Journal of Scientific Exploration, Vol. 11, No.2, pp. 181-192 1997 0892-3310/97

A Bayesian Maximum-Entropy Approach to Hypothesis Testing, for Application to RNG and Similar Experiments

P. A. STURROCK

Center for Space Science and Astrophysics, Varian 302G
Stanford University, Stanford, CA 94305-4060

Abstract — In assessing the results of RNG (random number generator) experiments, and in similar problems of the Bernoulli type, one needs to evaluate the proposition that the results are compatible with a specific hypothesis, such as the so-called "null hypothesis" that no extraordinary process is at work. This evaluation is often based on the "p-value" test according to which one calculates the probability of obtaining, on the basis of the specific hypothesis, the actual result or a "more extreme" result. Textbooks caution that the p-value does not give the probability that the specific hypothesis is true, and one recent textbook asserts "Although that might be a more interesting question to answer, there is no way to answer it." **A** Bayesian approach requires that we consider not just one hypothesis but a complete set of hypotheses. This may be achieved very simply by supplementing the specific hypothesis with the maximum-entropy hypothesis that covers all other possibilities in a way that is maximally non-committal. This procedure yields an estimate of the probability that the specific hypothesis is true. This estimate is found to be more conservative than that which one might infer from the p-value test.

Keywords: bayesian analysis – statistics – methodology

1. Introduction

A recent exchange of letters between Jefferys (1995a, 1995b), on the one hand, and Dobyns and Jahn (1995) and Dobyns (1995), on the other hand, raises a basic question concerning the analysis of results of experiments such as RNG experiments. The central point in the discussion is the standard procedure for evaluating the significance of an experimental result in terms of a "one-sided p-value." According to this procedure, one tests the "null hypothesis" that the result obtained is due to chance by evaluating the probability that, on the basis of chance, the experiment would yield the value actually obtained or a "more extreme" value.

This may seem a reasonable procedure if the probability distribution has a single peak, and the observed value is way out to one side of that peak. However, one can certainly construct experiments for which the expected distribution would have more than one peak. As an example, suppose one is told that a box contains two dice. One die has only one spot on one of its six faces; the other die has one spot on each of five of its six faces. The experiment is to draw a die, toss it 30 times, and count the number of spots. One expects that the number will either be close to 5 or close to 25. The probability distribution is shown in Figure 1.

Since a linear plot of probability does not give a clear indication of the departure of the probability from either zero or unity, it is helpful to introduce first the concept of "odds," defined by

$$\Omega = \frac{P}{1-P}, \tag{1.1}$$

where P is the probability, and then the concept of "log-odds," defined by

$$A = \log(\Omega) = \log\left(\frac{P}{1-P}\right). \tag{1.2}$$

The distribution of Figure 1 is shown as a distribution of log-odds in Figure 2.

It would be very surprising if the number were close to zero, or close to 30, but it would be much more surprising if the number were close to 15. Specifically there is probability 0.002 of getting zero spots, the same probability of getting 30 spots, but probability only 2 10^{-5} of getting 15 spots. If the number turned out to be, say, 10, how would one carry out ap-value test? If one were to carry out the sum of the probabilities expected of counts zero through 10, one would obtain almost 0.5 (in fact, one would get 0.497). The same would be true if the count were 20 and one were to count from 20 to 30.

It is clear from the preceding example that there are problems to which the p-value "tail test" would be inapplicable. Even if the p-value test is applica-

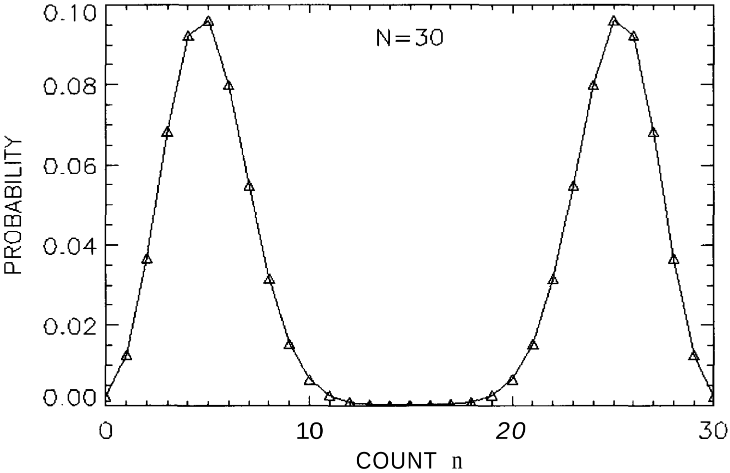


Fig. 1. The probability of getting *n* spots in 30 tosses of one of two dice, one of which has a spot on only one face, the other of which has a spot on each of five faces.

