Comparing P-values to Neyman-Pearson Tests

BU-1073-MA February, 1993

George Casella[1]
Martin T. Wells[2]
Cornell University

Key Words and Phrases: Hypothesis Testing, Decision Theory, Evidence

AMS 1980 Subject Classification: 62A99, 62C99

[1] Research supported by National Science Foundation Grant No. DMS89-0039

[2] Research supported by the U. S. Army Mathematical Sciences Institute at Cornell University

Summary

The classical hypothesis testing problem can be formulated as an estimation problem, using a particular loss function. Many forms of the loss function are quite reasonable, with one attractive candidate being squared error loss. Using this loss, we compare the risk function of the Neyman-Pearson test procedure and the p-value. For various testing problems the p-value has uniformly smaller risk than the Neyman-Pearson procedure. We conclude that the p-value is a viable measure of evidence and therefore the use of the p-value, rather than strict Neyman-Pearson "accept/reject" outcomes, is a reasonable decision theoretical action.

1. Introduction

In his classic book, "*Statistical Methods for Research Workers*," Fisher (1925) outlined a sequence of steps to test a statistical hypothesis. First of all an appropriate test statistic, say $t(\cdot)$ is selected, then the test statistic is evaluated at values of the sample, yielding $t(\mathbf{x})$. Next, compute a measure of how likely $t(\mathbf{x})$ was, that is, $p(\mathbf{x}) = P_{Ho}\left(t(\mathbf{X}) \geq t(\mathbf{x})\right)$, where the distribution of X is specified by the null hypothesis. Lastly, if $p(\mathbf{x}) > .05$ conclude that the deviation from the null hypothesis is not significant and if $p(\mathbf{x}) < .01$ conclude that the deviation is highly significant.

At about the same time, Neyman and Pearson (1928) proposed an alternate testing framework leading to two-point action spaces and level α tests. In the Neyman-Pearson paradigm, the experimenter is left only with the conclusion of acceptance or rejection of the null hypothesis. Fisher, to put it mildly, was not fond of the Neyman-Pearson procedure.

The theory of hypothesis testing, as applied today, is derived as a consequence of the Neyman-Pearson Lemma and results in decision procedures that are, for the most part, zero-one rules. Although these rules have various optimality features, serious criticisms have been leveled from many different directions. Bayesian critics (e.g. Berger, 1985) point out that two-point action spaces are too restrictive. The fact that the assessment of accuracy of the procedure is a pre-data assessment has been criticized by conditionalists (e.g. Kiefer, 1977) and Bayesians alike. An alternative to the strict Neyman-Pearson procedure is provided by the p-value. Originally, p-values, as developed by Fisher, were intended to be compared to an experimenter's preassigned level of significance. However, now the p-value has a life of its own (e.g. Berger and Selke, 1987; Casella and Berger, 1987; Berger and Delampady, 1987; Hwang et al., 1992). Since the p-value takes on values in the interval [0,1] rather than in the Neyman-Pearson two-point action space {0,1}, it is more readily thought of an estimate (or post-data assessment) than a decision.

The p-value has also been criticized. Bayesians have leveled many criticisms (e.g. Lindley, 1957; Berger and Selke, 1987; Berger and Delampady, 1988), most of which are based on the fact that, at the tails, the p-value may be much smaller than Bayesian posterior probabilities in the two-sided testing problem. In the one-sided problem these problems do not appear (Casella and Berger, 1987) since the p-value is a limit of Bayes rules. Bayesians are also quick to note that the p-value violates the likelihood principle (Berger and Wolpert, 1984). Classical Neyman-Pearson frequentists are also critical of the p-value since it has no foundation in frequentist theory.

In this article we formulate hypothesis testing as an estimation problem within a decision theoretic framework and compare the performance of the p-value and the Neyman-Pearson procedure. We shall primarily be concerned with testing a point null hypothesis for the Gaussian location problem, but also show how our results immediately extend to the important case of Student's t test. We do not discuss the one-sided problem since it is shown in Hwang et al. (1992) that the p-value is a generalized Bayes rule and hence is admissible. However, in the point null testing problem, it is shown in Hwang et al. that the p-value is inadmissible, therefore deserving of further study. See the work of Schaafsma, et al. (1989) and Gutmann (1989) for more discussions on the formulation of the testing problem as estimation.

