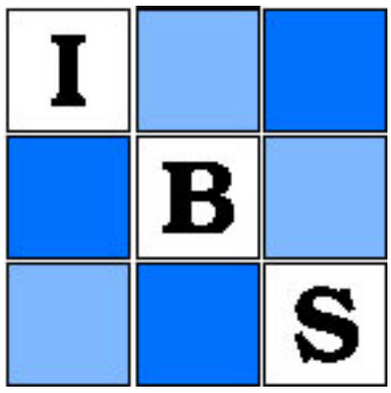

BIOMETRICS 53, 11–22
March 1997

The Behavior of the *P*-Value When the Alternative Hypothesis Is True*

H. M. James Hung,[1] Robert T. O'Neill,[2] Peter Bauer,[3] and Karl Köhne[4]

[1]Division of Biometrics I, CDER, Food and Drug Administration, HFD-710, Room 5062, 1451 Rockville Pike, Rockville, Maryland 20852, U.S.A.
[2]Office of Epidemiology and Biometrics, CDER, Food and Drug Administration, HFD-700, Room 15B31, 5600 Fishers Lane, Rockville, Maryland 20857, U.S.A.
[3]Institut für Medizinische Statistik, der Universtät Wien, Schwarzspanierstraße 17, A-1090 Vienna, Austria
[4]Institut für Medizinische Dokumentation und Statistik der Universtät zu Köln, D-50924 Cologne, Germany

SUMMARY

The *P*-value is a random variable derived from the distribution of the test statistic used to analyze a data set and to test a null hypothesis. Under the null hypothesis, the *P*-value based on a continuous test statistic has a uniform distribution over the interval $[0, 1]$, regardless of the sample size of the experiment. In contrast, the distribution of the *P*-value under the alternative hypothesis is a function of both sample size and the true value or range of true values of the tested parameter. The characteristics, such as mean and percentiles, of the *P*-value distribution can give valuable insight into how the *P*-value behaves for a variety of parameter values and sample sizes. Potential applications of the *P*-value distribution under the alternative hypothesis to the design, analysis, and interpretation of results of clinical trials are considered.

1. Introduction

The *P*-value is one of the most routinely used statistical measures of uncertainty, yet statisticians may in some situations (Goodman, 1992) disagree on its appropriate use and on its interpretation as a measure of evidence. The *P*-value is derived from the perspective of a test of hypothesis in which a test statistic is calculated from results of a given set of data and, under the assumption that the null hypothesis is true, the distribution of the test statistic is used to obtain the tail probability of observing that result or a more extreme result. Thus, the *P*-value is a measure of evidence against the null hypothesis. Because the *P*-value is based upon analysis of random variables, it itself is a random variable whose distribution, for continuous test statistics, is well known to be uniform over the interval [0, 1] under the null hypothesis. It is because of this fact that a cutoff for a *P*-value at, say 0.05, is used to control the chances that, for any given experiment, one of twenty *P*-values could be 0.05 or less, even when the null hypothesis is true. This concept, in the Neyman–Pearson theory of hypothesis testing, is known as the Type I error rate, which is a preexperiment error rate that determines the rejection region and is intended to control the overall frequency of making erroneous rejections of the null hypothesis.

It is of interest that the distribution of the *P*-value, when the null hypothesis is true, is uniform over $[0, 1]$ regardless of the sample size of an experiment, so there is no way to distinguish *P*-values derived from large studies from those derived from small samples, nor from studies well powered to detect a posited alternative hypothesis from those underpowered to detect that same posited alternative value. Other statistical measures, such as confidence intervals, may serve this purpose

* The views expressed in the article are not necessarily those of the U.S. Food and Drug Administration.

Key words: Power; Phyp plot; Meta-analysis; Significance.

in part, but there seems to be no direct relationship between the two concepts for illustrating the point of this article.

The magnitude of the P-value is, for most investigators, important to the interpretation and conclusions inferred from planned experiments and observational data. Since the P-value is a measure of evidence against the null hypothesis, it is informative to explore the magnitude of the P-value when the alternative hypothesis is true. This is the goal of this article. Under the alternative hypothesis, the distribution of the P-value becomes a function of the study sample size and the given value of the parameter (call it δ) in the alternative hypothesis. Moreover, in contrast to the P-value's uniform distribution when the null hypothesis holds (i.e., $\delta = 0$), the density of the P-value under any alternative hypothesis is markedly skewed, and this is especially the case as the sample size of the experiment increases, reflecting the increasing power of a study to detect (i.e., produce a P-value less than a prespecified level α, $0 < \alpha < 1$) any prespecified difference δ. In Section 2 the P-value density and distribution functions will be derived for a fixed point alternative δ. In Section 3 we show how the behavior of the P-value distribution depends on both the sample size for the experiment and the magnitude of the tested parameter δ in the alternative hypothesis. We shall explore the relationship of the power of the test statistic to the distribution of the P-value under the alternative hypothesis. An application will be presented in Section 5. Section 6 will be devoted to the situation where the value of δ is uncertain and the uncertainty can be quantified by a probability distribution on δ. Concluding discussions follow.

