

ON NEYMAN-PEARSON THEORY: INFORMATION CONTENT OF AN EXPERIMENT AND A FANCY PARADOX

Benito V. Frosini

1. INFORMATION CONTENT OF AN EXPERIMENT

1.1. Ordinal and cardinal information measures

Given two or more populations or random variables – univariate or multivariate – an experiment, usually consisting of drawing a random sample of elements or observations from *one* of these populations or variables, is aimed at providing useful information on the population or variable the sample comes from. In mathematical statistics it is common to think of an infinite family of random variables, indexed by a parameter – uni- or multi-dimensional – whose domain is unbounded; controls about assumptions and approximations are of course necessary when transferring the theoretical results to reality. The problem of comparing several possible experiments, concerning the ability of discriminating between the parent populations, naturally arises. As for many other decision problems, one would dispose of a real valued function, taking values on the real line, able to finely discriminate the possible experiments, projected in order to support the above decision problem. Such a solution is not practically feasible, at least if we want an order preserving function, applied to a large set of random variables which are comparable according to a widely accepted criterion.

One criterion devised by Frosini (1993, pp. 369-370) for the *ordinal* comparability of distributions, is based on likelihood functions (or LFs). Let $f(x; \theta)$ be the LF of θ given the sample x , and $R(x; \theta) = f(x, \theta) / \sup f(x, \theta)$ be the relative likelihood of θ given x ; for two experiments E_1 and E_2 (possibly the same experiment), with sample spaces S_1 and S_2 , and two samples $x_1 \in S_1$, $x_2 \in S_2$, the relative likelihood is $R_1(x_1; \theta)$ and $R_2(x_2; \theta)$ respectively. Then a comparison between LFs can be established by means of the following definition.

Comparison of likelihoods: An LF $f_1(x_1; \theta)$ with corresponding relative $R_1(x_1; \theta)$ is said to be *more informative* than the LF $f_2(x_2; \theta)$ with corresponding relative $R_2(x_2; \theta)$, if the following subset relation is satisfied:

$$\{\theta : R_1(x_1; \theta) \geq c\} \subset \{\theta : R_2(x_2; \theta) \geq c\} \quad (1)$$

for every $0 < c < 1$, with a proper subset relation valid for some c .

This definition is justified by the fact that the set of parameter values having plausibility $\geq c$, given the sample, belongs to a smaller neighbourhood of the maximum likelihood estimate (MLE) in the former case than in the latter case.

Generalizing this definition to every pair of samples x_1, x_2 , yielding the same MLE, we obtain a partial ordering of experiments.

Comparison of experiments: Based on the above reference, experiment E_1 is said to be *more informative* than experiment E_2 if relation (1) holds for every $0 < c < 1$, and every sample $x_1 \in S_1, x_2 \in S_2$ producing the same MLE, with strict relationship in some case.

Such an ordinal comparison – as well as other similar criteria – is of course applicable in rather special cases for a given sample size (Frosini, 1993, p. 370; Frosini, 1991), giving rise to a partial ordering of the kind “more informative than” for experiments sharing the same parameter space. Other partial orderings are feasible when embedded into a decision problem: “[the experiment] E is more informative than F , if to any decision problem and any risk function which is obtainable in F corresponds an everywhere smaller risk function obtainable in E ” (Torgersen, 1976, p. 186).

In most cases, nevertheless, it is possible to lean over simple functions, ensuring *cardinal* comparability; we shall briefly report two such information measures. The older one is Fisher’s Information; if $L = L(X; \theta)$ is the likelihood function for θ (uni-dimensional) given the random sample X produced by the experiment E , Fisher’s information is defined by

$$I_F = E_\theta \left\{ \left(\frac{\partial \log L}{\partial \theta} \right)^2 \right\} = -E_\theta \left\{ \frac{\partial^2 \log L}{\partial \theta^2} \right\} \quad (2)$$

which corresponds to the reciprocal of the variance of an efficient unbiased estimator. When θ is multi-dimensional, the generalization of (2) leads to an information matrix and inequalities between quadratic forms, scarcely useful for comparative purposes (Wilks, 1962, pp. 352 and 378).

The application of Fisher’s Information requires the fulfilment of very strong mathematical conditions, inside a given parametric model. Of course it is not applicable to any finite set of distributions.

Another well known measure of information is the Kullback-Leibler measure, especially aimed at discriminating between distributions belonging to a given set. If the hypotheses H_0 and H_1 imply probability distributions π_0 and π_1 , with densities f_0 and f_1 over the points ω of a space Ω , the mean information per observation from π_1 for discrimination in favour of H_0 against H_1 when H_0 is true is defined by

$$I(0:1) = E \left\{ \ln \frac{f_0(\varpi)}{f_1(\varpi)} \middle| H_0 \right\} \quad (3)$$

(Kullback, 1959, p. 6; 1983, p. 422). Although formal justifications and properties are well founded, this information measure lacks in meaning, where the values in the right side of (3) cannot be directly connected to probabilities or other characteristics of distributions.

1.2. The power of a test as an information measure

One promising property of the Kullback-Leibler information is the following inequality, which relates the definition (3) with the error probabilities α and β of a Neyman-Pearson test. Let us assume that the space Ω relates to n independent observations of a random variable X , $\varpi = (x_1, \dots, x_n)$, and consider a Neyman-Pearson test with E_0 the acceptance region of hypothesis H_0 and E_1 the acceptance region of hypothesis H_1 ($E_0 \cap E_1 = \emptyset$; $E_0 \cup E_1 = \Omega$). As usual, let α and β be the error probabilities; if H_0 is the null hypothesis, we can put

$$\alpha = P(\varpi \in E_1 \mid H_0); \quad \beta = P(\varpi \in E_0 \mid H_1).$$

If $I(0:1)$ refers to the space Ω just defined, the following inequality holds (Kullback, 1959, p. 74):

$$I(0:1) \geq \alpha \ln \frac{\alpha}{1-\beta} + (1-\alpha) \ln \frac{1-\alpha}{\beta} = F(\alpha, \beta). \quad (4)$$

