COMMENTS ON BERGER'S CRITICISM OF SOME CLASSICAL STATISTICAL PROCEDURES AND ON SANDVED'S CRITICISM OF BERGER'S CRITICISM

by Ragnar Norberg

Abstract

In his book "Statistical Decision Theory" James Berger cites some examples of classical statistical methods that, so it is claimed, may produce nonsensible conclusions. In a recent discussion Else Sandved argues that these methods, being in conflict with ancillarity considerations, are not likely to be used by classical statisticians. The present note gives further reasons why I think the cited examples are not killing to classical statistics and that examples of this kind cannot settle the controversy Bayes vs. Classical. A reply from Berger is appended.

1. Introduction

Berger's (1980) book is refreshening reading. The author gives the reader a part in his Bayesian revelations and lays open his reasons for converting to new convictions. Some of them were not convincing to me, however, and the following paragraph explains why. On the whole I share the points of view put forward by Sandved (1987), and I add here some on my own account. I urge to say that I find both classical and Bayesian models perfectly meaningful: they are appropriate mathematical formulations of different attitudes depending on philosophical positions and - I think - also to some extent on the nature of the problem. Paragraph 3 presents some personal opinions of mine on these matters and concludes that efforts to establish that Classical is absurd and Bayes is meaningful, or vice versa, are futile.

2. Two examples referred to by Berger

Example 1. Let $X_1,\ldots,X_n$ be a random sample from $\mathcal{U}(\vartheta-1/2, \vartheta+1/2)$. The pdf (probability density function) of the observations is

$$f_\vartheta(x_1,\ldots,x_n) = \begin{cases} 1, & \max x_i - ½ < \vartheta < \min x_i + ½ \ , \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

A convenient sufficient statistic is

$$(Y,Z) = (\max X_i - \min X_i, \tfrac{1}{2}(\min X_i + \max X_i)),$$

the range and the midrange. Its distribution is given by

$$Y \sim \mathcal{B}e(n-1,2),$$

$$Z|_Y \sim \mathcal{U}(\vartheta - \tfrac{1}{2}(1-Y), \vartheta + \tfrac{1}{2}(1-Y)). \qquad (2)$$

Thus Y is an ancillary statistic, that is, Y itself contains no information about ϑ, but its value decides how accurately ϑ can be determined by the point estimator $\hat{\vartheta} = Z$.

An obvious confidence interval based on (2) is

$$\hat{\vartheta} \pm \tfrac{1}{2}(1-\alpha)(1-Y). \qquad (3)$$

For each fixed Y its conditional confidence level is $1-\alpha$, and so this is also the unconditional confidence level.

Many other confidence intervals can be constructed. For instance, the marginal pdf of $\hat{\vartheta} = Z$ is

$$g_\vartheta(z) = \begin{cases} n(1 - 2|z-\vartheta|)^{n-1}, & |z-\vartheta| < \tfrac{1}{2}, \\ 0 & \text{otherwise}, \end{cases}$$

from which we obtain the $1-\alpha$ confidence interval

$$\hat{\vartheta} \pm ½(1 - \alpha^{1/n}). \tag{4}$$

Berger (1980, p. 19) proclaims (4) "The classical $1-\alpha$ confidence interval" and then denounces it: if $n = 25$, $\alpha = 0.056$, $\min X_i = 3.1$, $\max X_i = 3.2$ ($Y = 0.1$, $\hat{\vartheta} = 3.15$), then (4) turns out as (3.094, 3.206), which cannot reasonably by ascribed 95% confidence since "all that is really known about ϑ is that it lies somewhere between 2.7 and 3.6".

Already the quoted statement takes the sting out of the criticism since the interval (2.7, 3.6) is just the classical interval (3) with $\alpha = 0$, the 100% confidence interval based on (2). Clearly, "The classical confidence interval" is a frauditive acquisition. The interval (4) is based solely on $\hat{\vartheta}$, which is not a sufficient statistic. It is only to be expected that for some outcomes $X_1,\ldots,X_n$ it will produce conclusions in apparent conflict with "final precision" considerations that take also Y into account. The interval (3), which utilizes both $\hat{\vartheta}$ and Y, does not suffer from any such weaknesses. In the chosen numerical example it yields (2.72, 3.58), a completely sound 95% final precision statement.

