2

Hypothesis Testing: Criticisms and Alternatives

This chapter begins by distinguishing among different uses of hypothesis tests. It then summarizes the major criticisms that have been offered. Two alternatives to standard tests, the AIC and the BIC, are described, and the different criteria are applied to four examples. The examples show that model selection by the AIC, BIC, and classical hypothesis tests can lead to dramatically different conclusions.

2.1 HYPOTHESIS TESTING AND ITS DISCONTENTS

Classical hypothesis tests became the subject of severe criticism almost as soon as they began to be widely used: Berkson's (1938, 1942/1970) critiques are well known and are still cited today. The criticism has continued and even intensified since that time. Cohen (1994, p. 997), a psychologist, charges that hypothesis testing "has not only failed to support the advance of psychology as a science but also has seriously impeded it." Mason (1991, p. 343), a sociologist, holds that "much, perhaps most, use of statistical inference in the social sciences is ritualistic and even irrelevant. . . . Those asterisks that adorn the tables of our manuscripts are the product of ritual and little, if anything, more than that." McCloskey (1998, p. 111), an economist, proclaims that "statistical significance is bankrupt; all the 'findings' of the Age

of Statistical Significance are erroneous and need to be redone." Gill (1999, p. 647), a political scientist, speaks of "the insignificance of null hypothesis significance testing." The philosophers Howson and Urbach (1994, p. 50) assert that "the principles of significance testing . . . are simply wrong, and clearly beyond repair." Lindley (comment on Johnstone, 1986, p. 582), a statistician, maintains that "significance tests . . . are widely used, yet are logically indefensible."[1] Further examples of such views could easily be found.

Although their critics have been more vocal, hypothesis tests have some defenders. The basic argument in their favor was made by Mosteller and Bush (1954, pp. 331–332; see also Davis, 1958/1970, and Wallis and Roberts, 1956) 60 years ago: "The main purpose of a significance test is to inhibit the natural enthusiasm of the investigator." Without some accepted standard, researchers would be able to claim any parameter estimate with the expected sign as favorable evidence or, alternatively, dismiss any deviations from a preferred model as "trivial." Hypothesis tests have survived and spread in the face of criticism because they fill an important need. However, even if some standard is necessary, it is reasonable to ask whether conventional hypothesis tests provide the *best* standard, and this question will be explored in the following chapters.

2.2 USES OF HYPOTHESIS TESTS

Before considering the criticisms of hypothesis tests, it is necessary to distinguish the different purposes for which they are used. This section discusses the major ones, drawing particularly on Cox (1977; see also Anscombe, 1961, and Krantz, 1999).

2.2.1 Conclusions about Parameters of Interest

An important use of hypothesis tests, and the one that has received the most attention in statistical theory, is to evaluate theoretically based propositions

[1]Articles in statistics journals are often followed by short comments and replies. Comments will be included as separate entries in the bibliography if they have distinct titles, and otherwise indicated as in this reference.

about the value of a parameter. In most cases, the proposition is that the value of an independent variable x will have a particular association—positive or negative—with the outcome variable y after controlling for other variables. Sometimes the proposition of theoretical interest involves other aspects of the model such as functional form. The usual procedure is to fit an equation of the form $\hat{y} = \alpha + \beta x + \gamma_1 z_1 \ldots + \gamma_k z_k + e$, where x is the variable of interest and $z_1 \ldots z_k$ are other variables that might influence y.[2] The claim that x is associated with y implies that the null hypothesis $\beta = 0$ is false. Therefore, if the null hypothesis is rejected and the parameter estimate has the expected sign, that counts as support for the prediction derived from theory.

If a theory predicts an exact value for a parameter, it is possible to test the null hypothesis that it has that value. Accepting the null hypothesis then is a success for the theory in the sense that it is not refuted, although it does not provide positive evidence in favor of it. In the social sciences, it is rare for a theory to predict a specific nonzero parameter value, but sometimes a theory implies that there will be *no* association between two variables after appropriate controls are included.

2.2.2 Choice of Control Variables

In most research, there are many variables that might affect the outcome but that are not of theoretical interest. Omitting variables that affect the dependent variable will usually produce biased estimates of the parameters of interest, so the best strategy for minimizing bias would be to include every potential control variable. However, the inclusion of controls reduces the precision of the estimates of the parameters of interest, especially in small samples. As a result, it is necessary to have some procedure to decide which potential control variables should be included and which should be omitted. Hypothesis tests are often used to make these decisions: accepting the hypothesis that the coefficient for some variable equals zero means that it can be omitted. In contrast to conclusions about parameters of interest, the sign of the parameter estimate is not of interest for control variables: the only question is whether it is different from zero.

[2] I assume a linear model for simplicity of exposition.

2.2.3 Primary Structure

"Primary structure" (Cox, 1977, p. 52) involves parts of the model specification that define the parameters of theoretical interest. For example, many theories in the social sciences propose a relationship of the form "the larger the x, the larger the y." Hypotheses of this kind can be represented by:

$$y = \alpha + \beta f(x) + e \tag{2.1}$$

where $f(x)$ is any monotonically increasing function of x. The parameter of theoretical interest is β, and the nature of $f(x)$ is an issue of primary structure.

Simplicity of interpretation or ease of estimation are often important considerations in the choice of $f(x)$. In this example, one might begin with a linear model and test it against a quadratic model

$$y = \alpha + \beta_1 x + \beta_2 x^2 + e \tag{2.2}$$

Even if the hypothesis that $\beta_2 = 0$ could be rejected, a researcher might still prefer to use a linear regression as long as it provided a good approximation.

However, the precise specification of primary structure is sometimes important. For example, Myrskylä, Kohler, and Billari (2009) analyze the relationship between socioeconomic development and fertility in nations of the contemporary world (see Section 2.6.2 for a more detailed discussion). There is a strong negative association over most of the values of development, but they propose that at higher levels of development the relationship is reversed: that is, $\partial y/\partial x$ is positive when x is near the highest levels observed in the data. Conclusions on this point are sensitive to the exact specification of the relationship between x and y, so the issue is more important than it would be if the research question simply involved the general direction of association.

2.2.4 Secondary Structure

Secondary structure involves aspects of the model that are relevant to the estimation of parameters of interest but not to their definition. In the case of linear regression, the assumptions that the errors are independent, follow a normal distribution, and have constant variance are aspects of secondary structure. If these assumptions are incorrect, the parameter estimates will

still have the same interpretation, but ordinary least squares will be an inefficient method of estimation. As in questions of primary structure, the null hypothesis is often regarded as a convenient approximation rather than as a proposition that might be exactly true.

2.2.5 Goodness of Fit

The idea of a goodness-of-fit test is to assess the fit of a model to the data without specifying an alternative model (Anscombe, 1963). However, many goodness-of-fit tests can be understood as tests against a very general alternative. For example, the chi-square test for independence in a contingency table implicitly compares the model of independence to a "saturated" model in which each cell has its own parameter μ_{ij}. The defining feature of a goodness-of-fit test is that the alternative model is not adopted if the null hypothesis is rejected—rather, the result is taken to mean that it is necessary to continue searching for a satisfactory model.