$F(\alpha, \beta) = 0$ for $\alpha = 1 - \beta$; for fixed α , $F(\alpha, \beta)$ is monotonically decreasing for $0 \leq 1 - \beta \leq \alpha$, or $1 - \alpha \leq \beta \leq 1$, and monotonically increasing for $\alpha \leq 1 - \beta \leq 1$, or $0 \leq \beta \leq 1 - \alpha$. With increasing sample size, and maintaining a constant α , we expect a regular reduction of β , i.e. an increase in the power $1 - \beta$; the existence of an interval for $1 - \beta$ showing decreasing values of $F(\alpha, \beta)$ could be disturbing; however, we can observe perfect coherence (although $F(\alpha, \beta)$ is only a lower bound for the information measure) if we refer to unbiased tests, as in such cases $1 - \beta \geq \alpha$. Thus, as n increases - with fixed α - we are bound to observe an increase in the power $1 - \beta$ from the power $\geq \alpha$ calculable for $n = 1$, entailing increasing the lower bound $F(\alpha, \beta)$ for the Kullback-Leibler discrimination measure.

Now, resuming the observation that the values taken by this measure do not transmit any clear operational meaning, and also that no upper bound exists, for the same contest it is possible to take a step forwards, or perhaps backwards, leaning on the solid support of the power itself (or the complementary probability β of the type II error). On the other hand, as back as 1935 Neyman called the

attention to the “errors of the second kind” in order to establish a sensible evaluation of an experiment; and about twenty years later (Neyman, 1956, p. 290) he stressed that “the numerical values of probabilities of errors of the second kind are most useful for deciding whether or not the failure of a test to reject a given hypothesis could be interpreted as any sort of “confirmation” of this hypothesis”.

This same viewpoint was taken by Blackwell (1951), and mostly by Lehmann (1959). As Lehmann (1959, p. 75) writes, “Let $\beta(\alpha)$ and $\beta'(\alpha)$ denote the power of the most powerful level α test based on X and X' . In general, the relationship between $\beta(\alpha)$ and $\beta'(\alpha)$ will depend on α . However, if $\beta'(\alpha) \leq \beta(\alpha)$ for all α , then X or the experiment (f, g) [X has probability densities f and g under the null hypothesis H and the alternative K , respectively] is said to be more informative than X' ”.

In recent years I have been involved in the assessment of several epidemiologic studies, mostly for their relevance in civil and criminal cases; many of them have been published in qualified scientific journals. Of course they were not on the same footing on many respects, especially concerning sample sizes; the best summary for the assessment and comparison of the several “experiments” has been the power of these studies; “power quantifies the ability of a particular study to detect an excess risk that truly exists” (Beaumont & Breslow, 1981, p. 726). For the convenience of disposing of single numbers, as usual in occupational epidemiology I assumed a standard α equal to 0.01 or 0.05, and made a comparison of the situation of no excess risk for some causes of death (practically, what was known for the population at large) with a situation of double (or triple) risk, hypothesized for the particular sample of workers. Considerations of power are of utmost importance, because the probability α of the type first error can be fixed at will – usually at standard values – and through these one cannot obtain any grasp on the information content of the experiment.

In scientific research no preference should be given to any of the hypotheses, thus the equality $\alpha = \beta$ seems advisable. In such a case the problem of assessing the information content of the experiment leads to clearer solutions. For example, if the hypotheses are $H_0 = N(\mu_0, \sigma^2)$, $H_1 = N(\mu_1, \sigma^2)$, $\mu_1 > \mu_0$, α and β are equalized at

$$\alpha = \beta = P\left(Z > \frac{(\mu_1 - \mu_0)/2}{\sigma/\sqrt{n}}\right).$$

For $\mu_1 - \mu_0 = 10$, and the two values for $\sigma = 10$ and 20, Figure 1 shows the behaviour of the power ($1 - \beta = 1 - \alpha$) for sample sizes from 1 to 40. As the Kullback-Leibler information (3) equals, in the case at hand,

$$n(\mu_1 - \mu_0)^2 / (2\sigma^2) = 50n / \sigma^2,$$

the two Kullback-Leibler information curves are – respectively for $\sigma = 10$ and $\sigma = 20$ – $I(0 : 1) = 0.5n$ and $I(0 : 1) = 0.125n$. There seems to be no case for the choice of the power $1 - \beta$ instead of $I(0 : 1)$, in order to get a real and useful information about the experiments characterized by increasing values of n .

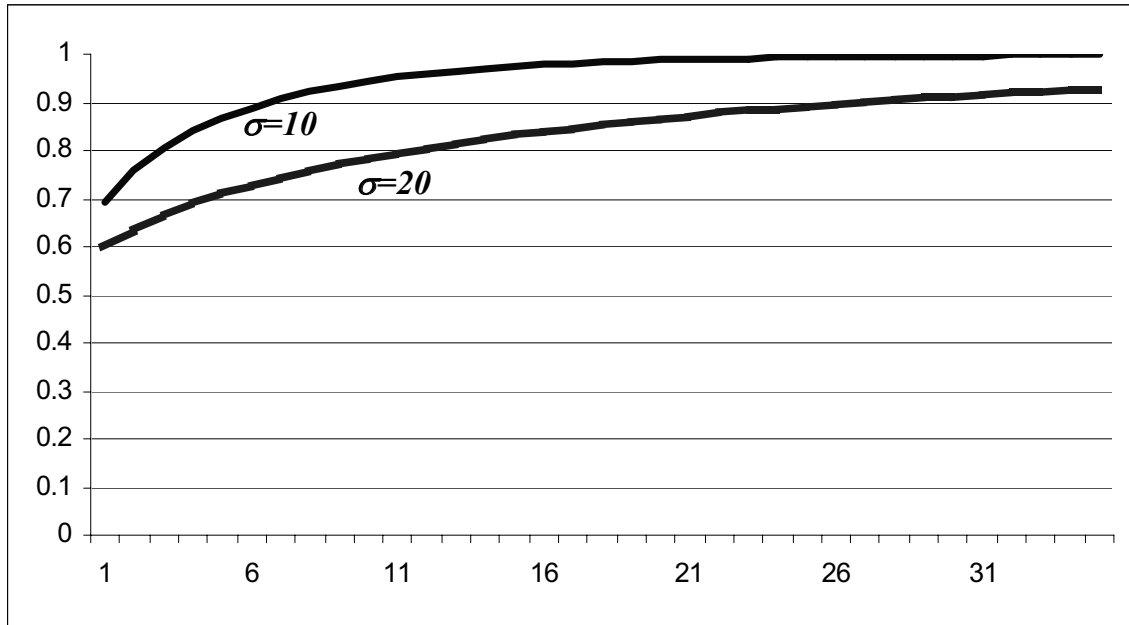


Figure 1 – Power values for a test regarding the mean of a normal variable, under two hypotheses for the standard deviation σ , for $n = 1, \dots, 40$.