2. Distribution of the *P*-Value for a Specified Parameter Value

As motivation, we begin by considering a one-sample experiment in which the response variable Y follows a Gaussian distribution with a mean parameter μ and a standard deviation σ. The comparative two sample experiment will be considered in Section 4. Let y_n be the sample mean of n independent observations of Y for the purpose of testing the hypothesis H_0: $\mu = 0$ versus H_1: $\mu > 0$ at a significance level α. The α level is a preexperiment Type I error rate used to control the probability that the observed P-value in the experiment of making an error rejection of H_0 when in fact H_0 is true is α or less.

The P-value is calculated from the distribution of the test statistic $T = \sqrt{n}y_n/\sigma$. We assume for simplicity that σ is given. In many applications, the assumption of known σ or distribution of Y is not necessary if samples are sufficiently large, because by replacing σ^2 with the sample variance the statistic T is approximately standard Gaussian under H_0. The one-sided P-value for testing H_0 against H_1 is derived by calculating the probability of the observed value or more extreme of the test T and has the value

$$p = 1 - \Phi(t), \tag{2.1}$$

where t is the realization of T and Φ is the standard Gaussian distribution function with density ϕ. For a given μ, T has a Gaussian distribution with mean $\sqrt{n}\mu/\sigma$ and variance one. Let $\delta = \mu/\sigma$. As shown in (A.1) in the Appendix, for a given δ and n, the density of the P-value is

$$g_\delta(p) = \phi(Z_p - \sqrt{n}\delta)/\phi(Z_p), \qquad 0 < p < 1, \tag{2.2}$$

where Z_p is the $(1-p)$th percentile of the standard Gaussian distribution. The Z_p is the value of the standard Gaussian random variable beyond which the tail probability is p. The distribution function of the P-value, given δ and n, is

$$G_\delta(p) = \int_0^p g_\delta(x)\,dx = 1 - \Phi(Z_p - \sqrt{n}\delta), \qquad 0 < p < 1.$$

The expected value $E_\delta(P)$ of the P-value is given by (A.2) in the Appendix and the variance $\mathrm{var}_\delta(P)$ can be derived similarly. Clearly, $G_\delta(p), E_\delta(P)$, and $\mathrm{var}_\delta(P)$ depend on δ and n.

3. Relationship of Sample Size and Power at a Specified Parameter Value to *P*-Value Distribution

For planned experiments designed to test the hypotheses H_0 versus H_1 described in Section 2, the sample size n is usually determined to detect an anticipated magnitude, $\mu = \mu^* > 0$, with power $1 - \beta$. The determination of sample size of the experiment is based on the well-known relation between n and μ^*,

$$n = \{\sigma(Z_\alpha + Z_\beta)/\mu^*\}^2, \tag{3.1}$$

where Z_ν is the $(1-\nu)$th percentile of the standard Gaussian distribution.

Consider an experiment designed to reject H_0 at $\alpha = 0.05$ and with power 90% when the expected effect size, say $\mu^* = \sigma/3$. Based on (3.1), the sample size required is approximately 80. The density function of the P-value for this sample size and the alternative value $\delta = 1/3$ is illustrated in Figure 1 as well as the densities for a variety of n. In Figure 2 the densities for a variety of assumed δ given the sample size $n = 80$ are presented. The corresponding cumulative distribution functions of P are depicted in Figures 3 and 4. Both the density function and the cumulative distribution function of the P-value are increasingly steep as the sample size n or the magnitude of δ increases. It is seen from Tables 1 and 2 that as either δ or n decreases, the expected value and the standard deviation increase and the percentiles shift towards one.

Note that the P-value distribution for the sample mean test is a function of $\sqrt{n}\delta$ (see equation (2.2)). Thus, if the true magnitude of μ is as anticipated (i.e., $\delta = \mu^*/\sigma$) and the sample size n is determined based on equation (3.1), then the distribution function and the characteristics (e.g., mean and percentiles) of the P-value for the sample mean test depend only on the power level $1-\beta$ via the value of $\sqrt{n}\delta$ when α is fixed. From Figure 5 and Table 3, one can see that as the power level increases, the mean and variance of the P-value become smaller, the percentiles shift towards zero, and the distribution function is steeper in shape. The information in Table 3 provides useful guidance to trial designers in selecting the power level for study. Table 3 illustrates that for a study planned with 90% power against any alternative, the probability is 50% that the P-value one will observe with a sample size chosen to maintain the power is no greater than 0.001, and it would be a rare occurrence (i.e., 5% chance) to observe a P-value greater than 0.10.

Two further observations are made here. First, the power characteristic of the sample mean test T for H_0 versus H_1 is closely tied to the P-value distribution. For a fixed effect size δ, the power of T can be written as

$$\begin{aligned} Q(\delta) &= \text{pr}\{T > Z_\alpha \mid \delta\} \\ &= \text{pr}\{P < \alpha \mid \delta\} \\ &= G_\delta(\alpha). \end{aligned}$$

That is, the power of T at a fixed effect size δ can be read from the P-value distribution $G_\delta(p)$ at $p = \alpha$. For example, from Table 1, for $n = 15$ and $\delta = 1/3$, the 25th percentile of the P-value distribution is about 0.025 [i.e., $G_{1/3}(0.025) = 0.25$]; therefore, the power of the sample mean test T performed at a 2.5% level of significance is only 25% at $\delta = 1/3$ when $n = 15$.

