respect to $\boldsymbol{\mu}$, the vector $\mathbf{z}_F$ is normal with mean $\bar{\mathbf{z}}^{**}$ and covariance matrix $\boldsymbol{\Sigma}(1+(T^{**})^{-1})$. Integrating with respect to $\boldsymbol{\Sigma}^{-1}$ as in Appendix 2, the distribution for $\mathbf{z}_F$ is Student

$$f(\mathbf{z}_F|\bar{\mathbf{z}},\mathbf{S},T)=f_S^k\left(\mathbf{z}_F|\bar{\mathbf{z}}^{**},\mathbf{S}^{**}\left(1+(T^{**})^{-1}\right)/\nu^{**},\nu^{**}-k+1\right).$$

The results needed for Chapter 6 are the conditional moments of $\mathbf{z}_F^I$ given $\mathbf{z}_F^J$ where the partition of $\mathbf{z}_F$ is $\mathbf{z}_F'=(\mathbf{z}_F^I,\mathbf{z}_F^J)$. Using the diffuse prior assumption that $\mathbf{S}^*=0$, $T^*=0$, $\nu^*=0$, the conditional moments are

$$E(\mathbf{z}_F^I|\mathbf{z}_F^J)=\bar{\mathbf{z}}^I+\mathbf{S}_{IJ}\mathbf{S}_{JJ}^{-1}(\mathbf{z}_F^J-\bar{\mathbf{z}}^J)$$

$$V(\mathbf{z}_F^I|\mathbf{z}_F^J)=(\mathbf{S}_{II}-\mathbf{S}_{IJ}\mathbf{S}_{JJ}^{-1}\mathbf{S}_{JI})(1+T^{-1})(T-k+1+k_J)^{-1}$$

where k_J is the number of elements in $\mathbf{z}_F^J$.

CHAPTER **4**

HYPOTHESES-TESTING SEARCHES

The first variety of specification search that we discuss corresponds to the familiar hypothesis-testing problem. We assume the existence of a set of M "models" or hypotheses of the form

$$H_i:\mathbf{Y}=\mathbf{X}_i\boldsymbol{\beta}_i+\mathbf{u}_i,\qquad i=1,\ldots,M,\qquad (4.1)$$

where $\mathbf{Y}$ is a $(T\times 1)$ vector of observable variables, $\mathbf{X}_i$ is a $(T\times k_i)$ matrix of observable explanatory variables, $\boldsymbol{\beta}_i$ is a $(k_i\times 1)$ vector of parameters, and $\mathbf{u}_i$ is a $(T\times 1)$ vector of unobservable disturbances assumed to be normally distributed with mean zero and variance-covariance matrix $\sigma_i^2\mathbf{I}_T$. The statistical problem is to determine which of these M models did, in fact, generate the data and at the same time to make inferences about the coefficient vectors $\boldsymbol{\beta}_i$.

The formal classical theory of hypothesis testing describes the decision problem of selecting an action from among a set of feasible actions. An action is either right or wrong, depending on the "true state of nature," and the statistician is interested in being wrong as infrequently as possible. The problem being considered in this chapter involves a set of M actions of the form "act as if hypothesis H_i were true" and a set of M states of nature of the form "hypothesis H_i is true." An error occurs when action i is chosen but hypothesis j $(i\neq j)$ is true.

If an action entailed the potential of some specific loss, hypothesis testing should be considered to be solving a decision problem. When the losses are not stated, it is difficult to interpret hypothesis testing as a solution to a decision problem. In fact, most researchers use statements such as "the hypothesis H_i is rejected at the .05 level of significance" to mean that the data cast doubt on H_i. The statement is *not* meant to suggest that because of the data evidence, it is undesirable to act as if H_i were true, regardless of the decision problem. Thus the *language* of hypothesis testing is used to summarize the data evidence. This chapter, therefore, largely ignores the decision-theory problem, but argues that classical hypothesis testing has led to greatly distorted data summaries.

In Chapter 3 we briefly reviewed the classical approach to hypothesis testing with a pair of nested hypotheses of the form

$$H_0: \mathbf{Y}=\mathbf{X}_J\boldsymbol{\beta}_J+\mathbf{u}$$
$$H_1: \mathbf{Y}=\mathbf{X}_I\boldsymbol{\beta}_I+\mathbf{X}_J\boldsymbol{\beta}_J+\mathbf{u}$$
$$=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$$

where $\mathbf{X}_I$ is a $T\times p$ observable matrix, $\mathbf{X}_J$ is a $T\times(k-p)$ observable matrix, $\mathbf{u}$ is a $T\times 1$ unobservable vector, distributed normally with mean zero and covariance $\sigma^2\mathbf{I}_T$, and $\mathbf{X}=[\mathbf{X}_I,\mathbf{X}_J]$, $\boldsymbol{\beta}'=[\boldsymbol{\beta}_I',\boldsymbol{\beta}_J']$. With $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $s^2=(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})/(T-k)$, the statistic

$$F=\frac{\mathbf{b}_I'\left(\mathbf{X}_I'\mathbf{X}_I-\mathbf{X}_I'\mathbf{X}_J(\mathbf{X}_J'\mathbf{X}_J)^{-1}\mathbf{X}_J'\mathbf{X}_I\right)\mathbf{b}_I}{ps^2},$$

conditional on the null hypothesis that $\boldsymbol{\beta}_I=\mathbf{0}$, has an F distribution with p and $T-k$ degrees of freedom. A large F is taken as evidence against the null hypothesis.

The F statistic has been written in Equation (3.10) in terms of error sums of squares as

$$F=\frac{(ESS_0-ESS_1)/p}{ESS_1/(T-k)} \tag{4.2}$$

where ESS_i is the error sum of squares associated with the ith hypothesis. In terms of R^2, it can be written as

$$F=\left(\frac{R_1^2-R_0^2}{1-R_1^2}\right)\left(\frac{T-k}{p}\right). \tag{4.3}$$

Formal testing of the hypothesis $\boldsymbol{\beta}_I=\mathbf{0}$ involves, first, selecting a significance level for the F test, say, α, then finding from a table the α point of the F distribution, and finally recording that the hypothesis is or is not

rejected at the α level, depending on whether F exceeds or falls short of this cutoff point. For example, the .05 point of the F distribution with 10 or more degrees of freedom ($T-k\geqslant 10$) varies from approximately five to one, depending on the actual number of degrees of freedom as well as on the number of restrictions p. Thus for moderate degrees of freedom, F values in excess of 5 would be regarded as evidence against the null hypothesis.

By referring to the foregoing formulas, we can see that although the R^2s of two equations may differ only in the tenth decimal place, an F may attain any arbitrarily large value if the degrees of freedom $T-k$ is large enough. In very large samples such as would be generated by surveys of individuals or firms, it thus turns out that almost any hypothesis of this form is rejected. To paraphrase a quotation of Berkson (1938),[1] since a large sample is presumably more informative than a small one, and since it is apparently the case that we will reject the null hypothesis in a sufficiently large sample, we might as well begin by rejecting the hypothesis and not sample at all.

This brings us to the first question of this chapter:

Problem 1. Is classical hypothesis testing at fixed level of significance a "good" way to summarize the evidence in favor of or against hypotheses of the form described above?

Our answer is decidedly negative—meaningful hypothesis testing requires the significance level to be a decreasing function of sample size. Incidentally, the argument that leads to this conclusion is not the same as Berkson's. He might have pointed out that it is practically certain that any series of real observations does not actually come from a regression process with $\boldsymbol{\beta}_I=\mathbf{0}$. If so, we do, indeed, want to reject the null hypothesis; and it is neither surprising nor unwarranted that a large informative sample leads to the rejection of the hypothesis. One, in those circumstances, should trouble himself not with the results of classical hypothesis testing but rather with the question of why he bothered to test an obviously false hypothesis in the first place. As it turns out, hypothesis testing does

[1] "I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. Having observed this, and on reflection, 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 the order of 200,000—the chi-square P will be small beyond any usual limit of signficance'."

have some validity, even when a restriction is practically certain to be false. Certainly false hypotheses are the subject of the next five chapters. In this chapter we concern ourselves with hypotheses that are not so trivially rejected; we are thus concerned with situations in which classical hypothesis testing has an unambiguously legitimate function. This distinction can be made quite clear from a Bayesian point of view: We are here considering hypotheses that are assigned positive probability; in later chapters we consider hypothesis testing in contexts when some of the hypotheses receive zero probability. The conclusion that the significance level should be a decreasing function of sample size is due largely (but not exclusively) to the fact that Bayesians who assign positive probabilities to hypotheses of this form summarize the evidence in ways that implicitly make the significance level a decreasing function of sample size.

This is a good place to indicate that I doubt that there are many instances when a regression hypothesis that involves a restricted parameter space would, in fact, be assigned positive probability. The things that we call models usually originate in other ways, and I consequently doubt the practical relevance of this chapter. This is, nonetheless, a useful topic to begin with, because it corresponds closely to the situation in which classical hypothesis testing is strictly relevant, and because many people think of specification searches exclusively in those terms.

The hypotheses we have just considered have an exceedingly simple form: the null hypothesis is a restricted version of the alternative. Practical specifications rarely have such a simple structure, and dealing with complicated sturctures of hypotheses is our second problem:

Problem 2. How should multiple hypotheses with a nonnested structure be treated?

I personally find the classical answer to this question hopelessly confusing and would prefer not to get too deeply involved in discussing it. A simple example illustrates some of the problems. Let the hypotheses be

$$H_0: \mathbf{Y}=\mathbf{x}\beta+\mathbf{u} \quad H_1: \mathbf{Y}=\mathbf{z}\gamma+\mathbf{u},$$

where $\mathbf{x}$ and $\mathbf{z}$ are vectors and β and γ are scalars. In this case it is possible to generate a .05-level test of H_0 by testing in the usual way the null hypothesis H_0 against the alternative $H_1^*: \mathbf{Y}=\mathbf{x}\beta+\mathbf{z}\gamma+\mathbf{u}$. Although this is a perfectly well-defined test, it completely ignores the fact that H_1^* is not the alternative hypothesis. It, furthermore, treats asymmetrically two hypotheses, H_0 and H_1, that are apparently symmetric.

Suppose, instead, that the hypotheses are treated symmetrically; in particular, suppose the equation with the higher R^2 is accepted. Then the conditional probability of error is $P, {}_0^2 < R_1^2 | H_0) = P(\mathbf{Y}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y} <$

$\mathbf{Y}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y}|H_0) = P((\mathbf{x}\beta+\mathbf{u})'(\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'-\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z})(\mathbf{x}\beta+\mathbf{u})<0|\beta)$, which depends in a complicated way on β, $\mathbf{x}$, and $\mathbf{z}$. (It can be shown that this probability varies from one-half to zero as β^2 varies from zero to infinity.) Traditionally, the test is set up so that this conditional probability of error is a number, not a function, and it proves difficult to interpret this test in terms of its error probabilities. This should not cause great consternation, since the relationship between the error probabilities of a test and the persuasiveness of the evidence in favor of the various hypotheses is indirect and poorly understood. Of course, this barely scratches the surface of a complex problem. The Bayesian approach yields such a straightforward answer with clear intuitive appeal that it hardly seems worthwhile to pursue further the classical approach. For more on the classical approach see the summaries by Dhrymes et al. (1972) and by Gaver and Geisel (1974).