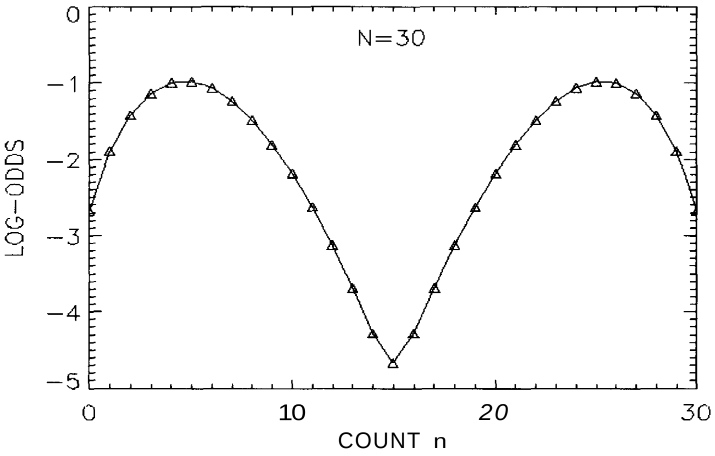


Fig 2. The same data as in Figure 1, displayed in terms of log-odds.

ble, it is not at all clear how it should be interpreted; this issue will be discussed further in Section 2. An alternative to the standard p-value test, one that arises naturally from Bayesian thinking, is proposed in Section 3. Further discussion is presented in Section 4.

2. Thep-Value Test

Even in simple problems with one or two well-defined "tails," how is the p-value test to be interpreted, and how is this interpretation to be justified? Suppose that one considers a well defined distribution over 101 possible outcomes, with only one peak, and one obtains 95 positive results where only 50 are expected, why should one add the probabilities of getting 95 or more? If one is going to go up 5 in one direction, why not go down 5 in the opposite direction also, so that one would be considering the probability of getting 90 successes or more?

Furthermore, how is one to interpret the p-value? If the p-value turns out to be .05, does that mean that the odds are 19 to 1 that the null hypothesis is incorrect? Utts (1996) is very clear on this point. She cautions that the p-value does *not* give the probability that the null hypothesis is true, and adds "Although that might be a more interesting question to answer, there is no way to answer it." That may be true of non-Bayesian statistics but, as we shall see, it is not true of Bayesian statistics.

The Bayesian approach offers a useful perspective on the p-value test. According to the Bayesian approach to scientific issues (see, for instance, Good, 1950; Jeffreys, 1931; Sturrock, 1994), it is essential to consider not a single hypothesis, but a complete set of hypotheses. If the p-value test were robust, it should not make much difference how one chooses to complete the set of hypotheses. If, on the other hand, our interpretation of the p-value test depends sensitively on which hypotheses are adopted to form a complete set, then we have cause for concern.

Let us examine this issue. We consider a specific hypothesis, *e.g.* that a coin is a fair coin. It is conventional to refer to this as the "null hypothesis," but this term makes sense only if one is considering also the consequences of another hypothesis (a "non-null" hypothesis), *e.g.* that the coin is biased by a specified amount. Since p-value test theory does not involve a second hypothesis, it seems gratuitous to refer to the only considered hypothesis as the "null hypothesis." For this reason, we will use the term "specific hypothesis" in place of the usual "null hypothesis." (Furthermore, in the next section we consider a "maximum entropy" hypothesis that is even more deserving of the term "null hypothesis" if we wish to use that term.) We designate the specific hypothesis as $H1$.

