Statistically incoherent hypothesis tests in auditing

D. J. Johnstone
University of Wollongong

UNIVERSITY OF WOLLONGONG

DEPARTMENT OF ACCOUNTANCY

STATISTICALLY INCOHERENT HYPOTHESIS TESTS IN AUDITING

by

David J. Johnstone

The University of Wollongong

1994 Working Paper Series

94/3

Abstract. A classical, Neyman-Pearson hypothesis test results in a decision (choice of action) justified not by any assessment of sample evidence, but by the pre-specified frequencies with which that procedure generates errors of the two possible types. By applying such a test in auditing, the hypothesis tested is accepted or rejected without the auditor having to consider whether the data observed confirms (in any degree), or disconfirms, that hypothesis. In contrast with the classical framework, the Bayesian approach is to evaluate the probability of the hypothesis tested conditional on the data observed, and then to make a decision on the basis of that revised probability. Decisions are thus *evidence-based* rather than *rule-based.* So as to compare the classical and Bayesian programs, a familiar test example is considered, and hypothetical data, which, on a classical view, marginally reject the auditee's stated account balance, are re-interpreted from a Bayesian, evidential perspective. The results of this comparison reveal that classical hypothesis tests in auditing do not have a consistent (from test-to-test) evidential basis, and, in Bayesian terms, are therefore "incoherent". Also, contrary to intuitive expectations, marginal rejection is found to imply evidence *in favor* of the auditee's stated balance. Asymptotically, an account balance which is rejected only marginally in a classical hypothesis test has an "objective" (not-dependent-on-prior) posterior probability arbitrarily close to one.

Keywords:

Auditing, Evidence, Hypothesis Test, Bayesian Inference, Belief Revision

1. Introduction

This paper contrasts the classical, Neyman-Pearson logic of hypothesis testing with the Bayesian program for statistical inference and decision making. Unlike previous studies in auditing comparing classical and Bayesian statistical methodology, primary consideration is given to foundational, philosophical and logical issues underlying the distinction and longstanding conflict between classical and Bayesian statistics. In particular, the question is asked of the result of a Neyman-Pearson hypothesis test, "What does it mean?", or more specifically, "What is the evidence so represented?". Comparison of the classical and Bayesian responses to this fundamental question reveals aspects of the classical paradigm which, to a Bayesian, are logically indefensible.

Comparative study of the foundations of classical and Bayesian statistical methodologies is anticipated in a chapter by Akresh et al. (1988) on "Audit Approaches and Techniques", in the American Accounting Association's agenda of *Research Opportunities in Auditing.* Amoung the identified research issues relating to audit sampling and decision theory is the following:

> What are the properties/advantages/disadvantages of Bayesian decision-theoretic models in auditing as compared to classical statistical models? Given that the goal is to improve audit decision making, how do these approaches compare in terms of sample sizes, defensibility, ability to aggregate with other souces of evidence, etc.? (p.50)

Each of the specifics mentioned in this passage, viz. the matters of sample size, evidence aggregation, and in particular the *logical defensibility* of classical decision theory in comparison with the Bayesian alternative, comes into focus once the *evidential content* of classical test results is questioned in the fundamental manner described above.

Previous research in auditing contrasting orthodox (i.e. "classical") and Bayesian statistical methods is mostly concerned with "practical" or technical matters - for example, there is a rich body of work correlating the frequentist coverage probabilities of Bayesian and classical confidence bounds in money unit sampling (e.g., Tsui et al. (1985), Dworin and Grimlund (1986), Smielianskas (1986) and Grimlund and Felix (1987)). While such cross-validation is of obvious theoretical and practical importance, many of the more foundational issues separating the classical and Bayesian formalisms have been given relatively little attention in the statistics-in-auditing literature, at least by comparison with the profile and history of philosophical debate in the disciplines of mathematical statistics and the philosophy of science.

There is sometimes the supposition in auditing that although practitioners employing classical statistical methods, including hypothesis tests, are formally commited to orthodox statistical concepts and terminology, they retain the status of "informal" or *de facto* Bayesians. Unfortunately, this ecumanical standpoint is generally untenable, as has been known to Bayesians since Jeffreys (1961, pp.359-60; 1st ed. 1939). More recently, following much related Bayesian criticism of the orthodox or classical paradigm (e.g., de Finetti (1975), Edwards et al. (1963); Good (1981), (1983); Lindley (1957), (1972); Pratt (1965) and Zellner (1971), (1984)), articles by Berger and Sellke (1987, p.136) and Berger and Delampady (1987, p.330) published in the *Journal of the American Statistical Association* and *Statistical Science* (also an ASA journal), and commended by various eminent discussants, conclude with the view that the results of classical hypothesis tests are so commonly and systematically disparate with Bayesian probability revision that the future of such tests in statistics is highly questionable. The mathematical theory of statistical sampling in auditing is very well developed, even by the standards of statistics proper, however the substance and severity of authorative Bayesian dissatisfaction with Neyman-Pearson methods, as seen in papers such as those cited above, is not similarly well represented in the applied literature in auditing.

To preclude unnecessary technical difficulty, the analysis below begins with consideration of the most well known of Neyman-Pearson hypothesis tests employed in auditing, that of the mean of a normal population. This is one of the tests by which Neyman and Pearson (1933, pp.153-6) exposited their logic of statistical hypothesis testing, and through which most students are introduced to the application of that theory in auditing. More importantly, the test of a normal mean, although technically straightforward, embodies all the paradigmatic characteristics of Neyman-Pearson logic and philosophy, and is therefore an appropriate starting point when the foundations of that paradigm are in question. Indeed, by avoiding the added distractions of tests which involve more difficult, or less familiar, applications of the same underlying logic, attention is restricted to issues of principle affecting all Neyman-Pearson hypothesis tests, regardless of technical idiosyncracy.

2. An Example Test

Assume a population $X \sim N(\theta,\sigma^2)$ of individual accounts (e.g., receivables) with unknown mean θ, and variance σ^2. To decide between the actions of accepting and rejecting the auditee's aggregate account balance, an auditor conducts a statistical hypothesis test of

$$H_0\colon |\theta| \leq m \quad \text{versus} \quad H_1\colon |\theta| > m,$$

where θ represents the population average error (i.e., the average of the errors in the auditee's accounts receivable) and m denotes the average error amount deemed material (intolerable) by the auditor. Following a textbook-standard procedure (e.g., Roberts (1978) p.45; Arens and Loebbecke (1981) pp.137-44; (1991) p.520), the auditor's decision rule is to accept H_0, and thus the auditee's stated balance, only if the two-sided $100(1-\beta)\%$ confidence interval estimated for θ lies *entirely* within the predetermined materiality bounds $[-m,m]$. Otherwise the stated balance is rejected. The effect of this decision rule is to set the minimum power of the test against H_1: $|\theta|>m$,

$$\min_{|\theta|>m} \ p(Reject\ H_0\,|\,\theta),$$

greater than $100(1\text{-}\beta/2)\%$. Having this power characteristic, the test described fits the definition of Duke et al. (1982, p.51) of a "negative" test.

Example Results. For the purpose of exposition assume known $\sigma=170$, and suppose the auditor tests

$$H_0\text{: } |\theta|\leq 75 \quad \text{versus} \quad H_1\text{: } |\theta|>75,$$

observing a sample mean $\overline{X}$ with n observations, where n is fixed before any data have been drawn or inspected. The classical two-sided $100(1\text{-}\beta)\%$ confidence interval for θ, the population mean, is $\overline{X}\pm z_{\beta/2}\sigma/\sqrt{n}$ (disregarding finite population correction). For $\beta=0.05$, $z_{\beta/2}=1.96$, etc. The rejection level, or "critical" value of the test statistic, henceforth denoted by $\overline{X}_c$, is $m-z_{\beta/2}\sigma/\sqrt{n}$. A result $\overline{X}=\overline{X}_c$ marginally rejects H_0 since the right-hand confidence limit, $\overline{X}_c+z_{\beta/2}\sigma/\sqrt{n}$, equals exactly the materiality limit m.[1] Letting $\beta=0.05$, consider the following three such results and their associated 95% confidence intervals:

$$(n=20, \overline{X}=0.4942) \quad \rightarrow \quad [-74.012, 75]$$
$$(n=50, \overline{X}=27.8784) \quad \rightarrow \quad [-19.243, 75]$$
$$(n=90, \overline{X}=39.8776) \quad \rightarrow \quad [4.755, 75].$$

For an auditor having observed one of these results, the problem is one of *interpretation*. Specifically, of what logical meaning or use is that observation? This question, which might arise in negotiation with a client, or in defence of the auditor's decision process in a court, is answered below, first from a classical frequentist-decision-theoretic perspective, and then from the less conventional, Bayesian standpoint. Contrary to what seems a "common sense" understanding of classical hypothesis tests, marginal rejection of H_0 is found to *confirm*, rather than disconfirm, that hypothesis. Indeed, with sufficiently large n, the probability of hypothesis H_0 conditioned on data marginally rejecting H_0 is close to one, whatever the assumed prior.

3. Classical Interpretation

Each of the supposed results leads the auditor to formally "reject" H_0 and thus to take accordant action, possibilities including marginal adjustment of the auditee's stated balance (Arens and Loebbecke (1981) pp.149-50) and, in the other extreme, qualification of the audited accounts (Kinney (1975) p.119). The "official" Neyman-Pearson rationale underlying this orthodox interpretation is that by acting strictly in accord with a predesignated decision rule, in one test after another, the frequencies of errors of types I and II built into that rule will almost certainly be achieved (approximately) over the long-run of applications. For instance, if the minimum power (1−"β risk") of a test procedure is 97.5%, then account balances will be accepted in no more than 2.5% of those test repetitions in which the stated balance is in fact materially incorrect. Note the order of the conditionality here - specifically the figure 2.5% represents the maximum probability of an acceptance given an incorrect balance, not the maximum probability of an incorrect balance given an acceptance. The former of these two probabilities is sometimes translated mistakenly as the latter (cf. de Finetti (1975) p.248).