Classical statistics would be in trouble if one could construct a problem where it seems impossible to find a classical solution that avoids absurdities with regard to final precision considerations. The present example does not serve that purpose.

But, there is a but. It is not clear that classical statisticians will unanimously prefer interval (3) to interval (4) in all situations. All those who adopt the ancillarity principle advocated by Sandved (1987) will. But there might be others who would base their choice of method entirely on some decision theoretic performance criterion, e.g. the expected length of a confidence interval. The expected lengths of (3) and (4) are, respectively,

$$l_3(\alpha,n) = (1-\alpha)\,\frac{2}{n+1}$$

and

$$l_4(\alpha,n) = 1 - \alpha^{1/n}.$$

It is easy to verify that for each $n \geq 3$ there exists an $\alpha(n) \in (0,1)$ such that

$$l_4(\alpha,n) \gtreqless l_3(\alpha,n) \quad \text{according as} \quad \alpha \lesseqgtr \alpha(n). \qquad (5)$$

Thus, as measured by expected length, (3) is better than (4) for small α, whereas for large α it is the other way around. For $n = 2$ (4) is the better for all $\alpha > 0$. The following table indicates how $\alpha(n)$ depends on n.

Table 1. The values of $\alpha(n)$ given by (5) for some n.

n	3	5	10	25	50	100
$\alpha(n)$	0.2500	0.2250	0.2137	0.2075	0.2054	0.2043

It must be admitted that classical statisticians who judge confidence intervals only by their expected lengths and are content with confidence levels less than 75%, are vulnerable to Berger's criticism. Paying regard only to average performance, they accept unreasonable conclusions in some sample points. It is, however, only fair to note that such outcomes are rare. Berger remarks to his numerical example that "it was unlucky to obtain such an uninformative sample". This is to say the least: the probability that $Y \leq 0.1$ in a sample of size 25 is $2.26 \cdot 10^{-23}$.

Now, to turn examples like these into compelling arguments in favour of Bayesianism, it is not sufficient to establish that some classical statisticians will sometimes present unreasonable answers. It must also be substantiated that all Bayesian answers will be reasonable.

One prominent Bayesian procedure is to use a noninformative prior, which in the present case is the (improper) uniform distribution $\mathcal{U}(-\infty,\infty)$. This leads to the posterior distribution $\mathcal{U}(\hat{\vartheta} - ½(1-Y)), \hat{\vartheta} + ½(1-Y))$. Thus, a Bayesian $1-\alpha$ credibility interval is (3), which was judged to be reasonable. It is a highest probability density credibility region, but so is any other subset of $(\hat{\vartheta} - ½(1-Y), \hat{\vartheta} + ½(1-Y))$ with measure $(1-\alpha)(1-Y)$ - it need not be an interval and it need not contain $\hat{\vartheta}$.

Another widely accepted Bayesian solution is to use a conjugate prior, which must be of the same form as (1), that is, a uniform distribution $\mathcal{U}(\vartheta', \vartheta'')$ (with $0 < \vartheta''-\vartheta' < 1$?). The corresponding posterior distribution is $\mathcal{U}(\max(\vartheta', \max X_i - ½), \min(\vartheta'', \min X_i + ½))$, provided that the indicated interval is not empty. If the Bayesian's prior opinion is that ϑ lies somewhere between $\vartheta' = 1.5$ and $\vartheta'' = 2.5$ and the outcome of the observations is as in Berger's example, then the posterior does not exist and nothing can be stated as to the whereabouts of ϑ. This absurdity I would not call "The Bayesian credibility interval". There are as many Bayesian solutions as there are priors, and it would be unfair to judge the Bayesian approach by the prejudices of one particular Bayesian whose judgement is poor. Almost as unfair as to judge classical statistics by one poor confidence interval, but only almost: after all, the crux of Bayesianism is to defend subjective beliefs ("a prior is a prior is a prior"), including the totally mistaken ones.

Example 2. An observation X which is $\mathcal{N}(\vartheta, \sigma^2)$ is to be observed, and it is desired to test H_0: $\vartheta = 0$ versus H_1: $\vartheta = 10$. The experimenter will be supplied with one of two possible measuring instruments to obtain the observation X. The first has $\sigma = 1$ (a new accurate instrument), while the second has $\sigma = 10$ (an old inaccurate instrument). The experimenter will receive the first instrument with known probability p and the second with probability 1-p, and he will know which instrument he has received.