In order to pursue Bayesian thinking on this problem, it is essential to supplement $H1$ with one or more additional hypotheses, so that all hypotheses are mutually exclusive and form a complete set. For present didactic purposes, we make the simplest assumption, that there is only one supplementary hypothesis that we call H2. We will assume that, if $H1$ is true, the probability of a certain event (say a coin coming up heads) is p_1 and that, if H2 is true, the probability is p_2. To be specific, we can consider the situation that a bag contains a fair coin for whichp, = 0.5, and a biased coin for whichp, is some known number different from 0.5. In using the p-value test to evaluate the probability that the specific hypothesis $H1$ is false, we are now evaluating the probability that the supplementary hypothesis H2 is true. If thep-value test were a robust test, it should not much matter what we choose for H2: then the relationship between the p-value and the post-probability that $H1$ is correct (the probability evaluated on the basis of prior knowledge and also on knowledge of the outcome of the experiment) would not depend sensitively on p_2. Let us see if this is the case.

We introduce $P(H1|I)$ to denote the probability that $H1$ is true, evaluated on the basis of "initial" information **I**, with a similar interpretation for $P(H2|I)$. Since **I** is assumed throughout, we ignore it where we can without loss of clarity, and write these probabilities as $P(H1|)$ and $P(H2|)$. We next suppose that an experiment is carried out in which one coin is drawn from the bag and tossed N times, yielding n heads. We denote by $P(n|N,H1)$ and $P(n|N,H2)$ the probability of this result based on the assumption that $H1$ is correct, or on the assumption that H2 is correct, respectively. Then the p-value, that we write as $V1$, is given by

$$V_1 = \sum_{r=O}^{n} P(r \mid H1) \text{ or } V_1 = \sum_{r=n}^{N} P(r \mid H1) \tag{2.1}$$

for n is close to zero and for n close to N, respectively. To be definite, we subsequently consider the former alternative.

We denote by $P(H1|n,N)$ the "post probability" that $H1$ is true, based on knowledge of the result of the experiment (and, implicitly, on the initial information I also), with a similar interpretation of $P(H2|n,N)$. Since we are told that the only relevant hypotheses are $H1$ and H2, it follows that

$$P(H1\, I) + \mathrm{P}(\mathrm{H2}\, I) = \mathrm{I},\quad P(H1 \mid \mathrm{n, N}) + \mathrm{P}(\mathrm{H2} \mid \mathrm{n, N}) = 1, \text{ etc.} \tag{2.2}$$

The Bayes (or Bayes-Laplace) theorem (see, for instance, Jaynes, 1989) tells us that

$$P(H1 \mid n, \mathrm{N}) = \frac{\mathrm{P(n \mid N}, H1)}{\mathrm{P(n \mid N)}} P(H1\, I), \tag{2.3}$$

where $P(n|N)$ may be evaluated from

$$P(n \mid N) = P(n \mid N, H1) P(H1\, I) + P(n \mid N, H2) P(H2\, I)\ . \tag{2.4}$$

We assume, for simplicity, that the prior probabilities of $H1$ and H2 are equal, so that $P(H1 I) = 0.5$ and $P(H2 I) = 0.5$. Then we see that

$$\frac{P(H1 \mid n, N)}{P(H2 \mid n, N)} = \frac{P(n \mid H1, N)}{P(n \mid H2, N)}. \tag{2.5}$$

Since H2 is equivalent to "not-$H1$," the left-hand side of (2.5) is the "odds" on $H1$, so that we can rewrite (2.5) as