Secondly, if the α-level sample mean test T has power $1-\beta$ at $\delta = \delta^*$, that is, $\text{pr}(T > Z_\alpha \mid \delta = \delta^*) = 1-\beta$, then

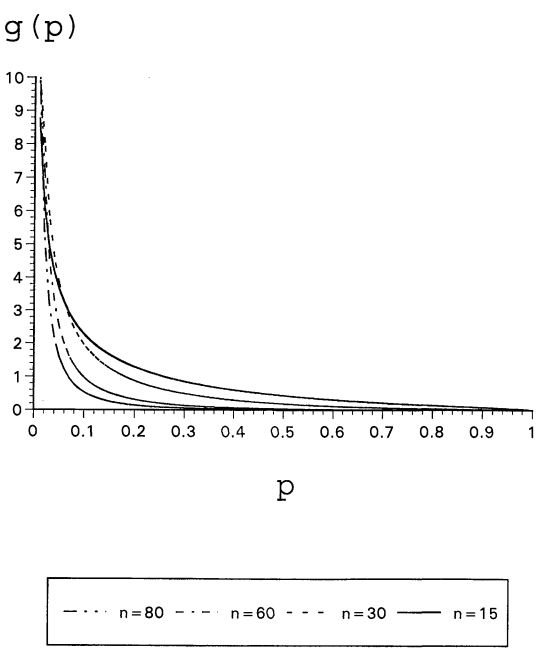

Figure 1. Density of the P-value for various n at $\delta = 1/3$.

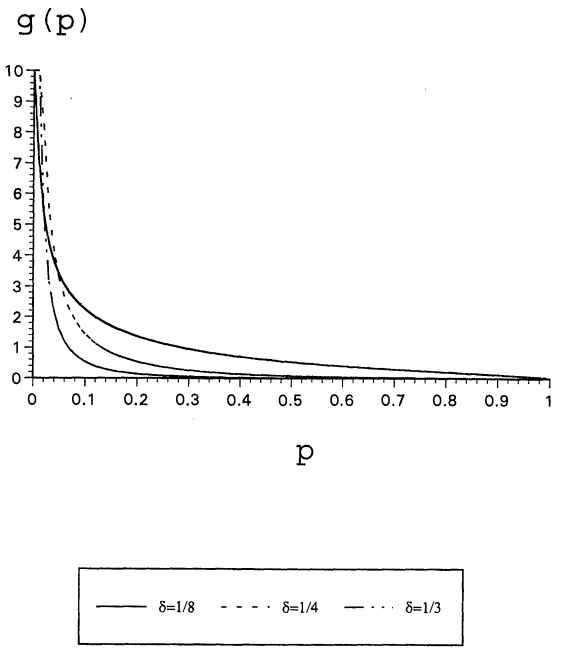

Figure 2. Density of the P-value at various δ for $n = 80$.

$$1 - \alpha = \text{pr}(T - \sqrt{n}\delta^* > -Z_\alpha \mid \delta = \delta^*)$$
$$= \text{pr}(P < \beta \mid \delta = \delta^*),$$

since $Z_\alpha - \sqrt{n}\delta^* = -Z_\beta$. This means that the type II error rate β of the α-level sample mean test T at a fixed point alternative $\delta = \delta^*$ is the $(1-\alpha)$th percentile of the P-value distribution. As seen in Tables 1 and 2, the 95th percentile reflects the Type II error rate (which is exactly equal to one minus power in the last column) of the 5%-level test T under each specified scenario.

4. Two Sample Scenario

We now consider the two-sample situation in which the target parameter is the difference in the mean of the response variable of interest between the two sample groups. Suppose that the response

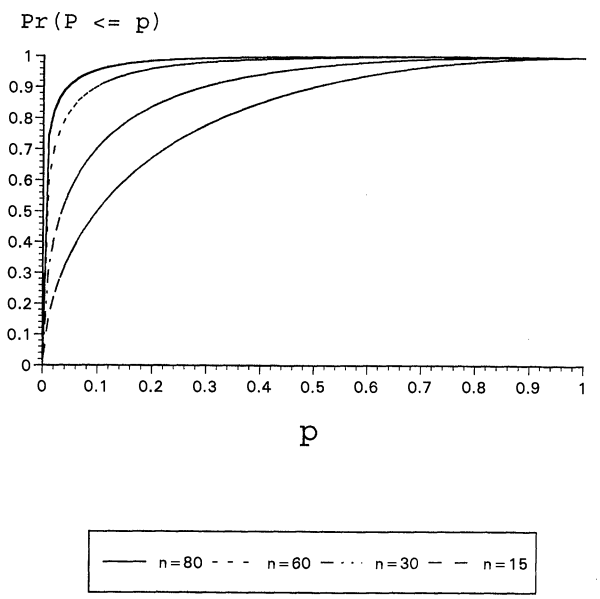

Figure 3. Cumulative distribution function of the P-value for various n at $\delta = 1/3$.