A class of level α tests is of the following form.

$$\text{Reject } H_0 \text{ if } X > K(\sigma), \tag{6}$$

where $K(\cdot)$ is chosen so that

$$p[1 - \Phi\{K(1)\}] + (1-p)[1 - \Phi\{K(10)/10\}] = \alpha. \tag{7}$$

Here Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Berger compares two special tests in the class:

Test 1: $K(\sigma) = K_1\sigma$, where K_1 is the upper α-fractile of $\mathcal{N}(0,1)$, i.e. reject with conditional level α for each instrument.

Test 2: $K(1) = K_2$, $K(10) = -\infty$, with K_2 determined by (7) (feasible if $p > 1-\alpha$), i.e. reject if $X > K_2$ for the accurate instrument, and always reject with the inaccurate instrument.

We quote from Berger (1980, p. 20): "It can be shown ... that for many values of α and p, Test 2 is more powerful than Test 1 ... hence a classical statistician concerned only with initial precision would recommend Test 2. One can imagine the reaction of the experimenter, who happens to get stuck with using the second instrument, when he is told by the statistician to ignore the experimental result and reject. If the experimenter is doing a long series of similar experiments ... it might be possible to convince him to use Test 2, but if he is involved in a one-time experiment he will be considerably less than enthusiastic about the advice. The experimenter in the latter case is interested only in final precision (which is the precision he can obtain using the measuring instrument he is given)."

Again Sandved (1987) points out that ancillarity considerations lead to conditional inference: the test should be designed separately for each instrument with regard to the conditional (final precision) probabilities of errors of types I and II, not the unconditional (initial precision) probabilities. Note that it does not necessarily follow that she would use Test 1 since she is not compelled to use the same level for both instruments. I shall return to this shortly.

The essence of Sandved's counter-argument is that initial precision criteria should not be used in the present situation. However, since initial precision considerations are just what Berger criticizes, it remains to see if they could be defended. Thus, let us take the position that the unconditional probabilities of erroneous decisions are all that matters. A typical situation would be that the test is to be performed repeatedly as an acceptance control of articles which may be either intact ($\vartheta = 0$) or defect ($\vartheta = 10$). The test is a sorting mechanism which is to be designed merely with regard to the fraction α of rejected intacts and the fraction β of accepted defects in the long run.

As a first step the problem can be stated as that of maximizing the power $1-\beta$ for a given level α. The observables are the standard deviation Σ of the measuring instrument and the measured value X, whose joint pdf at $(\Sigma,X) = (\sigma,x)$ is

$$f_{\vartheta}(\sigma,x) = \{pI_{\{1\}}(\sigma) + (1-p)I_{\{10\}}(\sigma)\}(2\pi\sigma^2)^{-½} \exp\{-(x-\vartheta)^2/2\sigma^2\},$$

where I_A is the indicator function of the set A. Using the Neyman-Pearson lemma, we easily find that the most powerful test is the following member of the class of tests given by (6) and (7).

Test 3: $K(1) = 5 + K_3$, $K(10) = 5 + 100K_3$, with K_3 determined by (7).