$$\Omega(H1 \mid n, N) = \frac{P(n \mid H1, N)}{P(n \mid H2, N)}. \tag{2.6}$$

On noting that

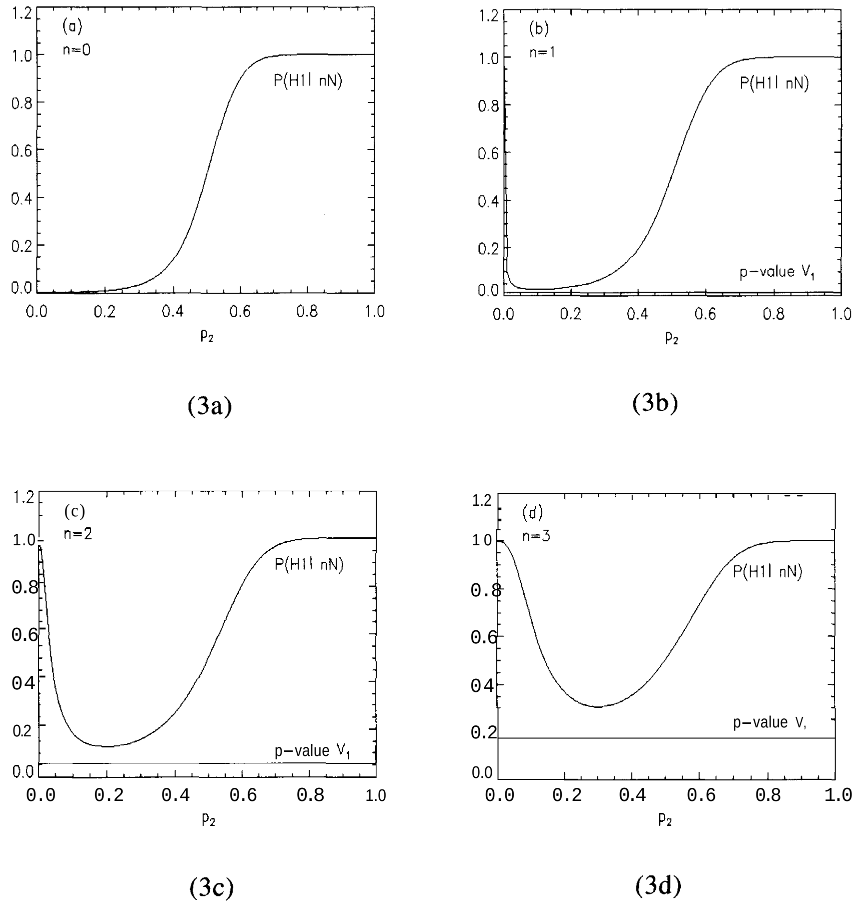
$$P(n \mid N, H1) = {}^{N}C_n p_1{}^n (1 - p_1)^{N-n}, \tag{2.7}$$

with a similar expression for $P(n|N, H2)$, we see that (2.6) becomes

$$\Omega(H1 \mid n, N) = \frac{p_1^{\,n}\left(1-p_1\right)^{N-n}}{p_2^{\,n}\left(1-p_2\right)^{N-n}}. \tag{2.8}$$

By using this expression and noting (2.2), we may calculate $P(H1|n,N)$ for given values of p_1 andp,. We can then compare this with **V„** evaluated from **(2.1)**and (2.7), and we can then examine conditions under which the Bayesian evaluation of the post probability of the null hypothesis agrees with the *p*-value.

Figure **3** has four panels giving both the p-value V_1 and the post-probability $P(H1|nN)$ for values **n = 0, 1, 2** and **3,** for the case thatp, = **0.5** and **N = 10,** and



(3a)

(3b)

(3c)

(3d)

Fig. 3. The upper curve is the post-probability that *H*1 is true, plotted as a function of p_2, and the lower curve (a straight line) is the p-value, for (a) n = 0, (b) *n* = 1, (c) n = 2, and (d) n = 3. Note that for n = 0 [case (a)], *V*1 = *0.001* so that the p-value line is indistinguishable from the axis.

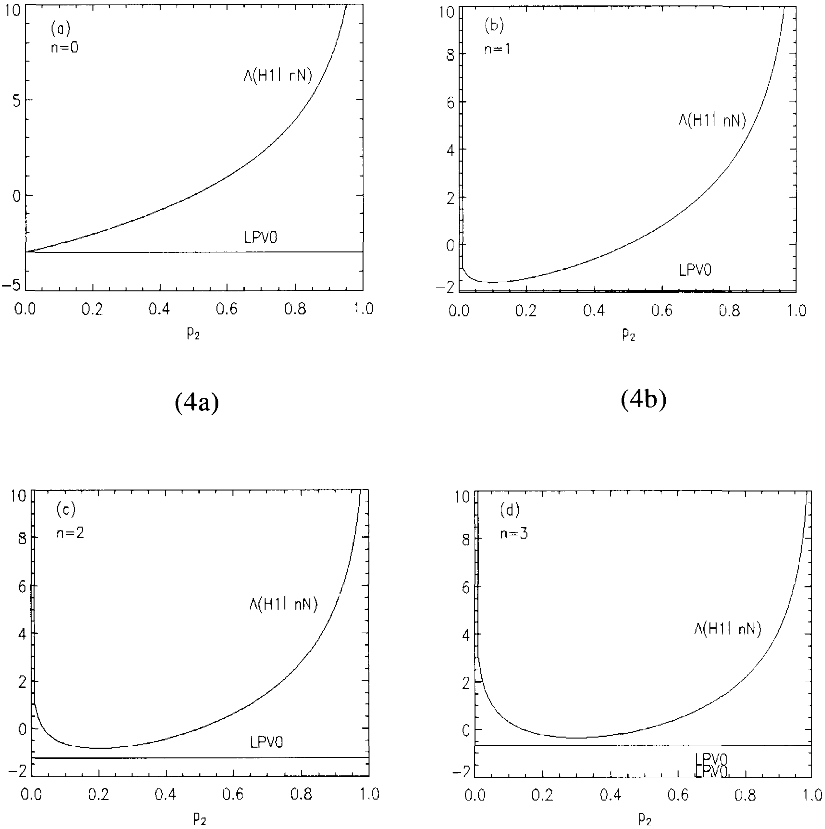
for all values of p_2 from 0 to 1. Figure 4 shows the same results expressed differently: it shows the post-log-odds $\Lambda(H1|n,N)$ and the "log-p-value-odds" **LPVO** defined analogously as