Ignoring assumptions (e.g., normality), actual error frequencies close to their theoretical or nominal values are "guaranteed" by the law of large numbers, provided, of course, that the practitioner complies in each test with the decision rule established before the test was run. No allowance may be made for the possibly narrow margin by which a particular result rejects or accepts

H_0, or for any countervailing factor such as an apparently "unlucky", unrepresentative or biased-looking random sample. If by occasionally and subjectively setting aside a decision rule stipulated before seeing the data – e.g., by treating a "just-over" rejection of H_0 as an acceptance, or by discounting an unstratified sample which, albeit drawn at random, has the appearance of being unrepresentative or "biased" [2] – the auditor introduces personal judgement and possible statistical bias, the nominal error frequencies attached to that decision rule no longer have the same objective meaning or relevance (cf. Kyburg (1974, p.221); Roberts (1978, p.43)). The "true" error frequencies of such a discretionary procedure are indeterminate, and may be better or worse than their nominal values. It is this latter possibility which underlies the joint tenets in frequentist (i.e., classical) statistics of predesignation and "no looking back".

4. Why a Bayesian Interpretation?

The orthodox or classical approach leads to a decision (choice of action) based *not on an assessment of evidence* contained within the data, but on the long-run average error frequencies ("operating characteristics") of the test procedure (decision rule) by which that choice is made:

> Neyman's school, followed strictly, maintains that there is no such thing as inconclusive evidence for hypotheses. We can only make decisions about hypotheses, following some pattern of decision-making with desirable characteristics. When we decide for an hypothesis, we do not do so because the evidence makes it credible. (Hacking (1973) p.490).

This instrumentalist philosophy is most clearly evident in Neyman's instruction on the interpretation of confidence intervals:

> The specific interval $\hat{\theta}-c<\theta<\hat{\theta}+c$, calculated for a confidence coefficient α, selected by the statistician to suit his purposes, gives an unambiguous answer: *act on the assumption that the unknown θ lies between the limits indicated.* If the consumer asks why he should do so, the answer is: if you behave that way, you will be right (approximately) in 100α per cent of cases. (Neyman (1971) p.80).

By comparison, the Bayesian approach is to think of a hypothesis test not as one of a long and perhaps hypothetical (imaginary) sequence, but as a logical procedure by which to gain information about the parameter θ; and hence to *support with evidence* a decision either to accept or reject the auditee's stated balance. Interpreted this way, hypothesis tests provide for both *inference and decision.* Data X are gathered and the probability distribution of θ conditional on that data, $p(\theta \mid X)$, is calculated. This distribution permits inferences concerning θ in the form, for example, of 95% "credible intervals" (the Bayesian correspondent of orthodox confidence intervals). Decisions might then ensue on the basis of $p(\theta \mid X)$, in conjunction with relevant loss functions, according to the criterion of minimum expected loss.

In basing decisions on a logical evaluation of evidence, $p(\theta \mid X)$, rather than on the hypothetical error frequencies of a test mechanism which might not be used again, the Bayesian paradigm is a model for what auditors purport to do. This apparent congruence between Bayesian methods and the audit process is well recognized (e.g., Beck et al. (1985), Kinney (1975) and Scott (1973; 1975)), but in practice, orthodox (non-Bayesian) statistical methods remain more generally accepted. Part of the reason for this is that Bayesian methods come to a posterior distribution for θ, $p(\theta \mid X)$, by way of Bayes' theorem, i.e., $p(\theta \mid X) \propto p(\theta)\, p(X \mid \theta)$, thus requiring a prior probability distribution for θ, $p(\theta)$. Prior distributions are seen by many as generally subjective (personal), and therefore of doubtful standing within the ideally "scientific" audit process.

This same argument is cited often by researchers in the social sciences when explaining why Bayesian methods have not been given greater application in published empirical research. It is widely conceded, however, that in applications where an "objective" (i.e., empirical, or deduced from theory) prior distribution is available, Bayesian methods should be employed (Kyburg

(1974) p.58). This concession traces to Neyman (e.g., 1957, p.19) and Fisher (e.g., 1973, p.17) who together convinced earlier generations of statistical theorists and applied statisticians of the general virtue, for the purposes of scientific inquiry, of eschewing altogether the methods of inverse probability (Bayesian statistics).

Rather than arguing the legitimacy of the subjectivist approach, or the associated argument that objectivist statistics are an illusion - their subjective aspects are merely "swept under the carpet" (cf. Good (1976), (1981, p.149); Barnett (1975, p.19); Savage (1961, pp.178, 183); (1962, p.53) - this paper takes the position that since Bayesian reasoning is considered the logical ideal, the auditor's intuitive interpretation of statistical results should be aided by a type of "what if" analysis, whereby posterior probabilities are found for H_0 assuming one possible prior and then another, covering as broad a range of possibilities as necessary or desired. By exploring the logical connotations of the data from various possible perspectives or starting points (priors), the auditor can develop intuitive "feel" for the direction and strength of evidence represented, and learn by experience which factors determine these evidential qualities, and the ways in which such relevant factors work and interact.

With some data sets it is found that the class of prior distributions under which a certain inference – e.g., $p(H_0 | X) > p(H_0)$ – remains valid, is so wide as to include practically all those which could possibly be considered reasonable (Edwards et al. (1963) pp.201, 210-11). That is, to come to any other conclusion one would have to manufacture a highly "personal" distribution of prior probability, which, apart from producing the desired posterior, would have no cause for being considered. In these circumstances of "practical objectivity" (inter-subjectivity), where the data tends to dominate or "swamp" the prior, any attempt to justify a rival inference would be hard pressed. To do so would involve, implicitly, either a perverse appeal to a tailor made and clearly inappropriate prior, or, alternatively, the abandonment of Bayes' theorem and of the fundamental probability axiom from which this theorem is deduced (i.e., the multiplication law).

By opting not to apply Bayesian methods, auditors and other practitioners pass up experience in inductive inferential reasoning which would extend and refine their abilities to assess evidence intuitively. As with deductive reasoning (e.g., mathematics), training and experience can greatly enhance human inductive inferential capabilities, particularly since it is known that in some situations data has an "objective" Bayesian interpretation quite disparate, at least in degree, from that usually attributed to it on the methods of analysis and principles of classical statistics (see, e.g., Edwards et al. (1963) pp.221, 225; Berger and Sellke (1987) pp.135-6, 138 and Berger and Delampady (1987) p.318). The results provided below represent a straightforward instance of this sometimes diametric disparity between Bayesian inference and conventional statistical practice.

4. Bayesian Interpretation

In this section posterior probabilities of H_0 are calculated, based on the three observations (marginal rejections of H_0) supposed in Section 2, and assuming various prior distributions for θ. For mathematical simplicity, normal prior distributions are presumed. This allows the use of standard Bayesian calculations for the normal mean (see below). The results described are not, however, dependent on priors of this parametric family. Rather, it is the general location and relative concentration of prior mass which affects results, not the particular mathematical or parametric "shape" of that distribution.

Priors (i) and (ii). The first two priors considered have mean $\mu=0$, prior (i) with variance $v^2=100^2$ and prior (ii) with variance $v^2=5000^2$. For graphical representation of these distributions, see Figure 6 (in appendix). Prior (i) is quite "informative" (low variance), at least compared with (ii) which is highly diffuse or "uninformative". Both (i) and (ii) indicate that the auditor's expected average population error is zero, however (i) implies far greater confidence in θ close to zero than does (ii). Distribution (ii) is very close to uniform and represents extremely vague prior knowledge about θ, thus implying little confidence in θ close to zero, or in any other particular narrow interval of the parameter space. Results of calculations for all priors, including (i) and (ii), are provided in Table 1. The formulae (1) and (2) with which these Bayesian results are calculated are provided below.

Table 1 about here

Note from this table that both for priors (i) and (ii) the posterior probabilities of H_0, $p(H_0 | \overline{X}_c)$, are all very high and much greater than their respective priors, $p(H_0)$, meaning that in each case H_0 is strongly confirmed. It is inferred, therefore, given any one of the supposed results, that the account balance tested is in fact *supported* rather than "rejected". This would be the logical conclusion of one who begins with a prior which either: (i) gives appreciable probability to θ in the null (immaterial) interval $[-m,m]$ relative to alternative θ values, or (ii) is quite diffuse, and therefore lets the data, in a sense, "speak for itself".

Priors (iii) and (iv). For diffuse priors like (ii), the posterior probability of H_0 is highly insensitive to the prior mean (location) μ. To demonstrate this, results are calculated for priors with the same high variance v^2=5000^2 as prior (ii) but with means greatly different from zero. Note that for each of the three supposed marginal rejections $\overline{X}_c$, $p(H_0 | \overline{X}_c)$ remains very high, despite $p(H_0)$ being very low, suggesting, as for priors (i) and (ii), that H_0 is strongly supported.

How Marginal Rejection Supports H_0. How is it that an observed sample mean which rejects H_0 (the auditee's balance) in a classical hypothesis test can actually greatly increase the Bayesian probability of that hypothesis? The answer to this question lies in the inherent conservativism of the "negative" type of hypothesis test, for which minimum power against $|\theta|>m$ is 0.975 or some similarly high figure. To institute such a high probability of rejecting H_0 when θ is barely material, the "critical" value of the sample mean, $\overline{X}_c$, must be set at $m-z_{\beta/2}\sigma/\sqrt{n}$, which is clearly *within* the null (immaterial) interval $[-m,m]=[-75,75]$, even with large n. This means that marginal rejection of H_0 occurs when the observed $\overline{X}$ is inside $[-m,m]$, which, of course, tends to support θ values in that null interval relative to alternative (material) values of the parameter.

