Here we are imagining that the analyzer knows all of the tests attempted. The situation is entirely different if one must resort to rounding up positive and negative results from the literature. The problem, often called the "file drawer problem," is that nonstatistically significant results may remain in file drawers, never to be published. Because these negative results would not be counted, a much higher proportion of statistically significant results would be found than actually exist. This is a good example of a canonical mistake.

Robert Rosenthal, a leader in the relatively recent area of "meta-analysis," discusses how one might "estimate the degree of damage to any research conclusion that could be done by the file drawer problem" (1987, 223). This attempt to estimate and subtract out the effect of studies remaining in file drawers is, in its intent, very much in the spirit of the error statistical program.

has no valid argument against postdesignation in such cases. (But see note 9.)

C. S. Peirce, that arch-predesignationist, actually anticipated the gist of the responses of the honest hunter. Perhaps this is not really surprising because Peirce in many ways seemed to have anticipated the appropriate construal of NP methods—a thesis to be pursued in chapter 12. Peirce discusses an important modification of the rule of predesignation, namely, when it is not necessary.

> Without any voluntary predesignation, the limitation of our imagination and experience amounts to a predesignation far within those limits; . . . thus . . . if the number of instances be very great . . . the failure to predesignate is not an important fault. . . . So that if a large number of samples of a class are found to have some very striking character in common, or if a large number of characters of one object are found to be possessed by a very familiar object, we need not hesitate to infer, in the first case, that the same characters belong to the whole class, or, in the second case, that the two objects are practically identical. (Peirce 2.740)

A Tactical NP Objection

One can articulate another kind of NP objection to hunting for correlations. Although it too can be seen to have its basis in what I called the error-probability principle (EPP), it is not an argument about the inadmissibility in principle of violating predesignation. The objection now is more practical: it is so difficult to figure out what the actual error probabilities are in postdesignated cases, this objection goes, that they are not recommended by this school of inference. Even adjusting the test specifications to get reasonable significance levels rests on very slippery assumptions or requires too-large sample sizes to be practicable. Nevertheless, new uses of computer-driven Monte Carlo simulations might get around these tactical criticisms. To their credit, GSSK, in their most recent work, appear to be heading in that direction.

9.4 THE CREATIVE ERROR-THEORIST

Some practitioners may feel dissatisfied at what I have provided thus far. They may grant that the NP theorist is correct to charge that the error probability guarantees of NP tests break down in hunting procedures. They may likewise grant the validity of calling for elaborate adjustments to significance levels, sample sizes, and so on to ameliorate the problem of high error probabilities in hunting procedures. But in point of fact they are still confronted with the realities of their inquiries, and being an honest hunter by the straight and narrow path still prevents them from doing the kinds of things that it seems they ought to be able to do with the evidence they have. If one keeps in mind that the ultimate goal is a severe argument from error, then it is possible to go further, while remaining within the NP school. The previous chapters give us a head start on this problem.

This takes us into the realm of informal error probability arguments, and hence from formal NP statistics into the broader realm of the error statistician. Consider a case in which a violation of predesignation results in the actual significance level being high (and thus, the severity being low). The honest hunter must report the actual significance level. So the statistical significance test analysis does not provide a severe test of the reality of some correlation. But that is not the only kind of analysis possible, even without getting more data. (Of course, if it were feasible to get more data, this might be desirable.) Even if none of the quantitative NP tools has anything more to offer (at least not at present), the error-statistician's tool kit certainly might. What is needed is an argument from error, an argument to rule out error. If it is remembered that that is the underlying rationale for NP inferences in the first place, then arriving at or approximating some such argument must be countenanced on error-severity grounds.

Attention to the inadmissibility argument spelled out above alerts the researcher that the significance test does not license a certain inference. It says nothing about other arguments that might be put forth. The error statistician requires only that it be able to be shown that the argument used is reliable in our sense. (The researcher cannot just say that he or she feels very strongly about the conclusion.) Nothing in what we have said precludes assessing the reliability of some other method that a researcher might custom-design. It might well be shown that the custom-designed method constitutes a reliable method for inferring a type of correlation. It might be shown to be a high severity use-construction rule (R-π), as defined in chapter 8.

Recall our discussion of the informal arguments from coincidence in chapter 3—how they justified inferring that an effect is real, that it will not go away. Hacking, for example, presented such an argument for dense bodies. The same was seen in Perrin's argument for coincidence for the causes of Brownian motion. At this stage of their arguments there was no need to formally define a test statistic with quantitative error probabilities. They could arrive at arguments that fairly well rule out the error of mistaking an artifact for a real effect. The informal calculation of severity mimics the formal one.