$$\mathbf{LPVO} = \log(\frac{\mathbf{V}}{\mathbf{1-V}}) \ . \tag{2.9}$$

We may note from these displays the following comparison of the p-value and post-probability —

(a) Thep-value coincides with the post-probability only for the special case that n = 0 and $p_2 = 0$.

(b) For all other cases, **the** p-**value** is smaller than the post probability.



(4a)

(4b)

(4c)

(4d)

Fig. 4. The same data as in Figure 3, displayed in terms of log-odds and LPVO (log-p-value-odds).

(c) The relationship between the p-value and the post probability is **very** sensitive to the choice of p_2. Consider, for instance, the case n = 0. If $p_2 = 0.1$, the post-odds is larger than the p-value-odds by a factor of only 3.9, but if $p_2 =$ 0.9, it is larger by a factor of 9 10^8.

(d) We see from Figure 3 that, even when one adopts the maximum-likelihood value of p_2 (namely *n/N*), the p-value is still smaller than the post-odds. The difference is small, with a logarithmic difference of 0.36, 0.40 and 0.32 for n = 1, 2, and 3, respectively, but that does not obviate that fact that the p-value is clearly misleading. Furthermore, it is inappropriate to give equal prior probabilities to the specific hypothesis and to the maximum-likelihood value of p_2, since the latter is determined by the result of the experiment!

We see that there is little reason to adopt the p-value method for hypothesis-testing, and plenty of reason to seek an alternative.

3. Maximum Entropy Hypothesis

In order to apply Bayesian methods, one must propose a complete set of mutually exclusive hypotheses. If one hypothesis is that one is tossing a fair coin, for which the probability of heads or tails is 0.5, one must specify one or more additional hypotheses so that we have a complete and mutually exclusive set. It is here proposed that we seek a hypothesis that is, in some sense, as unspecific as possible, so that it covers all other specific hypotheses. This can be viewed also as the "maximum entropy" or "maximum ignorance" hypothesis.

The maximum-entropy hypothesis, that we denote by "ME," must allow p_2 to adopt all possible values, except the value p,. It turns out that, when it comes to performing calculations, it makes no difference to the final formulae whether or not p_2 covers the value p,, so we can ignore that issue and simply consider that the maximum entropy hypothesis represents a distribution $D(p)$ of all probability values from 0 to 1.

The entropy of the distribution is given (see, for instance, Jaynes, 1957) by

$$H = -\int_0^1 dpD(p)\ln\left[D(p)\right], \tag{3.1}$$

where we now regard p simply as a parameter, and $D(p)$ is a probability distribution that represents our knowledge — or ignorance — about the parameter p. We now require that H take its maximum value subject to the constraint that

$$\int_0^1 dpD(p) = 1\,. \tag{3.2}$$

We require that $\delta H = 0$, *i.e.*

$$\int_0^1 dp[1 + \ln(D)]\delta D = 1 \tag{3.3}$$

for all $\delta D(p)$ subject to

$$\int_0^1 dp\delta D = 0 \, . \tag{3.4}$$

This clearly requires that D = constant, and (3.2) then shows that D = 1. Hence the maximum-entropy distribution is the uniform distribution.