Tests against a saturated model are not always possible—for example, in a standard regression model, the estimate of the variance is undefined if we fit a separate parameter for each observation. However, some specification tests are designed to detect a wide range of potential flaws in a model without specifying a serious alternative. For example, Davidson, Godfrey, and MacKinnon (1985) propose testing the hypothesis $\gamma = 0$ in the time series regression:

$$y_t = \alpha + \beta x_t + \gamma(x_{t+1} + x_{t-1}) + e \tag{2.3}$$

Because it is logically impossible for future values of x to influence y, Equation 2.3 would not be adopted if the hypothesis were rejected. Rather, the result would be taken to mean that the independent variable and the error term are correlated in the regression

$$y_t = \alpha + \beta x_t + e \tag{2.4}$$

Such correlation could result from a variety of causes, including omitted variables or errors in the measurement of x, so if the null hypothesis were rejected it would be necessary to carry out tests for the particular possibilities and modify Equation 2.4 in an appropriate fashion.

As this example illustrates, the distinction between a goodness-of-fit test and a test of a specific hypothesis is a matter of degree: a test may indicate a number of possible problems in the model without being completely general. Whether a given test should be regarded as a goodness-of-fit test thus depends to some extent on the purposes of the researcher. For example, serial correlation in the residuals of a time series regression may result from the omission of a relevant variable or misspecification of the functional form. If the hypothesis of no serial correlation is rejected, a researcher might estimate the model using a "correction" for serial correlation, or might conduct a wider search for different kinds of misspecification.

2.3 CRITICISMS OF CONVENTIONAL HYPOTHESIS TESTING

This section describes some of the most important criticisms of the conventional practices described in Section 1.3. The validity of the criticisms will not be considered here but rather left until after the discussion of the classical theory of hypothesis testing in Chapter 3. The first three criticisms involve the scope of hypothesis tests: they hold that there are situations in which conventional hypothesis tests cannot be applied or that there are important questions they cannot answer. The fourth and fifth involve ambiguity or paradoxes. The sixth, seventh, and eighth criticisms are more fundamental and raise questions about the logic of conventional hypothesis testing.

2.3.1 Sampling

Almost all textbook presentations of hypothesis testing involve random sampling from a population, or a process that can be repeated under identical conditions such as plays in a game of chance. In that case, *p*-values represent the proportion of random samples for which the test statistic would be greater than or equal to the observed value. However, hypothesis tests are often applied to data that do not represent random samples. For example, many statistical analyses involve the span of time for which data on the variables of interest are available. Such data could be regarded as a sample from the history of the unit (past and future) but cannot plausibly be regarded as a

random sample. Other research involves complete populations, for example, all nations in the contemporary world.

Some observers maintain that conventional hypothesis tests, even if they are valid in principle, can properly be applied only to data that represents a random sample from some population. Moreover, they hold that tests should be interpreted only in terms of sampling error, not in terms of general uncertainty about the parameter values. According to Morrison and Henkel (1969/1970, p. 186), "Significance tests . . . are not legitimately used for any purpose other than that of assessing the sampling error of a statistic designed to describe a particular population based on a probability sample." McCloskey (1996, p. 40) quotes this remark and adds that "no one disputed their declaration because it is indisputable. That is what 'statistical significance' means, mathematically speaking." Similarly, Schrodt (2014, p. 293) says that the use of significance tests on data that are not from probability samples requires "six-impossible-things-before-breakfast gyrations." As Schrodt's comment implies, there is disagreement on this point: some observers argue that conventional hypothesis tests can legitimately be applied to populations and nonrandom samples. This issue will be discussed in more detail in Section 3.1.

2.3.2 Credibility of Point Null Hypotheses

Many null hypotheses are not credible in principle. For example, standard variables such as gender, education, marital status, income, or ethnicity can be expected to have at least some slight association with almost any individual-level outcome that is of interest to social scientists. Turning to another unit of analysis, Mulligan, Sala-i-Martin, and Gil (2003) ask, "Do democracies have different public policies than nondemocracies?" If their question is taken literally, the answer is surely yes—the real question is whether the policies are "substantially" different. This point raises the question of what is learned from the test of a hypothesis that is almost certainly false. As Giere (1972, p. 173) puts it, "We know when we start that the null hypothesis is false. If our test fails to reject it, that only tells us we did not take a sample large enough to detect the difference, not that there is none. So why bother testing in the first place?"

As with the first criticism, there is disagreement on this issue. Some observers, such as Tukey (1991), maintain that all or almost all point null hypotheses are false. Others maintain that many or most point null hypotheses are true: for example, Sterne and Davey-Smith (2001, p. 227) estimate that about 90% of the null hypotheses tested in epidemiology are correct. The answer is likely to differ by field of study, but it is safe to say that many of the null hypotheses that are tested in the social sciences could be rejected as a matter of common sense without any statistical test.

2.3.3 Ranking of Models

A test of a point hypothesis can be regarded as a method for choosing between two models. The model represented by the null hypothesis can be regarded a special case of a larger model, and the null hypothesis involves the restriction that some of the parameters of the larger model have specific values. However, when more than two models are involved, a hypothesis test does not always lead to a clear decision in favor of one of them. This point was discussed in Section 1.3, using the example of the the regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

The full model may be tested against three alternative models, one imposing the restriction $\beta_1 = 0$, another imposing the restriction $\beta_2 = 0$, and the third imposing the restriction $\beta_1 = \beta_2 = 0$. It is possible that both $\beta_1 = 0$ and $\beta_2 = 0$ can be accepted while $\beta_1 = \beta_2 = 0$ is rejected. In this case, it would seem reasonable to say that there is no strong basis for preferring the model with $\beta_1 = 0$ over the model with $\beta_2 = 0$, but a researcher might nevertheless want to make a tentative choice. Conventional hypothesis tests do not provide a means to do this.

As a slightly more complex example, suppose that we are comparing two ways of representing social position: a single numerical measure of socioeconomic status (SES) and a set of five class categories. Hypothetical data for this situation are shown in Table 2.1. The column labeled p gives the number of independent variables (not counting the constant), and D represents the deviance. A test of the null hypothesis that the coefficients for the class dummies are all zero when SES is included gives a chi-square statistic of 3.0 with

TABLE 2.1. Hypothetical Fit Statistics from Alternative Definitions of Social Position

Model	*p*	*D*
Both	5	75.0
SES	1	78.0
Class dummies	4	78.0
Neither	0	100.0

4 degrees of freedom, for a *p*-value of .56. A test of the null hypothesis that the coefficient for SES is zero when class is included gives a chi-square of 3.0 with 1 degree of freedom ($p = .084$). Therefore, neither the class nor the SES model can be rejected in favor of the "full" model including both class and SES at the 5% level.

In this case, common sense suggests that the SES model is superior on the grounds that it fits equally well while using fewer parameters. The SES and class dummies models cannot be tested against each other using standard tests, because they are not nested. If a non-nested hypothesis test were used, neither model could be rejected against the other at the 5% level. One could say that the SES model should be preferred on the grounds that, in a test against the full model, the class model can be rejected at the 10% level and the SES model would not, but it is difficult to say how much evidence this amounts to.