Some more examples, linked with the epidemiological problems referred to above, have to do with a given cause of death, and with distributions of deaths of the Poisson type. We can start with a Poisson distribution $H_1 = \text{Poi}(7.5)$ (with parameter $\lambda = 7.5$), comparing it with the alternative $H_2 = \text{Poi}(15)$, then taking this as the null hypothesis and comparing it with $H_3 = \text{Poi}(30)$, and finally taking $H_4 = \text{Poi}(60)$ as the alternative to H_3 (each time doubling the risk). Two α values, $\alpha = 0.05$ and $\alpha = 0.01$, have been considered; as the Poisson distribution is discrete, exact values were obtained by randomization. Crossing the three tests with the two α values, we have obtained the corresponding β values in Table 1.

TABLE 1
 *β values for tests comparing Poisson distributions, according to $\alpha = 0.05$ and $\alpha = 0.01$
 (exact α values, and corresponding β values, obtained by randomization)*

	Poi(7.5) vs Poi(15)	Poi(15) vs Poi(30)	Poi(30) vs Poi(60)
$\alpha = 0.05$	0.2510	0.0584	0.0024
$\alpha = 0.01$	0.4659	0.1691	0.0130

Contrary to the case of normal distributions with same σ – examined above – the information measure $I(0 : 1)$ reveals its very nature of *directed* divergence

when applied to Poisson distributions; in fact, if $H_0 = \text{Poi}(\lambda_0)$ and $H_1 = \text{Poi}(\lambda_1)$, we obtain

$$I(0:1) = \lambda_0 \ln(\lambda_0 / \lambda_1) + \lambda_1 - \lambda_0$$

(Kullback, 1959, p. 73). For the three comparisons in Table 1 we can obtain the values for $I(0:1)$, respectively: 2.3014, 4.6028, 9.2056; reversing the order of null and alternative hypotheses, the calculation of $I(0:1)$ gives: 2.8972, 5.7944, 11.5888. A symmetric measure, suggested by Kullback, is simply the sum $J(0:1) = I(0:1) + I(1:0)$. Also in this case the usefulness to resort to the couple (α, β) appears crystal clear in order to appreciate the information contents of the experiments involved.

2. A FANCY PARADOX

2.1. Point null hypotheses

It is not surprising that Bayesian inference and Neyman-Pearson inference, being based on very distant assumptions and operational characteristics, can produce quite different results in specific contexts and problems; as the two approaches give different answers to different questions, it would be silly to make comparisons of the answers, without taking into account the fundamental gap in questions and assumptions. In my opinion, a sensible comparison can be made only with reference to: (1) a specific *real* context, taking into account all the information available, (2) a specific question that must be answered (e.g. within a case in court, the comparison of two drugs etc.), and (3) the persons that are expected to make use of the inferential conclusions. In particular, the relevance of points (1) and (2) must be evaluated with respect to the fundamental distinction between *model* and *reality*: when the inference is heavily founded on the specific assumptions of the model (which exists only in our minds), the theoretical conclusions may scarcely address the questions concerning the real problem.

A well known disagreement between theory and practice, unfortunately without due warning in most textbooks, regards *point null hypotheses*; for example: (a) the mean of a continuous random variable is equal to 5 (or another *exact* real number); (b) two or more random variables are independent; (c) the distribution of a certain characteristic is Normal, or Poisson etc. All such things have non-real existence; as for all other applications of mathematics to the real world, one must be careful to check the implications of such strong assumptions. A distressful implication of point null hypotheses results as an outcome of the *consistency property* usually required for all inference procedures (cf. Frosini, 2001, p. 374): it is well known that, if we take a sufficiently large sample, any point null hypothesis is bound to be rejected! This should not be surprising, as the chosen null hypothesis cannot be true, or at least its truth is impossible to recognize; as the information

increases, it is less and less likely to accept the null hypothesis, and this fact is absolutely correct, as the null hypothesis – taken literally – is certainly false (or practically certainly so).

All this means that classical tests of (point null) hypotheses are practically acceptable only for “small” sample sizes, where *small* is to be deemed according to the precision of the random variables involved; when the sample size is small, the sampling variability of the test statistic is generally so large as to dominate the imprecise (being *too precise*) specification of the null hypothesis. As we let the sample size increase, we must acknowledge the growing unsuitableness of the test procedure in order to answer the practical problem in the real world. Among the possible solutions: (1) avoid applications of such tests in case of large samples, and limit to estimation procedures; (2) restate the problem in more acceptable terms, e.g. by fixing intervals for parameters: $H_0 = \theta \in (a, b)$.

2.2. A Bayesian approach to point null hypotheses

One feature of the Neyman-Pearson inference for point null hypotheses is that no assessment is made of the probability attached to the null hypothesis – and let $H_0 : \theta = \theta_0$. Such a statement could sound obvious, as the N-P approach does not have recourse to (usually subjective) probabilities of hypotheses; however, the question is not so sharp, being as much obvious that a research worker in an empirical science, who judges appropriate the recourse to N-P approach for a specific problem, can nonetheless elicit subjective probabilities for the hypotheses at hand; and it is quite possible that the null hypothesis is not deemed as most likely. For example, in the quality assessment of an industrial product, or in the risk assessment connected to the exposure to a chemical compound, it is perfectly allowed that the *reference* parameter values under test are not the ones most likely (for the specific instance) according to the researcher’s opinion. Thus, the assumption – made by some Bayesian scholars, to be discussed in the sequel – that the point null hypothesis $H_0 : \theta = \theta_0$ is judged the most likely is just a “mathematical hypothesis” for dealing with a mathematical – not inferential – problem.

