Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio

Morris H. DeGroot

Journal of the American Statistical Association, Vol. 68, No. 344 (Dec., 1973), 966-969.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197312%2968%3A344%3C966%3ADWCNIA%3E2.0.CO%3B2-N

Journal of the American Statistical Association is currently published by American Statistical Association.

Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio

MORRIS H. DeGROOT*

Consider a problem in which a certain statistic X has a specified distribution function $F(x)$ if a given hypothesis H is true, and suppose that the hypothesis H is evaluated by calculating the tail area $1 - F(x)$ corresponding to the observed value x of the statistic X. Examples are given in which this tail area is equal to the posterior probability that H is true and in which it is equal to the likelihood ratio comparing H to a certain class K of alternatives. The purpose of these examples is to render the traditional statistical practice of calculating tail areas consonant with the principles of Bayesian statistics.

1. STATISTICAL METHODS BASED ON TAIL AREAS

The following situation is common in statistical problems: A sample of observations from a certain population is available to the statistician and it is desired to learn whether the probability distribution P of this sample belongs to a certain specified class C_0 of distributions. The class C_0 might contain just one distribution or it might contain more than one. In order to study the hypothesis H that $P \in C_0$, a real-valued statistic X is constructed from the sample with the following property:

If the hypothesis H is true and the distribution P actually does belong to C_0, then the statistic X has a certain specified probability density function (pdf) which we shall call f. If the distribution P does not belong to C_0, then for a wide class of distributions outside C_0, and possibly for all distributions outside C_0, the statistic X will tend to be larger than it would be when its pdf is f. More precisely, the distribution of X will be stochastically larger than the distribution for which the pdf is f, as defined in [7, p. 73].

It should be emphasized that I am considering here problems in which the pdf of X is the same for all distributions $P \in C_0$. In some problems f will be the exact small sample pdf of X and in others, such as χ^2 or Kolmogorov-Smirnov tests of goodness-of-fit, it will be an asymptotic approximation that is appropriate when the number of observations is sufficiently large.

We shall let F denote the distribution function (df) corresponding to the pdf f, so that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad -\infty < x < \infty.$$

Suppose now that the given sample of observations yields a particular value x of the statistic X. In order to assess the weight of evidence that the observed value x carries against the hypothesis H, it is standard statistical practice under the conditions being assumed here to calculate the tail area $1 - F(x)$ corresponding to the value x. The value of $1 - F(x)$ is also called the observed p-value or the observed significance level. A small value of $1 - F(x)$, e.g., a value less than 0.01 or 0.05, is typically regarded as strong evidence against the hypothesis H and, in general, the smaller the value of $1 - F(x)$, the more strongly one regards the evidence against H.

Because the tail area $1 - F(x)$ is a probability, there seems to be a natural tendency for a scientist, at least one who is not a trained statistician, to interpret the value of $1 - F(x)$ as being closely related to, if not identical to, the probability that the hypothesis H is true. It is emphasized, however by statisticians applying a procedure of this type that the tail area $1 - F(x)$ calculated from the sample is distinct from the posterior probability that H is true. In other words, although a tail area smaller than 0.01 may be obtained from a given sample, this result does not imply that the posterior probability of H is smaller than 0.01, or that the odds against H being true are 100 to 1 (or 99 to 1). Indeed, many statisticians who believe that it is meaningful to calculate the tail area in a given problem also believe that it is not meaningful to even attempt to assign a probability to the statement that H is true.

On the other hand, many statisticians who adhere to the Bayesian philosophy and believe that it is always possible (if not always practically feasible) to assign prior and posterior probabilities to a hypothesis H, do not believe that the calculation of a tail area is consonant with the principles of Bayesian statistics. To be true to their principles they must therefore resist the strong intuitive appeal of a procedure such as the χ^2 test of goodness-of-fit or the F test in the analysis of variance, and attempt to develop alternative procedures based on prior distributions over parameter spaces that are sometimes of high dimension (see, e.g., [1, Ch. 11] for a

* Morris H. DeGroot is professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, Penn. 15213. This research was supported in part by the National Science Foundation under grant GS-32514. The author is indebted to Bradley Efron and D.V. Lindley for their very helpful comments.