Priors (v) and (vi). We now consider priors appropriate when the auditor is confident of an average error size greater than m. Prior (v) has mean $\mu=90$ and variance $v^2=15^2$, and (vi) has $\mu=120$ and $v^2=25^2$. Both (v) and (vi) are informative rather than vague priors, (vi) indicating a higher expected average error than (v), but less confidence in θ about its expected value. Figure 6 (in appendix) depicts these distributions. It is found with both priors, (v) and (vi), that marginal rejection increases the probability of the "rejected" hypothesis – i.e., $p(H_0|\overline{X}_c)>p(H_0)$ – the more so the larger the sample size n (indeed if β is held at 0.05, then for $n=1000$, $p(H_0|\overline{X}_c)$ is about 0.94 for both (v) and (vi)).

Priors (vii) and (viii). The final class of priors considered allows for the unusual situation where the auditor is confident of a negative material error (understatement) in the auditee's stated balance. The priors used are the same as (v) and (vi) but with negative rather than positive means. With both priors (vii) and (viii), $p(H_0|\overline{X}_c)$ is high, even with n as low as 30. This is because the combination of prior mass to the left of the null interval, $[-m,m]$, and an observed mean, $\overline{X}_c$, towards the right of this interval, leads to a posterior distribution concentrated, between the two, mostly within $[-m,m]$. This is particularly so for large n as the weight of sample information pulls the prior mass more and more within the null interval.

Value of $p(H_0|\overline{X}_c)$ with large n. Given a normal population $X\sim N(\theta,\sigma^2)$ and a conjugate, normal prior distribution $\theta\sim N(\mu,v^2)$, the posterior distribution of θ, based on an observed sample mean $\overline{X}$, is normal with mean

$$\mu_1=\mu[\sigma^2/(\sigma^2+nv^2)]+\overline{X}[nv^2/(\sigma^2+nv^2)] \qquad (1)$$

and variance

$$v_1{}^2=\sigma^2v^2/(\sigma^2+nv^2). \qquad (2)$$

These standard Bayesian results (e.g., DeGroot (1986) pp.324-6) are the basis for the calculations tabulated above.

It follows from (1) that with large enough n the posterior distribution for θ will have a mean close to the sample mean $\overline{X}$. Similarly, the value of (2), the posterior variance, which can be written $(1/v^2+n/\sigma^2)^{-1}$, will be approximately σ^2/n, particularly with large prior variance v^2. Using these approximations, the posterior probability of H_0: $|\theta|\leq m$ given a result $\overline{X}=\overline{X}_c$, $p(H_0|\overline{X}_c)$, is, for large n, given by

$$\begin{aligned}
& p(-m\leq\theta\leq\overline{X}_c+z_{\beta/2}\sigma/\surd n \mid \overline{X}_c) \\
=\; & p(\theta\leq\overline{X}_c+z_{\beta/2}\sigma/\surd n \mid \overline{X}_c)-p(\theta<-m \mid \overline{X}_c) \\
=\; & \Phi(z_{\beta/2})-\Phi(-\{m+\overline{X}_c\}\surd n/\sigma) \\
\cong\; & \Phi(z_{\beta/2}) \text{ since } \Phi(-\{m+\overline{X}_c\}\surd n/\sigma)\cong 0, \\
\cong\; & 1-\beta/2,
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (e.g., $\Phi(-1.64)=0.05$, etc.). It is seen from this result that for a two-sided hypothesis test with minimum power against H_1: $|\theta|>m$ of 100(1-β/2)%=97.5%, $p(H_0|\overline{X}_c)$ will, for sufficiently large n, equal 1-β/2=0.975 (approximately), *whatever the prior.* With a uniform prior, the posterior is 1-β/2 for any n.

With informative (i.e., low v^2) priors favoring $\theta\in[-m,m]$, such as (i), $p(H_0|\overline{X}_c)$ may actually be higher than 1-β/2 for small n (see Table 1), but as n increases the information about θ represented in the prior becomes relatively less and less and the posterior probability of H_0 approaches 1-β/2, the value it would have for a uniform or uninformative prior. The same is true of priors with a spike at θ=0, as proposed by Edwards (1994) in comments on this paper.

The protocol of classical hypothesis tests requires that as n increases the more important error probability is allowed to fall from its preset level (e.g., 0.05 or 0.01) toward zero, thereby maintaining parity (i.e., an optimal trade-off) between the two error probabilities, or at least preventing the less important error probability from becoming the lower of the two (e.g., Cox and Hinkley (1974) pp.397-8; Kendall and Stuart (1979) p.197; Lehmann (1986) p.125).[3] Following this fundamental methodological requirement, the auditor must have $\beta\rightarrow 0$ as $n\rightarrow\infty$, the type II ("beta") error being regarded as the more important to avoid. Consequently,

$$\lim_{n\rightarrow\infty} \quad p(H_0|\overline{X}_c) \cong 1\text{-}\beta/2 \rightarrow 1,$$

and thus, for large enough n, H_0 is rejected by a result $\overline{X}_c$ when, whatever the prior, its Bayesian posterior probability is arbitrarily close to one.

This is similar to, but not the same as, "Lindley's paradox", examples of which, usually applying to point ($\theta=\theta_0$) rather than interval ($\theta_1\leq\theta\leq\theta_2$) null hypotheses, are well known in statistics.[4] Johnstone (1990) provides references on this paradox and example calculations. See also Berger and Mortera (1991) and Johnstone and Lindley (1993), and in accounting literature, Christie (1990, p.80). Contrary to what is found above, it has been considered previously that classical tests of *interval* null hypotheses are immune to problems of the like of Lindley's paradox (e.g., Casella and Berger (1987) p.133).

For one hoping to avoid any qualitative inconsistency between the classical and Bayesian interpretations of a result $\overline{X}=\overline{X}_c$, it is possible to engineer a prior distribution so as to inhibit or at least delay $p(H_0 | \overline{X}_c)$ increasing with n. This can be achieved by putting a "spike" of prior probability at θ marginally greater than m, but any such prior distribution would be obviously forced and unacceptable. For more reasonable priors, $p(H_0 | \overline{X}_c)$ tends to increase smoothly and rapidly as n increases. The intuitive reason for this is that the greater n, the greater the support offered by the sample for θ equal to its own mean, $\overline{X}_c \in [-m, m]$. Inevitably this volume of data will "swamp" relatively informative prior distributions like (v) and (vi), and so too even the most tendentious prior such as one with a spike at $\theta = m + \delta$ (δ small).

5. Apparent Incoherence

If $p(H_0 | \overline{X}_c)$ is very high, perhaps close to one, whatever the (reasonable) prior, a Bayesian auditor having observed $\overline{X}=\overline{X}_c$ may well accept H_0, and hence the auditee's balance, without alteration or further inquiry. From a conventional viewpoint, however, H_0 is "rejected" and some accordant action is necessary. If this action is merely a marginal adjustment to the stated balance, so as to bring it into the formal acceptance region of the test, then the practical cost and inconvenience to either auditor or auditee may not be much. Other possible responses, particularly that of adjusting the auditee's stated account balance to its point estimate, or perhaps some qualification of the accounts, will usually be of greater consequence.

The auditor's problem in these circumstances is not necessarily one of significant needless costs, but of justifying a decision procedure which formally "rejects" a balance with high probability of being materially correct, and yet in other, smaller tests, "accepts" balances which, subjectively at least, have appreciably lower probabilities of being correct in this sense. Consider, for example, the further possible result (n=30, $\overline{X}$=14.1), which yields a 95% confidence interval for θ, [-46.734, 74.934], lying wholly, albeit marginally, within [-m,m]=[-75,75], and therefore leads the auditor to accept H_0. Posterior probabilities of H_0 based on this assumed result are compared in Table 2 with those given marginal rejection of H_0 in the test with sample size n=90.

Table 2 about here

Depending on the prior used, $p(H_0 | \overline{X})$ is either the same (approximately) or lower for the (n=30, $\overline{X}$=14.1) marginal acceptance of H_0 as for the larger sample result (n=90, $\overline{X}=\overline{X}_c$=39.8776) which formally rejects H_0. Moreover, with priors (v)-(viii), according to which the auditor needs some strong reassurance before being persuaded from a contrary belief, the posterior probability that the stated balance is materially correct is much higher on the larger (n=90) test which rejects H_0 than on the smaller (n=30) test which accepts that hypothesis. Indeed, for the n=30 result, $p(H_0 | \overline{X})$ is only .481 given prior (v) or .432 with prior (vi), H_0 being *accepted,* whereas for the n=90 result, the minimum value of $p(H_0 | \overline{X}_c)$ across all priors (i)-(viii) is .688, and yet here H_0 is *rejected.* These outcomes appear anomalous and indicate that an intuitive or "common sense" evidential interpretation of conventional test results (e.g., "reject H_0" implies strong evidence against H_0) can be very misleading.

6. How Important is the Prior?

The paradoxical finding of Section 5, that the n=90 marginal rejection of H_0 implies at least the same weight of evidence in favor of that hypothesis as the

n=30 acceptance, is highly insensitive to the choice of prior, and is, therefore, quite "objective". Consider the difference between posterior probabilities:

$$p(H_0 \,|\, \overline{X}=14.1,\ n=30) - p(H_0 \,|\, \overline{X}=\overline{X}_c=39.8776,\ n=90), \qquad (3)$$

which is represented in Figures 1 and 2, in the forms of surface and contour plots, as a function of prior parameters μ and ν. The contour plot (Figure 2) is a "topographical map" of the surface plot, darker shades representing bigger negative values of the difference defined by (3).