Anyway, although accepting that $\theta = \theta_0$ is the most likely hypothesis, the main problem remains: is it reasonable that our researcher attaches a finite value (for example 1/2) to the probability $P(\theta = \theta_0)$, when $\theta = \theta_0$ is a point null hypothesis of the kind presented in section 2.1? No doubt that such an assessment is wholly unreasonable, given that the hypothesis $\theta = \theta_0$ is certainly false (or practically so). In principle, this fact is recognized also by some Bayesians; for example, Berger (1985, p. 148) writes: “... tests of point null hypotheses are commonly performed in inappropriate situations. It will virtually never be the case that one seriously entertains the possibility that $\theta = \theta_0$ *exactly* (cf. Hodges and Lehmann (1954) and Lehmann (1959)). More reasonable would be the null hypothesis that

$\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$, where $b > 0$ is some constant chosen so that all θ in Θ_0 can be considered “indistinguishable” from θ_0 . This last clarification by Berger must be evaluated, in our opinion, only as a possible instance of application, being true, in general, that the hypothesis

$$H_0 : \theta \in (\theta_0 - b, \theta_0 + b) \quad b > 0$$

must be given growing probabilities for increasing b values, i.e. by enlarging the set of parameter values.

Although maintaining a critical approach, Berger (1985, p. 149) works out the sharp approximation $H_0 : \theta = \theta_0$ with respect to $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, warning that “the approximation is reasonable if the posterior probabilities of H_0 are nearly equal in the two situations” (a rather strong condition). Following the approach – and using some results – laid out by Jeffreys (1939, 1948), Berger (1985, pp. 150-151) obtains some posterior probabilities, which can lead him to speak of *astonishing* comparisons with N-P approach. Starting from a prior probability distribution over the parameter values given by a positive probability π_0 attached to θ_0 , and a density $\pi_1 g_1(\theta)$ for $\theta \neq \theta_0$, with $\pi_1 = 1 - \pi_0$ and g_1 proper, the posterior probability of θ_0 given the observation $X = x$ with conditional density $f(x|\theta)$, is easily obtained:

$$\pi(\theta_0 | x) = \left[1 + \frac{1 - \pi_0}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)} \right]^{-1} \quad (5)$$

where $m_1(x)$ is the marginal density of X with respect to g_1 . To continue, it is enough to stick to the example worked out by Berger, and largely exploited in the Bayesian literature. Suppose a sample (X_1, \dots, X_n) is observed from a normal distribution $N(\mu, \sigma^2)$, σ^2 known. Reduction to the sufficient statistic \bar{X} leads us to consider an observation of the sample average \bar{X} from a normal $N(\mu, \sigma^2/n)$. The prior density g_1 is supposed normal $N(\mu, \tau^2)$ over $\theta \neq \theta_0$. In the special case $\mu = \theta_0$ the above formula takes a very simple appearance, where $z = \sqrt{n}|\bar{x} - \theta_0|/\sigma$ is the usual statistic employed for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ in the N-P approach:

$$\alpha_0 = \pi(\theta_0 | \bar{x}) = \left[1 + \frac{1 - \pi_0}{\pi_0} \cdot \frac{\exp\{0.5 \cdot z^2 [1 + \sigma^2/(n\tau^2)]^{-1}\}}{\{1 + n\tau^2/\sigma^2\}^{1/2}} \right]^{-1} \quad (6)$$

Making the further simplifying assumption $\tau = \sigma$, and putting $\pi_0 = 0.5$, Berger computes the values of $\alpha_0 = \pi(\theta_0 | \bar{x})$ for some couples (z, n) . For example, the row of the table corresponding to the standard value $z = 1.96$ (P-value 0.05) is the following:

n	1	5	10	20	50	100	1000
α_0	0.35	0.33	0.37	0.42	0.52	0.60	0.80

Berger (1985, p. 151) observes that “classical theory would allow one to reject H_0 at level $\alpha = 0.05$ But the posterior probability of H_0 is quite substantial, ranging from about 1/3 for small n to nearly 1 for large n . Thus $z = 1.96$ actually provides little or no evidence against H_0 (for the specified prior)”. With respect to this point, let it be sufficient to repeat that the above α_0 probabilities are correct only in a mathematical sense, but are completely devoid of inferential meaning – for any real problem – just because of the absurdity inherent in a prior distribution which attributes the value 0.5 to a specified real number, and a density integrating to 0.5 over the whole real line (excepting the fixed number). This and several related questions are treated in a paper by Shafer (1982), followed by a very interesting discussion.

About $\mu = \theta_0$ something was already said at section 2.1, but it is worth repeating that a point null hypothesis H_0 is simply a hypothesis of which the researcher wants to assess the conformity with respect to the available data (cf. Frosini, 2001, p.347). About the comparative treatment of the complementary hypotheses H_0 and H_1 , it can be said that the researcher is *impartial* (not *objective*) when he is able to establish the equality between the error probabilities α and β of I and II kind – when both hypotheses are simple – or between a sensible choice of such probabilities attached to representative hypotheses within H_0 and H_1 , in case of composite hypotheses. Impartiality or objectivity are concepts non applicable in a case in which we make a comparison of a hypothesis comprising only one real number θ_0 , against an alternative comprising an infinite interval of real numbers (in the case at hand, the whole real line excepting θ_0). In any case, the choice $\pi_0 = 1/2$ for the above problem appears totally arbitrary, although admitting for a moment the attribution of a positive probability to $\theta = \theta_0$. All that can be said, in general, about $\pi_0 = P(H_0)$, is that such probability is bound to increase with $b > 0$ (or at least not decrease) if H_0 refers to the interval $\Theta_0 = (\theta_0 - b, \theta_0 + b)$; with very small b , π_0 can reasonably assume positive values very near to zero.