In Section 2 we present some necessary preliminaries including loss function and decision theoretic formalizations. Section 3 contains the main results on the risk domination of the p-value over the Neyman-Pearson procedure in a large region of the parameter space. Section 4 contains further comments and conclusions.

2. Preliminaries

We consider the testing problem

$$H_0{:}\theta = 0 \quad \text{versus} \quad H_1{:}\theta \neq 0 \tag{2.1}$$

based on observing $X = x$, where X has a normal distribution with mean equal to θ and unit variance, $X \sim n(\theta,1)$. Let φ and Φ denote the p.d.f. and c.d.f. of X, respectively. The classical Neyman-Pearson procedure for testing (2.1) is given by the decision rule

$$\phi(x) = \begin{cases} 0 \text{ if } |x| \geq c \\ 1 \text{ if } |x| \leq c \end{cases} \tag{2.2}$$

where c is the $1-\alpha/2$ quantile of the standard normal distribution. If $X = x$ is observed, the p-value is defined by

$$p(x) = P_0(|X| \geq |x|) = 2(1-\Phi(|x|)). \tag{2.3}$$

The main goal is to compare the two procedures $\phi(\cdot)$ and $p(\cdot)$ in a decision theoretic framework. To do this we formulate the testing problem in (2.1) as the estimation of the set specified by H_0. That is, of estimating the parameter $\mathbb{I}_{\{\theta=0\}}(\cdot)$, where $\mathbb{I}_A(\cdot)$ is the indicator function of the set A.

Estimation of the indicator function, which we can write as $\mathbb{I}_{H0}(\theta)$, might at first seem to be an unusual parameter to estimate. However, this parameter is exactly what the experimenter is concerned about in an hypothesis testing problem. That is, the concern is whether the parameter θ is, or is not, in the set specified by H_0, and that is captured in the parameter $\mathbb{I}_{H0}(\theta)$. Other parameters can be considered, but none capture the essence of testing as this one does.

The performance of a decision rule δ is evaluated with respect to a loss function, $L(\theta,\delta)$ and a risk function $R(\theta,\delta) = EL(\theta,\delta(x))$. A decision procedure δ_1 is better than another decision procedure δ_2 if $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ with inequality for at least one value of θ.

If the loss function $L(\theta,\delta)$ is taken to be 0-1 loss, the usual Neyman-Pearson procedures result. Thus, classical testing can be equivalently formulated as point estimation starting from a loss function of the form

$$L(\theta,\delta) = |\theta-\delta|, \tag{2.4}$$

where we take the parameter θ to be $\mathbb{I}_{H0}(\theta)$. The standard decision-theoretic analysis would lead to Neyman-Pearson procedures as optimal. We could consider loss functions other than 0-1 loss, and for the problem at hand, squared error loss

$$L(\theta,\delta) = (\theta-\delta)^2, \tag{2.5}$$

is an eminently reasonable alternative. Squared error loss is one of the few loss functions for which the Bayes estimates of $\mathbf{I}_{H0}(\theta)$ are probabilities (note that this is not the case for 0−1 loss). For estimates of $\mathbf{I}_{H0}(\theta)$, probabilities are quite suitable, as they not only allow for a reasonable range of values, but also contain a built-in evidential assessment. Furthermore, the Bayes rules are a subset of the admissible rules, so we are working within a class of rules that, from a decision-theoretic view, are sensible. (For more details on various types of loss functions see Hwang, et al., 1992.)

Thus, the problem we focus on is the comparison of $\phi(x)$ of (2.2) and $p(x)$ of (2.3) through comparison of the risk functions $E_\theta\left(\mathbf{I}_{H0}(\theta) - \phi(X)\right)^2$ and $E_\theta\left(\mathbf{I}_{H0}(\theta) - p(X)\right)^2$.

One could also compare the risk function of various Bayes rules to the p-value and the Neyman-Pearson rule. However, since the behavior of the risk function of the Bayes rule would vary from prior to prior, interpretation of those results would be difficult.