© Journal of the American Statistical Association
December 1973, Volume 68, Number 344
Theory and Methods Section

discussion of testing hypotheses from the Bayesian point of view).

The purpose of this article is to present a few simple ideas which indicate how the calculation of tail areas can be made compatible with the principles of Bayesian statistics. These ideas, if successful, will serve the dual purpose of putting χ^2 tests, F tests, and other such procedures back into the repertory of the Bayesian statistician and of giving all statisticians the freedom that comes from being able to interpret the evidence exhibited in a tail area simply as a likelihood ratio or as a posterior probability.

The relationship between tail areas and likelihood ratios has also been considered by Good [4, 5, 6] and Efron [2]. Lindley [8] has shown that in many standard significance tests, the posterior distribution of the χ^2, F, or t statistic, when regarded as a function of the unknown parameters for given values of the observations in a large sample, will be approximately the same as the sampling distribution of the statistic. Therefore, the calculated tail area $1 - F(x)$ will have a Bayesian interpretation as a tail area of the posterior distribution. The approach taken in this article is somewhat different. The results that will be presented here pertain to an arbitrary statistic based on an arbitrary sample and are, therefore, both more abstract and more elementary.

2. THE TAIL AREA AS A LIKELIHOOD RATIO

We shall continue to consider a problem in which a statistic X has a specified pdf f when a certain hypothesis H is true. We shall assume also that a statistician observes the value x of X in a given sample and calculates the tail area $1 - F(x)$. If this tail area is small, the statistician regards the observed sample data as being strong evidence against H, presumably because he has in mind certain alternatives to H for which the observed value x is much more likely than it is under f.

However, these alternative distributions are often not explicitly specified. Rather, in many problems, they are simply vaguely conceptualized by the statistician as distributions that are stochastically larger than the distribution for which the pdf is f. In such a problem, the calculation of the tail area $1 - F(x)$ has the dual appeal of being both intuitively reasonable and easy to calculate.

Suppose, on the other hand, that the statistician can specify an explicit class of alternative pdf's f_θ indexed by a parameter θ whose value must belong to some given parameter space Ω. In such a problem it appears to be reasonable to calculate the following likelihood ratio $\lambda(x)$ for the observed value x:

$$\lambda(x) = \frac{f(x)}{\sup_{\theta \in \Omega} f_\theta(x)}.$$

If $\lambda = 0.01$, for example, it indicates that there is a value of θ for which the likelihood of the observed value x is 100 times larger than it is under H. (More properly, if the supremum in the denominator of $\lambda(x)$ is not actually attained at any point $\theta \in \Omega$, then this likelihood may be only $100 - \epsilon$ times larger than it is under H.)

We shall now present a specific family K_1 of alternative pdf's for which $\lambda(x) = 1 - F(x)$. Consider the interval Ω of all numbers θ such that $0 < F(\theta) < 1$. Corresponding to each value of $\theta \in \Omega$ we can define a pdf f_θ as follows:

$$f_\theta(x) = \begin{cases} \dfrac{f(x)}{1 - F(\theta)} & \text{for } x \geq \theta, \\ 0 & \text{for } x < \theta. \end{cases}$$

In words, if f is the pdf of some random variable Y, then f_θ is the conditional pdf of Y given that $Y \geq \theta$. The family K_1 of alternatives is defined to be the family of all pdf's of the form f_θ for $\theta \in \Omega$.