Figures 1 and 2 about here

The maximum value of (3) is only .0018, this difference occuring with prior parameters μ=-5.04 and ν=51.60, on which the posteriors are .9927 (n=30) and .9909 (n=90). Apart from a class of normal priors with about this same mean μ, all yielding posteriors near one for both the data, the difference defined by (3) is negative, the posterior based on the n=90 rejection being the higher, and clearly so for a large class of priors with ν between about 15 and 65. It is seen, therefore, at least for normal priors, that the posterior probability of H_0 is either practically the same, or higher, given the n=90 rejection as that conditioned on the n=30 acceptance.

The robustness or "objectivity" of this finding can be tested further by searching for priors of any form which might upset it. If these are hard to find, or unreasonable on other grounds, it should be acceptable, even to those who would not normally take a Bayesian approach, that the conclusion reached is not "subjective" or personal in the sense that it holds only for a narrow subclass of possible priors.

Since by Bayes' theorem $p(\theta \mid \overline{X}) \propto p(\theta)\, p(\overline{X} \mid \theta)$, the starting point when looking for priors which might make the difference between posteriors defined in (3) more than negligibly greater than zero is to compare the likelihood functions,

$$p(\overline{X} \mid \theta) = \exp[-n(\theta - \overline{X})^2/2\sigma^2],$$

of the two results. These are plotted in Figure 3.

Figure 3 about here

An informal approach to maximizing (3) is to distribute $p(\theta)$ *jointly between* values of θ within H_0 (i.e., $\theta \in [-75,75]$) for which the likelihood of the n=30 result is higher than that of the n=90 result,[5] *and* θ values outside H_0 where the opposite is true. As shown in Figure 3, the likelihood functions intersect at θ=30.442 and θ=75.091, which means that $p(\theta)$ must be concentrated on the sub-null interval $-75 \leq \theta < 30.442$, or better still on θ within this interval for which the difference between likelihoods is greatest (i.e., around θ=1.49), and on the alternative θ values $75 < \theta \leq 75.091$. There can be little or no prior mass on the intervening θ values (i.e., on the remainder of the null interval, $30.442 < \theta \leq 75$) because in this interval the n=90 likelihood is much the larger. Similarly, $p(\theta)$ cannot be large for alternative θ values other than those in the extremely narrow interval between 75 and 75.091, since outside this tiny segment of H_1 the n=30 likelihood is the higher of the two.

Note that prior distributions with only one of the two suggested modes or points of concentration will not lead to (3) being appreciably greater than zero. In particular, with priors massed only within $-75 \leq \theta < 30.442$, both posteriors will be approximately one. This has been seen already with the prior $N(-5.04, 51.60^2)$ which maximizes (3) for normal priors, the posteriors here both being greater than .99. Similarly, for priors which are essentially spikes in $75 < \theta \leq 75.091$, both posteriors will be zero or approximately zero. For example, assuming the prior distribution $\theta \sim N(75.0455, .03^2)$, which gives $p(H_0)$=.0647, the respective posteriors are .0651 (n=90) and .0649 (n=30).

These considerations reveal that to make the difference between posterior probabilities represented by (3) more than negligibly greater than zero, $p(\theta)$ must be bi-modal with a spike or sharp peak in the tiny interval $75<\theta\leq75.091$. A prior of this form is hardly likely to be considered reasonable - indeed any inference hinged on such a prior could not be taken seriously.[6] Here we have used the Bayesian device of excluding some particular inference on the basis that the prior required to come to such a conclusion is unsustainable. Any inference which is not explicitly Bayesian is liable to be "found out" with this form of hypothetico-deductive refutation.

7. No Possible Reconciliation

An auditee obliged to alter a stated balance when the evidence supporting that balance is strong - i.e., when $p(H_0 | \overline{X})$ is high over a wide class of priors - might attribute this imposition to the auditor's innate and justifiable conservativism. The perception of the auditor as one who requires almost deductive (100%) confirmation of H_0 before accepting the auditee's balance is appealing, yet cannot be sustained once an instance of H_0 being accepted when indeed its posterior probability is not high (for at least a subclass of reasonable priors) is seen to occur.

A possible orthodox defence of such apparently inconsistent behavior by the auditor is that the acceptance of H_0 when $p(H_0 | \overline{X})$ is relatively low might be explained by a difference in error costs between that test and any larger test in which H_0 is rejected when $p(H_0 | \overline{X}_c)$ is relatively high. That is, there might possibly be a reconciliation between the two results once error costs are considered, particularly since one test uses smaller n than the other, this in itself suggesting a lesser error cost, for one or other error type (or both), in the test with lesser *n*.

To examine this possibility, compare again the tests with sample sizes *n*=30 and *n*=90 discussed above. The power functions, $p(Reject\ H_0 \mid \theta,\ n)$, of these two tests are shown in Figure 4.

Figure 4 about here

Both tests are "negative" in that each has preset minimum power against $|\theta|>m$ of 97.5% (cf. Duke et al. (1982) p.51). The power functions of the tests over $|\theta|>m$ are practically equal, both tests giving very high priority to errors of type II (incorrect acceptance of H_0). Clearly, however, the smaller (*n*=30) test has a higher probability of rejecting H_0 when $|\theta|\leq m$; i.e., a higher probability of type I ("alpha") error (incorrect rejection of H_0).

The underlying difference between the two tests implied by this comparison is that in the test with smaller *n* there is less cost associated with incorrectly rejecting H_0 (the auditee's balance); errors of this type can therefore be tolerated more frequently. But this difference does not explain why H_0 is accepted in the smaller test with, for priors such as (v)-(viii), relatively low posterior probability, for if anything, a lower cost of incorrectly rejecting H_0 would suggest, *ceteris paribus*, that H_0 might have relatively high probability, rather than low probability, before being accepted.[5]

It appears, therefore, that the decisions compared, despite both being "by the book" from the classical viewpoint, do not have a consistent evidential basis. Rather, one hypothesis with either practically the same or smaller probability than another (depending on the prior) is accepted when the other, better supported, hypothesis is rejected, the only difference between the two tests being that the cost of a false rejection of the hypothesis tested is higher in the test where that hypothesis *is* rejected than in the one where it is not. Bayesians refer to such logical inconsistencies as "incoherence" and regard statistical coherence, or *mutually consistent* decisions and inferences, as the hallmark of rationality and the fundamental objective of mathematical reasoning (Lindley (1972) pp.3-10).

8. Possible Cross-Subsidies

Having observed a marginal rejection of H_0 in circumstances where $p(H_0 | \overline{X}_c)$ is "objectively" (independent-of-prior), or at least arguably, high, the auditor is left in an invidious position. Other evidence aside, she must either comply with her predesignated decision rule, reject H_0, and proceed to negotiation and perhaps further sampling, thus adding to the cost of the audit, or she can rely on a Bayesian interpretation, introducing prior beliefs formally into the analysis, and infer that the auditee's balance is acceptable. Either way she might be criticised. If she holds to conventional practice and rejects the stated balance, it could be argued that she has inflexibly and illogically overserviced the client, thereby giving rise to unnecessary costs, including perhaps those associated with lowering the auditee's reported net income. On the other hand, if she fails to act on a marginal rejection in qualitatively the same way as on a more extreme rejection in which the observed $\overline{X}$ is much outside $[-m,m]$, she will have introduced an element of personal discretion or subjectivity and cast aside the mathematical surety of specified error frequencies over the long-run. From this frequentist point of view, she is obliged to overlook considerations which appear relevant (subjectively) within the circumstances and requirements of a particular single test, so as to ensure (mathematically) good results *on average* over many tests. Note here the following remarks of Beck et al. (1990) pp.173-4:

> While such *ex post*-to-sampling modifications are permissible (even necessary) within a Bayesian framework, we recognize that they can be viewed as fundamentally inconsistent with classical hypothesis testing. Under classical hypothesis testing, the decision maker (auditor) precommits to reject the null hypothesis under specified conditions. Provided that the underlying assumptions are satisfied, precommitment enables the decision maker (auditor) to control statistically the frequency of inferential errors [incorrect decisions] associated with the sampling/estimation process. (square brackets mine)

If she abides strictly by her predesignated decision rule, an auditor can defend her behavior (decision) in a particular single test by appeal to the theoretical error frequencies of that rule of "inductive behavior" (Neyman's term) over a long run of different tests and different audits and auditees. Taking such an approach, all individual tests in a sequence of such tests conducted by a given auditor (or firm) are equally defensible, but by being treated as merely one of a sequence, those tests in which rejection of H_0 is only marginal give rise to costs higher than might have been, had such tests been treated severally (i.e., without reference to any other test, or to any envisaged sequence of tests). This implies a form of cross-subsidy, some clients bearing higher than necessary costs so as to allow the auditor to instantiate strict compliance with a forestated decision rule, thereby cementing the theoretical defence of her similarly mechanistic, precommited practice in the cases of other clients, past and future.

9. Confidence Intervals Rather Than Hypothesis Tests

By attributing any rejection of H_0, marginal or otherwise, the same qualitative meaning regardless of the sample size and other evidentially relevant factors, users of classical hypothesis tests make decisions without a consistent rational basis. To avoid such incoherent behavior, it would help if decisions were based on confidence measures rather than hypothesis tests. Consider, for example, the orthodox 100(1-β)% confidence interval for θ, assuming the model $X \sim N(\theta,\sigma^2)$, that is $\overline{X} \pm z_{\beta/2}\sigma/\sqrt{n}$. This interval is also the Bayesian maximum density credible interval for θ, provided either (*a*) the prior is at least locally uninformative,[8] or (*b*) *n* is large enough that the prior is in effect uninformative. Under these circumstances, it can be held *legitimately* that the probability that θ lies in $[\overline{X}-z_{\beta/2}\sigma/\sqrt{n}, \overline{X}+z_{\beta/2}\sigma/\sqrt{n}]$ is approximately $1\text{-}\beta$. From here, it is a straightforward and natural mathematical extension to calculate the probability, or level of inductive confidence, applying to the null interval H_0: $\theta\in[-m,m]$.