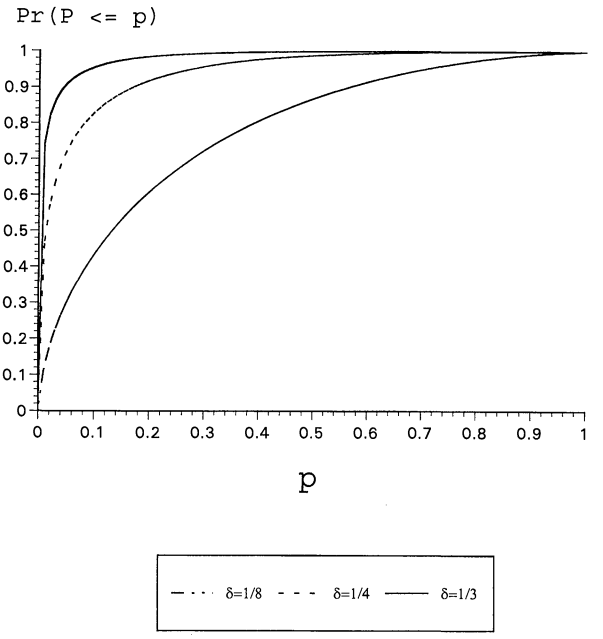

Figure 4. Cumulative distribution function of the P-value at various δ for $n = 80$.

variable Y follows a probability distribution F_1 with mean μ_1 and variance σ_1^2 in group 1 and a probability distribution F_2 with mean μ_2 and variance σ_2^2 in group 2, where F_1 and F_2 belong to the same family. The variance parameter σ_i^2 may be a function of the mean parameter μ_i $(i = 1, 2)$. An example is that Y is a binary outcome following a Bernoulli distribution with mean π and variance $\pi(1 - \pi)$, where $\pi = \text{pr}(Y = 1)$.

Let y_1 and y_2 be the sample means of m_1 and m_2 independent observations of Y from group 1 and group 2, respectively, for the purpose of testing the hypothesis

$$\text{H}_0\text{: } \mu_1 = \mu_2 \qquad \text{versus} \qquad \text{H}_1\text{: } \mu_1 > \mu_2$$

at a significance level α. The test statistic often employed is given by

$$\hat{T} = (y_1 - y_2)/(\hat{\sigma}_1^2/m_1 + \hat{\sigma}_2^2/m_2)^{1/2},$$

where $\hat{\sigma}_i^2$ is a consistent estimator of σ_i^2 from group i $(i = 1, 2)$, that is, $\hat{\sigma}_i^2$ converges to σ_i^2 in probability as m_i tends to infinity. When m_1 and m_2 are sufficiently large, the test $\hat{T}$ is asymptotically normal with mean

$$(\mu_1 - \mu_2)/(\sigma_1^2/m_1 + \sigma_2^2/m_2)^{1/2}$$

and variance one.

To use the formulas developed in the previous section, we need to define n and δ. Let

$$n = m_1 m_2/(m_1 + m_2),$$

that is, $2n$ is the harmonic mean of m_1 and m_2. Let

Table 1

Characteristics of the P-value distribution for various sample sizes n and the corresponding power at $\delta = 1/3$ when $\alpha = 0.05$

			Percentile							
n	Mean	S.D.	5th	10th	25th	50th	75th	90th	95th	Power at δ
15	0.181	0.207	0.002	0.005	0.025	0.098	0.269	0.496	0.638	0.36
30	0.098	0.148	0.0003	0.0009	0.006	0.034	0.125	0.293	0.428	0.57
60	0.034	0.077	<0.0001	<0.0001	0.0006	0.005	0.028	0.097	0.174	0.83
80	0.018	0.050	<0.0001	<0.0001	<0.0001	0.001	0.011	0.045	0.100	0.90

Table 2
Characteristics of the P-value distribution for various values of δ and the corresponding power at various values of δ for $n = 80$ and $\alpha = 0.05$

			Percentile							
δ	Mean	S.D.	5th	10th	25th	50th	75th	90th	95th	Power at δ
0.125	0.215	0.225	0.003	0.008	0.037	0.132	0.329	0.565	0.701	0.30
0.25	0.057	0.107	<.0001	0.0002	0.002	0.013	0.059	0.170	0.277	0.72
0.33	0.018	0.050	<.0001	<.0001	0.0001	0.001	0.011	0.045	0.100	0.90

$$\sigma = \{\lambda\sigma_1^2 + (1-\lambda)\sigma_2^2\}^{1/2},$$
$$\delta = (\mu_1 - \mu_2)/\sigma,$$

where $\lambda = m_2/(m_1 + m_2)$. Then for the given μ_1, μ_2, σ_1^2, and σ_2^2, T is asymptotically normal with mean $\sqrt{n}\delta$ and variance one. When the two groups have the same variance, σ is the common value of the standard deviation. If $m_1 = m_2 = m$, then $n = (1/2)m$. By following this convention, all the formulas in the previous sections are applicable to the two-sample scenario.

5. An Application

The P-value distribution under an alternative hypothesis has several applications. In the context of meta-analysis of several studies or analysis of a multicenter study, the P-value distribution may be used to explore the variability of the evidence as measured by the P-value against the null hypothesis when the same alternative parameter value is assumed for each of several studies or centers. This idea can also be applied to examine heterogeneity of observed treatment effects across subpopulations.

For the purpose of illustration, we select the example of meta-analysis taken by Fleiss (1993) in which the results of seven randomized studies of the effect of aspirin (versus placebo) in preventing death after a myocardial infarction were reviewed (see Table 4). The alternative hypothesis of interest is that the relative risk is greater than one, indicating that aspirin reduces mortality risk.