3. Risk Comparisons

In this section we compare the risk function of the p-value, $p(\cdot)$, and that of the Neyman-Pearson critical function, $\phi(\cdot)$, under the loss function in (2.5). It will be shown that, for common α levels, $R(\theta,p) \leq R(\theta,\phi)$ for a large region of the parameter space.

It is easy to verify that, for $\theta \neq 0$, the risk functions in question are given by

$$R(\theta,p) = 4 \int_{-\infty}^{\infty} (1-\Phi(|x|))^2 \varphi(x-\theta)\, dx$$

and

$$R(\theta, \phi) = \int_{-\infty}^{\infty} \mathbb{I}_{[-c,c]}(x)\, \varphi\, (x-\theta)\, dx$$

with $R(0,p) = \frac{1}{3}$ and $R(0,\phi) = \alpha$. Clearly, it follows that $R(0,p) \geq R(0,\phi)$ for all $\alpha \leq \frac{1}{3}$. However, for small $\epsilon > 0$, we also have that $R(\epsilon,p) < R(\epsilon,\phi) = 1-\alpha$ for all $\alpha < \frac{2}{3}$, and we are mainly concerned with values of $\theta \neq 0$. (Note that both $R(\theta,p)$ and $R(\theta,\phi)$ are symmetric functions of θ, hence it is sufficient to study the case where $\theta > 0$.)

Define the difference of the risk functions for $\theta \neq 0$ as

$$\Delta(\theta) = R(\theta,\phi) - R(\theta,p)$$

$$= \int_{-\infty}^{\infty} \{\mathbb{I}_{[-c,c]}(x) - 4(1 - \Phi(|x|))^2\}\, \varphi(x-\theta)\, dx \qquad (3.1)$$

Therefore, it is desired to show that $\Delta(\theta) > 0$. The following technical lemma will facilitate the proof of the main result.

Lemma 3.1. The function $\Delta(\theta)$ has two sign changes.

Proof. Applying the results of Brown et al. (1981) we know that $\varphi(x-\theta)$ is a strict variation reducing kernel of infinite order (SVR_{∞}) since the normal distribution is an exponential family. Hence, by Property 2.1 of Brown et al. (1981), the number of sign changes of $\Delta(\theta)$ is bounded by the number of sign changes of the integrand $\{\mathbb{I}_{[-c,c]}(x) - 4\left(1 - \Phi(|x|)\right)^2\}$. Since $\{\mathbb{I}_{[-c,c]}(x) - 4\left(1 - \Phi(|x|)\right)^2\}$ has sign changes at $\pm c$, it follows that $\Delta(\theta)$ has at most two sign changes. Since $\Delta(\theta)$ is symmetric about zero, there can only be zero or two sign changes. We will show that there are exactly two sign changes.

An elementary calculation shows

$$\lim_{\theta \to \infty} \frac{\varphi(y-\theta)}{\varphi(x-\theta)} = \begin{cases} 0 \text{ if } y < x \\ 1 \text{ if } y = x \\ \infty \text{ if } y > x \end{cases} \tag{3.2}$$

and, an application of L'Hopital's rule along with (3.2) yields

$$\lim_{\theta \to \infty} \frac{R(\theta, \phi)}{\Phi(c-\theta)} = 1 \ , \tag{3.3}$$

where the constant c defines the rejection region of the Neyman-Pearson test. Note that since

$$\frac{d}{d\theta} R(\theta,p) = 8 \int_{-\infty}^{\infty} \Big(1 - \Phi(|x|)\Big) \operatorname{sgn}(x)\varphi(x)\varphi(x-\theta)dx, \tag{3.4}$$

another application of L'Hopital's rule along with (3.4) shows that

$$\lim_{\theta \to \infty} \frac{R(\theta, p)}{\Phi(c-\theta)} = \infty \, . \tag{3.5}$$

Expressions (3.3) and (3.5) imply

$$\lim_{\theta \to \infty} \frac{R(\theta, \phi)}{R(\theta, p)} = 0 \ ,$$

hence, in the tail, ϕ dominates p. A similar arguement can be made for $\theta \to \infty$. Therefore the result that $\Delta(\theta)$ has two sign changes follows. □