The mathematical correspondence between orthodox and Bayesian confidence measures for sufficiently large *n*, or where the prior assumed is suitably uninformative, affords orthodox results *quasi*-Bayesian status. By observing the confidence coefficient applying to the θ-interval H_0: $\theta\in[-m,m]$, rather than merely whether H_0 is formally "accepted" or "rejected" at a sometimes quite arbitrary "critical" level, the auditor is able to make a judicial (evidence-based) rather than mechanical (rule-based) decision to either accept or reject the auditee's stated balance. Also, there is the possibility of incorporating other, perhaps extra-statistical, considerations into the decision process, as is normally imperative.

This argument is consistent with remarks of Loebbecke and Neter (1975, pp.39-40) on the respective uses in practical situations of hypothesis tests and confidence interval estimates:

> While there is generally a direct connection between the testing and [confidence interval] estimation approaches, the important distinction is in their uses. When a testing approach is utilized, a decision is made on the basis of a particular sample result. With the estimation approach, on the other hand, information about the magnitude of the characteristic of interest is obtained without leading directly to a decision. A decision may, indeed, be made after a number of characteristics have been estimated or after additional nonquantitative information has been considered; however, a decision is not made on the basis of one sample result alone. (square brakets mine)

These comments appear to support an approach based on an inductive or *inferential* interpretation of confidence intervals, as opposed to the strictly decision-theoretic approach of classical hypothesis testing, according to which each point X in the sample space translates without any possible subjective interpretation into a decision (action) - either the null hypothesis is rejected or accepted, not on any consideration of the strength of evidence, but simply according to whether that value of X falls on one side or the other of a predesignated partition or "critical" value, here denoted by X_c.

An intrinsic logical strength of confidence intervals over hypothesis tests is that as the sample size increases, or as sampling variation decreases (perhaps through better experimental design; e.g., stratification) confidence intervals tend to become "more Bayesian" and hypothesis tests "less Bayesian". In an auditing context, the divergence, as n increases, between the results of orthodox hypothesis tests and Bayesian inference is apparent in the *Lindley-like* paradox seen above (Section 4). On the other hand, because orthodox and Bayesian confidence levels tend to converge as n increases, the prior becoming relatively less and less informative, orthodox confidence intervals have more general relevance as the sample size becomes larger. Specifically, levels of confidence take on the meaning which users commonly give and require of them - that is, they become broadly interpretable as measures of inductive probability, or degrees of certainty, applying to the proposition that θ lies within the stated interval. This type of probability statement is strictly inadmissible within the strictures of orthodox (Neyman-Pearson) statistics (Arens and Loebbecke (1981) p.115), but is sanctioned by Bayesian theory, for a broader and broader spectrum of priors as n increases.

10. Conclusion

The evidential assessment of classical levels of significance levels is highly problematic,[10] and requires consideration of factors not conventionally considered relevant *ex post* (i.e., after the test has been run), including particularly the sample size n. In general, data of a fixed significance level (say 5%) provide evidence more and more *in favor* of H_0 the larger n (although without explicit Bayesian calculations a precise measure of support cannot be stated). Put another way, the posterior probability of H_0 given data X_c marginally significant at a given "critical" level, $p(H_0 | X_c)$, tends to increase with *n*, and in the auditors "negative" test structure, is shown to approach one, whatever the prior. As a consequence, orthodox conclusions based on results about X_c are generally internally inconsistent across tests of different sample sizes. Specifically, a result of about X_c with large *n*, marginally *rejecting* H_0, provides practically the same or greater evidence in favor of H_0 (depending on the prior) than a smaller sample result, also about X_c but marginally *accepting* that hypothesis. Hence the suggestion above of statistical or evidential "incoherence".

The practical consequences of such logical inconsistencies in conventional statistical decision making in auditing include possible cross-subsidies, where in commiting to the Neyman-Pearson, frequentist protocol, and its justification in terms of long-run average error frequencies, marginal rejection of a particular auditee's stated balance gives rise to costs which would not have been incurred had the same test result been viewed from a Bayesian (single-case) rather than classical (long-run average) standpoint. To avoid such unwarranted costs, it would help if auditors were made aware of the evidential relevance of the sample size from both *ex ante* and *ex post* perspectives. Pre-test (*ex ante*), the larger *n* the better, more information always being preferred to less (apart from the costs of sampling), irrespective of whether the logic employed is classical or Bayesian. However, post-test (*ex post*), marginal rejection of H_0 with larger *n*, rather than representing more reliable or stronger evidence against H_0, tends to more *support* that hypothesis. This is contrary to popular intuition (cf. Nelson et al. (1986) p.1301).

Unfortunately, it is usually the case that orthodox textbooks, excepting those cited in Section 4, make no comment on the *ex post* relevance, or otherwise, of *n*. Because of this perhaps, there is quite a common view that the sample size is irrelevant *ex post.*[11] This view holds only if users of hypothesis tests are obliged to maintain a strictly *aevidential,* frequentist-decision-theoretic (mechanistic) interpretation of tests, as promulgated by Neyman and Pearson, although subsequently, it is argued (Johnstone (1987)), recanted by both.

Following standard statistical textbooks, most expositions of hypothesis testing in auditing are ostensibly frequentist-decision-theoretic in philosophy, but difficulties arise when the results of such tests have to be aggregated with other (evidential) considerations so as to allow a broadly-based and defensible decision. Under obligation to provide a meaningful in-the-single-case (to client's, courts etc.) interpretation of orthodox results, even the most technically careful theorists occasionally lapse into an evidential, and therefore technically heretical, statement of results. For example, Arens and Loebbecke (1981, pp.133-34), having at first denied the legitimacy of any such interpretation (see above), make the strictly improper (or, alternatively, Bayesian) statement that the confidence coefficient attached to an orthodox confidence interval is the probability that the parameter estimated lies within that interval:

> The concept of the confidence level is the same for variables sampling as it is for attributes. It is a statement of the probability that the true population error value actually falls within the limits of the confidence interval.

This is an intuitive but unwarranted response to the question raised by Neyman (see Section 3) of why users might justifiably decide to act on the presumption that the unknown parameter lies within the estimated interval. Similarly, Loebbecke and Neter (1975, p.40) fall into an apparently inductivist, contra-frequentist interpretation of a test result as substantiating or confirming a stated account balance:

> When an auditor is examining an account by means of a single auditing procedure and the audit objective is to *substantiate the correctness of the account balance,* the testing approach may be the appropriate one. (my italics)

Neyman (e.g., 1950, pp.259-60) specifically disallows any intuitive, inferential understanding of the terms "accept" and "reject". On his strict, but received,[12] frequentist view, the result of a hypothesis test is merely an *automated decision* to act either as if H_0 were true (e.g., to accept the clients balance) or as if H_1 were true (Neyman (1976) pp.750-1). This action is taken not on the basis of the "evidence" in any sense, but because a predesignated decision rule (hypothesis test), with inbuilt low error frequencies, says to do so:

> Neyman-Pearson tests are presented as a kind of recipe... One simply fixes the size α of a test, finds the the most powerful test having a given size and then accepts or rejects. If asked why anyone should find it desirable to do this the rationale given by Neyman is this: If one 'behaves' in this way one will incorrectly reject h_0 not more than 100α percent of the time and incorrectly accept h_0 not more than 100β percent of the time, (for α, β the probabilities of type I and type II errors respectively). (Mayo (1981) p.196)

> To emphasise the theoretical difference between their program and the Bayesian one, Neyman-Pearson relied on the ordinary language meaning of such terms as *behavior* (in the technical term *'inductive behavior'*), as opposed to *inference*, and the characterization of their program as fundamentally *'decision theoretic.'* (Seidenfeld (1979) pp.14-5)

To some, Neyman's injunction against an inferential or cognitive interpretation of test results appears philosophically pedantic and overly "theoretical", but in fact his position is justified, indeed necessitated, by the logical deficiencies of classical methods when interpreted as methods of inference (rather than as mere decision rules), including, as discussed in this paper, their inherent evidential incoherence. Detailed support for this dismissal of Neyman-Pearson theory as a normative framework for inference (belief revision) in auditing, is provided in Johnstone (1994).

Auditing is commonly regarded as an inductive, evidential or judicial process (e.g., Arens and Loebbecke (1991) p.2), thus raising the issue of whether the application of statistical procedures which, on the express admissions of their proponents, are without evidential content, can be justified. Adding to this infirmity, the argument provided herein reveals that observed results close about the "critical" accept/reject partition in conventional hypothesis tests can often result in reversed decisions (actions) when the auditor takes an evidential (here meaning Bayesian) perspective, rather than the classical frequentist-decision-theoretic view. Because practical audit decisions are bound to be affected, the logical and philosophical foundations of classical hypothesis tests in auditing should be re-evaluated. If it is found, as suggested above, that there is a fundamental incompatibility between conclusions of the genre offered by this paradigm and those which auditors require, the apparent tendency of both theorists and practitioners to interpret hypothesis tests and confidence intervals in ways specifically disallowed by orthodox statistical texts will be better understood, and to some extent, vindicated.