As noted by Fleiss, the observed relative risks for the first five studies appear homogeneous, varying over the narrow interval from 1.21 to 1.43. The log relative risk is approximately normally distributed for sufficiently large group sample sizes. Thus, applying the formulas of Section 4, the

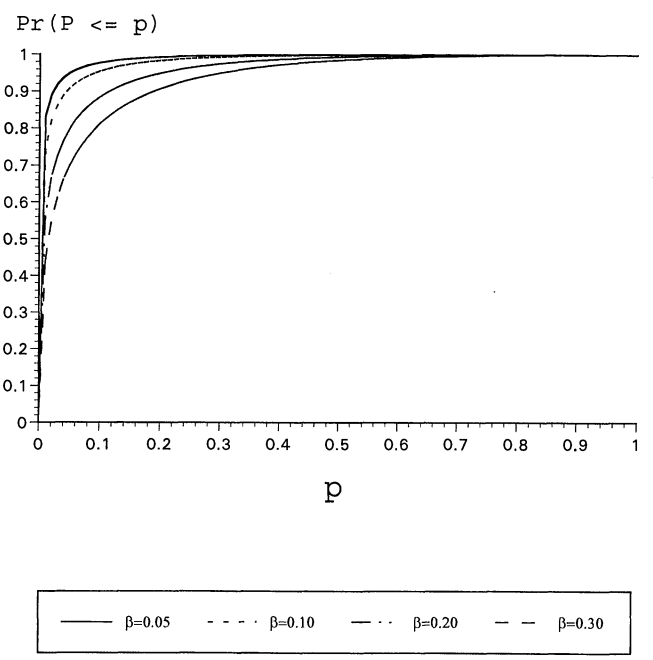

Figure 5. Cumulative distribution function of the P-value at various β chosen to detect a specified value of δ at $\alpha = 0.05$ and with sample size determined using equation (3.1).

Table 3

Characteristics of the P-value distribution at various power levels chosen for detecting any specified value of δ at $\alpha = 0.05$ and with sample size determined using equation (3.1)

Power (%)	Mean	S.D.	Percentile 5th	10th	25th	5th	75th	90th	95th
95	0.010	0.035	<0.0001	<0.0001	<0.0001	0.0005	0.004	0.022	0.050
90	0.018	0.050	<0.0001	<0.0001	0.0001	0.001	0.011	0.045	0.100
85	0.029	0.069	<0.0001	<0.0001	0.0004	0.004	0.022	0.080	0.150
80	0.039	0.085	<0.0001	<0.0001	0.0008	0.006	0.035	0.114	0.200
75	0.051	0.099	<0.0001	0.0002	0.001	0.010	0.050	0.150	0.250
70	0.063	0.113	<0.0001	0.0003	0.002	0.015	0.067	0.187	0.300

estimated value of δ ranges from 0.07 to 0.10 (see Table 4). Assuming that the expected effect size δ is 0.07 for each of the seven studies, we generate Figure 6 by accounting for the sample size of each study to illustrate the relative position of the observed P-value for each study in contrast to the 50th and 95th percentiles of the P-value distribution that would be expected for the respective sample sizes. Note that the 50th and 95th percentiles vary because of the different study sample size. The shaded region in this plot (we call it "P*hyp* plot") is the region of $P < 0.05$. This plot can be used to identify the study results where the observed P-values are improbable if the assumption is correct that each study has a common effect δ. From Figure 6, the observed P-value of the AMIS study stands far above the 95th percentile. This suggests the heterogeneity in the relative risk between the AMIS study and the first five studies.

Numerically, the relative risk observed in the ISIS-2 study also appears different from those of the first five studies. However, from Figure 6, the P-value for that study is well within the percentiles expected for a study of that sample size. The ISIS-2 gives an estimate of 0.04 for δ. Assuming that $\delta = 0.04$ for all studies, Figure 7 still suggests the heterogeneity in the relative risk between the AMIS study and the remaining studies.

One further remark is in order. An appealing feature of the P*hyp* plot method as compared with more complicated procedures for quantifying interstudy heterogeneity, such as chi-square tests for heterogeneity, is simplicity. The P*hyp* plot summarizes the result of a clinical trial in a single P-value statistic taking values between zero and one, and also provides percentile bounds that allow one to judge whether the observed P-value is in the expected range. The use of tests for outliers runs into the complication that no common underlying distribution exists independently of sample sizes of the studies or centers.

6. Distribution of the *P*-Value When the Tested Parameter Follows a Probability Distribution

In practice, the true value of δ is often not known with certainty and may vary from study to study. It is therefore appealing to view the effect size parameter δ as a random variable distributed over

Table 4

Results of seven placebo-controlled randomized studies of the effect of aspirin in preventing death after myocardial infarction

Study	Aspirin m_1	Aspirin p_1	Placebo m_2	Placebo p_2	n	$\log(rr)$	$\hat{\delta}$	One-sided p-value
MRC-1	615	0.0797	624	0.1074	310	0.2983	0.095	.047
CDP	758	0.0580	771	0.0830	382	0.3584	0.097	.029
MRC-2	832	0.1226	850	0.1482	420	0.1896	0.075	.063
GASP	317	0.1009	309	0.1230	156	0.1980	0.070	.191
PARIS	810	0.1049	406	0.1281	270	0.1998	0.074	.113
AMIS	2267	0.1085	2257	0.0970	1130	−0.1120	−0.038	.898
ISIS-2	8587	0.1828	8600	0.2000	4297	0.0899	0.044	.002

p_1, death rate of the aspirin group; p_2, death rate of the placebo group; m_1, sample size of the aspirin group; m_2, sample size of the placebo group; rr = relative risk = p_2/p_1; $n = m_1 m_2/(m_1 + m_2)$.