The result of Lemma 3.1 shows that $\Delta(\theta)$ has exactly two sign changes, which occur in a $-+-$ pattern. That is, in the tails ϕ dominates p, while for θ near the null hypothesis p dominates ϕ. For the case $\theta_0 = 0$ we carried out a numerical search for the solutions to $\Delta(\theta) = 0$. We found that the risk functions crossed at $\theta^* = \pm 10.25854$, that is, over 10 standard deviations from the null hypothesized value of θ. The value of the risk functions at θ^* equaled 1.1102×10^{-16}. Therefore, the risk domination of p by ϕ in the tail is somewhat insignificant. Hence for all practical purposes, we can conclude the p dominates θ.

Figure 1 shows the graph of $\Delta(\theta)$ for $\theta > 0$. It can be seen that the risk domination is quite substantial in the neighborhood of the origin, but the domination tapers off as θ increases. One way to interpret this is that the Neyman-Pearson procedure can not distinguish between subtle changes in the parameters near the null hypothesis. This has been one of the major criticisms leveled at the

Neyman-Pearson testing paradigm.

An application of Lemma 3.1 gives our main result by noting that $\Delta(\theta) > 0$ for all θ satisfying $0 < |\theta| < 10$ and $\alpha < 2/3$.

Theorem 3.1. *If* $X \sim n(\theta,1)$, **under the loss function in (2.5) the p-value dominates the Neyman-Pearson procedure for all $\alpha < 2/3$ in a significant portion of the parameter space ($0 < |\theta| < 10$).** □

The proof follows quickly from the fact that the risk of the p-value for $\theta \neq 0$ is a symmetric differentiable function, and that $\Delta(\theta)$ has two sign changes. The search for decision procedures which dominate the Neyman-Pearson procedure may be limited to this class, whose members include Bayesian posterior probabilities. If there is interest in testing the location hypothesis in (2.1) for a family of distributions which are a mixture of normal distributions, the same type of sign change argument may be used if the mixing measure is strictly positive. Examples of this type of distributions include the Student-t, double exponential, Cauchy and other elliptically symmetric distributions. (See Kariya and Sinha (1989) for more details on mixtures of normals.)

The most important consequence of the mixture formulation is that Theorem 3.1 remains valid for Student's t test. Suppose $X_1, ..., X_n$ are iid $n(\mu,\tau^2)$, both unknown, and it is desired to test

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0.$$

If we observe $\bar{X} = \bar{x}$ and $S^2 = s^2$, values of the sample mean and variance, the p-value is given by

$$p(\bar{x}, s^2) = P_{H_0}\Big(|\bar{X}/s| > |\bar{x}/s|\Big) \tag{3.6}$$

and the Neyman-Pearson (uniformly most powerful unbiased) test is the t-test

$$\phi(\bar{x}, s^2) = \begin{cases} 0 \text{ if } |\bar{x}/s| \geq c \\ 1 \text{ if } |\bar{x}/s| < c \end{cases} \tag{3.7}$$

Using the loss (2.5), a consequence of Theorem 3.1 is, for a significant portion of the parameter space,

$$R\Big((\mu, \sigma^2), p(\bar{X}, s^2)\Big) \leq R\Big((\mu, \sigma^2), \phi(\bar{X}, s^2)\Big) \tag{3.8}$$

as long as $\alpha < \frac{2}{3}$. Figures 2 and 3 show the graph of the risk difference for the t-test. It can be seen that the risk domination is greater for the t-test than for the z-test of Figure 1. Thus, the thickness of the tail plays an important role in the size of domination.

The result of Lemma 3.1, and hence Theorem 3.1, would hold with the p-value, $2(1 - \Phi(1 \times 1))$, replaced by any estimate of $\mathbb{I}_{\{\theta = 0\}}$ which takes values strictly in the open interval (0, 1). Therefore the results of this paper hold for variety of estimators of $\mathbb{I}_{\{\theta = 0\}}$, including Bayesian posterior probabilities.