11. Postscript: A Further Example

In comments on this paper, Aldersley (1994) replicated the calculations of Section 5 in the case of another test, more common in auditing practice. The test considered is that in attribute sampling of the population proportion ρ, where the number of errors, *k*, observed in a sample of size *n* has a binomial distribution with parameter ρ ($0 \leq \rho \leq 1$) and index *n*. The auditor tests the null hypothesis

$$H_0: \rho \leq m$$

against its alternative (complement)

$$H_1: \rho > m,$$

where m=0.15 (say) is the level at which the relative frequency of errors in the population is deemed material. Two possible test results are considered:

$$(n{=}30,\ k{=}1) \quad \rightarrow \quad \text{UCL} = .149$$
$$(n{=}!00,\ k{=}9) \quad \rightarrow \quad \text{UCL} = .152.$$

The classical decision rule (Arens and Loebbecke (1981) pp.78-9) is to reject H_0 if and only if the (one-sided) upper confidence limit [UCL] for ρ exceeds

m=0.15. On this rule, H_0 is accepted in the n=30 test and rejected in the n=100 test. It is found however that the Bayesian posterior probability of H_0 is as high or higher on the n=100 rejection as for the n=30 acceptance, again contra-indicating the classical results. Note that the decision rule described constitutes a "negative" test, since its effect is to fix the power function with respect to $\rho=m$. (The power of the test against other possible values of ρ depends on the sample size, n.)

The Bayesian calculations are as follows. Assuming a beta (conjugate) prior with parameters a and b $(a,b>0)$, the posterior distribution of ρ is also beta, with parameters $a+k$ and $b+n-k$, k denoting the number of errors observed in n trials (DeGroot (1986) pp.321-2). The posterior probability of H_0 is given then by the normalized incomplete beta function

$$\{\Gamma(a+b+n) / [\Gamma(a+k)\ \Gamma(b+n-k)]\} \int_0^{.15} \rho^{a+k-1}\ (1-\rho)^{b+n-k-1}\ d\rho. \qquad (4)$$

Letting each of the prior parameters a and b take any value between 0 and 50, thus allowing for the broadest possible class of prior distributions, the posterior probability of H_0 is found from (4) for each of the two sample observations supposed above. The difference between these posteriors

$$p(H_0\,|\,k=1,\ n=30) - p(H_0\,|\,k=9,\ n=100),$$

defined as a function of the prior parameters a and b, is plotted in Figure 5. It is seen from this figure that as in the test of the normal mean, the large sample rejection of H_0 leaves that hypothesis with at least the same posterior probability as does the smaller sample rejection, regardless of the values chosen for a and b. Furthermore, by the same argument as discussed above in the case of the test of the normal mean, it is not possible to reconcile these apparently anomalous results through consideration of the two tests' implicitly different error costs.

Figure 5 about here

12. Appendix

Figure 6 is a graphical representation of the prior distributions numbered (i), (ii), (v) and (vi) in Section 4 above.

Figure 6 about here

To illustrate the results of Bayesian analysis, Figure 7 shows the posterior distribution corresponding to each of the four numbered priors, assuming the possible observation (n=90, $\overline{X}$=39.8776).

Figure 7 about here

It can be seen by comparing Figures 6 and 7 that although the four priors represented are greatly different, the resulting posteriors are not much different. Such comparison of posterior distributions should be regarded as a method of sensitivity analysis, through which the auditor can assess the weight of sample evidence by noting the extent to which the conclusion or decision prompted by that evidence depends on the assumed prior. In cases of sufficiently informative data, the prior is practically irrelevant. Conversely, if the prior chosen makes a great difference, the data have relatively little information content.

Footnotes

1. By adding an arbitrarily small number δ to $\overline{X}=\overline{X}_c$, the resulting confidence interval, $(\overline{X}_c+\delta)\pm z_{\beta/2}\sigma/\surd n$, would lie partly outside $[-m,m]$ and hence H_0 would be rejected unambiguously. However, the results which follow, including particularly Bayesian posterior probabilities, would be unchanged to any required number of decimal places. For this reason, we can think of $\overline{X}_c$ as practically $\overline{X}_c+\delta$, and hence of $\overline{X}_c+z_{\beta/2}\sigma/\surd n$ as marginally greater than m, therefore rejecting H_0.

2. An "unlucky" random sample might include only or mostly very small accounts, or accounts which have not been transacted for a long period, therefore probably misrepresentating the broader population. A commonly overlooked advantage of stratification is that, with selection of appropriate strata, such samples cannot occur, and hence not only is there generally a reduction in sampling variation, and hence an increase in power, but also the practitioner goes some way towards avoiding what logicians (e.g., Seidenfeld (1979) pp.15, 55) have called "problems of the backward look"; see Johnstone (1988; 1989) for further explanation and references.

3. Note that the largest of the particular results considered has n of only 90. This test has minimum type I (less important) error probability, $\beta(\theta=0)=.026$, greater than its maximum type II (more important) error probability, $\beta(|\theta|=m)=.025$, hence satisfying the error probability priority requirement discussed above (the complete power function, $1-\beta(\theta)$, of this test is shown in Figure 4).

4. The Lindley paradox requires increasing n, but the limit of $p(H_0|\overline{X}_c) \cong 1-\beta/2$ found in this paper holds for uniform priors whatever the value of n, and depends on increasing n only in that it then becomes "objective", or, in other words, independent of the chosen prior (cf. Johnstone (1993)).

5. That is, $p(\overline{X}=14.1\,|\,\theta, n=30) > p(\overline{X}=39.8776\,|\,\theta, n=90)$.

6. The adage that it is hard to tell good fish but easy to tell bad fish applies also to prior probability distributions. Interestingly, although of little practical relevance, the prior which maximizes (3) is that with two spikes: $p(\theta=1.49)=p(H_0)=.324$ and $p(\theta=75+\delta)=p(H_1)=.676$. This leads to $p(H_0|\overline{X})$

equal to .752 (n=30) and .248 (n=90), and hence (3) equals .504. Use of such a prior entails a test of two point hypotheses. Also of "theoretical" interest only, the minimum value of (3) is -.997, arising with μ= -19436 and ν=316.819. Equally large negative differences arise for large positive μ.

7. What else implied by the power functions could rationalize the auditor's decisions? In the smaller test $p(H_0)$ might be relatively low, in which case a relatively high probability of type I error could be allowed. But a low prior on H_0 would reduce rather than increase $p(H_0 | \overline{X})$, thus not reconciling the two results. Or, looking at the larger test, its low type I error probability might be explained by a high prior $p(H_0)$, which would increase its $p(H_0 | \overline{X})$, therefore similarly accentuating the apparent inconsistency.

8. A prior is called "locally uninformative" or "locally uniform" if it is both (*a*) fairly flat in the θ-interval where the likelihood function is high, and (*b*) not large outside this interval (Box and Tiao (1973) p.23). Edwards et al. (1963, p.201) discuss reliance on such priors under the heading "Stable Estimation".

9. The following comments of Edwards et al. based in part on Lindley's finding (i.e., that for a point null hypothesis with non-zero prior probability, marginal rejection at any given significance level, no matter how small, tends to support that hypothesis strongly as n becomes sufficiently large) applies equally to the hypothesis tests used in auditing: "...evidence that leads to classical rejection of the null hypothesis will often leave a Bayesian more confident of that same null hypothesis than he was to start with." (Edwards et al. (1963) p.240)

10. In papers by Berger and Sellke (1987, pp.135-6) and Berger and Delampady (1987, pp.317-8), this difficulty is discussed as one of "calibration". The problem in these terms is to calibrate the statistical significance scale onto an evidence scale, or, in other words, to transform significance levels into measures of evidence. This transformation involves n, but other factors are relevant also, particularly the population variance σ^2, and, as revealed by Casella and Berger (1987*a*), the broad class of procedure in question (e.g., whether H_0 is a point, or interval such as $\theta \leq m$.). The general conclusion of the Berger et al. studies is that evidential calibration of significance levels presents such difficulties that even with the intuition developed with experience users cannot easily "learn to interpret" these conventional measures in terms of evidence. On the assumption that an *evidential* interpretation of tests is that which most users

require, Berger and Sellke (p.136) conclude in regard to statistical significance tests that "the future of the concept statistics is highly questionable".

11. Many users of statistical tests believe that because the sample size is taken into account in the "standardization" of the observed sample result - e.g., in the calculation of $z(\overline{X})=(\overline{X}-\theta_0)\sqrt{n}/\sigma$ - the resultant decision to either "accept" or "reject" H_0 has meaning independent of that subsumed n.

12. Kempthorne (1976, p.765) contends that virtually all statistical texts adopt Neyman's mathematical and philosophical framework. Also recognizing Neyman's profound influence on the theory and practice of statistics, his colleagues Lecam and Lehmann (1974, p.vii) make the following comments: "Neyman's publications span a period of fifty years. His early work has become so thoroughly part of the common statistical consciousness that it is now only rarely referenced and is no longer conceived as an individual contribution."

References

Akresh, D. A., Loebbecke, J.K. and W.R. Scott 1988. Audit Approaches and Techniques in Abdel-Khalik, A.R. and I. Solomon (eds) *Research Opportunities in Auditing: The Second Decade.* American Accounting Association: Auditing Section.

Aldersley, S. (1994) Comments. *Presented at the 12th Annual USC/Grant Thornton Audit Judgement Symposium,* Oxnard, California.

Arens, A.A. and J.K. Loebbecke. 1981. *Applications of Statistical Sampling to Auditing.* New Jersey: Prentice Hall.

_____ 1991. *Auditing: An Integrated Approach.* 5th edn. New Jersey: Prentice Hall.

Barnett, V. 1975. *Comparative Statistical Inference.* London: Wiley.

Beck, P.J., Solomon, I. and Tomassini, A. 1985. Subjective Prior Probability Distributions and Audit Risk. *Journal of Accounting Research.* (Spring) 23: 37-56.

Beck, P.J., Roberts, D.M. and I. Solomon. 1990. Discussion of A Reevaluation of the Positive Testing Approach in Auditing. *Auditing: A Journal of Theory and Practice* (Supplement) 9: 167-175.