Dwelling a little longer on formula (6), and evaluating it as an approximation with respect to a null hypothesis $H_0 : \theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$ with b very small (cf. Hill, 1982, p. 345), it is worth while making the following comments concerning α_0 as a function of $\pi_0, z, \gamma^2 = \tau^2 / \sigma^2$, and n :

(a) α_0 is increasing with π_0 : this is quite correct, in the sense that with a very small interval Θ_0 we should give the interval a very small prior probability π_0 , resulting with a small or very small posterior probability α (contrary to what was obtained with $\pi_0 = 1/2$);

(b) α_0 is decreasing with \varkappa : this is also correct, because a large standardized distance \varkappa (from θ_0) is an indication of data which do not comply (or agree) with the hypothesis H_0 ;

(c) $\gamma^2 = \tau^2/\sigma^2$ and n are strictly tied in determining the behaviour of α_0 ; as γ^2 , or n , or both, increase, α_0 increases too, at least from a certain point farther, becoming closer and closer to one.

This last phenomenon – and especially the one depending on large n values – results from a likelihood ratio $L(H_0, H_1)$ tending to infinity: “This is the phenomenon that Jeffreys (1939, 1948) discovered, and that was called by Lindley (1957) a paradox” (Hill, 1982, p. 345; Berger, 1985, p. 156). However this fact depends heavily on $\pi_0(\cdot)$ being degenerate at θ_0 , and is clearly the most outstanding feature of an absurd choice for the prior distribution.

From the points (a) and (b) above, it can be interesting to examine the iso- α_0 curves, such as those in Figure 2, obtained by joining points (π_0, \varkappa) with $\alpha_0 = 0.05$; the curves in Figure 2 are computed for $n = 20$ and two values for $\gamma^2 = \tau^2/\sigma^2 = 0.5$ and 20.

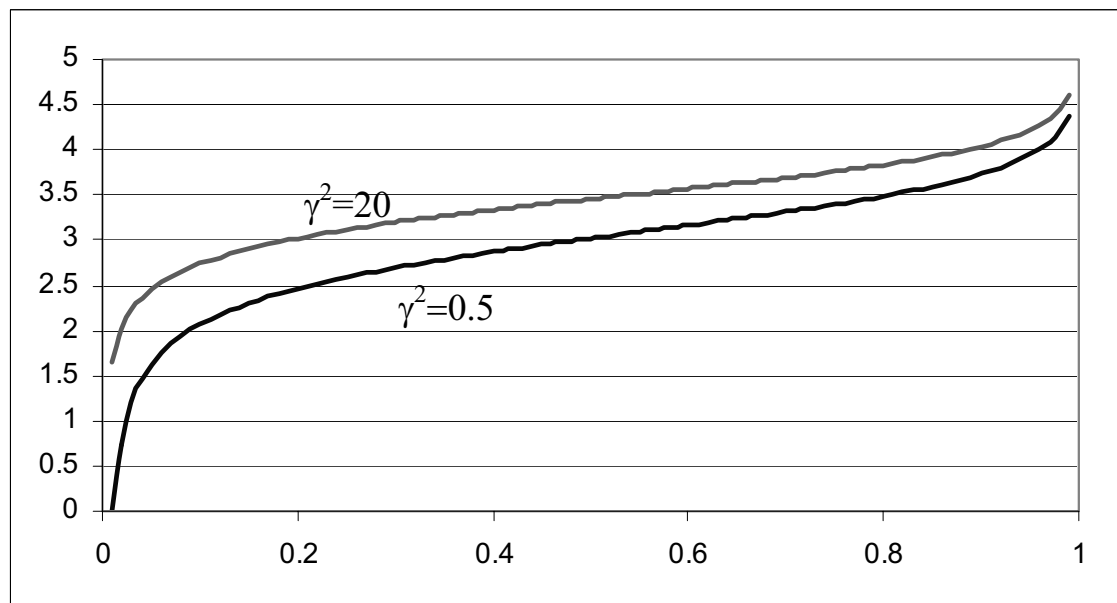


Figure 2 – Iso- α_0 curves of points (π_0, \varkappa) , given $\alpha_0 = 0.05$, $n = 20$ and two γ^2 values.

2.3. Sensible results from sensible hypotheses

Although never forgetting the essential distinctive traits between Bayesian and Neyman-Pearson approaches, some possibility of making a limited comparability exists, provided we drop the strict sharp hypothesis $H_0 : \theta = \theta_0$, and accept a null hypothesis of type $H_0 : \theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$. Just to simplify mathematics at the outset (or else, a translation could be applied later), let $\theta_0 = 0$, so that our hypotheses in the N-P approach are as follows:

$$H_0 : \theta \in \Theta_0 = [-b, b]; \quad H_1 : \theta \notin \Theta_0, \text{ or } \theta \in \Theta_1 = (-\infty, -b) \cup (b, \infty)$$

In the sequel H_i and Θ_i ($i = 1, 2$) will be used interchangeably.

As before (see Berger's example) the reference is to a random sample (X_1, \dots, X_n) from a normal $N(\theta, 1)$. As the boundary set of H_0 is given by $\omega = \{-b, b\}$, the unbiased test of H_0 against H_1 requires fixing the same power ψ of the test over the boundary (Lehmann, 1986, p. 134): $\psi(-b) = \psi(b)$ (probabilities of the sufficient statistic \bar{X} falling in the critical region, under $\theta = -b$ and $\theta = b$, respectively). Owing to symmetry, the critical region will be the union of \bar{x} values $(-\infty, -k) \cup (k, \infty)$, with k determined by putting (being $Z \sim N(0, 1)$):

$$P(\bar{X} < -k \mid -b) + P(\bar{X} > k \mid -b) = P(\bar{X} < -k \mid b) + P(\bar{X} > k \mid b) = \alpha$$

or

$$P(Z < \sqrt{n}(k - b)) + P(Z > \sqrt{n}(k + b)) = \alpha \tag{7}$$

By fixing, just as an example, $\alpha = 0.05$, the equation (7) has been solved for k , for the seven cases of $n = 1, 5, 10, 20, 50, 100, 1000$, and $b = 0.02, 0.10, 0.20, 0.40$. For each of these 28 cases, the power function – the fundamental tool of N-P approach – has been computed, according to the expression

$$\psi(\theta) = P(Z < -\sqrt{n}(k + \theta)) + P(Z > \sqrt{n}(k - \theta)); \tag{8}$$

two examples are reported in Figure 3.