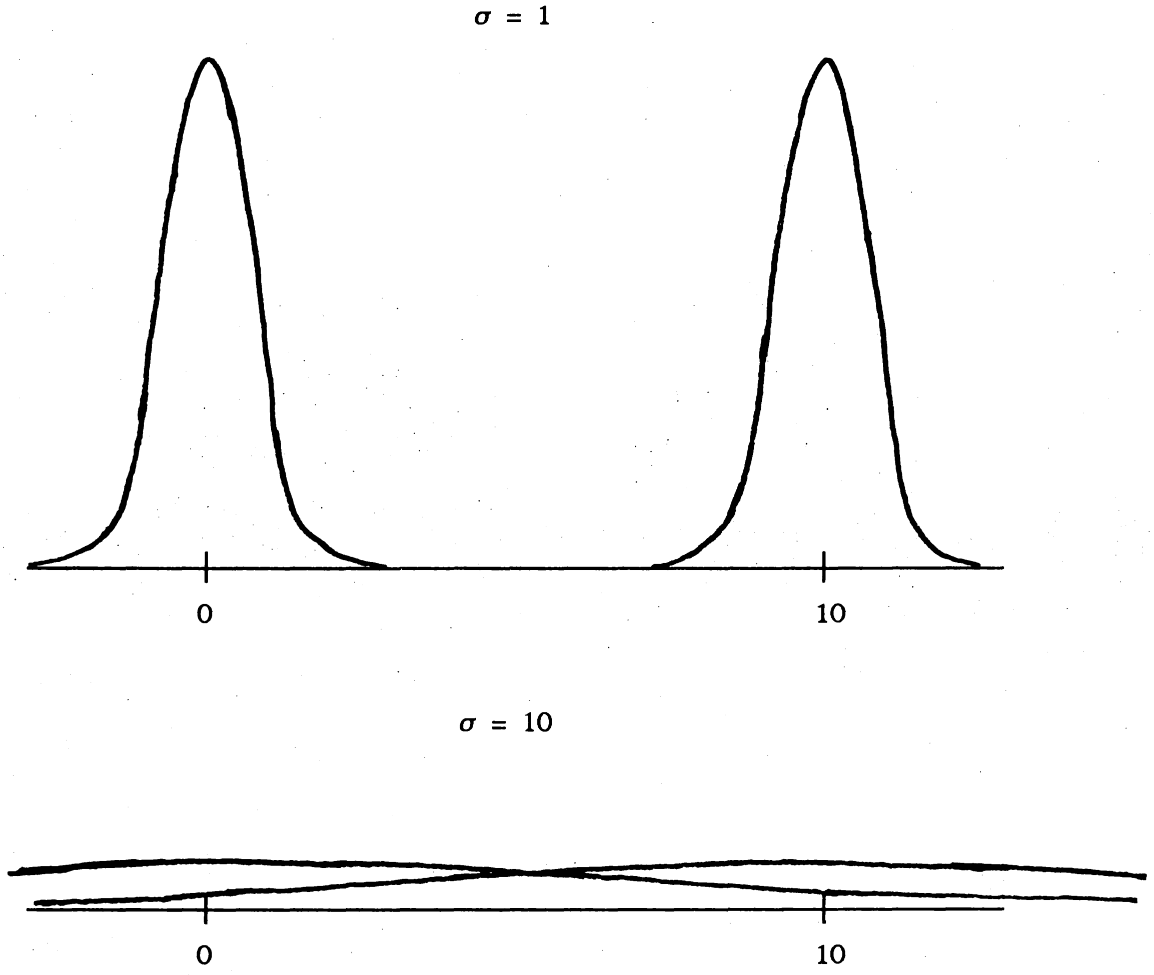
Test 3 will be recommended by classical statisticians concerned with initial precision, and it is this test that has to be examined and possibly critized from a final precision point of view. Table 2 shows, for each of the Tests 1 and 3, the conditional rejection limits $K(\sigma)$, the conditional probabilities of type I error, $\alpha(\sigma) = 1 - \Phi\{K(\sigma)/\sigma\}$, and of type II error, $\beta(\sigma) = \Phi\{(K(\sigma)-10)/\sigma\}$, and the unconditional probability of type II error, $\beta = p\beta(1) + (1-p)\beta(10)$, for some values of p and $\alpha = p\alpha(1) + (1-p)\alpha(10)$.

Table 2. Conditional and unconditional performance of Tests 1 and 3 for some values of α and p. Small numbers are written shortly by indicating the number of zeros between the decimal point and the first nonzero digit by a subscript, e.g. $._{2}31$ signifies 0.0031.