Anyone who has read any papers in applied econometrics has read statements of the form: "model A has performed the best; it has a high R^2, and all of its coefficients are the right sign and are statistically significant." Whatever is the meaning of the reference to the coefficients? Books on classical statistics do not suggest that the validity of an F test depends on the signs and statistical significance of the coefficients. Perhaps the author of this statement is thinking that there is no restriction that could be placed on this model that would not be rejected, but why the reference to "right" signs? This is a pretty obvious Bayesian problem, in that there is a priori information at least about the signs of the coefficients. This intuitively ought to have an impact on the hypothesis testing. Thus our third problem is:

Problem 3. How does the existence of a priori information about parameters influence the interpretation of evidence about models?

Parallel to the testing of hypotheses, we are interested in making inferences about parameters. There is a presumption that the ambiguity over the model should dilute the information about regression coefficients, since part of the evidence is spent to specify the model:

Problem 4. What estimates of the parameters and what measures of uncertainty should apply in a situation of uncertainty about the model?

Classical inference has little to say about this, although we review in the next chapter the pretesting literature that deals with estimation while testing. Again, the Bayesian solution is entirely straightforward. Among its conclusions are the fact that the ambiguity over the model is irrelevant for

inference about a coefficient if and only if the estimated coefficients and standard errors are the same for all specifications.

The last problem is similar:

Problem 5. What measure of overall confidence analogous to an R^2 should apply to a research effort which reports many different equations with different R^2s? We subsequently propose a special kind of average R^2. Again, there is no classical counterpart.

In the first two sections of this chapter, the problem of identifying the class of admissible tests is distinguished from the problem of selecting a particular test. Classical hypothesis testing concerns itself almost exclusively with the first problem, but it has nothing meaningful to say about the second. The rule of thumb quite popular now, that is, setting the significance level arbitrarily to .05, is shown to be deficient in the sense that from every reasonable viewpoint the significance level should be a decreasing function of sample size.

A few words may now be said in anticipation of the sections to follow, which describe in detail the Bayesian approach to hypothesis testing. By Bayes' rule, the relative posterior probabilities of two hypotheses can be written as

$$\frac{P(H_i|\mathbf{Y})}{P(H_j|\mathbf{Y})} = \left[\frac{P(\mathbf{Y}|H_i)}{P(\mathbf{Y}|H_j)}\right]\left[\frac{P(H_i)}{P(H_j)}\right]. \tag{4.4}$$

The second factor in brackets is the prior odds ratio in favor of H_i. The data-dependent term in the first set of brackets is the "Bayes factor."

The data are said to favor H_i relative to H_j if the Bayes factor exceeds one, that is, if the observed data $\mathbf{Y}$ is more likely under hypothesis H_i than it is under hypothesis H_j. The densities of $\mathbf{Y}$ implied by the hypotheses (4.1) are conditional on the parameters, $\boldsymbol{\beta}_i$ and σ_i^2, but may be straightforwardly "mixed" into a marginal density as

$$f(\mathbf{Y}|H_i) = \int_{\boldsymbol{\beta}_i}\int_{\sigma_i^2} f(\mathbf{Y}|H_i,\boldsymbol{\beta}_i,\sigma_i^2) f(\boldsymbol{\beta}_i,\sigma_i^2)\, d\sigma_i^2\, d\boldsymbol{\beta}_i \tag{4.5}$$

where $f(\boldsymbol{\beta}_i,\sigma_i^2)$ is the prior density. The conditional p.d.f. $f(\mathbf{Y}|H_i,\boldsymbol{\beta}_i,\sigma_i^2)$ for a particular value of $\mathbf{Y}$ is a likelihood function of $(\boldsymbol{\beta}_i,\sigma_i^2)$, and (4.5) defines $f(\mathbf{Y}|H_i)$ as a weighted or marginal likelihood.

The Bayes factor is to be contrasted with the likelihood ratio, which is used classically to summarize the data evidence. The likelihood ratio is

$$L(H_i,H_j) = \frac{\max\limits_{\boldsymbol{\beta}_i,\sigma_i^2} f(\mathbf{Y}|\boldsymbol{\beta}_i,\sigma_i^2,H_i)}{\max\limits_{\boldsymbol{\beta}_j,\sigma_j^2} f(\mathbf{Y}|\boldsymbol{\beta}_j,\sigma_j^2,H_j)}.$$

The Bayes factor averages the likelihood function over all values of $(\boldsymbol{\beta}_i,\sigma_i^2)$. The likelihood ratio evaluates the likelihood function at its maximum.

Any attempt to summarize the data evidence in favor of the hypotheses (4.1) leads to an irreconcilable index number problem of the following form. If $\boldsymbol{\beta}_i$ assumed one value, the data evidence could be said unambiguously to favor the ith hypothesis, but if $\boldsymbol{\beta}_i$ assumed another value, the data unambiguously cast doubt on H_i. Since H_i allows $\boldsymbol{\beta}_i$ to assume any value, the data evidence is necessarily ambiguous.

The classical solution to this dilemma seems most appropriate for testing a point null hypothesis against a composite alternative. The null hypothesis is regarded as the favorite; it is the one that is being "tested." If there is *any* way for the alternative to look as good as the null hypothesis, we should be worried about retaining the null as the favorite. Consequently, we identify the evidence *against* the null hypothesis in terms of the evidence in favor of the alternative at the value of $\boldsymbol{\beta}$ that makes the alternative appear best. The appropriate statement however is not that the alternative is favored. All that is said is that the alternative is *conceivably* favored.

There is a great tendency in practice to forget the all-important word "conceivably" in this sentence, and as a consequence, classical tests distort the data evidence. In the more common case when the null hypothesis is a composite hypothesis, classical tests usually also evaluate the data evidence at the parameter point that makes the null hypothesis appear best. The resulting statement about the evidence is: "If each hypothesis is allowed to 'put its best foot forward,' hypothesis H_j is favored." In practice, the qualifying phrase "if...forward," is often forgotten, and the data evidence may consequently be significantly distorted.

A Bayesian approach, in contrast, presupposes a prior distribution that can be used to weight the evidence at different values of the parameters. Thus instead of letting an hypothesis "put its best foot forward," the performance at all values of the parameters is considered. The apparent problem that then arises is the construction of a nonarbitrary weight function. Here and elsewhere, we take the position that a researcher is obligated to report as fully as possible the mapping of priors into posteriors. He should describe the data evidence as favoring hypothesis H_1 if the prior takes one form, and favoring H_2 if it takes another. He thereby avoids having to make a choice that rightly belongs to his readers: the choice of prior distribution.

4.1. Hypothesis Testing: A Judicial Analogy

The subject of hypothesis testing may be usefully introduced by an analogy. Based on the evidence presented, a judge and jury in a legal

proceeding decide whether a defendant should be set free or sent to jail. If they decide that the evidence favors the hypothesis of guilt, they accordingly send the defendant to jail. Otherwise he is set free. The assumption of innocence until proven guilty beyond a reasonable doubt explicitly favors the hypothesis of innocence. We refer to this favored hypothesis as the *null* hypothesis or H_0 and the hypothesis of guilt as the *alternative* hypothesis or H_1. The evaluation of the evidence and the decision either to free or jail the defendant is called a "test" of the null hypothesis *against* the alternative, and the decision is described as acceptance versus rejection of the hypothesis of innocence.

The more critical error—sending an innocent man to jail—is called an *error of the first kind* or *type I error*. Acceptance of the null hypothesis when it is in fact false—freeing a guilty man—is called an *error of the second kind* or *type II error*. Schematically we have

Actions

Hypotheses (States)		Set Free (accept H_0)	Send to Jail (reject H_0)
H_0:	Innocent		Type I error
H_1:	Guilty	Type II error	

If a man is innocent, we want to have a low probability of sending him to jail. Let this probability be α

$$\alpha = P(\text{jail}|\text{innocent defendant}).$$

Analogously, let

$$\beta = P(\text{set free}|\text{guilty defendant}).$$

Both α and β are defined before the judicial process commences. In effect, they predict the quality of the evidence and the ability of the court to process the evidence effectively. For example, a low value of α amounts to the prediction that if the defendant is innocent, the evidence will be so unambiguous and the process by which a verdict is rendered will be so perfect that with near certainty he will be justly found innocent.

The theory of hypotheses testing deals with defining procedures such that α and β are small. A typical choice set for α and β is depicted in Figure 4.1. Flipping a coin to decide whether to free or jail the suspect implies $\alpha = \beta = .5$. The line running from the point $(\alpha = 1, \beta = 0)$ to $(\alpha = 0, \beta = 1)$ represents the set of all such randomized decisions. The value $(\alpha = 0, \beta = 0)$ represents perfect evidence and a perfect procedure which is excluded

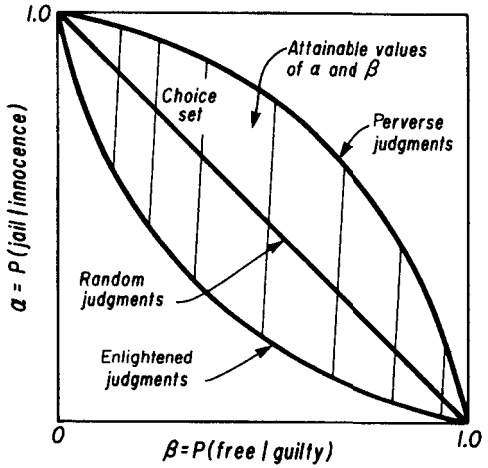

Fig. 4.1 Probabilities of error.

from the choice set in Figure 4.1. If $(\alpha = 0, \beta = 0)$ is not available, the perfect error point $(\alpha = 1, \beta = 1)$ is similarly not available, since if we could be sure of making an error, by doing the exact opposite we could be sure of *not* making an error. The curved line labeled "enlightened judgments" represents the best possible court procedures based on the available evidence. The curve labeled "perverse judgments" is just the mirror image of the "enlightened judgment" curve, involving the exact opposite action.

The choice of a courtroom procedure is usefully thought to involve two steps. The first step is to identify the set of procedures that involve enlightened use of the evidence, that is, those that make α and β as small as possible. The second step is to choose a particular procedure from among this set of admissible procedures. The former is a logical mathematical problem that admits a clear-cut uncontroversial solution; the latter is not. Let us consider the latter problem.