Of course, whether one is relying on a formal or an informal argument from error, one must be careful not to infer more than the argument warrants. Having knowledge of a real effect is not the same as having knowledge of an important effect; much less is it the same as knowing about its specific cause. The need for a piecemeal breakdown into arguments from error, and the need to limit the inference to what is strictly warranted by each argument, must be respected (chapters 5 and 6).[9]

The quantitative NP test procedures serve as canonical models of error—so do NP inadmissibility arguments. Canonical models of error, recall, are exemplars, of both admirably high and infamously low reliability. The "worst case" scenarios, as in the examples of Kish, Selvin, and Giere, are examples of the latter. They demonstrate how hunting expeditions can lead you terribly astray. The lesson is that the onus is on you to show how you are getting around that possibility. The situation is similar to what happens in retrospective analyses of correlations. Knowledge of how that can lead one astray has given rise to a host of procedures for reliably inferring correlations retrospectively. In a much less systematic way, individual researchers, especially in medicine, develop informal procedures for learning from correlations that are only noticed from "peeking" at the results of medical trials. Some oncologists hold that the major advances in cancer chemotherapy were made based on retrospective studies of small groups of patients (Greenspan 1982, 8). However difficult it may be to arrive at these approximate or qualitative error probabilities, if they are arrived at and are valid from a frequentist point of view, there are no error statistical grounds for condemning them.

9.5 CONCLUSION

The real NP argument against postdesignated significance tests is that if the same data are used both to construct as well as to test statistically

[9] Even if one manages to obtain a low actual level of significance and thereby pass the nonchance hypotheses severely, this severity does not carry over into particular interpretations of the correlation found. These distinct inferences (we may locate them in a model above the experimental hypothesis of a real correlation) introduce a distinct set of possible errors.

For example, in the social sciences, as Paul Meehl has steadily warned, genuine correlations can be found among nearly all variables "since in social science everything correlates with everything to some extent" (1990, 207). This fact—Meehl calls it the "crud factor"—introduces an important canonical error: when the crud factor is high, a test that takes evidence of a real correlation as evidence of a particular causal hypothesis would often pass causal hypotheses erroneously—the test would have poor severity. Meehl makes the intriguing proposal of estimathypothesized correlations, the actual probability of erroneously declaring a correlation to be genuine—that is, the actual significance level—differs from, and may be much greater than, the computed significance level. Using the computed significance level in postdesignation cases forces one to adopt a different interpretation of a significance level, one that conflicts with the intended interpretation and use of significance levels (as error probabilities).

The upshot of the real NP argument is a warning: since violating predesignation may alter the actual significance level (by altering the test procedure), it is invalid to report the results *in the same way* as if hypotheses were predesignated. Now the high actual significance level corresponds to low severity, for it means that there is a high probability of affirming the existence of a real correlation erroneously. Hence the justification for the NP warning is that if one fails to heed it, tests will be construed erroneously as having high severity.

One could suggest, as I believe GSSK do, that the computed significance level be calculated simply as a kind of measure of fit and then give some other report of the overall error probability. In getting these other reports, they also allude to various background considerations or "constraints" regarding which variables are or are not connected. If these other reports of reliability are valid from a frequentist point of view, then they may certainly be sanctioned and even welcomed by the error probability statistician. It is important to see, however, that this is not a defense of hunting in standard significance tests, quite the opposite. It is to grant that the NP inadmissibility argument necessitates some wholly other kind of test or analysis. I take it that this is what GSSK are really hunting for.

The examples we have considered of hunting for statistical significance should be viewed as canonical models of error—as classic ways of being led into clearly unreliable tests. Rather than viewing them as part of an utter prohibition of violating predesignation, they should be viewed as invitations to articulate creative arguments to substantiate reliability by other means. Such developments are quite in keeping with the statistical philosophy of E. S. Pearson:

> There is perhaps in current literature a tendency to speak of the Neyman-Pearson contributions as some static system, rather than as part of the historical process of development of thought on statistical theory which is and will always go on. (Pearson 1966d, 276)

ing the crud factor for given domains—something I hope will be pursued. Meehl's work provides an excellent source for building a tool kit of errors for social science research.

Considerations of the creative postdesignationist provide the groundwork for justifying a break with overly narrow construals of NP methodology. This break is not new but reflects sound uses of those procedures in much of scientific practice. What is still needed is a clear articulation of the associated error-statistical arguments. This is part of the larger task of setting out an adequate methodology of experiment, a task that requires domain-specific considerations and is beyond the scope of this book. Despite wide latitude for such a program, the one thing retained is the constraint—formal or informal—of error statistics or severity. This stands in marked contrast to the alternative program represented by the Bayesian Way. Deliberate disregard for this constraint, as will be seen in the next chapter, "frees" the Bayesian to view hunting and data snooping as irrelevant to the import of the evidence in hand.