As this example illustrates, conventional hypothesis tests do not produce a complete ranking of potential models. A model that cannot be rejected against any of the others can be regarded as acceptable, but often there will be more than one acceptable model. The example in Table 2.1 is a very simple one: as more parameters are considered, the chance of having more than one "acceptable" model will tend to increase. Moreover, the classification of models as acceptable or unacceptable will differ depending on the significance levels used to make the decisions, adding another level of ambiguity.

2.3.4 Flexible versus Inflexible Interpretation of Significance

The *p*-value is a continuous measure, and figures slightly above and below conventional standards of statistical significance represent almost the same

amount of evidence: as Rosnow and Rosenthal (1989, p. 1277) put it, "Surely God loves the .06 almost as much as the .05." This point suggests that the standards should be treated flexibly—for example, it would be reasonable to regard a *p*-value of .06 as providing *some* evidence against the null hypothesis. However, other observers recommend strict adherence to conventional levels on the ground that it reduces the influence of the investigator's hopes and expectations on conclusions. Allowing room for judgment in the interpretation of *p*-values introduces a bias in favor of conclusions that support prior beliefs, and rigid adherence to a conventional level of significance prevents such "gamesmanship" (Bross, 1971, p. 508).

2.3.5 Arbitrary Nature of Significance Levels

The general idea behind conventional tests is that a hypothesis should be rejected if the observed data would be unlikely if the hypothesis were true. Although this principle seems reasonable, it is not clear where to draw the line between "likely" and "unlikely." In the early development of hypothesis testing, a number of different conventions were proposed. At one time, a *t*-ratio of 3 (equivalent to an α level of about .0025) was often taken as the standard. According to one textbook (Waugh, 1943, p. 257), "Some call a statistical result 'significant' if it would arise by chance only once in 100 times. . . . Our point, three standard errors, errs if at all in requiring too much before the possibility of chance is ruled out. It is, however, the most commonly accepted point among American statisticians." A debate between Ross (1933) and Peters (1933) in the *American Journal of Sociology* involves the use of 3 standard errors as the standard for significance and anticipates some of the points made in later debates over the interpretation of "nonsignificant" results. A somewhat later textbook (Hagood and Price, 1952, pp. 323–324) reported that "in certain fields of research it has become conventional to use the '5-percent level of significance,' . . . in others to use the 'one-percent level of significance' . . . and still others to use the 'one tenth of one-percent level of significance.'" As time went on, the convention of 5% became firmly established across a variety of fields (see Leahey, 2005, for a historical account of this process in sociology).

Some observers argue that the widespread convergence on the 5% level

shows that it is not arbitrary. Bross (1971, p. 507) argues that the .05 standard has come to dominate because it works well in practice. For example, a level of 0.1% "is rarely attainable in biomedical experimentation. . . . From the standpoint of communication the level would have been of little value and the evolutionary process would have eliminated it." Cowles and Davis (1982, p. 557) maintain that early statisticians had adopted something close to the 5% level of significance as a standard even before the development of the theory of hypothesis testing, and they suggest that it represents a widespread disposition: "People, scientists and nonscientists, generally feel that an event which occurs 5% of the time or less is a rare event."

However, regardless of the merits of these arguments, the 5% convention is certainly not derived from statistical theory. Lehmann (1959, p. 61) observed that "it has become customary to choose for α one of a number of standard values such as .005, .01, or .05. There is some convenience in such standardization since it permits a reduction in certain tables needed for carrying out various tests. Otherwise there appears to be no particular reason for selecting these values." Yule and Kendall (1950, p. 472) were more direct: "It is a matter of personal taste."

2.3.6 Effects of Sample Size

Many observers have noticed that it is often easy to reject null hypotheses when the number of cases is large. In a sufficiently large sample, researchers who use hypothesis tests to guide model selection are often led toward complicated models containing many parameters that seem unimportant or difficult to interpret. In Berkson's (1938, pp. 526–527) frequently cited words: "I make the following dogmatic statement, referring for illustration to the normal curve: 'If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on an order of 200,000—the chi-square p will be small beyond any usual limit of significance.'" This issue is sometimes referred to as a distinction between "statistical" and "substantive" significance—in a large sample, effects that are too small to be of practical or scientific interest may be statistically significant.

In itself, this point is not an argument against the use of hypothesis tests. In the approach described in Section 1.2, if the null hypothesis is

rejected the investigator looks at the size of the parameter estimate. At this stage, some statistically significant parameter estimates may be set aside as being too small to be worth further discussion. Researchers who omitted the hypothesis test and immediately looked at parameter estimates would put themselves at risk of "explaining" effects that do not actually exist. As Wallis and Roberts (1956, p. 480) observed, "It is futile to speculate on practical significance unless the finding is statistically significant."

The connection between sample size and statistical significance does raise a more serious problem, however. There are two kinds of errors—rejecting a true null hypothesis (Type I error) and accepting a false null hypothesis (Type II error). If one uses a constant standard of statistical significance, the chance of Type II errors declines as the sample size increases, but the chance of Type I errors remains the same. Yet, if both types of error are important, it is desirable to make the chances of both decline. This would mean using a generous standard for statistical significance in small samples and making the standard increasingly stringent as the number of cases increases. Many observers have called for such adjustments: for example, Wolfowitz (1967/1980, p. 440) says that "the use of the conventional levels of significance (.05, .01, .001) without any regard for the power of the test is an absurdity which has been pointed out in many places." Yet there are no generally accepted principles for doing this: Kadane and Dickey (1980, p. 246) observe that "after all these years there is no satisfactory theory of how to decide what level of test to set, and most practitioners continue to use .05 or .01 in an automatic way, without justification." Sometimes ad hoc adjustments are made, such as using .10 as the standard for significance in a "small" sample, but these are open to the same objection as the flexible interpretation of statistical significance discussed in the previous section: they let researchers adjust the standards to increase the chance of reaching the conclusion that they prefer.[3]

[3]Hendry (1995, p. 490) proposes setting α equal to $1.6N^{-0.9}$, asserting that the rule strikes "a reasonable balance between the costs of Type I and Type II errors when the actual cost of either mistake cannot be assigned." However, he does not offer a theoretical argument for the formula; he merely says that it seems to match common practice. Even if his judgment is accurate on this point, it raises the question of whether common practice is correct.

2.3.7 Lack of Symmetry

Conventional hypothesis tests are asymmetrical: they cannot produce evidence in favor of the null hypothesis, only evidence against it. As Fisher (1937, p. 19) put it, "The null hypothesis is never proved or established. . . . Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

Whether the inability to produce evidence in favor of the null hypothesis should be regarded as a drawback is open to debate. Some observers see asymmetry as desirable or necessary. Tukey (1960, p. 426) proposes that a basic question in scientific research is whether or not we *can* reach a conclusion, and that in this sense "asymmetry can be essential." A nonsignificant test statistic tells us that we cannot reach a conclusion about whether a parameter is positive, negative, or zero. Popper (1959) offers a different argument for asymmetry: he holds that, as a matter of principle, scientific theories can never be confirmed, only refuted. However, in a conventional hypothesis test, the alternative hypothesis cannot be refuted: even if the parameter estimate is exactly zero, that result is compatible with the proposition that it has a small positive or negative value. Therefore, Popper's approach implies that the null hypothesis should represent the theoretical prediction.