The essential problem the court faces once the line of enlightened judgments is computed is that stricter interpretation of the given body of evidence and a greater tendency to send men to jail which could reduce the probability β of freeing guilty men necessarily increases the probability α of jailing innocent men. By assumption, it is desirable to have both α and β small. The choice dilemma is that reduction of one necessarily leads to an increase in the other. Actual choice can be said to depend on a preference function $U(\alpha, \beta)$ indicating numerically the level of satisfaction attained if the courtroom procedure yields probabilities α and β. Several "contour" lines of a typical preference function are indicated in Figure 4.2.

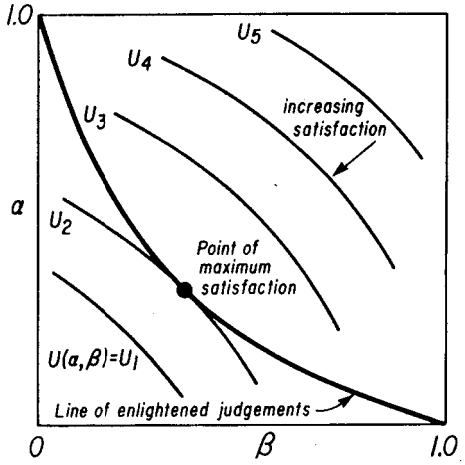

Fig. 4.2 Indifference curves of a preference function.

The line $U(\alpha,\beta)=U_1$ indicates all values of (α,β) that yield a level of satisfaction U_1. Increasing preference (lower α and β) is in the direction indicated, and maximum satisfaction on the enlightened judgment line occurs at the point indicated.

One way to alleviate the choice dilemma is to gather more evidence, thereby making *both* α and β smaller. Without any information, the line of enlightened judgments is just a straight line from corner to corner. As more and more evidence is accumulated, this line shifts in toward the origin. As this occurs, the point of maximum satisfaction also travels in toward the origin tracing out an *information expansion path* depicted in Figure 4.3. This represents the values of α and β that the court would actually choose depending on the amount of evidence that is amassed.

Three possible ways of selecting α and β have been suggested, and their preference functions and information expansion paths are indicated in Figure 4.4.

(1) Set $\alpha=.05$. The most commonly practiced procedure is to set $\alpha=.05$ and minimize β. The type I error is considered more important, and by setting $\alpha=.05$ it necessarily assumes a small value. Note the peculiar information expansion path that allows β to be infinitesimally small with α still at .05.

(2) Minimize the maximum of $l_1\alpha$ and $l_2\beta$, where l_1 is the loss associated with a type I error and l_2 is the loss associated with a type II error. Note that the expansion path moves continuously toward the origin and that an increase in the evidence is used to reduce *both* α and β. The relative probabilities are $\alpha/\beta=l_2/l_1$, and if type I loss l_1 is relatively great, the type I probability α is correspondingly relatively small.

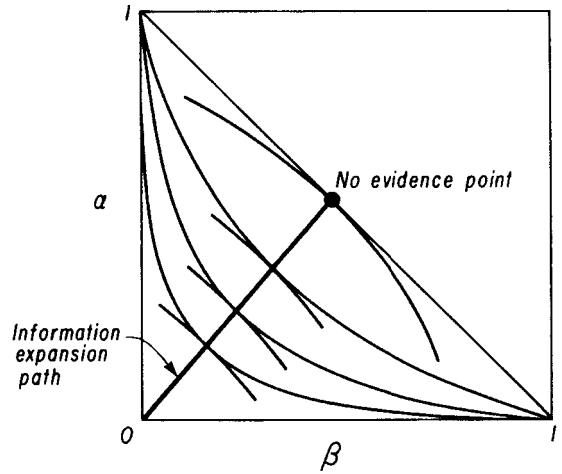

Fig. 4.3 Information expansion path.

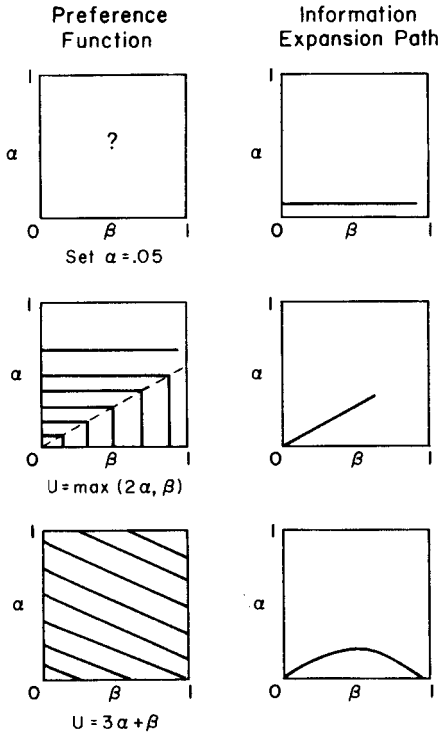

Fig. 4.4 Preference functions and information expansion paths.

(3) Minimize the expected loss. Let p be the probability that the defendant is innocent and $(1-p)$ the probability of guilt. The expected loss is $l_1\alpha p + l_2\beta(1-p)$. The slope of a typical contour or indifference curve depicted in Figure 4.4 is $-l_2(1-p)/l_1 p$. If either l_1 or p is relatively large, these lines are relatively flat, and α is set to a relatively small value. Thus type I error is avoided in the sense of setting α to some small number, if either the defendant is quite likely to be innocent or if the cost of sending an innocent man to jail is high.

It should now be clear that the choice of α and β is less than obvious. A simple argument attributed to Savage does imply that the indifference curves should be straight lines. Identify two points (α_0,β_0) and (α_1,β_1) between which you are indifferent. If restricted to select one of these two points, you must mean that you are willing to allow someone else to make the selection. Well, I will make the selection for you in the following way. Making use of a random device, I select (α_0,β_0) with probability π and (α_1,β_1) with probability $(1-\pi)$. As a result, your type I and type II error probabilities are, in fact, $(\alpha_0\pi+\alpha_1(1-\pi),\beta_0\pi+\beta_1(1-\pi))$, and you have revealed your indifference between this point and the two original points. By varying π, all probability couples on the line joining (α_0,β_0) to (α_1,β_1) can be shown to be on the same indifference curve. Needless to say, Bayesian indifference curves are straight lines.

The hypothesis-testing problem can be described more formally as follows. A sample outcome $\mathbf{z}$ (the testimony) is assumed to come from a sample space Z of all possible samples. The sample space is partitioned into a region of acceptance A and a region of rejection R, where R is the set of all $\mathbf{z}\in Z$ that would lead to rejection of the null hypothesis (jailing the defendant); and A is the complement of R. The corresponding error probabilities are

$$\alpha_R = P(\mathbf{z}\in R|H_0)$$

$$\beta_R = P(\mathbf{z}\in A|H_1).$$

In discussing the theory of hypothesis testing, we may first consider the purely mathematical problem of defining the set of admissible tests; we must be sure that the partition of the sample space into A and R leads to error probabilities on the line of enlightened judgments. Second, we must choose a particular test from among the set of admissible tests. A failure of the theory of classical inference is that it offers no meaningful comment on this second problem. And the rule "set $\alpha=.05$" regardless of sample size seems undesirable under close examination.

4.2 Testing a Point-Null Hypothesis Against a Point Alternative

This section deals briefly with testing a point-null hypothesis against a point alternative. The purpose of this material is to explain how the discussion in the previous section applies to a formal problem. The test depends on a sample of size T, $(Y_1, Y_2, \ldots, Y_T)$ from a normal distribution with mean μ and variance one. The null hypothesis is $H_0:\mu=0$, and the alternative is $H_1:\mu=1$. In terms of the distribution of the mean $\bar{Y}=\Sigma Y_t/T$, the hypotheses are

$$H_0: \bar{Y}\sim N(0,T^{-1})$$

$$H_1: \bar{Y}\sim N(1,T^{-1}).$$

These two distributions are graphed in Figure 4.5 for $T=1$.

Large values of $\bar{Y}$ favor the alternative hypothesis, and a typical decision function is

$$\text{if } \bar{Y}\geqslant c, \text{ reject } H_0$$

$$\text{if } \bar{Y}<c, \text{ accept } H_0$$

where c is some preassigned cutoff point.

The probabilities of error depend on the cutoff point c:

$$\alpha(c)=P(\bar{Y}\geqslant c|\mu=0)$$

$$\beta(c)=P(\bar{Y}\leqslant c|\mu=1).$$

Using a table of normal distribution, we may compute $\alpha(.5)=\beta(.5)=.31$, as depicted in Figure 4.5. If a smaller value of α is desired, c may be increased, say, to 1, which lowers α to .16, but this change in c also

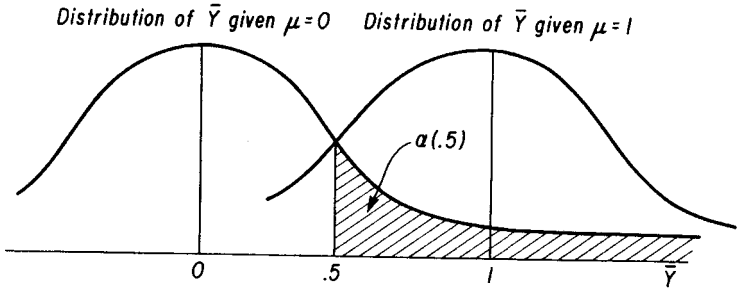

Fig. 4.5 Distributions given null and alternative hypotheses.

increases β to .5. As c is varied from $-\infty$ to $+\infty$, $\alpha(c)$ and $\beta(c)$ thus trace out a line of "enlightened judgments" as in Figure 4.1.

It is also easy to see that the line of enlightened judgments shifts in toward the origin as more evidence is accumulated, that is, as T grows. If T is four instead of one, the standard errors of the distributions become one-half. For $c=.5$, it is easy to calculate that $\alpha=\beta=.16$, compared to .31 before. Similar calculations apply to all values of c.

Traditionally, the cutoff point c is chosen so that $\alpha(c)=.05$. This implies a value of c equal to $1.65/\sqrt{T}$, which shifts toward zero as T increases. This contrasts with a Bayesian cutoff point, which shifts toward .5 as T increases. A Bayesian test selects H_0 if the expected loss from acting as if H_0 were true is less than the expected loss from other action. The probability that H_0 is true given the data $\bar{Y}$ is by Bayes' rule

$$P(H_0|\bar{Y})=\frac{f(\bar{Y}|H_0)P(H_0)}{f(\bar{Y}|H_0)P(H_0)+f(\bar{Y}|H_1)P(H_1)}$$

where $f(\bar{Y}|H_0)=(T/2\pi)^{1/2}\exp[-T\bar{Y}^2/2]$ and $f(\bar{Y}|H_1)=(T/2\pi)^{1/2}\exp[-T(\bar{Y}-1)^2/2]$. The expected loss from proceeding as if H_0 were true is $l_1P(H_1|\bar{Y})$, where l_1 is the loss if H_1 is true and action H_0 is taken; similarly, for H_1. Thus H_0 is to be rejected if

$$l_1P(H_1|\bar{Y})>l_0P(H_0|\bar{Y}),$$

that is, if

$$l_1f(\bar{Y}|H_1)P(H_1)>l_0f(\bar{Y}|H_0)P(H_0),$$

or if $T[\bar{Y}^2-(\bar{Y}-1)^2]>2\log[l_0P(H_0)/l_1P(H_1)]$ or if $2\bar{Y}-1>2T^{-1}\log[l_0P(H_0)/l_1P(H_1)]$. The term on the right-hand side converges to zero as T increases, and the region of rejection becomes $\bar{Y}>.5$.