- $\pi(\theta)$ the density of the prior distribution for $\theta \in (-\infty, \infty)$;
- $f(\bar{x} \mid \theta)$ the likelihood for the sample average;
- $\pi(\theta \mid \bar{x})$ the posterior density of θ given \bar{x} ;
- π_0 the prior probability of the null hypothesis H_0

$$\pi_0 = P(H_0) = \int_{H_0} \pi(\theta) d\theta = \int_{H_0} \pi_0 g_0(\theta) d\theta,$$

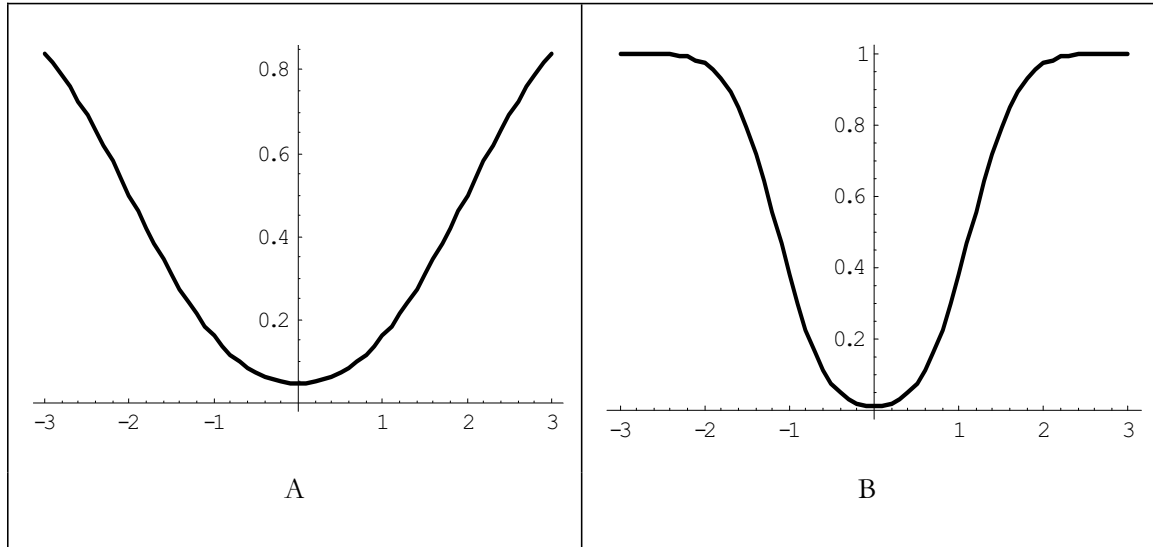


Figure 3 – Two examples of power function (8). A: $n = 1$, $b = 0.2$; B: $n = 5$, $b = 0.4$.

$g_0(\cdot)$ being a proper density over H_0 ;

– $\pi_1 = 1 - \pi_0$ the prior probability of the alternative hypothesis H_1

$$\pi_1 = P(H_1) = \int_{H_1} \pi(\theta) d\theta = \int_{H_1} \pi_1 g_1(\theta) d\theta,$$

$g_1(\cdot)$ being a proper density over H_1 ;

– $p_0 = \pi_0(\bar{x})$ the posterior probability of the null hypothesis

$$p_0 = P(H_0 | \bar{x}) = \int_{H_0} \pi(\theta | \bar{x}) d\theta \propto \int_{H_0} \pi_0 g_0(\theta) \cdot f(\bar{x} | \theta) d\theta;$$

– $p_1 = \pi_1(\bar{x})$ the posterior probability of the alternative hypothesis

$$p_1 = P(H_1 | \bar{x}) = \int_{H_1} \pi(\theta | \bar{x}) d\theta \propto \int_{H_1} \pi_1 g_1(\theta) \cdot f(\bar{x} | \theta) d\theta.$$

Four prior distributions have been chosen; the first three appear as direct generalizations of the one applied by Berger for an analogous problem (see section 2.2). The first has been built starting from a prior $Z \sim N(0, 1)$ over the whole parameter space $H_0 \cup H_1$, then rising the central part over the interval $[-b, b]$ by the multiplicative constants:

$$0.5/P(-0.02 \leq Z \leq 0.02) = 0.5/0.0159566 \text{ for } b = 0.02,$$

$$0.5/P(-0.10 \leq Z \leq 0.10) = 0.5/0.0796556 \text{ for } b = 0.10,$$

$$0.5/P(-0.20 \leq Z \leq 0.20) = 0.5/0.1585194 \text{ for } b = 0.20.$$

$$0.5/P(-0.40 \leq Z \leq 0.40) = 0.5/0.3108434 \text{ for } b = 0.40,$$

and lowering the tails from $-\infty$ to $-b$ and from b to ∞ by the multiplicative constants

$$\begin{aligned} 0.5/0.9840434 \text{ for } b = 0.02; & \quad 0.5/0.9203444 \text{ for } b = 0.10; \\ 0.5/0.8414806 \text{ for } b = 0.20; & \quad 0.5/0.6891566 \text{ for } b = 0.40. \end{aligned}$$

Figure 4A shows the prior density for the case $b = 0.40$. Such constants succeed in equalizing to $1/2$ the prior probabilities π_0 and π_1 . All the calculations presented in Table 2 employ this choice of the prior probabilities, just as the ones used by Berger in the example commented above.