Other observers see scientific research as a contest among competing explanations (e.g., Chamberlin 1890/1965; Anderson, 2008; Burnham and Anderson, 2002). This view implies that a test should be able to provide evidence in favor of one alternative over the others. If a test gave the relative odds of two models in light of the evidence, then a ratio of 1:1 is clearly the dividing line. Observers might disagree about whether a given ratio, say 3:1, should be regarded as weak, moderate, or strong evidence, but they could not disagree about the direction of the evidence. Thus, in contrast to conventional hypothesis tests, the standards for evaluating the test would not be completely arbitrary. Informally, researchers using conventional tests sometimes interpret large *p*-values as evidence in favor of the null hypothesis, but the choice of any specific level above which the *p*-value should count as evidence in favor of the null hypothesis is at least as arbitrary as the choice of a level for statistical significance, and no standard convention has been adopted.

From either point of view, using classical tests when theoretical proposi-

tions are represented by the alternative hypothesis means that a theory cannot be refuted, and so it may not be abandoned even after repeated failures to confirm its predictions. A theory that predicts the sign could be refuted by statistically significant estimates with the "wrong" sign, but as Merton (1945, p. 464) points out, many "theories" in the social sciences "indicate types of variables which are somehow to be taken into account rather than specifying determinate relationships between particular variables." In essence, they are claims that a certain class of factors is an important influence on an outcome. General claims of this kind can often be interpreted to accommodate parameter estimates with either sign, so that they can never be refuted by classical hypothesis tests (see Jackson and Curtis, 1972, for discussion of an example). In fact, the possibility of being among the first to find empirical support for a well-known theory will give researchers an incentive to keep trying, and given enough attempts, a few are likely to produce "significant" results simply by chance. As a result, a theory may endure in a sort of limbo—neither strongly supported nor clearly refuted.

2.3.8 Likelihood Principle

A *p*-value represents the probability that a test statistic would be greater than or equal to the observed value if the null hypothesis were true.[4] That is, it takes account of the probability of events that did not occur—values of the test statistic greater than the one that was observed. Some observers regard this as illogical: they hold that conclusions should be based only on the probability of the event that actually occurred. As Jeffreys (1961, p. 385) put it in a widely cited remark: "What the use of P[-values] implies . . . is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. . . . On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it." Similar passages appear in several of his other writings (Jeffreys, 1938, p. 194; 1980, p. 453), suggesting that he considered it to be an important point.

This idea that conclusions should be based only on the observed data has come to be known as the "principle of likelihood" (Savage et al., 1962,

[4]With a two-tailed *t*-test, the test statistic should be understood as $|t|$.

p. 17). In a large sample, the probability that a test statistic will have *any* specific value approaches zero, so the principle of likelihood implies that a test must be based on a comparison of models. The principle of likelihood raises complex issues, and is not universally accepted. However, it has an obvious intuitive appeal, so the fact that conventional hypothesis tests violate it is cause for concern.

2.4 IMPLICATIONS OF THE CRITICISMS

In recent years, many of the criticisms of conventional hypothesis testing have coalesced around a central point: that the conventional .05 standard is too weak, and consequently a large fraction of "statistically significant" results are spurious. If R is the ratio of true to false null hypotheses, then the odds that a test statistic that is significant at the .05 level represents a real relationship cannot be greater than 20:R. For example, suppose that, out of 1,000 null hypotheses, 100 are false and 900 are true, so that $R = 9$. We can expect to reject 45 ($.05 \times 900$) true null hypotheses. Even if all 100 of the false null hypotheses are rejected, about 30% (45 of 145) of the rejected null hypotheses will actually be true. Realistically, we will not always reject the null hypothesis when it is false, so the proportion of spurious "findings" will be higher. If statistically significant results are more likely to be published, and investigators have some flexibility to select the most impressive results (e.g., reporting only specifications in which the parameter of interest is statistically significant), the proportion will be higher still. Ioannidis (2005; see also Young and Kerr, 2011) presents a model of the research and publication process in which *most* published findings of statistically significant relationships are spurious. These concerns have been particularly prominent in medical research and in social psychology, where it has become common to speak of a "crisis of replication," but researchers in other fields—for example, Hauser (1995) in sociology—have expressed similar concerns.

If this diagnosis is accepted, the obvious response would be to raise the bar for statistical significance; for example, to use the .01 level as the standard. However, as discussed in Section 2.3.5, there is no obvious way to decide on a specific alternative, so any choice will be somewhat arbitrary and open to objection. In addition, one could argue that the standard should

depend on sample size, but as discussed in Section 2.3.6 there is no accepted rule for making an adjustment. Finally, because conventional hypothesis tests cannot produce evidence for the null hypothesis, it would still not be possible to dispose of ideas that had not been supported. For these reasons, some observers hold that the problem goes beyond the .05 standard and involves the basic logic of conventional hypothesis tests.

Another criticism of standard hypothesis testing is that it has a detrimental effect on the development of theory. If a theory predicts that a parameter has a particular sign, a significant parameter estimate with the expected sign can be claimed as favorable evidence. As a result, researchers have no incentive to develop theories that offer predictions about the exact value of a parameter; in fact, they have an incentive *not* to, since offering a precise prediction would make it easier for the theory to be refuted. The result is a proliferation of vaguely specified theories (Meehl, 1978).

This line of criticism could be seen as directed at the state of theorizing in the social sciences rather than at the procedure of testing. If a theory offered a precise prediction about the magnitude of a parameter, that prediction could be treated as the null hypothesis (see Section 2.3.7). Accepting the null hypothesis would mean that the theory had survived a challenge, and rejecting it would suggest that the theory needed to be revised. Therefore, one response is that theories should go beyond predicting the direction of association and offer more precise claims about magnitude or functional form. As Tukey (1969/1986, p. 728) put it, if "elasticity had been confined to 'When you pull on it, it gets longer!,' Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared." However, few theories in the social sciences offer precise predictions or appear to have much potential to be developed to the point of offering precise predictions in the near future, leaving open the question of how best to evaluate the kind of theoretically based hypotheses that we actually have.

2.5 ALTERNATIVES TO CONVENTIONAL TESTS

Several alternatives to conventional hypothesis tests have been proposed, but two are particularly popular. Both are "penalized likelihood criteria" of the form:

$$D + p\,f(N)$$

where p is the number of parameters estimated in the model and $f(N)$ is a nondecreasing function of sample size. Statistics of this kind impose a penalty for increased complexity, which is understood as the number of parameters in the model. Unlike conventional hypothesis tests, penalized likelihood criteria provide a complete ranking of all of the models that are applied to a given set of data. Akaike (1974/1998) proposed making $f(N) = 2$, yielding the AIC. Schwartz (1978) proposed making $f(N) = \log(N)$, yielding the BIC. A number of other model selection criteria have been proposed, but most of these are asymptotically equivalent to either the AIC or the BIC (Teräsvirta and Mellin, 1986).