4.3 Testing a Point-Null Hypothesis Against a Composite Alternative

The more difficult problem of testing a point-null hypothesis against a composite alternative is discussed in this section. A composite hypothesis is a set of values of the parameter vector, each of which determines a different data distribution. In testing a point-null hypothesis against a composite alternative, we ask the question "was the data more likely to have come from the null distribution or from one distribution selected from the set of distributions which comprise the alternative hypothesis?" Except in certain trivial cases, this question cannot admit an unambiguous answer.

An examination of a likelihood function illustrates the difficulty of testing composite hypotheses. Based on a sample from a normal distribution with unknown mean μ and known variance σ^2, we would like to test the null hypothesis $H_0:\mu=0$ against the alternative $H_1:\mu\neq0$. Letting $\mathbf{1}_T$ be a vector of ones and $\mathbf{Y}$ the vector of T independently drawn observations, the sampling distribution can be written as

$$f_N(\mathbf{Y}|\mathbf{1}\mu,\sigma^2\mathbf{I})=(2\pi\sigma^2)^{-T/2}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{1}_T\mu)'(\mathbf{Y}-\mathbf{1}_T\mu)\right].$$

If μ were equal to zero, the density function of the data would be $f_N(\mathbf{Y}|\mathbf{0},\sigma^2\mathbf{I})$. The likelihood function

$$L(\mu;\mathbf{Y})=\frac{f_N(\mathbf{Y}|\mathbf{1}\mu,\sigma^2\mathbf{I})}{f_N(\mathbf{Y}|\mathbf{0},\sigma^2I)}$$

formally summarizes the evidence in favor of some other value of μ in comparison with the hypothesized value $\mu=0$ by indicating how much more likely it is that the data were drawn from a distribution located at μ than from a distribution located at zero. By a simple manipulation we may write it as

$$L(\mu;\mathbf{Y})=\exp\left[-\frac{T}{2\sigma^2}(\mu-\bar{Y})^2\right]\exp\left[\frac{T}{2\sigma^2}\bar{Y}^2\right]$$

where $\bar{Y}$ is the sample mean $\mathbf{Y}'\mathbf{1}/T$.

This is a function that is symmetric around its maximum point $\mu=\bar{Y}$ where it assumes the value $\exp[\bar{Y}^2T/2\sigma^2]$. An example is graphed in Figure 4.6.

We need now to indicate whether the data favor or cast doubt on the null hypothesis $\mu=0$. In the ideal situation the likelihood function $L(\mu;\mathbf{Y})$ is either zero at $\mu=0$ or zero everywhere else, and we could unambiguously conclude in the former case against the value $\mu=0$, and in the latter case in favor of it. Less precise information that nonetheless incontrovertibly favors one hypothesis or the other is implied by a likelihood function that attains either its minimum or its maximum at $\mu=0$. Unhappily, the probability of these unambiguous outcomes is zero, and we are forced to deal almost always with the sort of ambiguous situation depicted in Figure 4.6, in which the null hypothesis looks better than the alternative at some values of μ but worse at others.

We argue subsequently that there is, in fact, no solution to this dilemma. Corresponding to three different statistical schools are three different approaches, each of which is discussed, each of which has apparent shortcomings.

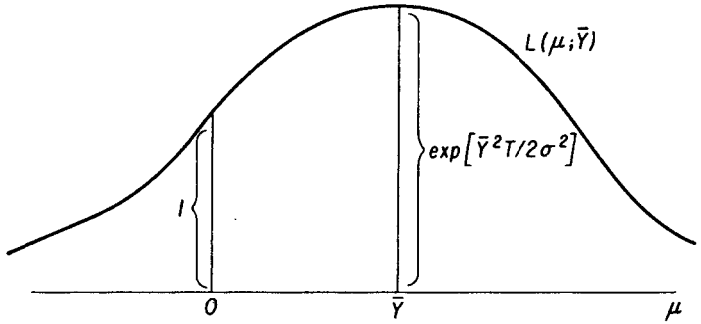

Fig. 4.6 Likelihood function $f_N(\mathbf{Y}|\mathbf{1}\mu, \sigma^2\mathbf{I})/f_N(\mathbf{Y}|\mathbf{0},\sigma^2\mathbf{I})$.

THE LIKELIHOOD SCHOOL

The likelihood approach is apparently straightforward. It is hypothesized that $\mu=0$. Suppose there were another value of μ for which the data were 100 times more probable than for $\mu=0$. Would not this shake your faith in $\mu=0$? This suggests that the right number to report against the null hypothesis is the likelihood ratio at the value of μ that makes the alternative appear best, namely, at $\mu=\bar{Y}$:

$$L_1=\max_{\mu} L(\mu;\mathbf{Y})=\exp\left[\frac{\bar{Y}^2T}{2\sigma^2}\right],$$

which by a series expansion truncated after the first term can be written approximately as

$$L_1\sim 1+\frac{\bar{Y}^2T}{2\sigma^2}$$
$$=1+\frac{z^2}{2}$$

where $z^2=\bar{Y}^2T/\sigma^2$. Thus for $z^2=2$ we can say that the data cast doubt on H_0 in the sense that there is a value of μ that is approximately twice as likely to have generated the given data.

SAMPLING THEORY SCHOOL

Sampling theory is concerned not with the likelihood function of μ given the data $\mathbf{Y}$ but rather with the sampling distribution of $\mathbf{Y}$ given the null hypothesis $\mu=0$. Statements such as the following are usually made. If the observed value of $\mathbf{Y}$ is in the tail of the distribution $f(\mathbf{Y}|\mu=0,\sigma^2)$ the data cast doubt on H_0 in the sense that something unlikely would have had to occur if H_0 were true.

Most commonly, tests are based on the sufficient statistic $\bar{Y}$ which is distributed normally with mean μ and variance σ^2/T. An indication of whether $\bar{Y}$ came from the tail of its distribution given $\mu=0$ is

$$S=\frac{\max_{\bar{Y}} f_{\bar{Y}}(\bar{Y}|\mu=0,\sigma^2/T)}{f_{\bar{Y}}(\bar{Y}|\mu=0,\sigma^2/T)}$$
$$=\exp\left[\frac{\bar{Y}^2T}{2\sigma^2}\right]=\exp\left[\frac{z^2}{2}\right]$$

which is seen to be identical to L_1 before. Thus for $z^2=\bar{Y}^2T/\sigma^2$ large we may say that the data cast doubt on $\mu=0$ both in the sense that there are other values of μ that are more likely to have generated the data and also in the sense that if μ were zero, an unlikely event occurred.

A more traditional description of the tail of the distribution is expressed in terms of the probability mass rather than in terms of the relative density. That is, before observing $\bar{Y}$, it is decided that a value of $\bar{Y}^2$ in excess of some arbitrary number, say, c^2, will cast doubt on H_0. The probability mass in the tail of the distribution beyond c^2 is called the significance level of the test,

$$\alpha(c)=P(\bar{Y}^2>c|\mu=0),$$

and if $\bar{Y}$ falls in the described region, the null hypothesis is said to be rejected at level α. It is customary to select c such that $\alpha(c)=.05$, and the familiar "region of rejection" is

$$|\bar{Y}|\geqslant\frac{1.96\sigma}{\sqrt{T}}$$

or in terms of z^2, $z^2\geqslant 1.96^2$. The corresponding probability statement is

$$P(z^2\geqslant 1.96^2|\mu=0,\sigma^2)=.05.$$

Thus z^2 in excess of 1.96 casts doubt on H_0 in the sense that given such a z we would be led to reject H_0 at the .05 level.

There is yet another "metric" for measuring the evidence against H_0. The P-value of a test is the level at which the data is "just significant"; at any significance level less that P the null hypothesis would not be rejected, whereas at any larger significance level it would be. Both the P value and S are graphed in Figure 4.7. Both indicate whether the observed value of $\bar{Y}$ is in the tail of its distribution or not, and both are (increasing) functions of z^2 only.

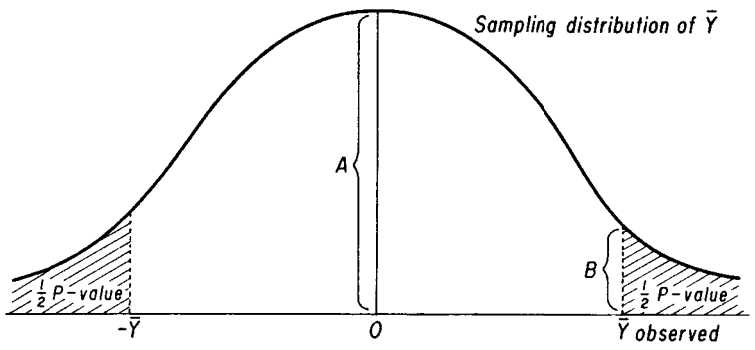

Fig. 4.7 Sampling distribution: P-value and $S = A/B$.

THE BAYESIAN SCHOOL

Bayesians, of course, apply their single commandment, Bayes' rule, and conclude that the null hypothesis is favored by the data if its posterior probability exceeds its prior probability (ignoring the loss function, for now). If the prior for μ is continuous, both the prior and the posterior probability of the point null hypothesis are zero, and the problem is uninteresting. One prior distribution that allocates probability π to the null hypothesis is

$$P(\mu) = \begin{cases} \pi & \mu = 0 \\ (1-\pi) f_N(\mu|0, h^{*-1})\, d\mu & \mu \neq 0, \end{cases}$$

where $f_N(\mu|0, h^{*-1})$ indicates a normal distribution located at the origin with variance h^{*-1}. In words, there is a spike of mass π at $\mu = 0$, and the rest is allocated normally over the line. By a straightforward application of Bayes' rule we have the posterior distribution of μ given the data mean $\bar{Y}$ as

$$P(\mu|\bar{Y}) \propto P(\bar{Y}|\mu) P(\mu),$$

and the odds ratio in favor of the alternative hypothesis is

$$\frac{P(H_1|\bar{Y})}{P(H_0|\bar{Y})} = \frac{\int_{\mu \neq 0} f_N(\bar{Y}|\mu, \sigma^2/T) P(\mu)}{f_N(\bar{Y}|\mu = 0, \sigma^2/T)\pi} = \int L(\mu; \bar{Y}) P(\mu)\, d\mu$$

$$= (1-\pi)\pi^{-1}\left(1 + \frac{T}{\sigma^2 h^*}\right)^{-1/2} \exp\left[z^2\left(1 + \frac{h^*\sigma^2}{T}\right)^{-1} \Big/ 2\right]$$

where $z^2 = \bar{Y}^2/(\sigma^2/T)$ and where we have made use of some results to

follow (alternatively, it requires some straightforward numerical integration). That is, the posterior odds ratio is the prior odds ratio times what is called the "Bayes factor"

$$B(h^*) = \left(1 + \frac{T}{\sigma^2 h^*}\right)^{-1/2} \exp\left[z^2\left(1 + \frac{h^*\sigma^2}{T}\right)^{-1} \Big/ 2\right]. \tag{4.6}$$

This, like the previous data summaries, is an increasing function of $z^2 = \bar{Y}^2 T/\sigma^2$; the larger z^2 is the more the data cast doubt on H_0. However, and most importantly, B is a function of sample size also. Let z^2 take on some arbitrarily large value so that, classically, we would say the data cast doubt on H_0. For sufficiently large T/σ^2, $B(h^*)$ can take on any arbitrarily small value, and rather than concluding that the data cast doubt on H_0, we would claim that they quite strongly favor H_0.