Case	α	p	Test	K(1)	$\alpha(1)$	$\beta(1)$	K(10)	$\alpha(10)$	$\beta(10)$	β
1	.05	.10	1	1.65	.05	$._{6}30$	16.5	.05	.74	.667
			3	5.11	$._{6}16$	$._{6}51$	15.9	.056	.72	.650
2	.05	.50	1		Same as for Case 1					.371
			3	5.08	$._{6}19$	$._{6}43$	12.8	.10	.61	.305
3	.05	.90	1		Same as for Case 1					.074
			3	4.95	$._{6}37$	$._{6}22$	0	.50	.16	.016
4	.05	.99	1		Same as for Case 1					$._{2}74$
			3	1.75	.04	$._{16}7$	-320	1^{*}	0^{**}	$._{16}69$
5	$._{2}31$	.99	3	5	$._{6}29$	$._{6}29$	5	.31	.31	$._{2}31$

$^{*}1-0._{224}55$ $\quad$ $^{**}0._{238}46$

Cases 1-3 do not comply with the requirement $p > 1-\alpha$ (hence Test 2 cannot attain level α). They are included here because they bring forth an aspect that has not been touched upon in the previous discussions by Cox (1958), Sverdrup (1966), Cornfield (1969), Berger (1980), and Sandved (1987), namely that it is not at all obvious that Test 1 is reasonable in general. Indeed, it seems rather unwise to reject H_0 at 5% level in the presence of the accurate instrument since the quality of the article can then be determined virtually with certainty, see Fig.1.


$\sigma = 1$

$\sigma = 10$

Fig.1. The densities of $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}(10,\sigma^2)$ for $\sigma = 1$ and $\sigma = 10$.

A rejection limit K(1) in the vicinity of 5 yields $\alpha(1) \approx \beta(1) \approx 0$. By lowering K(1) to 1.645 one increases the fraction of rejected intacts to 5% without gaining any essential decrease in the fraction of accepted defects. Test 3, based on initial precision, takes care of this viewpoint in Cases 1-3.

The picture turns out differently in Case 4, where $1-\alpha < p$. The explanation is concealed in the great unbalance between $\alpha = 0.05$ and $\beta = 6.9 \cdot 10^{-17}$, which implies a very special and, indeed, extreme judgement of the consequences of the two types of error. This point, which is crucial in classical as well as in Bayesian decision theory, has to be discussed more carefully.

A complete classical analysis of the testing problem would include the specification of a loss function, typically of the form

$$L(\vartheta,d) = \begin{cases} a & \text{if} \quad \vartheta = 0 \quad \text{and} \quad d = H_1, \\ b & \text{if} \quad \vartheta = 10 \quad \text{and} \quad d = H_0, \\ 0 & \text{otherwise,} \end{cases}$$

with a, b > 0. Test 3 is only the first step in the construction of a decision rule $\delta(x)$: it generates the set of admissible δ, one for each K_3. To fix K_3, one must introduce a criterion for evaluating the risk function,

$$\rho(\vartheta,\delta) = \begin{cases} a\alpha & \text{for} \quad \vartheta = 0, \\ b\beta & \text{for} \quad \vartheta = 10. \end{cases}$$

One such criterion is the minimax principle: minimize $\max(a\alpha, b\beta)$. It is easily verified that the minimax test for $a = b$ (rejecting an intact and accepting a defect are considered equally serious) is the one with $K_3=0$. This is Test 3 in Case 5 - reasonable, isn't it?

Test 3 in Case 4 is the minimax solution corresponding to the rather extreme choice $a/b = 1.4\cdot 10^{-15}$. It is practically the same as Test 2, and it is now seen under which circumstances the classical statistician would tell the experimenter to as well reject in the seldom case where the inaccurate instrument is chosen: that is when the false rejections cost practically nothing.

Another criterion consists in minimizing the weighted average risk,

$$\pi a\alpha + (1-\pi)b\beta \ ,$$

which leads to Test 3 with

$$K_3 = \frac{1}{10} \ln\left[\frac{\pi a}{(1-\pi)b}\right] \ . \qquad (7)$$

see e.g. Berger (1980). Mathematically this is just the Bayes solution with prior distribution $(\pi, 1-\pi)$ on (H_0, H_1). To the classical statistician the prior is simply a pair of weights attached to the two hypotheses in order to summarize the properties of the test in one single scalar quantity, which can be optimized. To the Bayesian the prior expresses prior beliefs. They may - and probably will - argue endlessly about their different interpretations of the prior, but they will agree upon which test to use if they happen to choose the same π. Suppose that $p = 0.99$, $a = b$, and they both pick $\pi = 7.38 \cdot 10^{-15}$ (again quite extreme). Then they would end up with Test 3 in Case 4 (in practice the same as Test 2) and unanimously tell the experimenter who is stuck with using the inaccurate instrument, to forget about X and reject H_0.