As the name suggests, the BIC is based on Bayesian principles (although not all Bayesians accept it, for reasons that will be discussed in Chapter 4). The general idea of the Bayesian approach to hypothesis testing is to obtain a statistic representing the weight of evidence in favor of one model over the other, known as the "Bayes factor." Although many observers find the idea of a statistic representing the weight of evidence appealing in principle, direct calculation of Bayes factors is usually difficult. The attraction of the BIC is that it offers a simple way of calculating an approximate Bayes factor. If the BIC for Model 1 is $B1$ and the BIC for Model 2 is $B2$, then the Bayes factor favoring Model 2 over Model 1 is $e^{(B1-B2)/2}$. For example, if $B1$ is 100 and $B2$ is 90, then the corresponding Bayes factor is $e^5 \approx 148$. This means that the observed data are 148 times more likely under the assumption that Model 2 is true. A Bayes factor is distinct from the prior probabilities of the models: two observers who disagreed about whether Model 2 was plausible in principle might nevertheless agree on the Bayes factor.

There is no single Bayes factor; instead, different definitions of the hypotheses will lead to different Bayes factors. This point will be discussed in detail in Chapter 4, but in this chapter attention will be confined to the BIC, which is much more widely used than other Bayesian tests. This is not simply because it is easy to calculate; advocates argue that the BIC can be justified as a default choice (Kass and Raftery, 1995).

The BIC and AIC imply very different standards for the inclusion of extra parameters. The AIC implies that a parameter should be included if the absolute value of its t-ratio is greater than $\sqrt{2}$, which is equivalent to using a

p-value of about .15 in a two-tailed test. As a result, if parameters are added or removed one at a time, more complex models will be chosen when the AIC is used as a criterion than when conventional hypothesis tests are used. The BIC implies that a parameter should be included if the absolute value of the *t*-ratio is greater than $\sqrt{\log(N)}$. If $N = 100$, this is about 2.15, slightly higher than the conventional critical value for a two-tailed test at the .05 level. If $N = 1{,}000$, the value is 2.63, and if $N = 10{,}000$, it is 3.03. As a result, the BIC tends to favor simpler models than conventional hypothesis tests or the AIC, especially when the sample size is large. Moreover, unlike conventional hypothesis tests, it can provide evidence in favor of the null hypothesis. This is part of the attraction of the BIC for some observers: by raising the bar for statistical significance and allowing for evidence in favor of the null hypothesis, it seems to offer a solution to the "crisis of replication" (Hauser, 1995; Wagenmakers, 2007).

For both the AIC and the BIC, the "break-even" value increases in proportion to the degrees of freedom. In a conventional chi-square test, the critical value increases at a decreasing rate: for example, the 5% critical values for 1, 5, and 10 degrees of freedom are 3.84, 11.1, and 18.3, respectively. Therefore, as the number of degrees of freedom in a test increases, the α corresponding to the AIC break-even point declines toward zero. For example, in a test with 7 degrees of freedom, the break-even value for the AIC will be approximately equal to the .05 critical value; in a test with 16 degrees of freedom, it will be approximately equal to the .01 critical value. The "conservatism" of the BIC relative to standard hypothesis tests is enhanced in tests with multiple degrees of freedom. As a result, model selection using conventional hypothesis tests, the AIC, and the BIC can lead to very different conclusions. The examples considered in the next section will illustrate these differences.

2.6 EXAMPLES

This section will consider the application of conventional hypothesis tests, the AIC, and the BIC to four examples. No small group of examples can adequately represent the whole range of research in the social sciences, but these datasets are quite diverse and use hypothesis tests for a number of different purposes.

2.6.1 Economic Growth

The first example involves the selection of control variables through the use of data on economic growth in 88 nations between 1960 and 1996 compiled by Sala-i-Martin, Doppelhofer, and Miller (2004). The dataset includes measurements of 67 variables that have been suggested as influencing economic growth. The goal of their research was to give an overview, ranking each variable in terms of the strength of evidence that it had some effect on growth. However, for this example, I will focus on one variable—"ethnolinguistic fractionalization," or the probability that two randomly selected members of the nation will not speak the same language—and treat the others as merely potential control variables. Some observers have argued that fractionalization reduces economic growth, either by leading to political unrest or by making it more difficult to organize economic activity (Alesina, Devleeschauwer, Easterly, Kurlat, and Wacziarg, 2003). The values of the fractionalization measure range from zero in South Korea to 0.89 in Tanzania. The dependent variable is average annual growth in per-capita GDP, which ranges from –3.18% (Zaire) to 6.91% (Singapore).

Table 2.2 shows selected statistics from seven models.[5] The first model includes all potential control variables, while the others use stepwise regression to decide on the control variables. Models 2–4 begin with only the measure of fractionalization and use levels of .05 (Model 2), .10 (Model 3), and .15 (Model 4) to add or remove control variables. Models 5–7, like Models 2–4, use standards of .05, .10, and .15, respectively, but they begin by including all independent variables. The estimated effects of fractionalization differ widely among the models, ranging from 0.30 in Model 7 to –1.14 in Model 2. The estimate in Model 2 is statistically significant by conventional standards ($p = .012$), while the estimate in Model 3 is in the range that is sometimes treated as significant ($p = .055$). Given the range of the independent and dependent variables, an estimate of –1.14 is certainly large enough to be of interest: it would imply that the difference in fractionalization between Korea

[5]The figures for the AIC and BIC given here are $N \log(SSE) + 2p$ and $N \log(SSE) + p \log(N)$. The likelihood of a model with normally distributed errors depends on the variance of the error distribution, which is unknown and has to be estimated from the data. This formula uses the maximum likelihood estimate of the variance—a formula using the unbiased estimate would give somewhat different figures.

TABLE 2.2. Estimated Effects of Fractionalization on Economic Growth

Model	Estimate	Standard error	*p*	*df*	R^2	*MSE*	AIC	BIC
1	–0.02	1.41	67	20	.913	1.362	–779.7	–613.7
2	–1.14*	0.44	10	77	.788	0.868	–814.7	–789.9
3	–0.90	0.46	11	76	.792	0.862	–814.5	–787.2
4	–0.02	0.50	17	70	.837	0.734	–823.9	–781.7
5	–0.24	0.47	13	74	.824	0.749	–825.2	**–793.0**
6	0.19	0.49	15	72	.839	0.702	**–829.3**	–792.1
7	0.30	0.56	27	60	.876	0.650	–828.1	–761.2

Note. Bold type indicates the best fitting model according to that criterion.
*Indicates that the estimate is significantly different from zero at the .05 level.

and Kenya produced a difference of about 1% in annual growth rates, which would come to about 50% over the whole period. However, the estimates in the other models are not statistically significant—the *t*-ratios are all well under 1.0. Thus, conclusions about the relationship, or absence of a relationship, between fractionalization and growth depend on the choice of control variables.