4. Conclusions

It follows from Theorem 3.1 that even though the p-value has no roots in frequentist theory the practitioner who uses the p-value rather than the strict reject/accept paradigm is using a superior (in terms of risk) procedure. Since the journals in applied fields are flooded with p-values, not with the outcomes of Neyman-Pearson tests, our results give some formal statistical setting in which the p-value is an acceptable measure. These results also take a step toward resolving the dilemma of whether one should report p-values or dichotomous outcomes of test procedures.

In the introduction we laid out the opposing approaches of Fisher and Neyman-Pearson. Part of that conflict arose because Fisher did not base his measures on values of the parameter under the alternative hypothesis, while Neyman and Pearson had a clear alternative in mind in which to discuss results on power. The results here somewhat confound these arguments in that, under the values of the null hypothesis, the Neyman-Pearson procedure is better, while if considering value of the alternative hypothesis the p-value is better. This runs contrary to the original argument between Fisher and Neyman-Pearson.

A somewhat startling result in Hwang et al. (1992) is that any Bayes rule for the hypothesis in (2.1) is unable to dominate the p-value. Therefore, until a reasonable dominating procedure is found, the p-value seems to be a viable assessment of the plausibility of the null hypothesis.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag, New York.

Berger, J.O., and Wolpert, R.W. (1984). *The Likelihood Principle*, 2nd edition. IMS Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

Berger, J.O. and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (with discussion). *J. Amer. Statist. Assoc. 82,* 112-122.

Berger, J.O. and Delampady, M. (1987). Testing Precise Hypotheses (with discussion), *Statistical Science 2,* 317-352.

Brown, L.D., Johnstone, I.M. and MacGibbon, K.B. (1981). Variation Diminishing Transformations: A Direct Approach to Total Positivity and Its Statistical Applications. *J. Amer. Statist. Assoc. 76,* 824-831.

Casella, G. and Berger, R.L. (1987). Reconciling Evidence in the One-Sided Testing Problem (with discussion). *J. Amer. Statist. Assoc. 82,* 106-111.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Gutmann, S. (1989). Loss Functions for P-Values and Simultaneous Inference. Technical Report No. 43, Statistics Center, Massachusetts Institute of Technology, Cambridge, MA.

Hwang, J.T., Casella, G., Robert, C., Wells, M.T., and Farrell, R.H. (1992). Estimation of Accuracy in Testing. *Ann. Statist. 20*, 868-881.

Kariya, T. and Sinha, B.K. (1989). *Robustness of Statistical Tests.* San Diego: Academic Press

Kiefer, J. (1977). Conditional Confidence Statements and Confidence Estimators (with discussion). *J. Amer. Statist. Assoc. 72,* 789-808.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses,* 2nd edition. New York: John Wiley.

Lindley, D.V. (1957). A Statistical Paradox. *Biometrika 44,* 187-192.

Neyman, J. and Pearson, E.S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika 20A,* 175-240, 263-295.

Schaafsma, W., Tobloom, J., Van der Menlen, B. (1989). Discussing Truth or Falsity by Computing a Q-Value. *Statistical Data Analysis and Inference*, Y. Dodge, ed. Elsevier Science Pub., B. V., North Holland.

Figure 1: Risk functions $E_\theta L(\mathbf{I}_{H0}(\theta),\delta) = E_\theta(\mathbf{I}_{H0}(\theta) - \delta)^2$, for H_0: $\theta = 0$, where θ is the mean of a normal distribution with known variance. The solid line is the risk of the p-value and risks of Neyman-Pearson tests, for different α levels, are given by the other lines. Values are $\alpha = .01$ (long dashes), $\alpha = .05$ (dots), $\alpha = .1$ (short dashes). For comparison with the p-value we also include $\alpha = .68$ (closely spaced dots).