Berger, J.O., and M. Delampady. 1987. Testing Precise Hypotheses (with discussion) *Statistical Science.* 2: 317-52.

Berger, J.O., and J. Mortera. 1991. Interpreting the Stars in Precise Hypothesis Testing. *International Statistical Review.* 59: 337-53.

Berger. J.O., and T. Sellke 1987. Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence (with discussion). *Journal of the American Statistical Association.* 82: 112-39.

Box, G.E.P. and G.C. Tiao. 1973. *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

Casella, G., and R.L. Berger. 1987a. Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association.* 82: 106-111.

_____ 1987b. Rejoinder in Berger. J.O., and T. Sellke Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence. *Journal of the American Statistical Association.* 82: 133-35.

Christie, A. 1990 Aggregation of Test Statistics: An Evaluation of the Evidence on Contracting and Size Hypotheses. *Journal of Accounting and Economics.* 12: 15-36.

Cox, D.R., and D.V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman and Hall.

de Finetti, B. (1975) *Theory of Probability.* Vol. 2. London: Wiley.

DeGroot, M.H. 1986. *Probability and Statistics.* 2nd edn. Reading, Mass.: Addison-Wesley.

Duke, G.L., J. Neter and R.A. Leitch. 1982. Power Characteristics of Test Statistics in the Auditing Environment: An Empirical Study. *Journal of Accounting Research.* 20 (Spring): 42-67.

Dworin, L. and R.A. Grimlund. 1986. Dollar-Unit Sampling: A Comparison of the Quasi-Bayesian and Moment Bounds. *The Accounting Review.* 61 (January): 36-57.

Edwards. W. 1994. Number Magic, Auditing Acid, and Materiality -- A Challenge for Auditing Research. *Paper presented at the 12th Annual USC/Grant Thornton Audit Judgement Symposium,* Oxnard, California.

Edwards, W., Lindman, H. and L.J. Savage. 1963. Bayesian Statistical Inference for Psychological Research. *Psychological Review.* 70 (May):193-242.

Fisher, R.A. 1973. *Statistical Methods and Scientifiic Inference.* 3rd edn. New York: Hafner Press.

Good, I.J. 1976. The Bayesian Influence, or How to Sweep Subjectivism Under the Carpet. In Harper, W.L. and C.A. Hooker, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* Vol. 2. Dordrecht, Holland: D. Reidel. Pp. 125-74.

______ 1981. Some Logic and History of Hypothesis Testing in J.C. Pitt (ed.) *Philosophy in Economics.* Dordrecht, Holland: D. Reidel.

______ 1983. *Good Thinking: The Foundations of Probability and its Applications.* Minneapolis: University of Minnesota Press.

Grimlund, R.A. and W.L. Felix. 1987. Simulation Evidence and Analysis of Alternative Methods of Evaluating Dollar-Unit Samples. *The Accounting Review.* (July) 62: 455-479.

Hacking, I. 1973. Propensities, Statistics and Inductive Logic. In Suppes, P., Henkin, L., Joja, A. and G.C. Moisil (eds) *Logic, Methodology and Philosophy of Science IV.* Amsterdam: North Holland. Pp. 485-500.

Jeffreys, H. 1961. *Theory of Probability.* 3rd ed. London: Oxford University Press.

Johnstone, D.J. 1987. On the Interpretation of Hypothesis Tests Following Neyman and Pearson. In Viertl, R. (ed.) *Probability and Bayesian Statitsics.* New York: Plenum. Pp.267-77.

______ 1988. Hypothesis Tests and Confidence Intervals in the Single Case. *The British Journal for the Philosophy of Science.* 39: 353-360.

______ 1989. On the Necessity for Random Sampling. *The British Journal for the Philosophy of Science.* 40: 443-457.

______ 1990. Sample Size and the Strength of Evidence: A Bayesian Interpretation of Binomial tests of the Information Content of Qualified Audit Reports. *Abacus.* 26: 17-35.

______ 1993. A Statistical Paradox in Auditing. *Abacus.* 30: 44-49.

______ 1994. Foundations of Neyman-Pearson Hypothesis Tests in Auditing. *Working Paper,* Graduate School of Business, University of Sydney.

Johnstone, D.J., and D.V. Lindley. 1993. Bayesian Inference Given Data 'Significant at α': Tests of Point Hypotheses. *Theory and Decision* In print.

Kendall, M.G., and A. Stuart. 1979. *The Advanced Theory of Statistics,* Vol. 2, 4th edn. London: Griffin.

Kempthorne, O. 1976. Of What Use Are Tests of Significance and Tests of Hypothesis. *Communications in Statistics - Theory and Methods.* A5, 8: 763-77.

Kinney, W.R. 1975. A Decision Theory Approach to the Sampling Problem in Auditing. *Journal of Accounting Research.* (Spring) 13: 117-32.

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

Lecam, L and E.L. Lehmann. 1974. J. Neyman on the Occasion of His 80th Birthday. *Annals of Statistics.* 2: vii-xiii.

Lehmann, E.L. 1986. *Testing Statistical Hypotheses.* 2nd edn. New York: Wiley.

Lindley, D.V. 1957. A Statistical Paradox. *Biometrika* 44: 187-92.

______ 1972. *Bayesian Statistics, A Review.* Philadelphia, Penn: Society for Industrial and Applied Mathematics.

Loebbecke, J.K. and J. Neter. 1975. Considerations in Choosing Statistical Sampling Procedures in Auditing (with discussion) *Journal of Accounting Research.* 13 (Supplement): 38-69.

Mayo, D. 1981. Testing Statistical Testing. In Pitt, J.C. (ed) *Philosophy in Economics.* Dordrecht, Holland: D. Reidel. Pp. 175-203.

Nelson, N., Rosenthal, R. and R.L. Rosnow. 1986. Interpretation of Significance Levels by Psychological Researchers. *American Psychologist* 41 (November): 1299-1301.

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

______ 1957. 'Inductive Behavior' as a Basic Concept of the Philosophy of Science. *Review of the International Statistical Institute.* 25: 7-22.

______ 1971. Foundations of Behavioristic Statistics. In Godambe, V.P. and D.A. Sprott (eds.) *Foundations of Statistical Inference.* Toronto: Holt, Rinehart and Winston. Pp. 1-19.

______ 1976. Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena. *Communications in Statistics - Theory and Methods.* A5, 8: 737-51.

Neyman, J and E.S. Pearson 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions Royal Society. A,* 231: 289-337. Reprinted in Neyman, J and E.S. Pearson 1967. *Joint Statistical Papers of J. Neyman and E.S. Pearson.* Pp.140-185. Berkeley and Los Angeles: University of California Press.

Pratt, J.W. 1965. Bayesian Interpretation of Standard Inference Statements (with discussion). *Journal of the Royal Statistical Society. B,* 27, 169-203.

Roberts, D.M. 1978. *Statistical Auditing.* New York: AICPA.

Savage, L.J. 1961. *The Subjective Basis of Statistical Practice.* Ann Arbor: University of Michigan.

______ L.J. 1962. *The Foundations of Statistical Inference: A Symposium.* New York: Wiley.

Scott, W.R. (1973) A Bayesian Approach to Asset Valuation and Audit Size. *Journal of Accounting Research.* (Autumn) 11: 304-330.

______ (1975) Auditor's Loss Functions Implicit in Consumption-Investment Models. *Journal of Accounting Research.* (Supplement) 13: 98-125.

Seidenfeld, T. 1979. *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher.* Dordrecht, Holland: D. Reidel.

Smielianskas, W. 1986. A Note on a Comparison of Bayesian with Non-Bayesian Dollar-Unit Sampling Bounds for Overstatement Errors of Accounting Populations. *The Accounting Review.* (January) 61, 118-128.

Smith, C.A.B. 1965. Personal Probability and Statistical Analysis. *Journal of the Royal Statistical Society. A,* 128, 469-499.

Tsui, KW., Matsumura, E.M. and K.L. Tsui. 1985. Multinomial-Dirchlet Bounds for Dollar-Unit Sampling in Auditing. *The Accounting Review.* (January) 60, 76-96.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

______ 1984. *Basic Issues in Econometrics.* Chicago: University of Chicago Press.

Table 1

Posterior Probabilities Given Marginal Rejection of H_0

		Assumed Sample Result		
Prior	$p(H_0)$	n=20, $\overline{X}$=·4942	n=50, $\overline{X}$=27.8784	n=90, $\overline{X}$=39·8776
(i) $N(0, 100^2)$	·547	·965	·981	·980
(ii) $N(0, 5000^2)$	·012	·951	·975	·975
(iii) $N(10^4, 5000^2)$	·002	·951	·974	·975
(iv) $N(-10^4, 5000^2)$	·002	·952	·974	·975
(v) $N(90, 15^2)$	·159	·417	.575	·688
(vi) $N(120, 25^2)$	·036	·335	.566	·707
(vii) $N(-90, 15^2)$	·159	·420	.922	1·000
(viii) $N(-120, 25^2)$	·036	·339	.967	1·000

Table 2

Comparison of Posterior Probabilities of H_0

		Assumed Result	
Prior	$p(H_0)$	$n=30$ $\overline{X}=14{\cdot}1$ *Accept* H_0	$n=90$ $\overline{X}=39{\cdot}8776$ *Reject* H_0
(i) $N(0, 100^2)$	·547	·980	·980
(ii) $N(0, 5000^2)$	·012	·973	·975
(iii) $N(10^4, 5000^2)$	·002	·972	·975
(iv) $N(-10^4, 5000^2)$	·002	·974	·975
(v) $N(90, 15^2)$	·159	·482	·688
(vi) $N(120, 25^2)$	·036	·433	·707
(vii) $N(-90, 15^2)$	·159	·636	1·000
(viii) $N(-120, 25^2)$	·036	·654	1·000

