SOME RESULTS ON ROBUSTNESS IN TESTING

by

Herman Rubin[1]
Purdue University

Technical Report #86-47

Department of Statistics
Purdue University

1986

[1] Research supported by the National Science Foundation under Grant DMS-8401996.

*

SOME RESULTS ON ROBUSTNESS IN TESTING

by

Herman Rubin[1]
Purdue University

ABSTRACT

We discuss some of the problems of robustness with respect to prior assumptions in testing. Some of these problems have been treated previously from the standpoint of prior robustness. We extend these results and show that even the more desirable uniform posterior robustness can frequently be obtained, and how one can sometimes still get prior robustness using what remains of uniform posterior robustness if it sometimes does not hold.

INTRODUCTION

There are several approaches which have been made to the problems of robustness and testing from the Bayesian, decision-theoretic, and classical viewpoints. We attempt to unify these viewpoints, and to some extent we show that this is possible. In particular, we claim that the Bayesian position should be modified, and that with this modification it is frequently feasible in testing problems to achieve robust results from this standpoint, which includes the decision-theoretic approach. We give some examples which show that in many reasonable cases this can be accomplished, but that care must be taken, both from the prior and the posterior standpoint. At the end of this paper, we discuss those aspects of the classical approach which should be included. We only consider the fixed-sample fixed-experiment problem in this paper.

To carry out this discussion, it is necessary to properly define the terms used. We now proceed to do this.

A *testing problem* is a problem in which one of two actions, which we shall call *accept* and *reject*, must be taken.

The *decision-theoretic* approach to a problem is that the evaluation of a procedure is to be based on its consequences in *all* states of nature. In any particular state of nature, the assessment of the consequences is determined by the probability distribution of the results. This assessment is often described by a loss function.

The *posterior Bayesian* position is that the procedure should be based entirely on the posterior distribution based on the observation.

The *axiomatic decision-theoretic approach* is that certain consistency (coherence) axioms are to be satisfied. This leads to a *prior Bayesian* position.

[1] Research supported by the National Science Foundation under Grant DMS-8401996.

The *robustness* of a procedure is the extent to which its properties do not depend on those assumptions which we do not wish to make.

Let us approach the testing problem with these points in mind. Let Ω be the collection of possible states of nature, Θ the set of those states (the *null hypothesis*) for which it is desired to accept, and Φ the set of those states for which rejection is preferable. We assume that $\Omega = \Theta \cup \Phi$, although except for the inferential Bayesian this assumption is irrelevant.

Now the inferential Bayesian would say that all that is wanted is the posterior probability of Θ. We maintain that this ignores too much. The consequences of a wrong action can vary substantially with the value of the parameter. It is unlikely that one would be disturbed at saying that a coin is fair if its bias is 10^{-50}, but this is not the case if the probability of heads is 0.9. Thus not only the probability, but also the distribution, must be taken into account.

The usual way that the posterior decision-theoretic Bayesian makes the decision is to integrate the loss function over Θ and over Φ with respect to the posterior, and to accept (reject) if the Θ (Φ) integral is larger; if they are equal, the actions are indifferent. This can be summarized by stating that the posterior decision-theoretic Bayesian acts on the values of the integrals only.

Now what does the axiomatic prior Bayesian (coherent decision-theorist) do? Whatever axiom system is used, the conclusion is that the procedure to be used is that which minimizes the integral of the risk with respect to the prior measure, that is, the expression

$$\rho(\xi, \delta) = \int \rho(\omega, \delta)\, \xi(d\omega),$$

where $\rho(\omega, \delta) = E_\omega(L(\omega, \delta(X)))$.

Note that the loss function L and the prior ξ enter only through their product, and if we define the *weight measure* W by $W(d\omega, a) = L(\omega, a)\xi(d\omega)$, it is only the weight measure that counts. In fact, in Rubin [1987], it is argued that the separation is not operational and coherence does not require its existence.

The prior Bayesian approach considers minimizing the prior Bayes risk. The posterior Bayesian instead uses the posterior Bayes risk. Of course, Bayes' Theorem tells us that these give the same procedure, assuming that the prior Bayes risk is finite, if the weight measure is a countably additive positive measure. Note that the finiteness of the prior measure is unimportant for this result; however, without countable additivity, the Radon-Nikodym Theorem is quite complicated, and the usual procedure of using pointwise limits of Radon-Nikodym derivatives for sequences of countably additive measures can be finagled to yield any result.