For each value of $\theta \in \Omega$, we shall let F_θ denote the df corresponding to the pdf f_θ. Then $F_\theta(x) \leq F(x)$ for $-\infty < x < \infty$ and, if $\theta_1 < \theta_2$, then $F_{\theta_2}(x) \leq F_{\theta_1}(x)$ for $-\infty < x < \infty$. In other words, each distribution in the class K_1 is stochastically larger than the distribution for which the pdf is f, and furthermore, the distributions in the class K_1 form a stochastically increasing family of distributions, as defined in [7, p. 273]. Therefore, the class K_1 may be a reasonable class of alternative distributions to keep in mind in a problem for which a tail area calculation is regarded as a reasonable procedure.[1] Nevertheless, with due regard for these desirable properties, our main motivation in introducing the particular class of alternatives K_1 here is to illustrate how the relation $\lambda(x) = 1 - F(x)$ can be obtained.

For any observed value x such that $f(x) > 0$, we now have

$$\sup_{\theta \in \Omega} f_\theta(x) = \frac{f(x)}{\inf_{\theta \leq x}[1 - F(\theta)]} = \frac{f(x)}{1 - F(x)}.$$

It follows therefore that $\lambda(x) = 1 - F(x)$. Thus, for the given class K_1 of alternatives, the likelihood ratio $\lambda(x)$ will be equal to the tail area $1 - F(x)$ calculated under the hypothesis H. Hence, if the observed value x corresponds to a tail area of 0.01, it is indeed correct to state that there is an alternative distribution for which x is 100 times more likely than it is under H.

In a Bayesian framework, a prior pdf $\xi(\theta)$ is assigned to θ and the overall likelihood, or average likelihood, of the alternative class K_1 for the observed value x is then defined as follows:

$$L_\xi(x) = \int_\Omega f_\theta(x)\xi(\theta)d\theta.$$

Therefore, the likelihood ratio $\lambda_\xi(x)$ determined by x will be

$$\lambda_\xi(x) = \frac{f(x)}{L_\xi(x)}.$$

If b denotes the left end point of the interval Ω (it is

[1] Good [3, pp. 93–94], very briefly mentions a family of alternatives similar to K_1 when discussing a χ^2 test of the hypothesis that a particular die is fair.

possible that $b = -\infty$), we have

$$\lambda_\xi(x) = \left[\int_b^x \frac{\xi(\theta)}{1 - F(\theta)}\,d\theta\right]^{-1} > 1 - F(x).$$

Thus, $\lambda_\xi(x) > 1 - F(x)$ and the value of $\lambda_\xi(x)$ may actually be very much greater[2] than $1 - F(x)$, depending on the prior weight that was assigned to values of θ much smaller than x. For this reason, in Bayesian discussions of statistics, the statement is often made that an observed tail area of 0.01 typically corresponds to a likelihood ratio somewhat larger than 0.01 and is therefore not necessarily strong evidence against the hypothesis H (see, e.g., [1, p. 240]).

The completion of the Bayesian approach in this example would utilize the prior probabilities π and $1 - \pi$ that H and K_1 are true, respectively, and would yield the posterior odds, or overall likelihood ratio, $\pi f(x)/(1 - \pi)L_\xi(x)$. However, these considerations should not obscure the fact that $\lambda(x) = 1 - F(x)$ and, hence, that a tail area of 0.01 does indicate the existence of an alternative in K_1 that is 100 times more likely than H.

3. THE TAIL AREA AS A POSTERIOR PROBABILITY

We shall now present another class K_2 of alternatives which, like the class K_1, also comprises a stochastically increasing family of distributions. For any given positive integer θ, let the pdf g_θ be defined as follows:

$$g_\theta(x) = (1 + \theta)F^\theta(x)f(x) \quad -\infty < x < \infty.$$

The class K_2 of alternatives is defined to be the class of all pdf's of the form g_θ for $\theta = 1, 2, \cdots$.

If we let G_θ denote the d.f. corresponding to the pdf g_θ, then $G_\theta(x) = F^{1+\theta}(x)$ for $-\infty < x < \infty$ and $\theta = 1, 2, \cdots$. It can again be seen that for any positive integer θ, $G_\theta(x) \leq F(x)$ for $-\infty < x < \infty$ and, furthermore, the distributions in the class K_2 form a stochastically increasing family of distributions. The class K_2 is therefore another reasonable class of alternative pdf's to keep in mind in a problem for which a tail area calculation is a reasonable procedure.