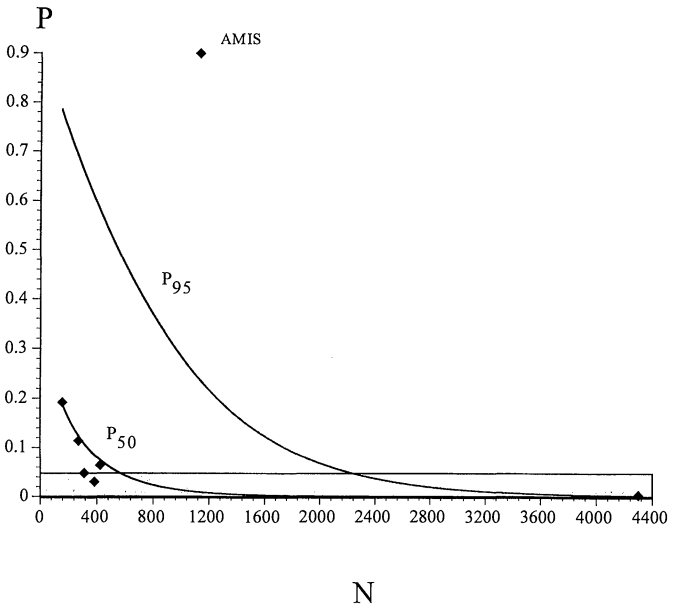

Figure 6. P*hyp* plot for the meta-analysis example (assuming $\delta = 0.07$).

a certain fixed range. We choose normal, uniform, and lognormal distributions for the effect size δ to study the distribution of the P-value. The mathematical derivation is given in the Appendix.

In what follows let p range over the interval $[0, 1]$. When δ is normally distributed with mean ζ and variance ω^2, the density of the P-value is

$$g(p) = [\omega(n+\omega^{-2})^{1/2}]^{-1} \times \exp\left\{-\frac{1}{2}\left[(\zeta/\omega)^2 - (\sqrt{n}Z_p + \zeta/\omega^2)^2/(n+\omega^{-2})\right]\right\}. \tag{6.1}$$

The distribution function of the P-value is

$$G(p) = 1 - \Phi\{(Z_p - \sqrt{n}\zeta)/(\omega^2 n + 1)^{1/2}\}. \tag{6.2}$$

If δ is uniformly distributed over the interval $[a, b]$, the marginal density function of the P-value is

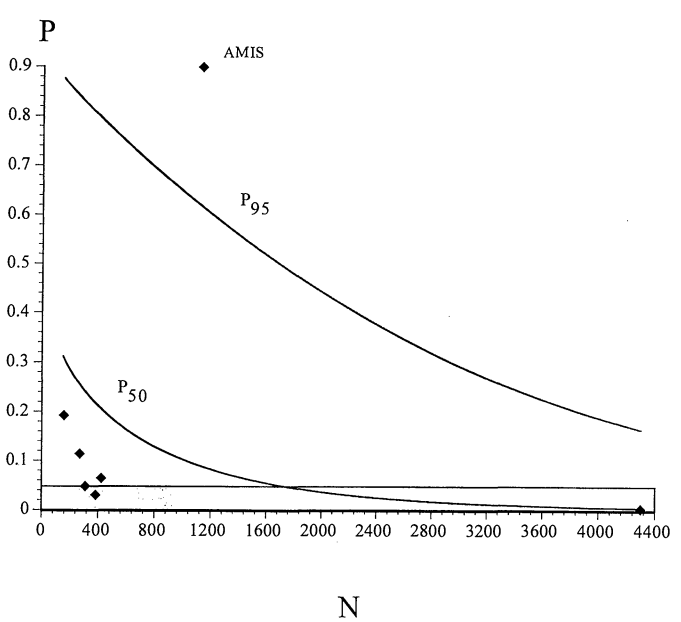

Figure 7. P*hyp* plot for the meta-analysis example (assuming $\delta = 0.04$).

$$g(p) = (2\pi/n)^{1/2}\exp(0.5Z_p^2)\{\Phi(\sqrt{n}b - Z_p) - \Phi(\sqrt{n}a - Z_p)\}/(b-a),$$

and the distribution function is

$$G(p) = \{\sqrt{n}(b-a)\}^{-1}\int_{Z_p}^{\infty}\{\Phi(v-\sqrt{n}a) - \Phi(v-\sqrt{n}b)\}\,dv.$$

For the lognormal case where $\log(\delta)$ is Gaussian with mean θ and variance ν^2, the marginal density of the P-value is

$$g(p) = \int_{-\infty}^{\infty}\phi(Z_p - \sqrt{n}e^{\nu u+\theta})\phi(u)/\phi(Z_p)\,du,$$

derived by use of the transformation $u = (\log(\delta) - \theta)/\nu$. The corresponding distribution function is given by

$$G(p) = \int_{-\infty}^{\infty}\Phi(\sqrt{n}e^{\nu u+\theta} - Z_p)\phi(u)\,du.$$

The mean and variance of the P-value for each of these assumed distributions of the parameter δ are derived in the Appendix. The average power function can be obtained similarly.