ROBUSTNESS

We now wish to consider the problem of robustness. Usually in considerations of robustness we consider the ratio of the consequences of the candidate procedure with respect to those of the optimal procedure. The ordering of the expressions we are comparing is unaffected by adding a (signed) measure to W, or equivalently a function of ω alone to the loss function L. If we allow the resulting weight measure or loss function to become negative, ratios which are intuitively close to 1 can be made arbitrarily close to 0 or ∞, and if we let the quantities become negative, can even reverse their order. Thus it is common to consider the regret form of the problem, i.e., to have the best action for a given ω to have loss or weight 0. This makes the ratio as extreme as possible, assuming non-negative losses.

We consider a procedure to have good prior robustness if the ratio of the prior risk of the procedure to that of the optimal procedure is close to 1. For a particular value x of the observation, we can similarly define posterior robustness at that value. If a procedure has *uniform* posterior robustness, it has prior robustness. It is possible for a procedure to have good posterior robustness except on a "small" set of observations and still be prior robust, but it is necessary to show that the exceptional set is small enough, and this may not be trivial. In all of the cases in which we can obtain prior robustness results with respect to the full likelihood Bayes procedure and in which the likelihood functions corresponding to the observations and to the candidate procedure can be reasonably calculated, this must be true. This must also be the case for those situations, such as the use of non-parametric procedures, in which prior robustness results may even be fairly easy, but for which, at present, there is no reasonable method to obtain the corresponding likelihood function.

We give a simple example, which the reader can easily check, in which it is not difficult to obtain prior robustness, while uniform posterior robustness results cannot be obtained. Let us consider testing that the mean of a two-dimensional normal random variable is 0. Assume that the loss is 1 for a wrong decision, and that the prior measure can be approximated by a point mass of q at 0, and a uniform density of $1/(2\pi)$ on the alternative. Then for covariance matrix vI, the odds ratio is $qe^{-\|x\|^2/2v} : v$, and for any values of q and v with q considerably larger than v, an uncertainty in q of a substantial factor will yield a range of $\|x\|^2$ for which the posterior procedure can switch from strongly in favor of acceptance to strongly in favor of rejection. Both classical and Bayes procedures say that one should reject if $\|x\|^2 > h$, for which the prior risk ρ is $r(h) = (qe^{-h/2v} + h)/(2\pi)$. The optimal value of h is $2v \log(q/2v)$, for which the risk is $r = (2v + h)/(2\pi)$. Thus if $q = 1$ and $v = 0.0001$, and if the user acts on the assumption of some other value of q, the prior risk does not increase by more than 10% if the assumed q is anywhere in the range (0.367, 4.45); if $v = 0.0000005$, the 10% range is (0.262, 10.9). Nevertheless, since q is a factor in the posterior odds, it is clear that there is a range of values of $\|x\|$ for which, depending on the value of q, the Bayes procedure can strongly favor acceptance or rejection, and thus for which the posterior risk can vary considerably.

Why consider prior robustness at all? There are several reasons. First, for the prior Bayesian, this is what counts. Since I have been a prior Bayesian for most of my professional life, this is what led me to work in the direction of prior robustness starting more than twenty years ago. Second, it is relatively easy to obtain prior robustness results. Since we are looking at the integrated weight measure, or the integral of loss with respect to the prior, we only need to consider the marginal distribution of the results for each state of nature. This can be done even in situations in which the likelihood functions are currently unavailable at present, such as most "non-parametric" procedures. Third, prior robustness results can hold in cases in which uniform posterior results are not known, such as for non-parametric tests, or even in cases, such as the preceding example, where posterior robustness results do not hold; in fact, before the research which led to this paper and to the definition of uniform posterior robustness, I was unaware that the latter alternative was even available. Fourth, in most problems in which likelihood function calculations can be made, one obtains posterior robustness over much of the observation space; this is useless for comparing procedures. In the example above, if prior robustness does not hold, the fact that over much of the sample space the posterior action does not depend on which "reasonable" prior is assumed is irrelevant; it is only those parts of the sample space which contribute significantly to the risk which matter.

PREVIOUS RESULTS

In several previous papers, the author has considered the problem of testing from the prior decision-theoretic standpoint. In Rubin and Sethuraman [1965], the problem of testing a point null against a composite alternative was considered assuming that the rejection loss could be considered homogeneous of some degree in the neighborhood of the null hypothesis, and the rejection prior could be approximated by some multiple of Lebesgue measure there. Explicit prior robustness results (how well does the procedure approximate the risk of the Bayes procedure *before* the observations are taken) are obtained. It is indicated there that there is considerable robustness with respect to the parameters. We have just given an example of this. This paper also has some results on the consequences of using procedures not even based on the likelihood function. For some specific testing problems, the author has followed this up in Rubin [1971] and Rubin [1972] for the efficiencies of some standard non-parametric procedures.