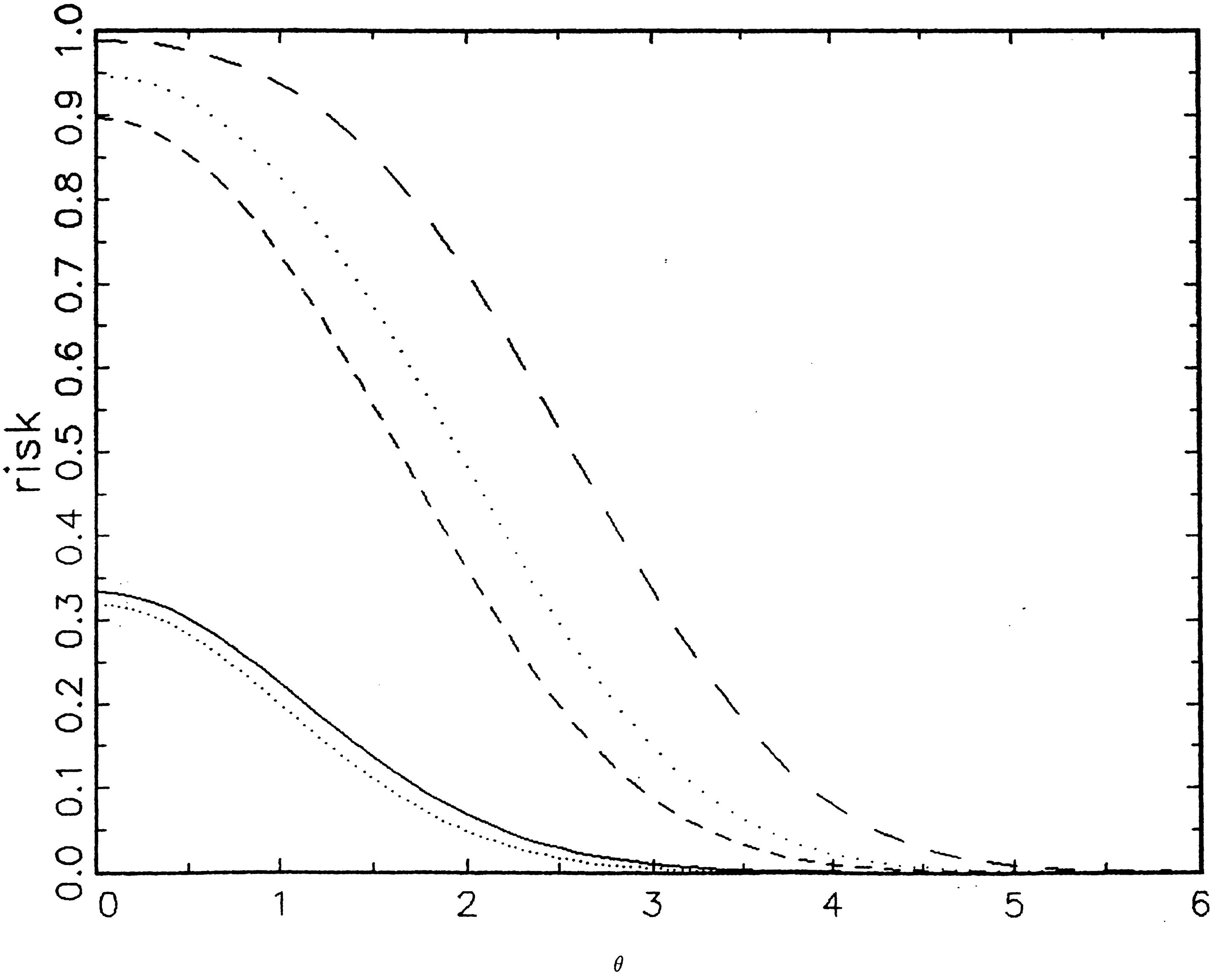


Figure 2: Risk functions $E_\theta L(I_{H_0}(\theta), \delta) = E_\theta(I_{H_0}(\theta) - \delta)^2$, for H_0: $\theta = 0$, where θ is the mean of a normal distribution with unknown variance, and the test statistic has Student's t distribution with 5 degrees of freedom. The solid line is the risk of the p-value and risks of Neyman-Pearson tests, for different α levels, are given by the other lines. Values are $\alpha = .01$ (long dashes), $\alpha = .05$ (dots), $\alpha = .1$ (short dashes). For comparison with the p-value we also include $\alpha = .68$ (closely spaced dots).