The AIC chooses Model 6 as the best, while the BIC chooses Model 5. However, the difference between the BIC statistics for Models 5 and 6 is only 0.9, implying odds of only about 1.5:1 in favor of Model 5 over Model 6. When using conventional hypothesis tests, the decision depends on the prior choice of a standard of significance. Although .05 is the usual standard, it is possible to make arguments for a less stringent level. First, the sample is relatively small, so it may be difficult to reject a false null hypothesis. Second, if the goal is to estimate the effect of fractionalization on economic growth, the potential of bias from omitting a relevant variable might be regarded as more serious than the loss of efficiency from including an irrelevant one. However, neither of these arguments is definitive, so individual judgment plays a role.

A second issue in the standard approach is how to choose between models using the same level of significance but different starting points. For example, if we adopt a standard of .05, then it is necessary to decide between Models 2 and 5. The models are not nested, but it is possible to fit a model that includes all independent variables that appear in either model. With respect to Model 2, the hypothesis that the coefficients of the additional variables in the larger model are all zero can be rejected ($p = .004$); with respect

to Model 2, the hypothesis that all of the additional coefficients are zero cannot be rejected ($p = .12$). Therefore, Model 5 could ultimately be chosen over Model 2.

In a general sense, all three approaches yield the same conclusion: the data do not provide clear evidence for the claim that ethnolinguistic fractionalization influences economic growth. However, there is a difference in the conclusions suggested by the BIC and those of classical tests. Using conventional hypothesis tests, the ultimate conclusion is that the data do not contain enough information to tell us much about the relationship between fractionalization and economic growth. In Model 5, the confidence interval for the coefficient ranges from –1.16 to +0.68; that is, it ranges from a substantial negative effect to a substantial positive effect. The BIC, in contrast, leads to a more definite conclusion. The difference in BIC values between Model 5 and a model omitting ethnolinguistic fractionalization can be computed by using the t-ratio and sample size: $\left(\frac{.24}{.47}\right)^2 - \log(88) = -4.2$. This difference gives odds of $e^{(4.2/2)} = 8.2$ in favor of the model that omits fractionalization. That is, the BIC implies that we have fairly strong evidence in favor of the null hypothesis that ethnolinguistic fragmentation has *no effect whatsoever* on economic growth.

An important point that applies to all three methods of model selection is that conclusions depend on the models that are considered. For example, if we limit attention to the three models that were obtained by starting from the regression of growth on fractionalization (that is, Models 2–4), then the BIC would choose Model 2. The t-ratio for ethnolinguistic fractionalization in that model is 2.59, implying odds of about 3:1 in favor of the proposition that fractionalization affects economic growth.[6] This point means that when it is not practical to estimate every possible model it is necessary to think about the strategy of searching for models.

2.6.2 Development and Fertility

The second example also relates to a comparison of nations, but it involves only two variables, the total fertility rate and scores on the Human Development Index (HDI), which is a combination of measures of health, education, and per-capita GDP. In 2005, the HDI ranged from 0.3 to 0.97, with a mean

[6]The break-even t-value for 88 cases is 2.11.

of 0.71. It has long been known that higher levels of development go with lower levels of fertility. Myrskylä, Kohler, and Billari (2009), however, argue that this relationship is not monotonic—that, after development exceeds a certain point, further increases in development are associated with increases in fertility. This hypothesis goes beyond saying that the relationship between development and fertility is nonlinear: it holds that there is a reversal of direction that occurs within the range of development found among nations today. At the same time, it does not imply a precise model for the relationship. Thus, the question is one of primary structure; that is, choosing a model in order to estimate theoretically meaningful parameters.

Table 2.3 shows fit statistics from several models of the relationship. The first is a linear regression, which can be taken as a baseline. The second is the spline model proposed by Myrskylä et al. (2009), in which there are two distinct linear relationships, one that holds when the HDI is less than or equal to 0.85, another that applies when the HDI increases beyond 0.85. Because the value of 0.85 was chosen after examination of the data, it should be regarded as another parameter rather than as a constant. The model therefore involves three parameters (excluding the intercept): the two slopes and the point dividing the range in which each slope applies. The third and fourth models are polynomials including powers of the HDI. Model 5 is a nonlinear regression given by

$$y = \alpha + \beta_1 x + \beta_2 x^{\gamma} + e;\ e \sim N(0, \sigma^2)$$

It is clear that the relationship between fertility and development is nonlinear: all criteria favor at least one of the nonlinear models over the linear

TABLE 2.3. Models for the Relationship betweeen Development and Fertility

Model	Form	R^2	p	df	MSE	AIC	BIC
1	Linear	.783	1	138	0.585	614.36	620.22
2	Spline (.85)	.810	3	136	0.519	599.81	611.54
3	Quadratic	.807	2	137	0.524	600.14	**608.94**
4	Cubic	.810	3	136	0.519	599.80	611.54
5	$\alpha + \beta_1 x + \beta_2 x^{\gamma}$	.811	3	136	0.515	**598.73**	610.46

Note. Bold type indicates the best fitting model according to that criterion.

regression. However, the lowest value of the AIC occurs for the nonlinear regression (Model 5), while the lowest value of the BIC occurs for the quadratic polynomial (Model 3). More precisely, according to the BIC the evidence favors Model 3 over Model 5 by exp[(610.46 – 608.94)/2] = 2.1. These models have very different implications on the point of theoretical interest. In the quadratic model, the relationship between development and fertility is negative over the entire range of HDI values. At HDI = 0.96, the estimated value of $\partial y \partial x$ is about –3. In Model 5, the relationship changes direction for HDI values above 0.915, and $\partial y \partial x$ is equal to about 0.9 at HDI = 0.96. Thus, conclusions about the Myrskylä, Kohler, and Billari hypotheses differ, depending on whether one uses the AIC or BIC.

In the classical approach, the linear model can be rejected against the quadratic model. The quadratic model cannot be rejected against the cubic model at conventional levels of significance (the *t*-ratio for the cubed term is 1.51, giving a *p*-value of .13). The quadratic model is a special case of Model 5, implying the restriction $\gamma = 2$. An *F*-test involving Models 3 and 5 gives a value of 3.39 (1, 136 *df*), which has a *p*-value of .068. This is in the gray area that might be treated as weak support for Model 5 or as grounds for accepting the quadratic model. The spline model cannot easily be tested against any of the others, but it has a larger mean square error than Model 5 and the same number of degrees of freedom, so there is no reason to prefer it.

To summarize, if the AIC is used as the model selection criterion, the results support the proposition that the effects of development on fertility change direction; if conventional hypothesis tests are used, they are ambiguous; and if the BIC is used, they count against it. None of the approaches can be interpreted as strongly favoring one model over the others, so it could be said that all agree in the sense of suggesting that more evidence is needed before we can offer a firm conclusion. Nevertheless, they lead to different conclusions about the most fundamental point—the direction of the evidence provided by the data.