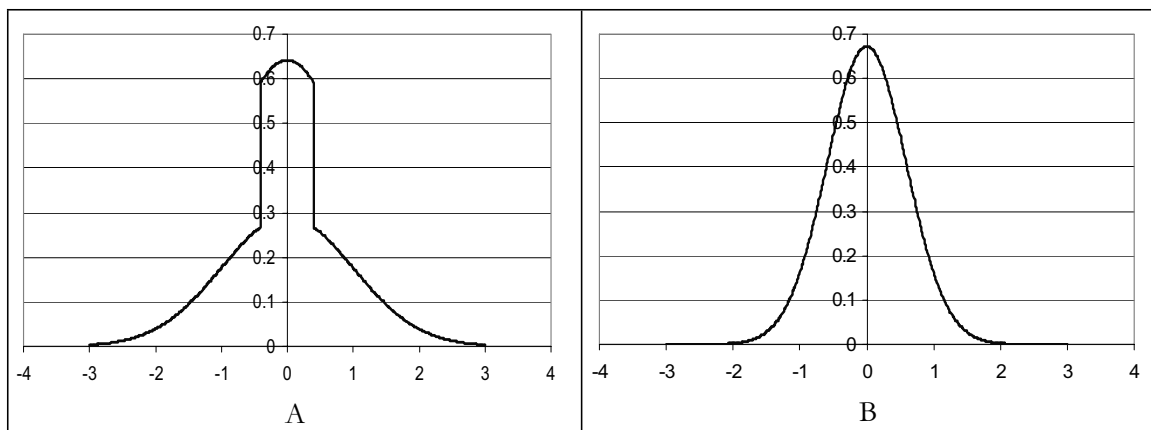


Figure 4 – A: I type prior with $b = 0.4$. B: IV type prior with $b = 0.4$.

Quite similar applications have been made for priors II and III, the II prior being built starting from the density of a normal $N(0, \tau^2 = 0.04)$, and the III prior derived from the density of a normal $N(0, \tau^2 = 25)$ (of course, the multiplicative constants have been changed accordingly). The IV density used in the calculations of Table 2, instead, is simply a normal density, chosen so that the integral between $-b$ and b is fixed at $\pi_0 = 1/2$. Figure 4 B shows such a density for $b = 0.40$.

While the values of the posterior probabilities of H_0 in Table 2, for $b = 0.02$ and the first prior distribution, are – as expected – very near to the values computed by Berger for the case of the point null hypothesis (see section 2.2), the other parts of the table show quite different values and behaviours of the posterior probabilities, denying the presumed paradox advocated by the Bayesians.

All the above calculations were made by the use of the prior probability $\pi_0 = 1/2$, whereas it should be reasonable to calibrate such probabilities according to the width of the interval $\Theta_0 = (-b, b)$. In case one would change π_0 , the final result for the posterior probability is simply obtained as a function of the value already calculated for $\pi_0 = 1/2$. Calling I_0 and I_1 the integrals proportional to p_0 and p_1 :

TABLE 2

Posterior probabilities $p_0 = P(H_0 | \bar{x})$ for the null hypothesis $H_0 : \theta \in [-b, b]$, $b = 0.02, 0.10, 0.20, 0.40$, concerning the mean of a normal distribution $N(\theta, 1)$, calculated on the basis of a sample size n , of an observed sample mean $\bar{x} := 1.96/\sqrt{n}$ (P-value 0.05), and assuming four prior distributions for the parameter θ (see text), all admitting a prior probability $\pi_0 = 1/2$ for the null hypothesis

n	1	5	10	20	50	100	1000
$b = 0.02$							
I prior	0.3495	0.3285	0.3638	0.4201	0.5140	0.5883	0.7514
II prior	0.4853	0.4388	0.3994	0.3545	0.3118	0.3060	0.3790
III prior	0.0724	0.1395	0.1849	0.2415	0.3317	0.4069	0.5971
IV prior	0.4995	0.4973	0.4948	0.4895	0.4747	0.4521	0.2533
$b = 0.10$							
I prior	0.3412	0.3091	0.3300	0.3579	0.3828	0.3861	0.3802
II prior	0.4799	0.4166	0.3624	0.2958	0.2096	0.1599	0.0947
III prior	0.4416	0.6111	0.6749	0.7228	0.7566	0.7637	0.7645
IV prior	0.4870	0.4417	0.3965	0.3311	0.2284	0.1614	0.0711
$b = 0.20$							
I prior	0.3281	0.2673	0.2583	0.2481	0.2351	0.2289	0.2210
II prior	0.4698	0.3747	0.2959	0.2078	0.1170	0.0774	0.0343
III prior	0.4306	0.5662	0.6002	0.6137	0.6158	0.6155	0.6151
IV prior	0.4521	0.3311	0.2533	0.1807	0.1162	0.0896	0.0589
$b = 0.40$							
I prior	0.2935	0.1789	0.1502	0.1326	0.1199	0.1146	0.1073
II prior	0.4318	0.2501	0.1531	0.0824	0.0341	0.0184	0.0050
III prior	0.3914	0.4402	0.4402	0.4387	0.4375	0.4369	0.4362
IV prior	0.3542	0.1807	0.1283	0.0966	0.0744	0.0655	0.0541

$$I_0 = \int_{-b}^b 0.5 \cdot g_0(\theta) f(\bar{x} | \theta) d\theta; \quad I_1 = \int_{-\infty}^{-b} + \int_b^{\infty} 0.5 \cdot g_1(\theta) f(\bar{x} | \theta) d\theta$$

the following equalities hold, with respect to p_0 computed with $\pi_0 = 0.5$, and to p_0^* computed with a generic $0 < \pi_0 < 1$:

$$p_0 = I_0 / (I_0 + I_1)$$

$$p_0^* = \pi_0 I_0 / [\pi_0 I_0 + (1 - \pi_0) I_1]$$

$$p_0^* = \left[1 + \frac{1 - \pi_0}{\pi_0} (p_0^{-1} - 1) \right]^{-1}.$$

For example, by the choice of $\pi_0 = 0.1$, for the cases $b = 0.02$ and $b = 0.10$ and the first prior distribution, one can simply obtain from Table 2 the following values for the posterior probabilities:

$b = 0.02$							
n	1	5	10	20	50	100	1000
p_0^*	0.056	0.052	0.060	0.074	0.105	0.137	0.215
$b = 0.10$							
n	1	5	10	20	50	100	1000
p_0^*	0.054	0.047	0.052	0.058	0.064	0.065	0.064

In conclusion, if one sticks to a Bayesian approach for the above problem, however with a sensible choice of hypotheses and related prior distributions, sensible results would follow. No wonder – on the contrary – that (practically) absurd assumptions can yield absurd or embarrassing outcomes.