This is a version of the Lindley (1957) paradox. It represents a sharp disagreement between classicists and Bayesians over the interpretation of evidence. Classicists claim for this problem that the evidence against H_0 can be fully summarized in terms of z^2 alone. Bayesians would be in general agreement that the sample size matters, as well, in the sense that the larger is the sample size the greater must be z^2 to constitute convincing evidence against H_0. Is there a resolution to this controversy? I think there is, and I think it is quite clear that the Bayesians are right.

Consider first the significance level approach. Corresponding to a .05-level test is a power curve $\phi(\mu) = P(z^2 \geq c^2(.05)|\mu)$ indicating the probability of rejecting the null hypothesis for particular values of μ. An example is graphed in Figure 4.8. The reader may convince himself that ϕ is a symmetric function around $\mu = 0$, converging to one as μ increases.

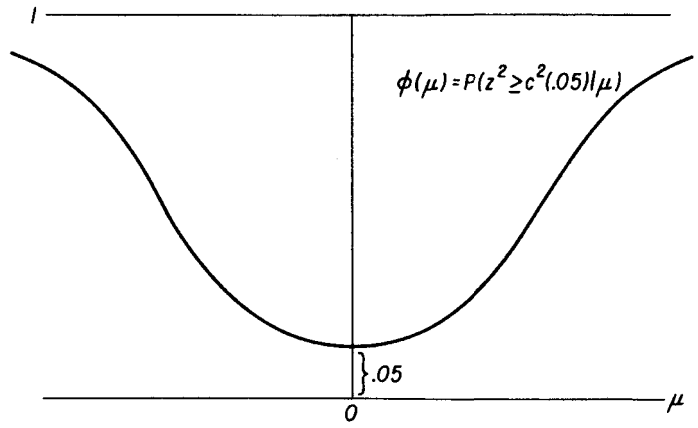

Fig. 4.8 Power curve.

As sample size increases, this power function gets steeper and steeper around $\mu=0$, and $[1-\phi(\mu)]$ which is the probability of making an "error of the second kind"—accepting a false null hypothesis—converges to zero at any value of $\mu\neq 0$. In contrast, the significance level—the probability of making an error of the first kind—stays forever at .05. The result is perfect error control against an error of the second kind but "mediocre" (.05) error control against an error of the first kind. The null hypothesis, however, is meant to be the favored hypothesis in the sense that incorrectly rejecting it is a more serious error than incorrectly accepting it. Fixed-level hypothesis testing, in contrast, clearly favors the alternative hypothesis—more and more so as sample size grows. The cure for this problem is obvious—the significance level must be made a decreasing function of sample size. Thus the conclusion we reached previously from the Bayesian viewpoint (that the interpretation of z^2 as evidence against H_0 depends on sample size) can also be reached within the confines of classical hypothesis testing. This still leaves arbitrary the particular function of sample size that the significance level should be set to. Although the Bayes factor $B(h^*)$ is a precise function of sample size, it depends on a somewhat arbitrary prior distribution. What is clear from both viewpoints is the fact that the interpretation of z^2 should depend on sample size.

As should be expected, a similar argument can be made from the likelihood standpoint. Suppose that the likelihood function of Figure 4.6 were discontinuous at μ_1 as in Figure 4.9. If we let the evidence against H_0 be summarized by $L(\mu_1;\bar{Y})$ we would conclude against H_0, even though for every other value of μ the null hypothesis is favored. It seems doubtful that we would really conclude against H_0 in this instance—some reference would be made to the a priori probability of μ_1, particularly to the fact that the function approximates its modal value on a "zero volume" set. But, essentially, the same thing happens to the likelihood function as sample size increases—it gets steeper and steeper around its mode and approximates the modal value in an ever-decreasing region. This suggests that instead of using the modal value as an indicator of the evidence against H_0 we might use the average value within a small fixed-size region around the mode, say, $\bar{Y}\pm\varepsilon$

$$L_2=\int_{\mu=\bar{Y}-\varepsilon}^{\bar{Y}+\varepsilon}(2\varepsilon)^{-1}L(\mu;\mathbf{Y})\,d\mu$$

$$=(2\varepsilon)^{-1}\exp\left(\frac{z^2}{2}\right)\int_{x=-\varepsilon}^{\varepsilon}\exp\left[-\frac{1}{2\sigma^2}\frac{x^2}{T}\right]dx.$$

The integral in this expression is the area under a normal curve. Making use of the fact that as σ^2/T gets small, essentially all of the probability is

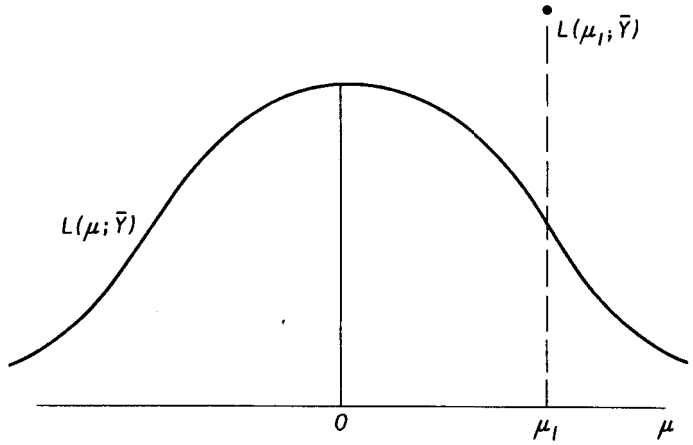

Fig. 4.9 Likelihood ratio: a peculiar example.

within an ε of its mean, we may write L_2 approximately as

$$L_2\propto\frac{1}{2\varepsilon}\left(\frac{2\pi\sigma^2}{T}\right)^{\frac{1}{2}}\exp\left[\frac{z^2}{2}\right], \tag{4.7}$$

a decreasing function of sample size T. Incidentally, as sample size increases, the Bayes factor $B(h^*)$ converges to

$$B(h^*)=h^{*\frac{1}{2}}\left(\frac{\sigma^2}{T}\right)^{\frac{1}{2}}\exp\left(\frac{z^2}{2}\right),$$

not unlike L_2.

Before proceeding, it may be useful to point out that the testing of a composite hypothesis is an index number problem. An index is a single-valued function of a set of numbers $(x_1,\ldots,x_n)$ that in some sense captures the essential aspects of the entire set. Weighted averages are commonly used as indexes, say, $I(x_1,\ldots,x_n)=\Sigma_{i=1}^{n}w_ix_i$, $\Sigma^i w_i=1, w_i\geqslant 0$. Theories of index numbers usually imply weight values w_i that are known only approximately, but this has not inhibited the practical construction of indexes. For example, consumer price indexes have weights that depend theoretically on unobservable marginal utilities.

An alternative index—the maximum of the class $I^*(x_1,\ldots,x_n)=\max_i x_i$—has to my knowledge never been proposed in the context of the theory of index numbers. It clearly misrepresents the class of numbers except in the special case when most other members of the class attain or nearly attain the maximum.

The problem of characterizing the evidence in favor of the foregoing hypotheses is precisely an index number problem. Instead of having a single number for each hypothesis, we have a continuous set of numbers. Interestingly enough, a classical analysis implicitly solves this index number problem by taking the maximum of the class, since a test of H_0 versus H_1 is based on the likelihood ratio $L_1 = \max_\mu L(\mu; \mathbf{Y})$. This solution to the index number problem seems inappropriate in general, but it is increasingly inappropriate as sample size increases, since as sample size increases the likelihood function becomes steeper and steeper and approximates its maximum in an ever decreasing region. In contrast, for a small sample size the likelihood function is relatively flat and the maximum *is* representative of wide regions.

A natural alternative is a weighted-average index, $\int L(\mu; \mathbf{Y}) w(\mu) \, d\mu$, $w(\mu) \geqslant 0$, $\int w(\mu) = 1$. A prior distribution suggests itself as a weight function $w(\mu)$. It assigns low weight to relatively implausible values, thereby indicating that the performance of the "model" with a priori unlikely parameters is less important than the performance of the model with a priori likely parameters.

To put this another way, the hypotheses become hypotheses *cum* distributions; $H_1 : \mu \sim w(\mu)$. Each composite hypothesis is thereby mixed into a simple hypothesis, in the sense that each specifies *one* data distribution; $H_1 : \mathbf{Y} \sim \int f_N(\mathbf{Y} | \mathbf{1}\mu, \sigma^2 \mathbf{I}) w(\mu) \, d\mu$. Data are interpreted to favor one hypothesis relative to another if the data were more likely to have come from one of these distributions than from another.

This way of putting it highlights the fact that the Bayesian approach, rather than solving the composite hypothesis-testing problem, in fact transforms it into the simple hypothesis-testing problem of discriminating among completely specified distributions. It can be a useful approach only if we can identify personally and publicly acceptable weight functions and/or if these weight functions do not matter "too much." The difficulty of selecting weights has not inhibited the use of weighted-average indices in other problems, nor shall it necessarily inhibit their use for this problem.

4.4 Weighted Likelihoods: Conjugate Priors

Bayes factors for testing regression models are now discussed. The Bayes factor in favor of the ith hypothesis relative to the jth hypothesis is $P(\mathbf{Y}|H_i) / P(\mathbf{Y}|H_j)$. It is convenient here to compute the marginal data density $P(\mathbf{Y}|H_i)$ for some hypothesis, and the subscript i may be suppressed. The model is taken to be

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad \mathbf{u} \sim N(\mathbf{0}_T, h^{-1}\mathbf{I}_T), \tag{4.8}$$

where $\mathbf{Y}$ is a $(T \times 1)$ observable vector, $\mathbf{X}$ is a $(T \times k)$ observable matrix, $\boldsymbol{\beta}$ is a $(k \times 1)$ unobservable parameter vector, and $\mathbf{u}$ is a $(T \times 1)$ unobservable error vector distributed normally with mean vector zero and variance-covariance matrix $h^{-1}\mathbf{I}_T$ with h a scalar precision parameter.