The point is that the Bayes solutions coincide with the optimal classical solutions (varying π in (7) generates all K_3 between $-\infty$ and $+\infty$), and so any criticism of the limiting Test 2 hits Bayes just as severely as it hits Classical.

By way of conclusion, the criticism based on the present example is easily countered. In fact, the counter-argument - the equivalence between Bayes solutions and admissible solutions in the present case - is easily compiled from Berger (1980).

3. To believe or not to belive - that is the question

Basic in any statistical model is the pdf of the data X for a fixed state ϑ of the nature, the so-called likelihood function,

$$f_\vartheta(x), \qquad x \in \mathcal{X}, \quad \vartheta \in \mathcal{T}. \tag{8}$$

The classical view of f as a long term frequency rests on the objectivistic conception of the world as a system governed by general laws, uniformity and reproducability. Bayesian thinking appears less standardized at this point. It seems that some Bayesians are willing to accept the frequency interpretation, some refuse it and refer to ideas of symmetry and the like, and yet some are not so specific and just take f for granted. However, even though different notions lie dormant in puristic minds, the likelihood is not a point at issue in the debate between Classical and Bayes: in practice there will be agreement as to the functional form of f.

The controversy turns on the legitimacy of the prior distribution. The classical model is given by (8) alone, hence the data X and its relation to ϑ through the likelihood is the only basis for inference. The Bayesian model extends the classical one by including also an a priori pdf,

$$g(\vartheta), \qquad \vartheta \in \mathcal{T}, \tag{9}$$

accommodating knowledge/beliefs prior to and independent of the data. In the extended model (8) - (9) the parameter becomes a random variable Θ, and inference about it is based on the joint pdf

$$f_{\vartheta}(x)\ g(\vartheta), \qquad x \in \mathcal{X},\ \ \vartheta \in \mathcal{T}. \tag{10}$$

As regards the mathematics, the difference between the two approaches is that classical statisticians confine themselves to conditional inference, given $\Theta = \vartheta$. The resulting methods will, of course, remain valid/correct in the extended model, but will in general not be optimal there since they disregard the information contained in the marginal distribution (9). Thus, from a Bayesian point of view, the classical statisticians are not on the wrong scent, but they commit an error of default.

This objection is valid only to those who allow subjective beliefs to be explicitly moulded into the model. To a classical statistician it is void since she insists that the inference be based exclusively on the objective facts (and indeed holds X and (8) to be objective entities). One cannot talk a non-believer into some beliefs on the grounds that believers draw stronger conclusions.

In a scientific context it is clearly seen that classical and Bayes are rooted in different philosophical positions. Classical statistics, based on a frequency interpretation of probability, fits perfectly into the widely accepted objectivistic paradigm of natural science, whereas Bayes conflicts with it. It is not obvious to me how perfectly Bayes accords with the subjectivist school within the social sciences, but it is clear that Classical conflicts with it. At any rate, the dividing line between objectivism and subjectivism extends far beyond the area of statistics, and I think statisticians should not invest too much energy in trying to settle today a dispute that will prevail as long as intellectuals inhabit the earth (hopefully, yet some few years). What they can do with their mathematical models is to set a surveyable stage for the discussion.

In the non-scientific context the problems are of a different nature. Consider a clear-cut decision problem which requires action to be taken here and now on basis of the information you have - you cannot lean back in your arm-chair awaiting for evidence to accrue. For instance, you are an actuary in an insurance company, and one day you are asked to fix a premium for the insurance of a stamp collection. You order an outprint of the risk statistics for stamp collections. It turns out there is no such statistics because this stamp collection is the first one to be insured. You will be out of business if you tell the owner of the collection to come back in some twenty years when the company has got some statistics. What you do is, of course, to make a skilled guess (I say "of course" since I am sure you must agree). If you should formalize your subjective assessment, then you would make Bayesian statistics. You might now object that you would not do that because you are not a Bayesian (or would not like to be one). Then I would assert that the difference between you and a Bayesian is not that you abstain from subjective judgements - you make them, all people do all the time, more or less consciously. The difference is only that you are unwilling to quantify your subjective opinions - you abstain from formulating mathematically something that is real to you and play an important role in your practical life. Personally, I consider it a great advantage that the elements of subjective assessments, which are inevitable in some situations, can be laid open and subjected to discussion by means of the Bayes formalism.