Suppose now that the pdf of the statistic X is either f or some pdf in the class K_2. It can be seen from the definition of g_θ that $g_0 = f$. Therefore, the pdf of X is assumed always to have the form g_θ. Under the hypothesis H, the value of θ is 0, and under any alternative in K_2, the value of θ is a positive integer.

Now consider the prior distribution of θ, and suppose that the prior probability that $\theta = k$ is proportional to a_k for $k = 0, 1, 2, \cdots$; i.e., suppose that

$$\Pr(\theta = k) \propto a_k \qquad k = 0, 1, 2, \cdots.$$

Then the posterior probabilities will satisfy the following relation:

$$\Pr(\theta = k|x) \propto a_k g_k(x) \qquad k = 0, 1, 2, \cdots.$$

Specifically, the posterior probability $\Pr(\theta = 0|x)$ will be specified by the equation

$$\Pr(\theta = 0|x) = \frac{a_0 g_0(x)}{\sum_{k=0}^{\infty} a_k g_k(x)}.$$

We shall now show that there is a prior distribution of θ such that for any observed value x, the posterior probability $\Pr(\theta = 0|x)$, which is the posterior probability that H is true, is simply equal to the tail area $1 - F(x)$. The prior distribution which yields this result is as follows:

$$\Pr(\theta = k) \propto \frac{1}{k + 1} \quad k = 0, 1, 2, \cdots.$$

Since the sum of the values $1/(k + 1)$ is infinite, these prior probabilities cannot be normalized so that their sum is one. In this sense, the prior distribution is an improper distribution. The use of an improper prior distribution such as this one is well established in Bayesian statistical procedures (see, e.g., [1, Ch. 10]).

The posterior distribution derived from these prior values will be a proper probability distribution. Specifically, the posterior probabilities will have the form

$$\Pr(\theta = k|x) \propto F^k(x)f(x) \quad k = 0, 1, 2, \cdots.$$

It may be assumed that $0 < F(x) < 1$ for the observed value x, since x will satisfy this relation with probability 1. Then

$$\sum_{k=0}^{\infty} F^k(x) = 1/[1 - F(x)],$$

and it follows that the posterior probabilities are

$$\Pr(\theta = k|x) = [1 - F(x)]F^k(x) \quad k = 0, 1, 2, \cdots.$$

In particular, $\Pr(\theta = 0|x) = 1 - F(x)$.

Thus, if the statistician is willing to consider the class K_2 of alternatives, together with the improper prior distribution over H and K_2 specified here, then quite simply, the tail area $1 - F(x)$ is the posterior probability that H is true.

It is not being suggested that this particular prior distribution is necessarily the appropriate one to use in a given problem. In accordance with the Bayesian approach, each statistician will have his own subjective prior distribution for θ. However, the simple and appealing form of the prior distribution specified here may make it a convenient approximation in some problems.

4. ENLARGING THE CLASS OF ALTERNATIVES

We shall now enlarge the class K_2 of alternatives so that it contains all pdf's of the form g_θ for all positive numbers θ, and not just positive integers. Let K_3 denote this enlarged class of pdf's. It is still true that each distribution in K_3 is stochastically larger than the distribution for which the pdf is f, and the distributions in K_3 still form a stochastically increasing family of distributions.

We can now determine the value of the likelihood ratio $\lambda(x)$ when this enlarged family of alternatives is used. For

[2] Good [4, 5, 6] has suggested that in many problems, $\lambda_\xi(x) = \gamma[1 - F(x)]$ where γ lies in the range $10/3 < \gamma < 30$.

any observed value x, it can be found by elementary differentiation that $\sup_\theta g_\theta(x)$ is attained when θ has the following value $\hat{\theta}$:

$$\hat{\theta} = \frac{1}{\log [1/F(x)]} - 1.$$

We shall assume that the observed value x is large enough so that $\hat{\theta} > 0$. Therefore, $\hat{\theta}$ belongs to the parameter space of possible values of θ.