2.6.3 Comparative Social Mobility

The third example involves social mobility in the United States and Britain (Long and Ferrie, 2013). The data are tables of father's occupation by own occupation for samples of 19th-century British men, 20th-century British

men, 19th-century American men, and 20th-century American men. Occupation is classified into five categories: higher white-collar, lower white-collar, farmer, skilled and semiskilled manual, and unskilled manual.[7] The sample includes 9,304 men. The table can be analyzed using what Erikson and Goldthorpe (1992; see also Xie, 1992) call a "uniform difference" model:

$$\log(\hat{n}_{ijk}) = \beta_{ik} + \beta_{jk} + \phi_k \gamma_{ij} \tag{2.5}$$

where i represents father's occupation, j represents own occupation, k represents the combination of nation and time, and n is the number of cases that have a given combination of characteristics. For example, n_{111} is the number of 19th-century American men with professional jobs whose fathers also had professional jobs. The observed values of n are assumed to follow a Poisson distribution with parameter $\hat{n}$. The β parameters represent the marginal totals of men in different occupations at each time and place. The association between occupations is represented by the product of the γ_{ij} parameters, which represent the pattern of assocation between different occupations, and the ϕ_k parameters, which represent the strength of the association.

An appealing feature of this model is that it makes it possible to describe differences in mobility in simple terms: a higher value of ϕ means a stronger association, or less mobility. It is possible that the model will not hold; that the pattern of association will differ in more complex ways over times or places. In terms of the classification in Section 2.2, this means that one of the purposes of model selection in this example is to decide on primary structure.

Table 2.4 shows fit statistics from several models. The column headed "*df*" gives the number of degrees of freedom remaining after fitting the model; the number of parameters in each model is 100 – *df*. Model 3 represents the model in Equation 2.5: a common pattern but different amounts of mobility. The estimated values of ϕ are 1.48 for 19th-century Britain, 1.26 for 20th-century Britain, 0.61 for 19th-century America, and 0.90 for 20th-century America. The basic conclusion is that the gap between the nations

[7]Long and Ferrie considered several different classifications. The data used here are taken from Long and Ferrie (2008, Tables A-2-1 and A-2-2).

TABLE 2.4. Fit Statistics for Models of Social Mobility in the United States and Britain

Model	Association	Deviance	*df*	AIC	BIC
1	No association	1,747.3	64	1,819.3	2,076.3
2	No differences in association	268.9	48	372.9	744.1
3	Different amounts	108.1	45	218.1	**610.7**
4	Amount by nation, pattern by time	48.1	30	**188.1**	687.8
5	Amount by time, pattern by nation	69.0	30	209.0	708.7
6	Different patterns	0	0	200.0	913.8

Note. Bold type indicates the best fitting model according to that criterion.

became smaller in the 20th century, mostly because of increased inheritance of status in the United States. This is the best fitting model according to the BIC.

The AIC, however, favors Model 4, in which the *pattern* of mobility differs between the 19th and 20th centuries. In terms of Equation 2.5, this means that there are two sets of γ parameters, one for the 19th century and one for the 20th. Within each century, it is possible to compare the nations: as with Model 4, social mobility is greater in the United States in both the 19th and 20th centuries, but the national differences are smaller in the 20th century. However, this model does not allow one to say that social mobility increased or decreased over time; all one can say is that the pattern of mobility changed. Using conventional hypothesis tests, only Model 6 can be accepted at the .05 level of significance. This is a "saturated" model, fitting one parameter for every data cell, so that it reproduces the data perfectly but leaves no degrees of freedom. In this example, it can be understood to mean that each nation at each time has a qualitatively different pattern of social mobility. The *p*-value for Model 4 is about .02, so some investigators might argue in favor of accepting it, particularly given the large sample size. Model 3 can be rejected at any reasonable level of significance against both Models 4 and 6.

In the first two examples, the BIC and classical tests agreed that there was room for doubt about which model should be preferred. In this case, however, the conclusions are completely different. According to the BIC, the odds in favor of Model 3 against Model 6 are $e^{\frac{913.8-610.7}{2}}$, or about 10^{64}, although the *p*-value of Model 3 is about .00000004. The odds in favor of Model 3 against Model 4 are also very strong, about 10^{17}. As a result, the BIC implies that, if one of these models is correct, it is almost certainly Model 3.

The difference reflects two features of the data: first, the sample is large, and second, the tests involve multiple degrees of freedom. Both of these factors increase the divergence between the BIC and classical tests. In a sample of 139 cases, like the fertility example, an additional parameter must reduce the deviance by at least 4.93 in order to reduce the BIC; in a sample of 9,304, it must reduce the deviance by at least 9.14. For tests of a single parameter, these figures correspond to *t*-ratios of about 2.2 and 3.0, or *p*-values of about .03 and .004. The differences grow as the number of degrees of freedom increase: with a sample of 9,304, a model including three extra parameters must reduce the deviance by at least 27.4 to reduce the BIC, which corresponds to a *p*-value of about 5×10^{-6}.

Putting statistical issues aside, this example illustrates a feature of the BIC that many researchers find appealing (Hauser, 1995; Xie, 1999). In some cases, the BIC favors a model with a straightforward interpretation, while classical tests lead to a much more complex model. In this case, Model 3 permits a direct answer to questions about whether there is more or less mobility in different societies; Model 6 tells us that there are differences in the patterns of mobility but does not give any guidance on how they should be described.

2.6.4 Race and Voting Choices

The final example involves the 2004 U.S. presidential election, using data from the Edison–Mitofsky election day exit poll. The models are binary logistic regressions of the choice between the two major candidates, George W. Bush and John Kerry. In contrast to the first three examples, which all focus on specific hypotheses, this example involves an inductive or exploratory analysis. The general question is whether there are interaction effects involving race.[8] The survey contains information on a number of other potential influences on vote: age, family income, gender, marital status, state of residence, and size of community. More information on the variables is provided in Table 2.5. Although the exit poll does not contain as many demographic

[8]The race variable is a dichotomous division between blacks and all others. For convenience, I will sometimes refer to the nonblack group as "whites."

TABLE 2.5. Description of Variables, 2004 Exit Poll

Variable	Values
Categorical variables	
Race	Black Nonblack
Community	City over 500,000 City 50,000–499,999 Suburb City 10,000–49,999 Rural
Sex	Male Female
Marital status	Married Not married
State	
Covariates	
Family income	1 = under $15,000 . . . 8 = $200,000 or more
Age	1 = 18–24 . . . 9 = 75 and over

variables as some other election surveys, its large sample size provides more power to detect any interactions that might exist.

The baseline model includes the main effects of all variables but no interaction effects. Adding interactions between race and all of the other independent variables adds 47 parameters and reduces the deviance by 166.0. The 5% critical value for a chi-square distribution with 47 degrees of freedom is 64, and the 0.1% critical value is 82.7, so the hypothesis that the effects of all variables are the same among blacks and whites can be definitively rejected.

The next step in an investigation using conventional tests is to consider intermediate models in which there are interactions involving some subset of the independent variables. Table 2.6 shows the relevant parameter estimates and standard errors from the model including interactions (estimates involving state differences are omitted to save space).