As discussed in the previous section, a Bayesian approach requires this hypothesis to be complemented with a prior distribution for the parameters $(\boldsymbol{\beta}, h)$. This distribution is used to marginalize out the parameters, and the hypothesis is treated as if it specified a unique distribution

$$f(\mathbf{Y}) = \int\int f(\mathbf{Y}|\boldsymbol{\beta}, h) f(\boldsymbol{\beta}, h) \, d\boldsymbol{\beta} \, dh. \tag{4.9}$$

The posterior probability of this hypothesis is, by Bayes' rule, proportional to the prior probability times the density (4.9) evaluated at the observed value of $\mathbf{Y}$. For obvious reasons, this is called a marginal likelihood.

Beginning with the simplest case, suppose the process variance h^{-1} is known, and let the prior for $\boldsymbol{\beta}$ be normal with mean $\mathbf{b}^*$ and variance $\mathbf{H}^{*-1}$. By inspection of Equation (4.8), $\mathbf{Y}$ is a linear combination of normals which is itself normal with mean $E(\mathbf{Y}) = \mathbf{X}\mathbf{b}^*$ and variance $V(\mathbf{Y}) = \mathbf{X}\mathbf{H}^{*-1}\mathbf{X}' + h^{-1}\mathbf{I}_T$:

$$f(\mathbf{Y}) = (2\pi)^{-T/2} |V(\mathbf{Y})|^{-1/2} \exp\left[-\tfrac{1}{2} hQ\right] \tag{4.10}$$

where

$$hQ = (\mathbf{Y} - \mathbf{X}\mathbf{b}^*)' V^{-1}(\mathbf{Y})(\mathbf{Y} - \mathbf{X}\mathbf{b}^*).$$

This expression can be rewritten by defining $\mathbf{N}^* = h^{-1}\mathbf{H}^*$ and by observing that

$$V^{-1}(\mathbf{Y}) = \left(\mathbf{X}\mathbf{H}^{*-1}\mathbf{X}' + h^{-1}\mathbf{I}_T\right)^{-1} = h\left(\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{N}^*)^{-1}\mathbf{X}'\right)$$

$$|V(\mathbf{Y})|^{-1} = h^T |\mathbf{N}^*| |\mathbf{N}^* + \mathbf{X}'\mathbf{X}|^{-1}.$$

Letting $\mathbf{N} = \mathbf{X}'\mathbf{X}$ and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, and after some manipulation, the quadratic form Q can be written as either

$$Q = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \mathbf{b}^*)(\mathbf{N}^{*-1} + \mathbf{N}^{-1})^{-1}(\mathbf{b} - \mathbf{b}^*) \tag{4.11}$$

or,

$$Q = (\mathbf{Y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{Y} - \mathbf{X}\mathbf{b}^*) - (\mathbf{b} - \mathbf{b}^*)'\mathbf{N}(\mathbf{N}^* + \mathbf{N})^{-1}\mathbf{N}(\mathbf{b} - \mathbf{b}^*). \tag{4.12}$$

By inspection of (4.10), the data favor the hypothesis in question if Q is small. Q is written in (4.11) as the minimum error sum of squares plus a factor that depends on the difference between the least-squares estimate $\mathbf{b}$ and the prior location $\mathbf{b}^*$. It is apparent that a model is to be judged in

terms of not only its R^2 but also by the "plausibility" of its estimates. In fact, (4.12) describes Q as an error sum of squares using the prior location as the estimate minus a factor also depending on the difference between $\mathbf{b}$ and $\mathbf{b}^*$. Thus Q is a number between $(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})$ and $(\mathbf{Y}-\mathbf{Xb}^*)'(\mathbf{Y}-\mathbf{Xb}^*)$, which suggests that a model's performance should be judged not in terms of the acceptability of $\mathbf{b}$ per se, but rather in terms of the R^2 when a more acceptable coefficient vector is employed.

Of course, the case of known process variance, h^{-1}, is of little practical interest. If we use a conjugate prior with uncertain h, these results generalize straightforwardly. Let $(\boldsymbol{\beta}, h)$ have the distribution $f_N(\boldsymbol{\beta}|\mathbf{b}^*, h^{-1}\mathbf{N}^{*-1})\, f_\gamma(h|s_1^2, \nu_1)$ where f_N is a multivariate normal distribution with mean $\mathbf{b}^*$ and variance matrix $h^{-1}\mathbf{N}^{*-1}$ and f_γ is a gamma distribution with location and scale parameters s_1^2 and ν_1. The predictive density then becomes a multivariate Student function

$$f(\mathbf{Y}) = \int\int f(\mathbf{Y}|\boldsymbol{\beta}, h) f(\boldsymbol{\beta}, h)\, d\boldsymbol{\beta}\, dh$$

$$= k(\nu_1, T)\left|\frac{\mathbf{M}}{s_1^2}\right|^{1/2}\left(\nu_1 + \frac{Q}{s_1^2}\right)^{-(\nu_1+T)/2} \tag{4.13}$$

where

$$\mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{N}^* + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$|\mathbf{M}| = |\mathbf{N}^*||\mathbf{N}^* + \mathbf{X}'\mathbf{X}|^{-1}$$

$$k(\nu_1, T) = \nu_1^{\nu_1/2}\left(\frac{\nu_1}{2} + \frac{T}{2} - 1\right)!/\pi^{T/2}\left(\frac{\nu_1}{2} - 1\right)!$$

Note that Equations (4.13) and (4.10) are somewhat similar monotonic functions of Q.

Leamer (1970) provides an approximation to the marginal likelihood when the normal prior for $\boldsymbol{\beta}$ is independent of h; Dickey (1971), for the more general case of a Student prior for $\boldsymbol{\beta}$, writes the marginal likelihood as a one-dimensional integral.

4.5 Weighted Likelihoods: Diffuse Priors

The critical defect of a Bayesian analysis of data is that prior distributions are both personally difficult to specify and also subject to variation among interested people. As a consequence, a Bayesian analysis based on any particular prior is of little interest. Instead, a researcher is obligated to report as well as possible the mapping from priors to posteriors, thereby servicing a wide readership and also highlighting those features of the prior

that are critical in this sense. Posterior probabilities based on a particular prior are of interest *if* the particular prior is characteristic of a class of interesting priors and if the posterior implied by this prior is "essentially the same as" the posterior for all priors in the class.

For example, in the usual analysis of the linear-regression model the posterior distribution corresponding to the diffuse prior ($\boldsymbol{\beta}$ uniform) is of interest because there is a wide, easily identifiable class of priors that are dominated by the data and that, therefore, lead to posteriors essentially the same as the posterior corresponding to the diffuse prior. Unfortunately, this is not also true for posterior probabilities of composite hypotheses, since various priors that are all relatively diffuse and "noninformative" lead to radically different posterior probabilities.

Making use of Equation (4.13), for example, the Bayes factor in favor of H_i relative to H_j is

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} = \frac{k(\nu_{1i}, T)}{k(\nu_{1j}, T)}\frac{|\mathbf{N}_i^*|^{1/2}}{|\mathbf{N}_j^*|^{1/2}}\frac{|\mathbf{N}^*{}_i + \mathbf{X}_i'\mathbf{X}_i|^{-1/2}}{|\mathbf{N}_j^* + \mathbf{X}_j'\mathbf{X}_j|^{-1/2}}\frac{s_{1i}^{-T}}{s_{1j}^{-T}}\frac{\left(\nu_{1i} + Q_i/s_{1i}^2\right)^{-(\nu_{1i}+T)/2}}{\left(\nu_{1j} + Q_j/s_{1j}^2\right)^{-(\nu_{1j}+T)/2}} \tag{4.14}$$

This formula involves the ratio $|\mathbf{N}_i^*|^{1/2}/|\mathbf{N}_j^*|^{1/2}$. As we let the matrices $\mathbf{N}^*$ become small to reflect decreasing information about the coefficients, this ratio converges to the indeterminate ratio of two zeroes, which can take on any number between zero and infinity, depending on the assumed speeds of convergence.

We may also set ν_1 to zero, to let the prior for h be diffuse. Three obvious ways of making our prior for $\boldsymbol{\beta}$ diffuse ($\mathbf{N}^* \to \mathbf{0}$) are to define $\mathbf{N}^* = \delta\mathbf{I}_k$ or $\mathbf{N}^* = \sqrt[k]{\delta}\,\mathbf{I}_k$ or $\mathbf{N}^* = (\sqrt[k]{\delta})\mathbf{X}'\mathbf{X}$ and let the scalar δ go to zero. As the reader may verify, these lead to three different limiting expressions:

$$f(\mathbf{Y}|H_i)/f(\mathbf{Y}|H_j) = \begin{cases} 0 & \text{if } k_i > k_j \\ \left(\dfrac{|\mathbf{X}_j'\mathbf{X}_j|}{|\mathbf{X}_i'\mathbf{X}_i|}\right)^{1/2}\left(\dfrac{ESS_j}{ESS_i}\right)^{T/2} & \text{if } k_i = k_j \\ \infty & \text{if } k_i < k_j \end{cases}$$

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} = \left(\frac{|\mathbf{X}_j'\mathbf{X}_j|}{|\mathbf{X}_i'\mathbf{X}_i|}\right)^{1/2}\left(\frac{ESS_j}{ESS_i}\right)^{T/2}$$

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} = \left(\frac{ESS_j}{ESS_i}\right)^{T/2}$$

where $ESS_i = \mathbf{Y}'(\mathbf{I} - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i')\mathbf{Y}$. Each of these expressions is representative of the posteriors corresponding to a mathematically well-defined class of diffuse priors, but no class has an unambiguous claim to representing "vague" prior information. Thus the class of posteriors corresponding to diffuse priors is not well defined, and the posterior corresponding to a particular kind of diffuseness is of reduced interest. This contrasts with a posterior for the coefficient vector $\boldsymbol{\beta}$, which is the same under all (limiting) definitions of diffuseness.