Technically speaking, the situation described above is one were no X is at hand and a likelihood (8) cannot be made part of the analysis. The question is whether you are willing to perform an analysis at all. If yes, your only possibility to do so is to pick a prior (9).

Suppose now instead that the company had a substantial portfolio of insured stamp collections and you were supplied with ample statistics. Then you would, perhaps, just proceed in the standard classical way, confident that the data will speak for themselves. In that case you must ask yourself: is this classical attitude of yours ad hoc, determined by the incidental circumstance that (8) is now very informative and you, therefore, can afford to forget about your a priori insights? If you found it necessary to rely on (9) in the former case with no observations, when did you switch to a different philosophy? What would you do if you had only some scanty data, not very informative? Would you then combine the elements (8) and (9) and perform an analysis based on (10)?

Here is my own position. Faced with a typical decision situation with no data I have to rely on my subjective judgement. I find it perfectly meaningful to formalize my deliberations in the Bayesian framework and, further, think this is the most honest thing to do - play with open cards. I would stick to this philosophy also in the presence of "objective" information. As the amount of data increases, the relative influence of the prior diminishes, and in the end I can dispence with it and join the classical statisticians in a study of the likelihood alone. (Our interpretations of θ and ϑ, respectively, will still be different, but our practical conclusions will be the same).

Whether to use a prior or not is a philosophical question, depending to some extent on the nature of the statistical problem. It has to be decided prior to the choice of the mathematical framework - prior to specifying or not specifying a prior - and can certainly not be settled by excercises of the kind discussed in the previous section. It is highly desirable that statisticians reflect on their philosophical positions and, in particular, seek to clarify the thought contents of Classical and Bayes. I am convinced that both are useful in their respective fields of application. At any rate, classical statisticians and Bayesians should rejoice at the fact that probability theory has proved able to serve both parties well in their attempts to express their ideas in precise mathematical terms.

Finally, I will mention the empirical Bayes scheme because it - mistakingly, I think - is held by some statisticians to be a synthesis of Classical and Bayes. Robbins (1955, 1964), who coined the term "empirical Bayes", had in mind situations where the same decision problem presents itself repeatedly and independently. Let the problems be labeled by i = 1,2,... To each problem i there is associated an unobservable "interest parameter" θ_i and an observable X_i. The pairs (θ_i, X_i), i = 1,2,..., are taken as i.i.d. random elements with common distribution given by (8) - (10). Thus, the decision problems are independent in the stochastic sense, but they are still related since all θ_i stem from the same distribution. Such an assumption is appropriate when θ_i characterizes the i-th selection from a heterogeneous population of units that are basically of the same kind, but not identical. A typical example is acceptance control of batches of items from a certain manufacturing process: X_i is the proportion of defectives in a sample from batch No. i (the "quality" of the sample), θ_i is the proportion of defectives in the batch (the "quality" of the batch), and g represents the "quality" of the manufacturing process itself. Suppose I batches have been controlled so far and we are to assess the quality of batch No. I+1. The former observations X_i, i = 1,...,I, give partial information on the corresponding θ_i-values and, thereby, on their common density g, which generated the current θ_{I+1}. This piece of information can be combined with the current observation X_{I+1} to give an improved estimate of θ_{I+1}. As I increases, g will typically be consistently estimated, and in the limit g and the Bayes estimate of θ_{I+1} based on X_{I+1} will become known. Then the mathematics of the current estimation problem will coincide with that of a genuine Bayesian analysis with prior density g.

One might say that the empirical Bayes model framework takes care of those situations where the a priori insights stem from previous collateral experiments. The model is, however, entirely classical since all probabilities are given a frequency interpretation. It may be helpful to note that the empirical Bayes model framework includes as special cases the random effects models in the analysis of variance. What is peculiar to the empirical Bayes formulation is not the model, but the purpose of the analysis: empirical Bayes centres on estimating each latent θ_i, whereas ANOVA centres on estimating the distribution, viz. the mean and the variance components.