For the purpose of exploring how the distribution of the P-value, its first two moments, and percentiles change as the distribution of δ changes, we assign values to the parameters of the δ-distribution so that the expectation of δ is $1/3$. The following scenarios are considered:

1. δ is fixed at $1/3$.
2. δ is uniformly distributed on the interval $[0, 2/3]$; thus, δ has mean $1/3$ and variance $1/27$.
3. δ is Gaussian with mean $1/3$ and variance $1/27$.
4. δ is lognormal with mean $1/3$ and variance $1/27$.
5. δ is uniform on $[-2/3, 4/3]$ with mean $1/3$ and variance $1/3$.
6. δ is Gaussian with mean $1/3$ and variance $1/3$.
7. δ is lognormal with mean $1/3$ and variance $1/3$.

It can be seen from Figure 8 and Table 5 that the degree of uncertainty about the true value of δ affects the distribution of the P-value. The larger the variability of δ, the larger the expected value, the spread, and the percentiles of the P-value will be. The distributional characteristics of the P-value also depend on the distribution of δ. For instance, the P-value distribution obtained from an asymmetric δ-distribution (e.g., lognormal) can be quite different from that based on a symmetric one. The last column of Table 5 suggests that when δ for the potential medical environments of the trial is roughly uniformly distributed over the interval $[0, 2/3]$ for a "randomly" selected trial with a sample size of 80, the 5%-level sample mean test has an expected power of 72%. The

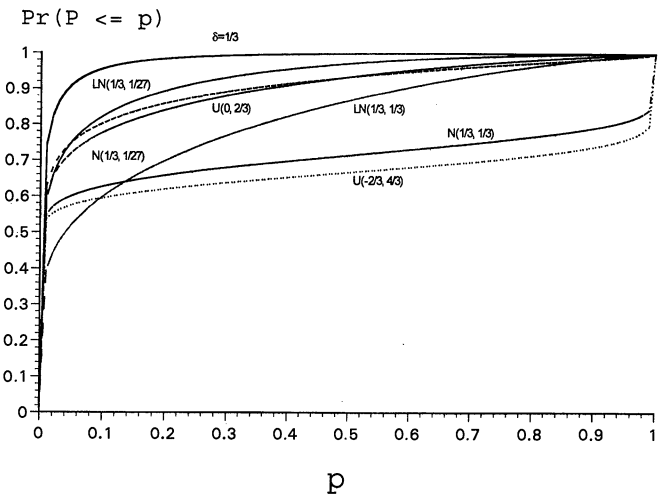

Figure 8. Cumulative distribution function of the P-value when the alternative δ follows various known distributions for $n = 80$.

Table 5

Characteristics of the P-value distribution when the alternative δ follows various known distributions and average power over the δ-distribution for $n = 80$ and $\alpha = 0.05$

δ distribution	Mean	S.D.	Percentile							Average power
			5th	10th	25th	50th	75th	90th	95th	
1/3	.018	.050	<.0001	<.0001	.0001	.001	.011	.045	.100	.90
U(0, 2/3)	.095	.196	<.0001	<.0001	<.0001	.001	.072	.364	.587	.72
N(1/3, 1/27)	.090	.204	<.0001	<.0001	<.0001	.001	.051	.334	.615	.75
LN(1/3, 1/27)	.064	.141	<.0001	<.0001	<.0001	.003	.050	.214	.377	.75
U(−2/3, 4/3)	.333	.437	<.0001	<.0001	<.0001	.001	.932	>.999	>.999	.57
N(1/3, 1/3)	.289	.411	<.0001	<.0001	<.0001	.001	.714	>.999	>.999	.60
LN(1/3, 1/3)	.174	.248	<.0001	<.0001	.0002	.041	.270	.586	.750	.52

U, uniform; N, normal; LN, lognormal.

concept of the expected power here is equivalent to "pretrial prediction" of a positive test decision in a Bayesian framework (Spiegelhalter, Freedman, and Parmer, 1994). The concept of expected power may be useful in the planning phase of a clinical trial.

8. Discussion

Several authors have recognized that distinguishing among a set of P-values generated from an unknown subset of true null hypotheses and an unknown subset of false null (or true alternative) hypotheses is a challenging effort. Schweder and Spjotvoll (1982) present a graphical procedure, called a P-value plot, which gives an overall view of the test statistics where it is possible to estimate the number of hypotheses that ought to be rejected. Parker and Rothenberg (1988) consider a similar problem, but present an approach that uses a mixture of several distributions to model the set of P-values (or test statistics) to characterize the expected P-value outcomes when multiple statistical tests have been carried out. Their approach is intended to distinguish P-values generated from false positive tests from those generated from true positive tests. The approach models one set of distributions for P-values consistent with a failure to reject the null hypothesis, while the other distributions in the mixture represent results inconsistent with the null hypothesis.

While all these authors use the fact that the distribution of the P-value from a statistical test performed on a true null hypothesis is uniform between 0 and 1 and that for a statistical test of a false null hypothesis the P-value would tend to be near 0, none of these authors take account of the sample size from which the P-value is generated nor the relationship of the P-value to the power of the test at a specific parameter value in the alternative hypothesis space. An exception is the work of Dempster and Schatzoff (1965) who considered the "expected significance level" defined as the expected value of the P-value when a particular alternative is true. They proposed estimation techniques based on Monte Carlo simulation. We believe the knowledge of the assumed alternative δ and the sample size of a given experiment are both necessary when looking into the problem of the distribution of the P-value under the alternative more closely. We also believe that, because the P-value is a random variable taking values in the interval $[0, 1]$, one can judge the consistency of the observed P-values of a set of studies against a common alternative hypothesis relative to the expected P-value (or percentiles) of each study when each study's sample size is properly accounted for.