We can see from another viewpoint the problem with the diffuse priors by examining the distribution of $\mathbf{Y}$. The variance of $\mathbf{Y}$ is

$$V(\mathbf{Y}) = \mathbf{X}V(\boldsymbol{\beta})\mathbf{X}' + V(\mathbf{u}).$$

Those linear combination $\psi'\mathbf{Y}$ that are orthogonal to the columns of $\mathbf{X}$, $\psi'\mathbf{X} = \mathbf{0}$, have variance independent of $V(\boldsymbol{\beta})$

$$\begin{aligned} V(\psi'\mathbf{Y}) &= \psi'\mathbf{X}V(\boldsymbol{\beta})\mathbf{X}'\psi + \psi'V(\mathbf{u})\psi \\ &= \psi'V(\mathbf{u})\psi. \end{aligned}$$

Thus as $V(\boldsymbol{\beta})$ explodes, the predictive density on the $T-k$ dimensional subspace of $\mathbf{Y}$ defined by $\psi'\mathbf{X} = \mathbf{0}$ maintains a finite variance. An interpretation of this might be that predictions about joint events in the full T-dimensional space are called off, but predictions on certain subspaces are still on. The subspace over which predictions are proper clearly depends on $\mathbf{X}$. Since we can compare the predictive performance of two models in a nonarbitrary way only if the two models are predicting the same events, it will prove impossible to choose models without informative priors. This statement holds, incidentally, even when the number of explanatory variables is the same in all models, since although the dimensionalities of the predictions are the same, the prediction spaces are different unless $\mathbf{X}$ is the same for all hypotheses.

This dilemma does have a potential escape. Instead of seeking diffuse priors, let us find dominated priors, that is, let us explore the behavior of the marginal likelihoods as sample evidence accumulates. As T, Q, and $\mathbf{X}'\mathbf{X}$ grow, Equation (4.14) is well approximated by

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} \cong c\left(\frac{|\mathbf{X}_j'\mathbf{X}_j|}{|\mathbf{X}_i'\mathbf{X}_i|}\right)^{1/2}\left(\frac{ESS_j}{ESS_i}\right)^{T/2} \tag{4.15}$$

where c is a constant that does not change with sample size. Like the prior odds ratio, this constant will come to be dominated by the other terms. By the same argument, if the explanatory variables come from a stationary process, $\mathbf{X}'\mathbf{X}/T$ will converge to a constant, and a useful approximation is

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} \cong cT^{(k_j-k_i)/2}\left(\frac{ESS_j}{ESS_i}\right)^{T/2}. \tag{4.16}$$

Equation (4.15) has one obvious defect. It is not invariant to scale transformations of the explanatory variables. By a suitable change in the units of measurement, $|\mathbf{X}_j'\mathbf{X}_j|/|\mathbf{X}_i'\mathbf{X}_i|$ can be made to favor any hypothesis (if each has at least one explanatory variable not found in any of the others). On the other hand, an argument can be made that some adjustment should be made for the variability of the explanatory variable set, as is done by this term. If there is a constant term in the regression, then $|\mathbf{X}'\mathbf{X}|$ is proportional to the determinant of the matrix of moments about the means of the nonconstant variables or the generalized variance of the explanatory variable set. Thus, by Equation (4.15), a model that enjoys a richly variable set of explanatory variables is expected to yield a smaller error sum of squares than a model with a poorer set of explanatory variables.

A formula that is both invariant to scale transformations and adjusts for the variability of the explanatory variable set is

$$\frac{f(\mathbf{Y}|H_i)}{f(\mathbf{Y}|H_j)} = c\left(\frac{|\mathbf{R}_j|}{|\mathbf{R}_i|}\right)^{1/2} T^{(k_j-k_i)/2}\left(\frac{ESS_j}{ESS_i}\right)^{T/2} \tag{4.17}$$

where $\mathbf{R}_i$ is the matrix of correlation coefficients of the explanatory variables. Note that $|\mathbf{X}'\mathbf{X}| = T^k|\mathbf{R}|\prod_{l=1}^{k}\hat{\sigma}_l^2$ where $\hat{\sigma}_l^2$ is the sample estimate of the variance of the lth explanatory variable. Equation (4.17) thus involves the assumption that $|\mathbf{N}^*|/\prod_l\hat{\sigma}_l^2$ is constant across models.

Which of these many formulas are we then to choose? The following properties do seem desirable:

a. There must be no arbitrary constants.
b. The posterior probability of a model should be invariant to linear transformations of the data.
c. There should be a degrees-of-freedom adjustment; of two models that both yield the same ESS, the one with the fewer explanatory variables should have the higher posterior probability.
d. A model with a richly variable explanatory variable set should be expected to yield a smaller ESS than one with highly collinear data.

Of these properties only (d) seems to be open to serious question, because by a linear transformation any set of explanatory variables can be made to be orthogonal. As a consequence, (d) is in conflict with (b). The formula that satisfies (a), (b), and (c) is Equation (4.16), currently this author's favorite. If there is an obvious parameterization such that $|\mathbf{N}^*|/\prod_l\hat{\sigma}_l^2$ is constant across models, than Equation (4.17), which does adjust for the variability of the explanatory variable set, is preferred.

It should be emphasized that the dominance notion is not really a solution to the problem, even though it leads unambiguously to formulas

such as Equation (4.16), since in order to apply these formulas we need to know whether the sample does, in fact, dominate the prior. For any proper prior, no matter how diffuse, there are observations that make Equation (4.16) a very poor approximation to (4.14). The only way to know if the particular sample does, in fact, dominate the prior is to specify fully the prior, but having done that Equation (4.14) applies.

For other discussions of improper priors for this problem see Jeffreys (1961), Thornber (1966), Geisel (1969), Lempers (1971), Zellner (1971), and Dagenais (1972).

To conclude this section, we have seen that various reasonable definitions of diffuseness lead to rather different posterior probabilities of composite hypotheses. This makes posterior probabilities computed from any particular formula less interesting. This author has the personal opinion that the problem is of academic interest only, since a prior that allocates positive probability to subspaces of the parameter space but is otherwise diffuse represents a peculiar and unlikely blend of knowledge and ignorance. Parenthetically, what often appears to be choice among potentially true models is, in fact, the choice of a simple model that works well for some decisions. In other cases, hypothesis testing is used to introduce into the analysis uncertain prior information about parameters. These as well as other specification searches are discussed in detail in subsequent chapters.

4.6 Conclusion

To conclude, let us reconsider the five problems mentioned in the introduction, this time providing answers:

Answer 1. Classical hypothesis testing at a fixed level of significance increasingly distorts the interpretation of the data against a null hypothesis as the sample size grows. *The significance level should consequently be a decreasing function of sample size.*

Under one definition of diffuseness we saw that the posterior odds ratio in favor of the alternative hypothesis is the prior odds times the factor

$$B=\left(\frac{ESS_0}{ESS_1}\right)^{T/2} T^{(k_0-k_1)/2}.$$

We say that the evidence favors the alternative hypothesis if $B>1$, which can be written in terms of the F value defined in (4.2) as

$$F>\frac{T-k}{p}(T^{p/T}-1)$$

where $p=k_1-k_0$ is the number of restrictions. Table 4.1 provides these critical values as a function of sample size T, the degrees of freedom $T-k$ and the number of restrictions p. For comparison, the critical value of the F test at the .05 level are also included, and we observe the general result that a Bayesian with this kind of prior requires much larger F values as sample size increases.

Answer 2. There is nothing special about complex, nonnested structures of hypotheses. The posterior probability of an hypothesis H_i is $P(H_i|Y)=P(Y|H_i)P(H_i)/\Sigma_j P(Y|H_j)P(H_j)$ regardless of the structure of the hypotheses.

Answer 3. The existence of prior information about the parameters influences hypothesis testing in the sense that a hypothesis is to be judged at a priori likely values of the parameters as well as at those values favored by the data.

However, there does not seem to me to be a proper prior distribution that would lead to the common procedure of discounting an R^2 in four increasing steps, depending on whether a parameter is significantly different from zero and the right sign, insignificant and the right sign, insignificant and the wrong sign, and (worst of all) significant and the wrong sign. Although the normal priors discussed previously inadequately capture information about signs, they lead one to hope that his estimates are insignificantly different from his prior mean. Take the case when $\boldsymbol{\beta}$ is known to be $\mathbf{b}^*$ and the prior for h is diffuse, $f(h)\propto h^{-1}$. The marginal likelihood is then proportional to

$$f(\mathbf{Y}|H)\propto\int h^{T/2-1}\exp\left[-\frac{1}{2}h\left(ESS+(\mathbf{b}^*-\mathbf{b})(\mathbf{X}'\mathbf{X})(\mathbf{b}^*-\mathbf{b})\right)\right]dh$$

$$=\left[ESS+(\mathbf{b}-\mathbf{b}^*)'(\mathbf{X}'\mathbf{X})(\mathbf{b}-\mathbf{b}^*)\right]^{-T/2}$$

$$=ESS^{-T/2}\left(1+F\frac{k}{T-k}\right)^{-T/2}$$

where F is the F statistic for testing $\boldsymbol{\beta}=\mathbf{b}^*$. The larger is this F value, the less is $f(\mathbf{Y}|H)$ and the less likely is the data to have been generated by this model. Thus you are hoping for insignificance, not significance.

The "bigger is better" philosophy embedded in the usual procedure would seem to require an improper prior that says "bigger is more likely." For example, given the sample mean $\bar{Y}$ which is distributed normally with mean μ and variance σ^2/T and a prior for μ that is uniform between zero and M, the Bayes factor in favor of the hypothesis $\mu=0$ versus the