In Rubin [1971] partial results were obtained for the two-sided problem of testing an imprecise null hypothesis. This is usually the case in testing a "point null"; it is rare that the "point null" can even be true. (We do not test that the speed of light in a vacuum is constant–we test that the distribution of the results in two experimental situations is "the same".) In this paper, the prior assumptions were stated in terms of a weight measure. The results here are also in the form of prior robustness. The specific formulation of the problem in that paper is to assume that the null hypothesis is $\omega = 0$, and to make the difference of the densities of the weight measures to be of the form $(h(\omega) - c\omega^2)d\omega$, where h was taken to be a multiple of a probability density. This is the two-sided form we shall treat here.

The intuitive idea behind this is that we can consider a state of nature as composite; in the neighborhood of the "point null" there is a measure, which we assume finite, which gives the weight for accepting the null, and that the weight for rejection is proportional to the square of the parameter with respect to Lebesgue measure. We assume that the acceptance measure has a density and note that there are three scalings available if we assume normal observations. We can scale the weight measure, we can scale the parameter, and we can scale the variance of the observations given the state of nature. Only the latter scaling provides any conceptual difficulty, but if we look at our observations as normal with a given variance, rather than corresponding to a given sample size, this problem vanishes.

In that paper, it was found numerically that if h is normal or double exponential, that if the the standard deviation of the observations and the location of the sign change of the weight difference differ by an order of magnitude, one may get an "easy" robust procedure. If h is concentrated it can be treated as a point null, and if the observations are concentrated it is only needed to estimate the parameter and see whether the estimate does or does not fall in the acceptance region. However, if these two parameters are comparable, no prior robust procedure was found. These conclusions hold for a wide variety of forms of h. This led to the current research.

The one-sided case is apparently more difficult, but this approach still works. Here we assume that we have a sufficiently large sample that the difference between the acceptance and rejection loss functions can be taken to be linear, that is, $L(\omega,\text{accept})$ - $L(\omega,\text{reject}) = k\omega - b$. We assume that the usual formulation of the hypothesis makes the division point 0, but we may need a constant term because of the difference of cost of treatments, unknown side effects, etc. We shall assume that the prior in the neighborhood of the null can be approximated by a multiple of Lebesgue measure plus a "bulge" which is concentrated near 0. Thus, after normalization, we can treat the weight measure as $dW(\omega) = (\omega - b)(d\omega + \rho(d\omega))$, where ρ is that portion of the prior corresponding to the belief that the parameter may be approximately 0. This gives a "bulge" in the prior. Now the Bayes procedure depends on the location of the zero, the mass of the bulge, the shape of the bulge, and the standard deviation of the (assumed normal) likelihood function. We do not know any quick way of obtaining the procedure. This means that the methods used in the previous papers, producing a reasonably simple procedure and comparing its risk to the Bayes risk, does not yield to easy analysis.

RESULTS

Since the prior assumptions are summarized in the weight measure, we must summarize the posterior results in the same manner. Let f be the likelihood function and g the prior weight measure. Then the posterior weight measure is summarized by $Q(d\omega) = f(\omega)g(d\omega)$. We should accept if $Q(\Omega) > 0$ and reject otherwise. The analog of the posterior probability of the acceptance region Θ in the parameter space is $V = Q(\Theta)/|Q|(\Omega)$, and the posterior regret of any procedure is proportional to V if we reject and $1 - V$ if we accept. Thus if, through error, we use U instead of V, we only err if U and V are on opposite sides of 0.5. However, the increase in the normalized regret is bounded by $|U - V|$. Hence if we can show that U and V are uniformly close, we have a good procedure. If this can be done uniformly for all values of the observations, we say that we have *uniform posterior robustness.*

Now in the one-sided case V is very difficult to handle. However, the contributions of the flat part of the prior and of the "bulge" are added; thus, if we can get uniform posterior robustness for both parts, we have uniform posterior robustness.

To give a simple example of the possibility of uniform posterior robustness, consider testing the point null $\omega = 0$ with loss 1 for improper rejection against the alternative prior density $(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}\omega^2/\tau^2}$ with squared error loss for improper acceptance. If the data is normal with variance 0.01, the maximum difference between the decision versions of the posterior probabilities is less than 0.018 for $\tau \in (1, \infty)$. Note that the robustness does not hold when the assumptions are expressed in terms of the prior odds ratio which is τ, but in terms of the weight measure, or in terms of the prior measure, locally normalized, and the loss. Even if the sample variance is 0.1, the maximum difference is less than 0.104. If the rejection loss is 1, in which case we are comparing posterior probabilities, the maximum difference is less than 0.051 for variance 0.1. Note that we are only talking about differences *and not ratios* of the normalized posterior measures. In the last problem, if the rejection loss is a constant 1000, the maximum difference is 0.378 for variance 0.1, and τ is important for the decision problem, but is 0.047 for variance 0.01; however the ratios are the same.