I conclude that the empirical Bayes set-up does not abolish the controversy Classical vs. Bayes. This controversy presents itself here as in every other statistical inference problem. Classical statisticians will base their analyses entirely on the above model, whereas Bayesians will extend the model by introducing a prior distribution on the space of densities g. Again your attitudes can be introspectively examined by considering the cases with (i) no previous experiments $(I = 0)$, (ii) a large number of previous experiments $(I \rightarrow \infty)$, (iii) some, but not overwhelmingly many previous experiments (the typical intermediate case).

Further discussions of the aspects treated here are found e.g. in Deeley and Lindley (1981) and Norberg (1979).

References

Berger, J.O. (1980). Statistical Decision Theory. Springer Verlag, New York. (Revised and extended 1985.)

Cornfield, J. (1969). The Bayesian outlook and its application. Biometrics 25, 617-657. (With discussion.)

Cox, D.R. (1958). Some problems connected with statistical inference. Ann. Math. Statist. 29, 947-971.

Deeley, J.J. and Lindley, D.V. (1981). Bayes empirical Bayes. J. Amer. Statist. Association 76, 833 - 841.

Norberg, R. (1979). The credibility approach to experience rating. Scand. Actuarial J. 1979, 181 - 221.

Robbins, H. (1955). An empirical Bayes appoach to statistics. Proc. Third. Berkeley Symposium on Math. Stat. and Prob., Vol.1, 157 - 163. University of California Press.

Robbins, H. (1964). The empirical Bayes approach to statistical problems. Annals of Math. Statist. 35, 1 - 20.

Sandved, E. (1987). Some critical remarks to the two books
James O. Berger: Statistical Decision Theory 1980,
James O. Berger: Statistical Decision Theory and Bayesian

Analysis 1985. Statist. Research Report No.2, 1987, Inst. of Math., Univ. of Oslo.

Sverdrup, E. (1966). The present state of the decision theory and the Neyman-Pearson theory. Rev. Internat. Statist. Inst. 34, 309-333.

Appendix: reply from James Berger

In a letter of March 15th, 1988, Berger writes:

Dear Professor Norberg:

Thank you for the copy of your comments on my and Sandved's criticisms. I am enclosing a copy of a letter I sent to Sandved in response.

I think it is important to distinguish between particular procedures and general methodologies. Frequentists have become good at appending a variety of adhoc principles (ancillarity being one of the first) to the basic frequentist position to try to prevent absurdities. None of these additional principles works well all of the time, however. More troubling is that there is no way of knowing when these additional principles will or will not work. The incredible complexity that a frequentist gets involved with in trying to produce flexible enough versions of conditional analysis is illustrated very well by the articles of Kiefer on the subject (referred to in my book).

In contrast, for a Bayesian the only question as to whether he has done a good analysis is - has he used a reasonable prior? Actually, Bayesians tend to work towards analyses or presentations valid for a variety of reasonable priors. But, in any case, anyone can judge their acceptance of the result by examining the prior (and model, etc.) or choosing their own. When I see a frequentist analysis I have no basis for judging it, since it takes deep investigation to judge what kind of conditioning is necessary and possible in a given case. Your analyses provide good illustrations of the point. It is simply much harder to consider and evaluate all the options from the frequentist side.

Note, by the way, that I feel there are many examples where no sensible frequentist analysis exists. I refer to a couple in my letter to Sandved. Others occur in sequential settings (see, e.g., the enclosed paper).

I question your claim on p. 17 that classical statistics fits perfectly into the objectivistic paradigm of natural science. Bayesians have claimed all along that frequentist methods are also subjective, and insidiously so. The sequential and "objectivity" papers enclosed reflect some of these arguments.

The rest of your paper I, of course, wholeheartedly agree with. Note that I too see uses for frequentist methods, as described in the second edition of my book. These uses have to do with approximating or selecting among Bayesian analyses, however.

Thank you for your paper. Your analyses of the two examples are definitely the most interesting I have seen.

Best wishes,

Jim Berger

James Berger

In handwriting he adds:

"I have no complaint with your representation of my views, except possibly that the intent of my "conditioning" examples seems to have been erroneously percieved as themselves arguing for Bayesian analysis. My original intent was simply to use them to show that some type of conditional (or likelihood) analysis is necessary. ... It is the purpose of the rest of the book (not these examples) to argue that the Bayesian approach to conditioning is best. I will admit to not having been clear on this, however."