Istituto di Statistica
Università Cattolica del Sacro Cuore di Milano

BENITO VITTORIO FROSINI

REFERENCES

- I. J. BEAUMONT, AND N. E. BRESLOW, (1981), *Power considerations in epidemiologic studies of vinyl chloride workers*, “American Journal of Epidemiology”, 114, pp. 725-734.
- D. BLACKWELL, (1951), *Comparison of experiments*, “Proceedings of the Second Berkeley Symposium, Mathematics Statistics Probability”, pp. 93-102.
- J. O. BERGER, (1985), *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer, New York.
- B. V. FROSINI, (1991), *On the definition and some justifications of the likelihood principle*. “Statistica”, 51, pp. 489-503.
- B. V. FROSINI, (1993), *Likelihood versus probability*, “Proceedings ISI 49th Session”, Vol. 2, Firenze, pp. 359-375.
- B. V. FROSINI, (2001), *Metodi Statistici*, Carocci, Roma.
- B. M. HILL, (1982), *Comment on the article by G. Shafer*, “Journal of the American Statistical Association”, 77, pp. 344-347.
- J. L. HODGES, AND E.L. LEHMANN, (1954), *Testing the approximate validity of statistical hypotheses*, “Journal of the Royal Statistical Society”, series B, 16, pp. 261-268.
- H. JEFFREYS, (1939, 1948), *Theory of Probability*, Oxford University Press, Oxford.
- S. KULLBACK, (1959), *Information Theory and Statistics*, Wiley, New York.
- S. KULLBACK, (1983), *Kullback Information*, in *Encyclopedia of Statistical Sciences*, Wiley, New York, vol. 4, pp. 421-425.
- E. L. LEHMANN, (1959), *Testing Statistical Hypotheses*, Wiley, New York.
- D. V. LINDLEY, (1957), *A statistical paradox*, “Biometrika”, 44, pp. 187-192.
- J. NEYMAN, (1935), *Contribution to the discussion of the paper by F. Yates*, “Supplement Journal of the Royal Statistical Society”, 2, pp. 235-241.
- J. NEYMAN, (1956), *Note on an article by Sir Ronald Fisher*, “Journal of the Royal Statistical Society”, 18, pp. 288-294.
- G. SHAFER, (1982), *Lindley's paradox* (with discussion by D.V. Lindley, M.H. DeGroot, A.P. Dempster, I.J. Good, B.M. Hill and R.E. Kass), “Journal of the American Statistical Association”, 77, pp. 325-351.
- E. N. TORGERSEN, (1976). *Comparison of statistical experiments* (with discussion by S. Johansen, B. Lindqvist, J. Hilden and O. Barndorff-Nielsen), “Scandinavian Journal of Statistics”, 3, pp. 186-208.
- S. S. WILKS, (1962), *Mathematical Statistics*, Wiley, New York.

RIASSUNTO

Sulla teoria di Neyman-Pearson: Informazione contenuta in un esperimento, e un paradosso fantasioso

Questo articolo tratta due argomenti collegati con la teoria di Neyman-Pearson sulla verifica di ipotesi. Il primo argomento riguarda l'informazione contenuta in un esperimento; dopo un breve accenno alla comparabilità ordinale degli esperimenti, vengono considerate dapprima le due misure di informazione più note, quella proposta da Fisher e quella proposta da Kullback-Leibler. Almeno per i casi più comuni, in cui si richiede di eseguire una comparazione di due esperimenti alla volta, emerge la superiorità della coppia (α, β) delle due probabilità di errore nell'impostazione di Neyman-Pearson, a causa del chiaro significato operativo di tali indici.

Il secondo argomento riguarda il c.d. paradosso di Jeffreys, o di Lindley; nel caso di un'ipotesi nulla puntuale si può mostrare che, se associamo una probabilità positiva a tale ipotesi, nell'impostazione bayesiana dell'inferenza le probabilità a posteriori possono assumere valori molto contrastanti con le probabilità di errore dell'impostazione di Neyman-Pearson. Viene argomentato in questo articolo che tali risultati sono prodotti semplicemente a causa delle assunzioni assurde che sono state fatte nell'impostazione bayesiana; è infatti mostrato, al contrario, che partendo da assunzioni ragionevoli riguardo a ipotesi intervallari (non puntuali) si possono ottenere probabilità a posteriori perfettamente compatibili con l'impostazione di Neyman-Pearson (sia pure tenuto conto che tali comparazioni richiedono molta cautela, dato che le due impostazioni a confronto sono radicalmente diverse sia rispetto alle assunzioni di partenza sia rispetto agli scopi dell'inferenza).

SUMMARY

On Neyman-Pearson Theory: Information Content of an Experiment and a Fancy Paradox

Two topics, connected with Neyman-Pearson theory of testing hypotheses, are treated in this article. The first topic is related to the information content of an experiment; after a short outline of ordinal comparability of experiments, the two most popular information measures – by Fisher and by Kullback-Leibler – are considered. As far as we require a comparison of two experiments at a time, the superiority of the couple (α, β) of the two error probabilities in the Neyman-Pearson approach is easily established, owing to their clear operational meaning.

The second topic deals with the so called Jeffreys – or Lindley – paradox: it can be shown that, if we attach a positive probability to a point null hypothesis, some «paradoxical» posterior probabilities – in a Bayesian approach – result in sharp contrast with the error probabilities in the Neyman-Pearson approach. It is argued that such results are simply the outcomes of absurd assumptions, and it is shown that sensible assumptions about interval – not point – hypotheses can yield posterior probabilities perfectly compatible with the Neyman-Pearson approach (although one must be very careful in making such comparisons, as the two approaches are radically different both in assumptions and in purposes).