It is then found that

$$\lambda(x) = \frac{f(x)}{\sup_{\theta>0} g_\theta(x)} = eF(x) \log \frac{1}{F(x)}.$$

In accordance with our assumption that the observed value x is large, we shall assume that the tail area $\epsilon = 1 - F(x)$ is small. Then as a first-order approximation, obtained by ignoring terms of order ϵ^2 and smaller terms, we have $\lambda(x) = e[1 - F(x)]$.

Therefore, in this model, if the observed value x corresponds to a tail area of 0.01, there is an alternative distribution under which the likelihood of x is $100/e$ times as large as it is under H. If a prior pdf $\xi(\theta)$ is assigned to θ, then the value of the likelihood ratio $\lambda_\xi(x)$ will be larger than the value of $\lambda(x)$ which has just been derived. Hence, we again see that a tail area of 0.01 is not necessarily strong evidence against the hypothesis H. The actual likelihood ratio $\lambda_\xi(x)$ based on x for comparing H with K_3 must be larger than $(0.01)e$ and could be much larger.

5. CONCLUDING REMARKS

In summary, the weight of evidence against the hypothesis H that is implied by a small tail area $1 - F(x)$ depends on the model and the assumptions that the statistician is willing to adopt. If he assumes that the alternatives to H lie in the class K_1, then the likelihood ratio calculated from x is actually $1 - F(x)$. If he assumes that the alternatives to H lie in the class K_2, and he assigns a special improper prior distribution, then the posterior probability of H is actually $1 - F(x)$. Finally, if he assumes that the alternatives to H lie in the class K_3, then the likelihood ratio calculated from x is $e[1 - F(x)]$. A bigger class of alternatives could be formed by combining all the alternatives considered in this article and considering the class $K_4 = K_1 \cup K_3$. However, this class will not be studied further here.

Other classes of alternatives with properties similar to K_1 and K_3 can be developed as follows: For each value of θ in an interval Ω, let $\varphi_\theta(t)$ denote a pdf on the unit interval $0 < t < 1$. Then for each value of $\theta \in \Omega$, it follows that $\varphi_\theta[F(x)]f(x)$ will be a pdf over the real line. If K denotes the class of alternatives containing all these pdf's for all values of $\theta \in \Omega$, then both the likelihood ratio $\lambda(x)$ and the ratio $\lambda_\xi(x)$ for a given prior pdf ξ will depend on the observed value x only through the tail area $1 - F(x)$. Both the classes K_1 and K_3 are obtained by special choices of the function φ_θ.

It would be interesting to learn whether any one of the classes K_1, K_2, or K_3, or some other class of the form K, can be considered as a "natural" class of alternative distributions for the statistic X, in the sense that this class of distributions can be derived from some "natural" assumptions, when H is not true, about the joint distribution of the random variables in the sample from which X is computed.

[*Received June 1972. Revised April 1973.*]

REFERENCES

[1] DeGroot, Morris, H., *Optimal Statistical Decisions*, New York: McGraw-Hill Book Co., 1970.

[2] Efron, Bradley, "Does an Observed Sequence of Numbers Follow a Simple Rule? (Another Look at Bode's Law)," *Journal of the American Statistical Association*, 66 (September 1971), 552–59.

[3] Good, I.J., *Probability and the Weighing of Evidence*, New York: Hafner Publishing Co., 1950.

[4] ———, "Saddle-Point Methods for the Multinomial Distribution," *Annals of Mathematical Statistics*, 28 (December 1957), 861–81.

[5] ———, "Significance Tests in Parallel and in Series," *Journal of the American Statistical Association*, 53 (December 1958), 799–813.

[6] ———, "A Bayesian Significance Test for Multinomial Distributions (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 29 (1967), 399–431.

[7] Lehmann, E.L., *Testing Statistical Hypotheses*, New York: John Wiley and Sons, Inc., 1959.

[8] Lindley, D.V., *Introduction to Probability and Statistics from a Bayesian Viewpoint; Part 2, Inference*, London: Cambridge University Press, 1965.