We may now evaluate $P(n|N,ME)$ from

$$P(n \mid N, ME) = \int_0^1 dp \ {}^N C_n p^n (1-p)^{N-n} . \tag{3.5}$$

We may evaluate this integral by noting that

$$(1 - p + xp)^N = \sum_{n=0}^{N} x^n \, {}^N C_n p^n (1-p)^{N-n} . \tag{3.6}$$

On integrating both sides over p from 0 to 1, and separating coefficients of x^n, we obtain

$$\int_0^1 dp \ {}^N C_n p^n (1-p)^{N-n} = \frac{1}{N+1} , \tag{3.7}$$

that leads to the result

$$P(n \mid N, ME) = \frac{1}{N+1} . \tag{3.8}$$

This shows that the post-probability distribution of the maximum-entropy hypothesis is the uniform distribution.

It is now straightforward to find the post-probability of $H1$ (that $\boldsymbol{p = p,)}$ by using (2.2) and (2.7). If we make the assumption (“I”) that $H1$ and $\boldsymbol{MI}$ have equal prior probabilities, so that

$$P(H1 \mid I) = P(MI \mid I) = 0.5 \ , \tag{3.9}$$

we finally obtain the following expression for the odds on $H1$:

$$\Omega(H1|I) = \frac{(N+1)!}{n!(N-n)! \, p_1{}^n (1 - p_1)^{N-n}} \tag{3.10}$$

4. Discussion

We may now return to the simple problem discussed in Section 2, for which p , = 0.5 and N = 10. For all values of n, we may estimate the post odds, given by (3.10), and hence obtain the post-probability from

$$P = \frac{\Omega}{1+\Omega} \tag{4.1}$$

We can also evaluate thep-value V,. Figure 5 shows both the post-probability of H1, and the p-values, computed for each "tail." Figure 6 presents the same data in the form of the post-log-odds and the log-p-value-odds *LPVO* introduced earlier. The same data are listed in Table 1. We see that the post odds and the *LPVO* behave very similarly, but the former is about an order of magnitude larger than the latter. In general, the post-log-odds will be larger than the *LPVO* by a factor of order N. We see, therefore, that the present "maximum-entropy" test is more conservative than the p-value test.

We may now return to the bimodal distribution presented (see Figures 1 and 2) in Section 1. There is now no problem is using the same method that we have applied to a simple unimodal distribution. We may use the same maximum-entropy hypothesis to be the considered alternative to the bimodal model. Then, if we get zero spots, the post-odds on the bimodal model is 31 x 0.002, *i.e.* 0.06. If we get exactly 15 spots, the post-odds on the bimodal model is 31 × (2 10^{-5}), *i.e.* 0.0006.

In examining the p-value test in Section 2, we saw cause for concern in the fact that, for the cases we examined, thep-value is smaller than the post-probability even when the specific hypothesis is compared with the hypothesis that the probability has its maximum-likelihood value. In examining the maximum-entropy analysis, we find that it is not subject to the same objection.

TABLE 1
Copmarison of p-Values and Post Probabilities

n	p-Value Left Tail	p-Value Right Tail	Post Probability
0	0.001		0.011
1	0.011		0.097
2	0.055		0.326
3	0.172		0.563
4	0.377		0.693
5			0.730
6		0.377	0.693
7		0.172	0.563
8		0.055	0.326
9		0.011	0.097
10		0.001	0.011

Forpl = 0.5 and N = 10, this table lists the p-values for each tail and the post-probability for *n* = 0 to 10

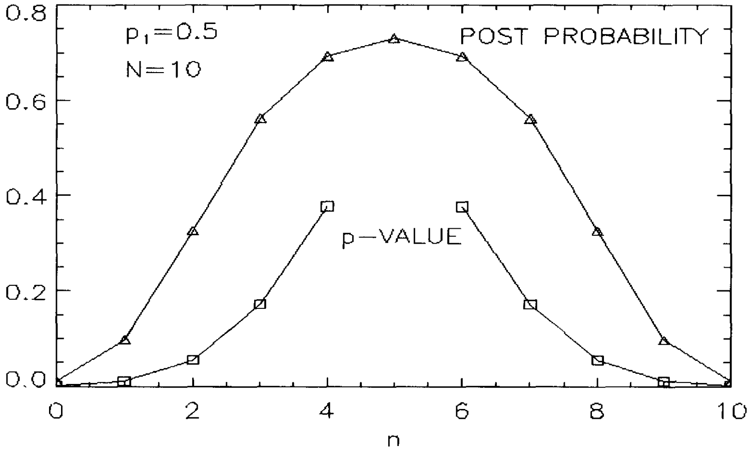


Fig. 5. The upper curve is the post-probability of H1, as estimated by the maximum-entropy test. The lower curves show the p-values for the left-hand tail and the right-hand tail.