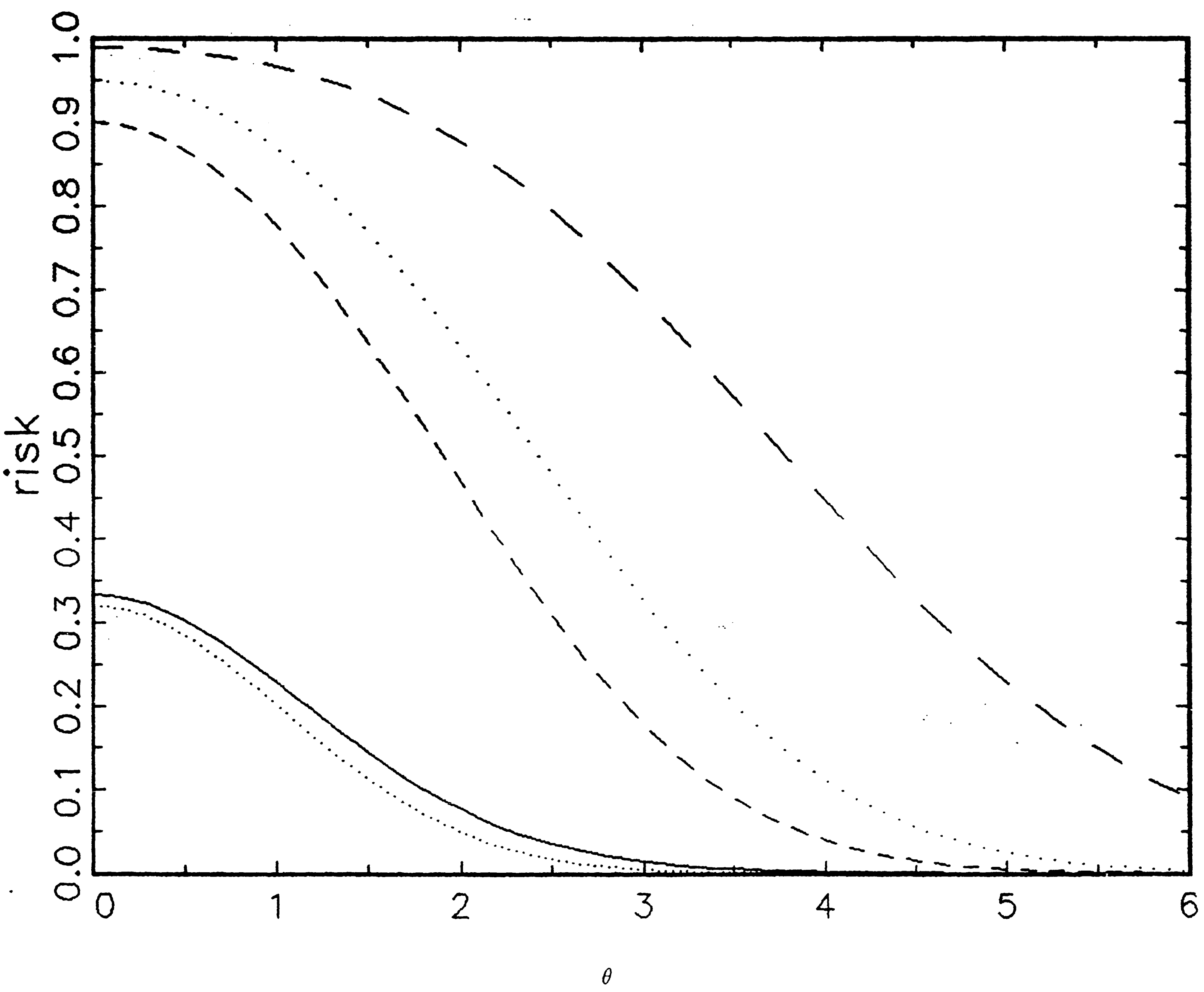
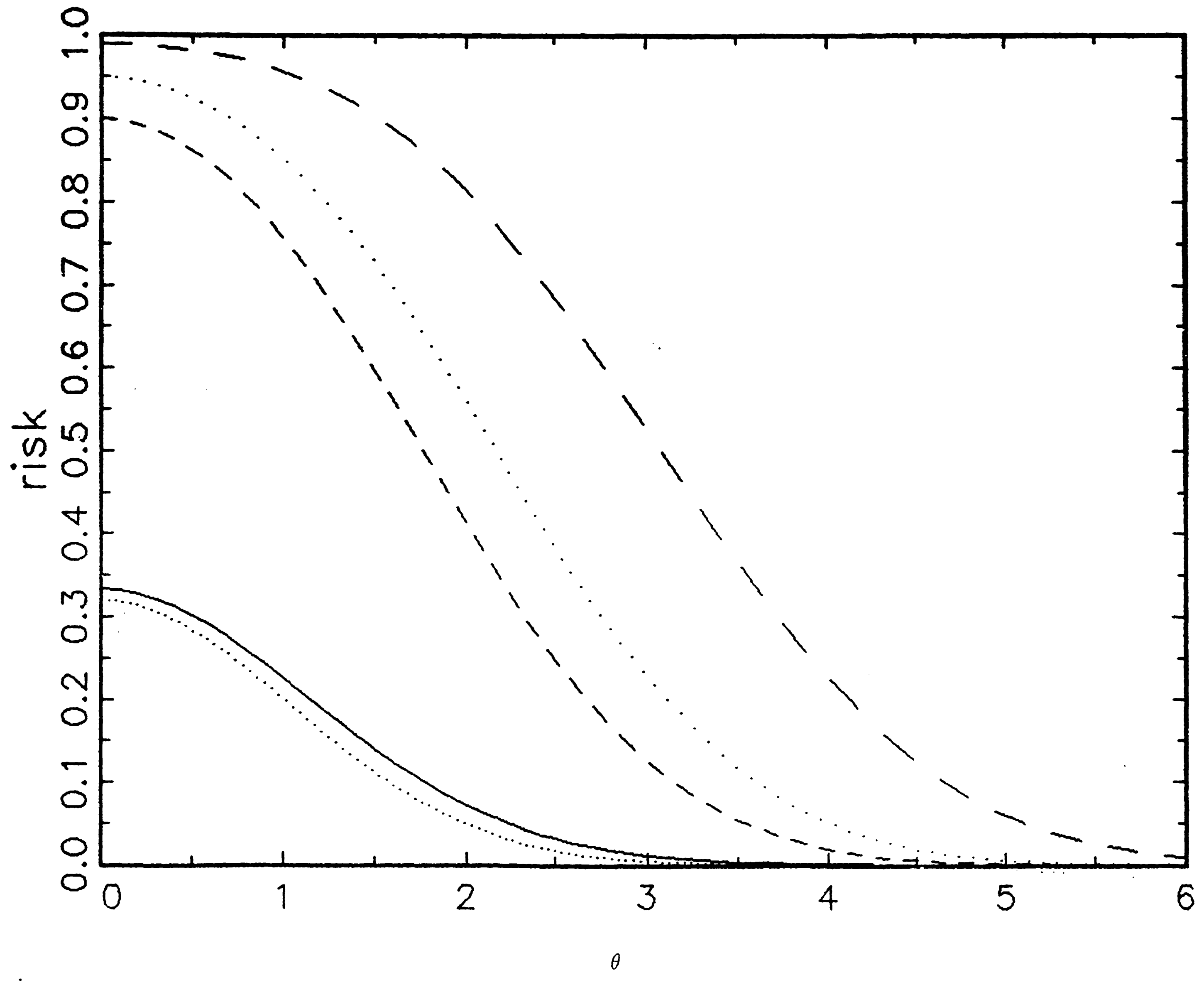


Figure 3: Risk functions $E_\theta L(I_{H0}(\theta),\delta) = E_\theta(I_{H0}(\theta) - \delta)^2$, for H_0: $\theta = 0$, where θ is the mean of a normal distribution with unknown variance, and the test statistic has Student's t distribution with 10 degrees of freedom. The solid line is the risk of the p-value and risks of Neyman-Pearson tests, for different α levels, are given by the other lines. Values are $\alpha = .01$ (long dashes), $\alpha = .05$ (dots), $\alpha = .1$ (short dashes). For comparison with the p-value we also include $\alpha = .68$ (closely spaced dots).


risk
θ