Table 4.1

Bayesian and Classical Critical Values of the F test

	$T-k=$	1	2	3	4	5	10	50	100	1000
	$k=1$	0.41	0.88	1.24	1.52	1.74	2.44	4.01	4.68	6.93
	2	0.44	0.83	1.14	1.39	1.60	2.30	3.95	4.64	6.92
	3	0.41	0.76	1.04	1.28	1.48	2.18	3.89	4.60	6.91
$p=1$	4	0.38	0.70	0.96	1.19	1.38	2.07	3.83	4.57	6.91
	5	0.35	0.64	0.89	1.11	1.29	1.98	3.78	4.53	6.90
	10	0.24	0.46	0.65	0.83	0.99	1.62	3.53	4.37	6.87
	20	0.16	0.30	0.44	0.57	0.69	1.20	3.13	4.07	6.81
	5% point[b]	161	18.5	10.1	7.8	6.6	4.96	4.03	3.94	3.85
	$k=1$	0.50	1.08	1.50	1.81	2.04	2.73	4.17	4.78	6.95
	2	0.54	1.00	1.36	1.63	1.86	2.57	4.10	4.75	6.94
	3	0.50	0.90	1.23	1.49	1.70	2.42	4.04	4.71	6.94
$p=2$	4	0.45	0.82	1.12	1.36	1.57	2.29	3.98	4.67	6.93
	5	0.41	0.74	1.02	1.26	1.46	2.17	3.92	4.63	6.93
	10	0.27	0.51	0.73	0.92	1.09	1.75	3.66	4.46	6.90
	20	0.17	0.32	0.47	0.61	0.73	1.27	3.23	4.15	6.84
	5% point[a]	200	19.0	9.55	6.94	5.79	4.10	3.18	3.09	3.00
	$k=1$	0.61	1.33	1.83	2.17	2.42	3.08	4.34	4.90	6.97
	2	0.67	1.22	1.63	1.93	2.17	2.87	4.27	4.86	6.97
	3	0.61	1.08	1.45	1.74	1.97	2.69	4.20	4.82	6.96
$p=3$	4	0.54	0.97	1.30	1.57	1.80	2.53	4.13	4.78	6.96
	5	0.48	0.87	1.18	1.44	1.66	2.40	4.07	4.74	6.95
	10	0.31	0.57	0.81	1.01	1.20	1.87	3.79	4.56	6.92
	20	0.18	0.35	0.51	0.65	0.79	1.35	3.33	4.24	6.86
	5% point[b]	216	19.2	9.28	6.59	5.41	3.71	2.79	2.70	2.61
	$k=1$	0.75	1.66	2.25	2.62	2.88	3.48	4.52	5.01	7.00
	2	0.83	1.50	1.97	2.30	2.55	3.22	4.44	4.97	6.99
	3	0.75	1.31	1.73	2.04	2.29	3.00	4.37	4.93	6.99
$p=4$	4	0.66	1.15	1.53	1.83	2.07	2.81	4.30	4.89	6.98
	5	0.58	1.02	1.37	1.66	1.89	2.65	4.23	4.85	6.97
	10	0.35	0.64	0.90	1.13	1.32	2.05	3.92	4.66	6.94
	20	0.20	0.38	0.54	0.70	0.84	1.43	3.43	4.33	6.88
	5% point[b]	225	19.25	9.12	6.39	5.19	3.48	2.56	2.46	2.38
	$k=1$	0.93	2.10	2.79	3.20	3.45	3.95	4.70	5.13	7.02
	2	1.05	1.86	2.40	2.76	3.01	3.63	4.62	5.09	7.02
	3	0.93	1.60	2.07	2.41	2.67	3.36	4.54	5.05	7.01
$p=5$	4	0.80	1.38	1.81	2.13	2.39	3.13	4.47	5.00	7.00
	5	0.69	1.21	1.60	1.91	2.16	2.93	4.40	4.96	7.00
	10	0.39	0.73	1.01	1.25	1.47	2.23	4.07	4.76	6.97
	20	0.21	0.41	0.59	0.75	0.90	1.53	3.55	4.42	6.91
	5% point[b]	230	19.30	9.01	6.26	5.05	3.33	2.40	2.30	2.22

[a] T = number of observations, k = number of parameters, p = number of restrictions being tested.

[b] Classical critical value at the .05 level of significance.

hypothesis $\mu>0$ is

$$\frac{f(\bar{Y}|\mu=0)}{\int_0^M f(\bar{Y}|\mu,\sigma^2)M^{-1}d\mu}$$

$$=\frac{(2\pi\sigma^2/T)^{1/2}\exp\left[-\bar{Y}^2/(2\sigma^2/T)\right]}{\int_0^M(2\pi\sigma^2/T)^{1/2}\exp\left[-(\bar{Y}-\mu)^2/(2\sigma^2/T)\right]M^{-1}d\mu}$$

$$=\frac{(2\pi\sigma^2/T)^{1/2}\exp\left[-z^2/2\right]}{M^{-1}P^*(0\leqslant\mu\leqslant M)}$$

where z^2 is the square of the normal statistic for testing $\mu=0$, $z^2=\bar{Y}^2/(\sigma^2/T)$, and $P^*(0\leqslant\mu\leqslant M)$ is the posterior probability that $0\leqslant\mu\leqslant M$ given an improper prior for μ that is uniform on the whole line. The numerator of this Bayes factor is unambiguously a decreasing function of z^2. The denominator, however, may either increase or decrease with z^2 depending on whether $\bar{Y}$ is in the interval $0\leqslant\bar{Y}\leqslant M$ or not. It may, nonetheless, be approximately true that the Bayes factor in favor of $\mu>0$ is greatest if $\bar{Y}$ is positive and z^2 large, relatively great for $\bar{Y}$ positive, small for $\bar{Y}$ negative, and especially small if $\bar{Y}$ is significantly negative.

Answer 4. Measures of location and dispersion of a parameter vector β_i follow necessarily from the probability function

$$f(\beta_i|\mathbf{Y})=\sum_j f(\beta_i|\mathbf{Y},H_j)f(H_j|\mathbf{Y}).$$

The interpretation of this simple formula is not entirely trivial. We have yet to identify the slightly confusing p.d.f. $f(\beta_i|\mathbf{Y},H_j)$ for $i\neq j$. It summarizes opinions about a parameter in the ith model given that the jth model generated the data. Conditional on the jth model the data come from the distribution $f(\mathbf{Y}|\beta_j,H_j)$ and therefore contain information only about β_j. It may, nonetheless, be the case that β_i and β_j are a priori correlated, and we would then obtain information about β_i, or in terms of probability functions $f(\beta_i|\mathbf{Y},H_j)\neq f(\beta_i|H_j)$.

For example, consider the hypotheses $H_1:\mathbf{Y}=\mathbf{x}\gamma+\mathbf{z}\alpha+\mathbf{u}$ and $H_2:\mathbf{Y}=\mathbf{x}\gamma+\mathbf{w}\delta+\mu$, with parameter vectors $\beta_1=(\gamma,\alpha)$, $\beta_2=(\gamma,\delta)$, and with the first parameters thus perfectly correlated. In this case the probability function for $\gamma=\beta_{11}=\beta_{12}$ is the mixture $f(\gamma|H_1,\mathbf{Y})f(H_1|\mathbf{Y})+f(\gamma|H_2,\mathbf{Y})f(H_2|\mathbf{Y})$ where $f(\gamma|H_i,\mathbf{Y})$ is the usual posterior distribution for γ given model H_i. Letting $\pi_i=P(H_i|\mathbf{Y})$, $m_i=E(\gamma|\mathbf{Y},H_i)$ and $V_i=V(\gamma|\mathbf{Y},H_i)$

the mean and variance of γ are

$$E(\gamma|\mathbf{Y}) = \sum \pi_i m_i$$

$$V(\gamma|\mathbf{Y}) = E(\gamma^2|\mathbf{Y}) - \left(\sum \pi_i m_i\right)^2$$

$$= \sum \pi_i (V_i + m_i^2) - \left(\sum \pi_i m_i\right)^2$$

$$= \sum \pi_i V_i + \left[\sum \pi_i m_i^2 - \left(\sum \pi_i m_i\right)^2\right]$$

$$= \sum \pi_i V_i + \sum \pi_i \left(m_i - \sum \pi_i m_i\right)^2$$

where the last term in the brackets is the variance of a discrete probability function that allocates probability π_i at location m_i. The point we wish to draw attention to is that although the mean is a mixture of the means from each of the regression equations, the variance exceeds a weighted average of the variances by an amount that depends on the variability of the estimates across equations. Thus although the several regressions may individually yield highly accurate estimates, if those estimates are very different and if given the data there remains considerable ambiguity about the model, the result may be considerable uncertainty about the parameter.

Next consider hypotheses that have no common parameters, and assume furthermore that the parameters are completely independent across hypotheses. Thus if the data are generated by the first model, no information can be gathered about coefficients in other models; in the foregoing notation we must have $f(\boldsymbol{\beta}_i|\mathbf{Y}, H_j) = f(\boldsymbol{\beta}_i|H_i)$. In order to apply the formula we must, of course, also determine $f(\boldsymbol{\beta}_i|H_j)$. The reader may verify that without saying so in the previous example we have set $f(\boldsymbol{\beta}_i|H_j) = f(\boldsymbol{\beta}_i)$ for all j. In words, our prior information about the coefficient $\boldsymbol{\beta}_i$ is independent of the hypothesis that applies. It is more natural in this second situation to assume that $f(\boldsymbol{\beta}_i|H_j)$ is a degenerate probability function that assigns all probability to a point. For example, consider the pair of hypotheses $H_1: \mathbf{Y} = \mathbf{x}\gamma + \mathbf{u}$, $H_2: \mathbf{Y} = \mathbf{z}\delta + \mathbf{u}$. We may either say that the distribution of δ given H_1 assigns all probability to the value zero, or we can define a new parameter β called the effect of $\mathbf{z}$ on $\mathbf{Y}$ which is zero given the first hypothesis. In either case, the distribution of the parameter is a mixture of the origin with weight $1 - P(H_i|\mathbf{Y})$ and the conditional posterior $f(\boldsymbol{\beta}_i|H_i, \mathbf{Y})$ with weight $P(H_i|\mathbf{Y})$. If we prefer to have $f(\boldsymbol{\beta}_i|H_j) = f(\boldsymbol{\beta}_i)$ we will obtain a posterior distribution for $\boldsymbol{\beta}_i$ that is a mixture of the prior $f(\boldsymbol{\beta}_i)$ with weight $1 - P(H_i|\mathbf{Y})$ and the conditional posterior $f(\boldsymbol{\beta}_i|H_i, \mathbf{Y})$ with weight $P(H_i|\mathbf{Y})$. The resultant increased uncertainty about the parameter, due to uncertainty about the model, is obvious.

Answer 5. A researcher who uses more than one model can report the overall effectiveness of his research in terms of the average marginal likelihood

$$f(\mathbf{Y}) = \sum_i f(\mathbf{Y}|H_i) P(H_i).$$

Assuming equal prior probabilities of M different models and the diffuse prior result (4.16), this becomes

$$f(\mathbf{Y}) \propto M^{-1} \sum_{i=1}^{M} T^{-k_i/2} (ESS_i)^{-T/2}.$$

We can transform $f(\mathbf{Y}|H_i)$ into an R^2 by the formula

$$f(\mathbf{Y}|H_i) \propto T^{-k_i/2} (1 - R_i^2)^{-T/2}$$

which solves to

$$(1 - R_i^2) \propto \left[\frac{T^{-k_i/2}}{f(\mathbf{Y}|H_i)}\right]^{2/T}.$$

The same transformation may be applied to $f(\mathbf{Y})$ to obtain a "grand" R^2

$$(1 - R^{*2}) = \left[\frac{T^{-k^*/2}}{f(\mathbf{Y})}\right]^{2/T}$$

$$= \left[M^{-1} \sum_{i=1}^{M} T^{(k^* - k_i)/2} (1 - R_i^2)^{-T/2}\right]^{-2/T}$$

where k^* is a (fictitious) average k and R^{*2} is the grand R^2. Assuming equal k_i or ignoring the $T^{k^* - k_i}$ terms we have the "overall" R^2 as

$$R^{*2} = 1 - \left[\frac{\sum_{i=1}^{M} (1 - R_i^2)^{-T/2}}{M}\right]^{-2/T}.$$

This formula is intended to penalize specification searches, since estimated models with low R^2s tend to lower the grand R^2. The penalty is not as great as you might imagine, however. Grand R^2s are reported in Table 4.2 which makes use of the assumptions that the best model yields an R^2 equal to .9, and that all the others yield identical R^2s. The lowest grand R^2 in the table is .43, which requires 99 models with zero R^2s and a very small