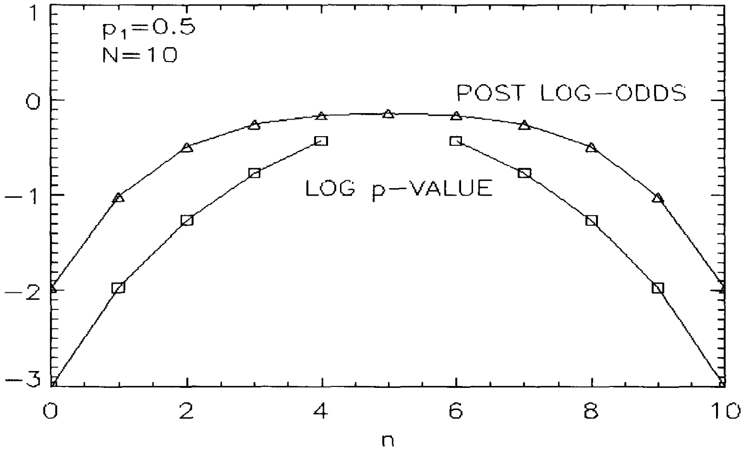


Fig. 6. The same data as in Figure 5, displayed in terms of log-odds and LPVO (log-p-value-**odds).**

Considering once more the case that $p_1 = 0.5$ and N = 10, we find that for n = 0, 1, 2 and 3, the post-odds on $H1$ are 0.011, 0.11, 0.48 and 1.29, respectively. When we take p_2 to be the maximum likelihood value (0, 0.1, 0,2 and 0.3, respectively), we obtain a post-odds of 0.001, 0.026, 0.15, and 0.44, respectively. Hence the maximum-entropy analysis yields estimates of the post-probability that are (appropriately enough) more conservative than the values that would be obtained by identifyingp, with the maximum-likelihood value.

Finally, we may note that (unlike the p-value test that only gives evidence against the considered hypothesis), the maximum-entropy test yields evidence in favor of the considered hypothesis when that is appropriate. For instance, for the simple problem that $p_1 = 0.5$ and N = 10, we see from Table 1 that the experiment favors the considered hypothesis if it yields n = 3, 4, 5, 6 or 7. The experiment yields evidence against the considered hypothesis if n = 0, 1, 2, 8, 9, or 10. Utts (1996) may well be correct when she writes, concerning the p-value test, that "(although) the probability that the null hypothesis is true... might be a more interesting question to answer, there is no way to answer it." We see, however, that a simple Bayesian approach to the same problem does yield an estimate of the probability that the considered hypothesis is true.

References

Dobyns, Y. H., and Jahn, R. G. (1995). ***Journal of Scientific Exploration*** (Letters to the Editor), 9, **122.**

Dobyns, Y. H. (1995). ***Journal of Scientific Exploration*** (Letters to the Editor), 9,597.

Good, I. J. (1950). ***Probability and the Weighing of Evidence.*** London: Griffin.

Jaynes, E. T. (1957). Information theory and statistical mechanics I. ***Physics Review,*** 106,620.

Jaynes, E. T. (1989). ***Papers on Probability, Statistics and Statistical Physics. Ed. R. D.*** Rosenkrantz, Dordrecht: Kluwer, 216.

Jefferys, W. H. (1995). ***Journal of Scientific Exploration*** (Letters to the Editor), 9, 121.

Jefferys, W. H. (1995). ***Journal of Scientific Exploration*** (Letters to the Editor), 9,595.

Jeffreys, H. (1931). ***Scientific Inference.*** Cambridge University Press.

Sturrock, P.A. (1994). Applied scientific inference. ***Journal of Scientific Exploration,*** 8, 491.

Utts, J. (1996). Seeing Through Statistics. Duxbury Press.