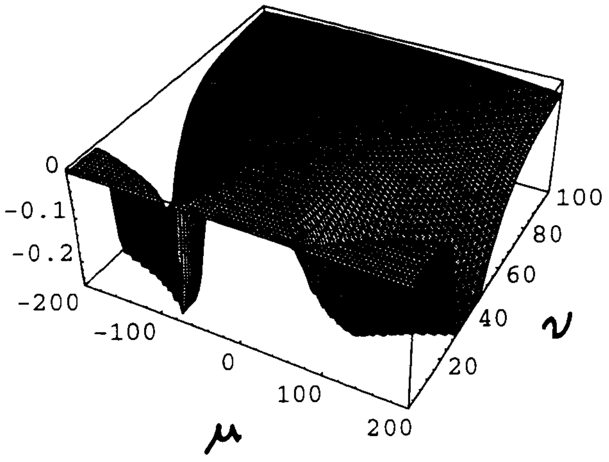

Figure 1
Surface Plot of Difference Between Posteriors

$$p(H_0|\overline{X}=14.1,\, n=30) - p(H_0|\overline{X}=\overline{X}_c=39.8776,\, n=90)$$

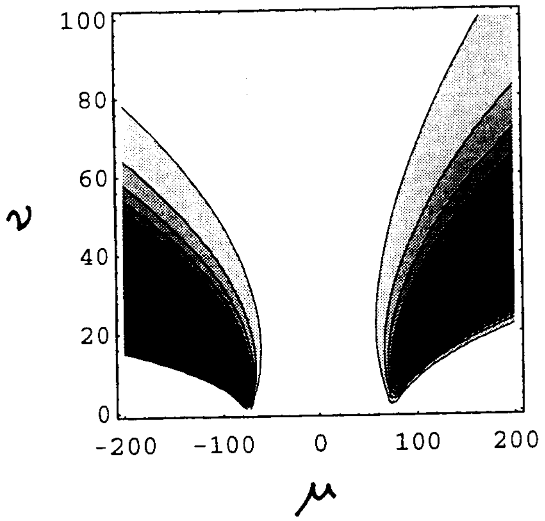

Figure 2
Contour Plot of Difference Between Posteriors

$p(H_0|\overline{X}=14.1, n=30) - p(H_0|\overline{X}=\overline{X}_c=39.8776, n=90)$

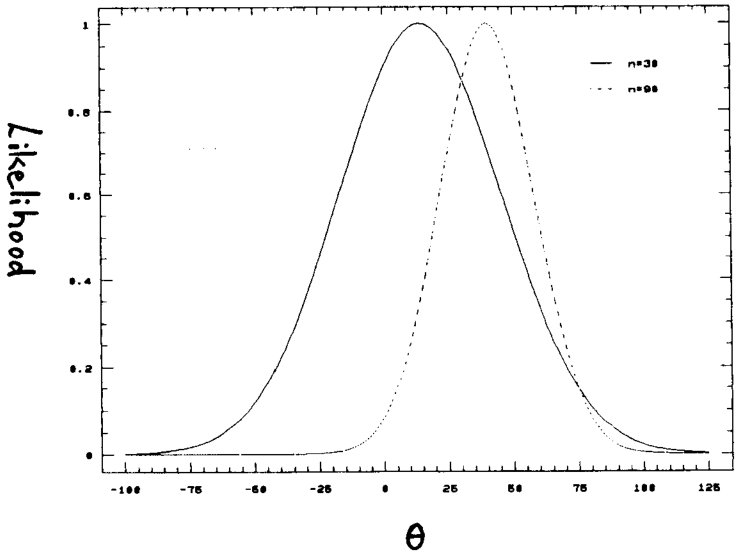

Figure 3
Likelihood Functions

$p(\overline{X}=14.1|\theta, n=30)$ and $p(\overline{X}=39.8776|\theta, n=90)$

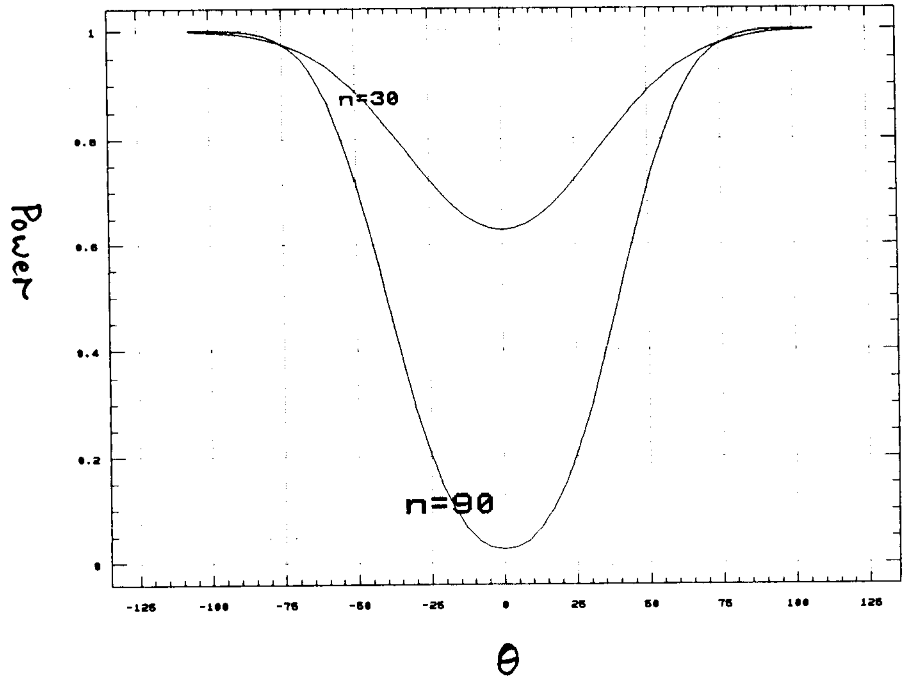

Figure 4
Power Functions of Tests with n=30 and n=90

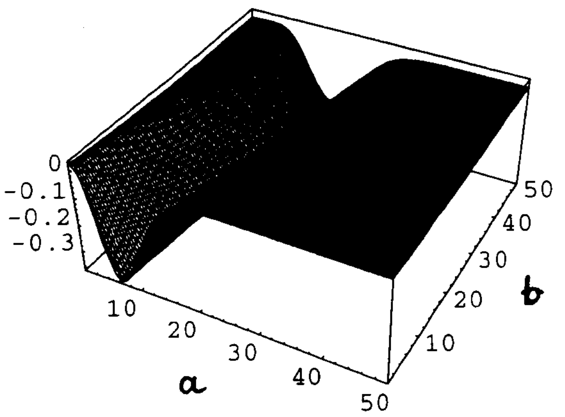

Figure 5
Contour Plot of Difference Between Posteriors
$p(H_0|k{=}1, n{=}30) - p(H_0|k{=}9, n{=}100)$

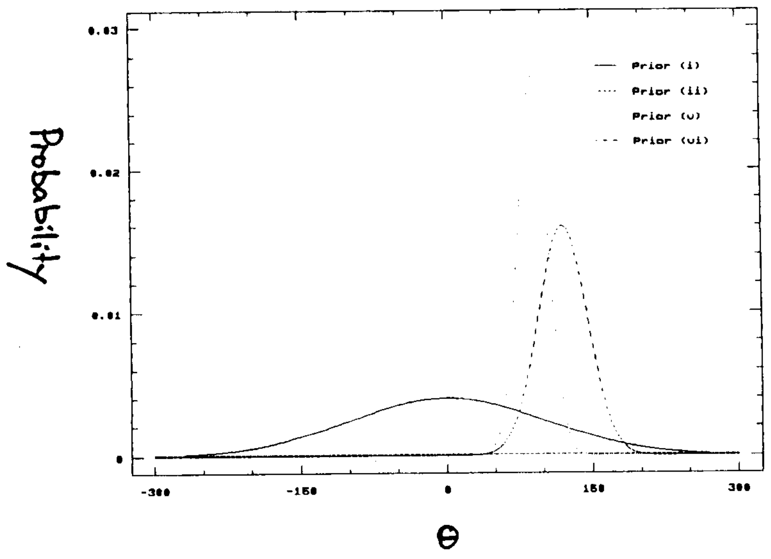

Figure 6
Prior Probability Distributions *(i), (ii), (v) and (vi)*

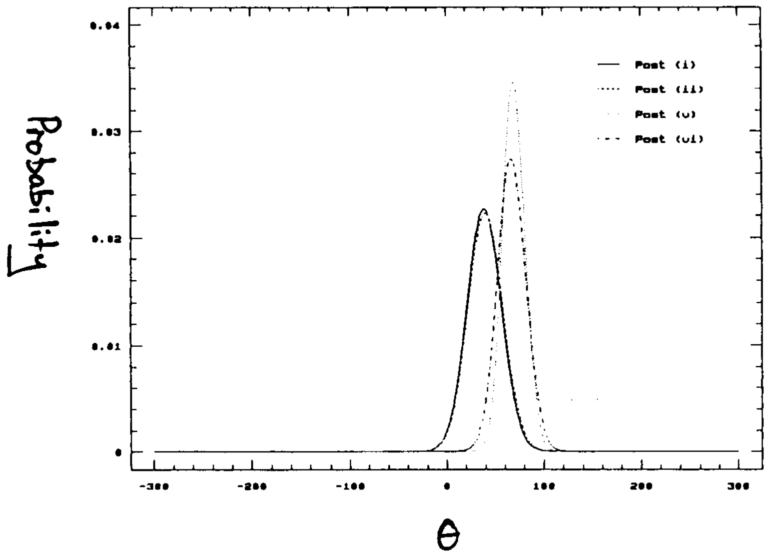

Figure 7
Posterior Probability Distributions *(i), (ii), (v) and (vi)*