The estimates for income and marital status are almost the same among blacks and whites, and the null hypothesis can obviously be accepted. This conclusion, however, does not mean that there is positive evidence that the effects are the same or even that any differences are "small." Examination of the standard errors shows that there is a wide range of uncertainty: it is cer-

tainly possible that the effects are the same, but it is also possible that they are substantially larger or smaller among blacks—that is, there is not enough evidence to make any strong claim one way or the other. The BIC, however, gives very different conclusions: odds of about 100:1 in favor of the hypothesis of no difference.

The estimated effects of gender are considerably smaller for blacks than for whites, but the hypothesis of no difference cannot be rejected using conventional tests—the *t*-ratio is about 1. The BIC finds strong evidence in favor of the hypothesis of no difference (odds of 70:1).

Table 2.7 shows fit statistics from a number of models. Models 3–5 consider state, community, and age interactions assuming that the effects of income, gender, and marital status do not differ by race. The difference in deviance between Models 3 and 4 is 2.8, which is in between the 5% and 10%

TABLE 2.6. Selected Parameter Estimates and Standard Errors from Model 2

	Nonblack	Black	Difference
Male	.258***	.061	.197
	(.042)	(.195)	(.199)
Age	–.020*	–.100*	.080
	(.010)	(.051)	(.052)
Income	.094***	.096	–.002
	(.013)	(.064)	(.065)
Married	.490***	.473*	.017
	(.047)	(.212)	(.217)
Community			
Large city	–.500***	–1.659**	
	(.096)	(.576)	
City	–.283***	.119	
	(.082)	(.433)	
Suburbs	–.146*	–.186	
	(.071)	(.438)	
Town	.067	–.347	
	(.103)	(.393)	
Rural	.000	.000	

Note. Standard errors in parentheses. Statistical significance is indicated by * = .05; ** = .01; *** = .001.

TABLE 2.7. Fit Statistics for Models of Democratic versus Republican Vote, Interactions with Race

Model	Association	Deviance	*df*	AIC	BIC
1	None	13,902.9	56	14,014.9	14,428.2
2	State, community, income, age, sex, married	13,736.9	103	13,942.9	14,703.2
3	State, community, age	13,738.0	100	**13,938.0**	14,676.1
4	State, community	13,740.8	99	13,938.8	14,669.5
5	State	13,752.8	95	13,942.8	14,644.0
6	State (nonblack only)	13,843.2	56	13,955.2	**14,368.5**

Note. N = 11,862 for all models. Bold type indicates the best fitting model according to that criterion.

critical values. The difference in deviance between Models 4 and 5 is 12.0, which falls in between the 5% and 1% critical values for a chi-square distribution with four degrees of freedom. The difference in deviance between Models 1 and 5 is about 150 with 39 degrees of freedom, which is well beyond the 0.1% critical value. Thus, a researcher using a 10% level would choose Model 3, a researcher using a 5% level would choose Model 4, and a researcher using a 1% level would choose Model 5. However, most researchers do not strictly follow a single significance level, so the usual conclusion would be that there is weak evidence that the effects of age differ by race, reasonably strong evidence that the effects of community differ, and strong evidence that the effects of state differ.

The lowest value of the AIC occurs for Model 3, followed closely by Model 4. Thus, researchers using the AIC and standard hypothesis tests would come to similar conclusions. The BIC, however, strongly favors a model including no interactions (Model 1) over Models 2–5.

The final alternative, Model 6, holds that there are differences by state in the voting choices of nonblacks, but not among blacks. The relationships among Models 1, 5, and 6 can be understood by considering β_j and γ_j, the set of parameters representing the effects of state among blacks and nonblacks. Model 5 imposes no restriction on the values of these parameters. Models 1 and 6 are both nested in this model: Model 1 imposes the restriction $\beta_j = \gamma_j$ for all j, while Model 6 imposes the restriction that $\beta_j = 0$ for all j. As seen previously, Model 6 can be rejected against Model 5: they differ by 90.4 in deviance and 39 in degrees of freedom, and the 0.1% critical value for a chi-

square distribution with 39 degrees of freedom is 72.0. The BIC, however, strongly favors Model 6 over Model 5.

This example, like the preceding one, illustrates the tendency of the BIC to favor simpler models than conventional hypothesis tests and the AIC, particularly when the number of cases is large. The example also illustrates a more subtle point—the ambiguity in the meaning of "a simple model." In terms of the number of parameters, Models 2 and 6 are equally complex. However, Model 2 contains only main effects, while Model 6 involves an interaction between state and race, so in some sense Model 2 might be regarded as simpler. Investigators analyzing this kind of data often do not even consider possibilities like Model 6.

2.7 SUMMARY AND CONCLUSIONS

This chapter has described the various purposes for which hypothesis tests are used and the major criticisms that have been made against that practice. It introduced two alternatives to conventional tests, the AIC and the BIC, and illustrated their application to four examples. The examples show that the different methods of model selection can lead to very different conclusions. This difference was particularly evident in the social mobility example: the BIC strongly supported the hypothesis that all four cases had the same pattern of mobility, while conventional tests strongly rejected that hypothesis. Even in cases where the methods favored the same model, their implications could differ in important ways. In the economic growth example, the BIC implies fairly strong evidence against the proposition that ethnolinguistic fractionalization affects growth, while a classical test merely says that we cannot rule out the possibility of no effect.

As a result, the strategy of using both the BIC and conventional hypothesis tests, as advocated by Xie (1999), is not viable. In many cases, when two alternative statistics have been proposed, it is informative to consider both. To take a simple example, the mean and median are both measures of central tendency; if there is a substantial difference between them, that tells us something about the distribution of the variable. However, the AIC, the BIC, and classical tests do not provide different information about the fit of the model;

rather, they all involve a comparison of deviance to degrees of freedom and simply propose different standards for interpreting the same information. Therefore, it is necessary to examine the rationale for each criterion in detail and to decide which is most persuasive.

RECOMMENDED READING

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1000. —One of the best known of recent critiques of conventional significance testing.

Cox, D. R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology, 14*, 325–331. —Gives a clear account of conventional significance tests and argues that they are useful in scientific research.

Giere, R. N. (1972). The significance test controversy. *British Journal for the Philosophy of Science, 23*, 170–181. —A review of Morrison and Henkel (1970) that provides a useful guide to the issues debated in that book.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 696–701. —Provides a model of the research and publication process that suggests that a majority of statistically significant "findings" are spurious.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44*, 1372–1381. —Discusses the different uses of hypothesis tests and concludes that "we need a foundational theory that encompasses a variety of inferential goals."

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine. —Collects a variety of articles from the 1940s through the 1960s. The ones by Lipset, Trow, and Coleman; Kendall; Davis; Morrison and Henkel; Lykken; Meehl; and Berkson are particularly informative.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301. —A balanced review of the controversy over significance testing, with many references to the literature.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed., Chap. 16). New York: Dover. —Discusses the question of testing null hypotheses that are almost certainly false.

Wolfowitz, J. (1967/1980). Remarks on the theory of testing hypotheses. In *Selected papers*. New York: Springer-Verlag. (First published in *New York Statistician, 18*, 1–3.) —Criticizes the practice of hypothesis testing; especially noteworthy because Wolfowitz was an important contributor to the theory.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com