With the approach we are taking, it is only required to "approximate" the Q-measure for the acceptance and rejection subsets of Ω. The approximation need not be too good in those cases in which the decision analog of the odds ratio is overwhelming, provided the approximation ratio is strongly in the same direction. Accordingly, we find it necessary to estimate the convolution of a normal distribution with a prior measure for the acceptance and rejection sets in the two-sided problem, and to estimate the integral of a linear function with respect to the product of the densities in the one-sided case.

It is sometimes easy to estimate such an integral, but frequently approximating

$$\int_{-\infty}^{\phi} f(\theta) g(X|\theta)\, d\theta,$$

where f is a prior density multiplied by a simple function (such as 1 or θ) and g is the normal density of the observation X, can be quite difficult. For instance, if the scales of the two densities are comparable, so that the integral does not have a convenient expansion about the mode of the concentrated density, either prior or observational, then we need a sophisticated approach. If the prior density is normal, there is no problem. Can we try to use this? Intuitively, the integral "should be" concentrated in the neighborhood of the crossing point ϕ. If this is the case, the prior density could be approximated by that multiple of the normal density which matches it and its first derivative at ϕ. Sometimes this gives a good approximation. In other situations, and this is easily seen for the Cauchy, we may often succeed by placing the remaining mass of the prior, which we assume centered at 0, as a point mass at 0. This does not always give a good approximation to the integral, but in the cases which are of importance in testing, the approximation is frequently good where it is of most importance.

Even when uniform posterior robustness cannot be obtained, we claim that the approach can still be used. We may be able to show that failure of posterior robustness is sufficiently rare that, at least for the prior Bayesian, it is unimportant. For the posterior Bayesian, we can identify those situations in which the assumptions made are inadequate, and thus inform the user that the data observed is such that more *must* be specified by non-statistical means to enable a sound decision to be made.

ADDITIONAL PROBLEMS

Notice that we have only considered robustness with respect to the prior assumptions about the parameters, assuming that the likelihood function can be taken to be that of a normal distribution to the approximation needed, which is frequently true. If the likelihood function is known, but is not normal, similar results are common, but we have not obtained any general results in such cases.

However, the typical non-decision-theoretic treatment of robustness ignores these problems and considers only the use of possibly inefficient procedures. We can and *must* combine these two aspects of robustness. *It is still frequently true that the likelihood function of the "robust statistic" that the classicist would use can be approximated by that of a normal distribution.* In fact, often the robust statistic has a more normal likelihood function that the original problem. The normal approximation need only be shown for those deviations which matter, and are likely to involve excessive deviations. In many cases, it will not be possible to obtain the local limit theorems which seem to be needed to approximate the density, but the necessary interchange of the order of integration to convert to integrating the probability of deviations may be adequate.

Unfortunately, for many tests, such as the Kolmogorov-Smirnov test, it is difficult to obtain even an approximate likelihood function under the alternative. As it is shown in Rubin and Sethuraman [1965], and also in Rubin [1972], it is still possible to obtain prior robustness results. We believe that it may still be possible, in the case that the type I error of asymptotically good tests is very small, to achieve posterior robustness, but we have not yet succeeded in this.

ACKNOWLEDGEMENT

Research supported by the National Science Foundation under Grant DMS-8401996.

REFERENCES

[1] Rubin, H. [1970] Decision-theoretic evaluation of some non-parametric methods. *Nonparametric Techniques in Statistical Inference*, edited by M. L. Puri. London, Cambridge University Press, 1970, pp, 579-583.

[2] Rubin, H. [1971] A decision-theoretic approach to the problem of testing a null hypothesis. *Statistical Decision Theory and Related Topics*, edited by S. S. Gupta and J. Yackel. New York and London, Academic Press, 1971, pp. 103-108.

[3] Rubin, H. [1972] On large-sample properties of certain nonparametric procedures. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1972, vol. 1, pp. 429-435.

[4] Rubin, H. [1987] A weak system of axioms for "rational" behavior and the non-separability of utility from prior, to appear in *Statistics and Decisions*

[5] Rubin, H. and Sethuraman, J. [1965] Bayes risk efficiency. *Sankhyā Ser. A* **27** (1965) pp. 347-356.