One final remark concerns the fact that the P-value distribution under the alternative hypothesis depends on the distribution of the test statistic T used. In this paper, we consider the case that T is approximately normal, which is reasonable for many practical applications. Future work is needed to deal with the nonnormal cases.

Acknowledgements

The authors wish to thank the Associate Editor and two referees for their constructive comments, which improved the presentation of this paper.

Résumé

Le degré de signification p est une variable aléatoire déduite de la distribution de la statistique utilisée pour analyser des données et tester une hypothèse nulle. Sous cette hypothèse nulle p,

basé sur une statistique de test continue, a une distribution uniforme sur $[0, 1]$, indépendamment de la taille de l'échantillon. A contrario la distribution de p sous l'hypothèse alternative dépend à la fois de la taille de l'échantillon et de la vraie valeur ou de l'étendue des vraies valeurs du paramètre testé. Les caractéristiques, telles que la moyenne et les percentiles, de la distribution des valeurs de p peuvent apporter des éclaircissements appréciables sur le comportement des valeurs de p pour divers valeurs des paramètres et des tailles d'échantillons. Des applications potentielles de la distribution du degré de signification p sous l'hypothèse alternative sont considérées pour la conception, l'analyse et l'interprétation des résultats d'essais cliniques.

References

Dempster, A. P. and Schatzoff, M. (1965). Expected significance levels as a sensitivity index for test statistics. *Journal of the American Statistical Association* **60,** 420–436.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2,** 121–145.

Goodman, S. N. (1992). A comment on replication, *P*-values and evidence. *Statistics in Medicine* **11,** 875–879.

Parker, R. A. and Rothenberg, R. B. (1988). Identifying important results from multiple statistical tests. *Statistics in Medicine* **7,** 1031–1043.

Schweder, T. and Spjotvoll, E. (1982). Plots of *P*-values to evaluate many tests simultaneously. *Biometrika* **69,** 493–502.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials (with discussions). *Journal of the Royal Statistical Society, Series A* **157,** 357–416.

Received April 1995; *revised February* 1996; *accepted May* 1996.

Appendix

At first a fixed effect size δ is considered. The test statistic T is assumed to follow a standard normal distribution under the null hypothesis and a normal distribution with density $\phi_{\delta\sqrt{n}}$, mean $\delta\sqrt{n}$, and unit variance under the alternative hypothesis. The P-value is a one-to-one transformation of the test statistic, $P = 1 - \Phi(T)$. By using $dT/dP = -\{\phi(\Phi^{-1}(1-P))\}^{-1}$, the density of the P-value for fixed δ is

$$g_\delta(p) = \phi_{\delta\sqrt{n}}(Z_p)/\phi(Z_p), \tag{A.1}$$

where Z_p is the $(1-p)$th percentile of the standard normal distribution. Note that in principle the arguments up to here are general and do not depend on a specific assumption of the distributional form of the test statistic (as long as the distributions under the null and alternative hypotheses are completely specified). By using $v = Z_p$ and $dp = -\phi(v)\,dv$, the expectation of the P-value for a fixed δ is given by

$$\begin{aligned} \mathrm{E}_\delta(P) &= \int_0^1 p\phi(Z_p - \sqrt{n}\delta)/\phi(Z_p)\,dp \\ &= \int_{-\infty}^{\infty} \Phi(-v)\phi(v - \sqrt{n}\delta)\,dv. \end{aligned} \tag{A.2}$$

The second moment can be similarly obtained.

Consider that δ is a random variable with density $h(\delta)$. The marginal density of the P-value is then given by

$$g(p) = \int_{-\infty}^{\infty} g_\delta(p)h(\delta)\,d\delta.$$

With a normal δ with mean ζ and variance ω^2, the transformation $\delta' = \delta(n + 1/\omega^2)^{1/2}$ leads to equation (6.1). $G(p)$ given in (6.2) is derived using $v = Z_p$ and $u = v(\omega^2 n + 1)^{-1/2}$. Using $\delta' = \delta(n + 1/\omega^2)^{1/2}$ and $u = (v - \zeta\sqrt{n})(\omega^2 n + 1)^{-1/2}$, we obtain

$$\mathrm{E}(P) = \int_{-\infty}^{\infty} \Phi(-v(\omega^2 n + 1)^{1/2} - \sqrt{n}\zeta)\phi(v)\,dv, \tag{A.3}$$

$$\mathrm{E}(P^2) = \int_{-\infty}^{\infty} [\Phi(-v(\omega^2 n + 1)^{1/2} - \sqrt{n}\zeta)]^2 \phi(v)\, dv. \tag{A.4}$$

Likewise, for δ being uniformly distributed on $[a, b]$,

$$\mathrm{E}(P) = \{\sqrt{n}(b-a)\}^{-1} \int_{-\infty}^{\infty} \Phi(-v)\{\Phi(v - \sqrt{n}a) - \Phi(v - \sqrt{n}b)\}\, dv. \tag{A.5}$$

For $\log(\delta) \sim \mathrm{N}(\theta, \nu^2)$,

$$\mathrm{E}(P) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(-v - \sqrt{n}e^{\nu u + \theta})\phi(v)\phi(u)\, du\, dv. \tag{A.6}$$

The second moment of the P-value can be similarly derived.