SOME THOUGHTS ON STATISTICAL INFERENCE[1]

By E. S. Pearson

University College, London

1. Introduction. A few weeks ago, before leaving England, I found some notes of various talks which I had given on a visit to the United States paid 30 years ago. In a lecture which I delivered here, at Cornell, in early May 1931 I seem to have used some words which it is perhaps rather bold of me to quote today before a Meeting of the Institute of Mathematical Statistics. Still, I will do it because I suppose that after all I am the same kind of person now as I was then! I used words like these: "I sometimes think that Statistics is becoming far too mathematical, and that it is a relief to turn to the many simple, unsolved problems which can be discussed in terms only of means and standard deviations."

It is evident from the context that the problems I was thinking of were concerned with what I might call the philosophy of statistical inference, whose principles and relationships can often be discussed most clearly in terms of simple situations. When I was here in 1931 the work of Neyman and myself was in an early stage; we spoke of the class of admissible alternative hypotheses and we were deriving tests using the likelihood ratio principle. But the idea of the power function and of the uniformly most powerful test was still in embryo, coming to birth at meetings contrived here or there in Europe or in correspondence carried on between Warsaw and London.

I must confess that the older I get, the more difficult I find it to be positive in this matter of statistical inference, but I have felt that as you have invited me to address you here on what is nearly the 30th anniversary of an earlier visit, I should try to formulate some of my thoughts on the relation between the Neyman-Pearson theory and fresh views on inference that are current today. I do this the more readily because I believe rather strongly in the value of emphasising continuity as well as differences in statistical philosophy. I am convinced that if we can only get to the bottom of the way in which similar situations are tackled by different approaches, all I believe lying within the broad path of development of our subject, our understanding will gain in richness—gain in a way which can never happen if we waste energy in trying to establish that we are right and the other fellow is wrong!

2. Some historical reflections on the development of the Neyman-Pearson theory. Allow me therefore to start with a few historical remarks. There is perhaps in current literature a tendency to speak of the Neyman-Pearson contributions as some static system, rather than as part of the historical process of development of thought on statistical theory which is and will always go on.

Received July 28, 1961; revised December 12, 1961.

[1] This article contains the substance of an invited paper read before the Regional Meeting of the Institute of Mathematical Statistics held at Cornell University, April 21, 1961.

Neyman and Pearson were after all very much persons of their time. They built on things which they found in the middle 1920's:

(a) The way of thinking which had found acceptance for a number of years among practising statisticians, which included the use of tail areas of the distributions of test statistics.

(b) The classical tradition that, somehow, prior probabilities should be introduced numerically into a solution—a tradition which can certainly be traced in the writings of Karl Pearson and of Student, but to which perhaps only lip service was then being paid.

(c) The tremendous impact of R. A. Fisher. His criticism of Bayes' Theorem and his use of Likelihood.

(d) His geometrical representation in multiple space, out of which readily came the concept of alternative critical regions in a sample space.

(e) His tables of 5 and 1% significance levels, which lent themselves to the idea of choice, in advance of experiment, of the risk of the "first kind of error" which the experimenter was prepared to take.

(f) His emphasis on the importance of planning an experiment, which led naturally to the examination of the power function, both in choosing the size of sample so as to enable worthwhile results to be achieved, and in determining the most appropriate test.

(g) Then, too, there were a number of common-sense contributions from that great practising statistician, Student, some in correspondence, some in personal discussion.

What Neyman and I experienced, as no doubt do the exponents of any new line of thought on inference, was a dissatisfaction with the logical basis—or lack of it—which seemed to underlie the choice and construction of statistical tests. We found this not only in the theoretical work of what was then called the Biometric School, but also in some of R. A. Fisher's writing, in so far as we could follow its underlying philosophy. We tried therefore to develop a set of principles having a mathematical basis which it seemed to us led to a rational choice of statistical procedures when faced with certain types of problem in the analysis and interpretation of data. Put in another way, we were seeking how to bring probability theory into gear with the way we think as rational human beings. No doubt because the scope of application of statistical methods was much narrower in those days, the emphasis which we gave to certain types of situation may now seem out of balance.

We were certainly aware that inferences must make use of prior information and that decisions must take account of utilities, but after some considerable thought and discussion round these matters we came to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities, that our line of approach must proceed otherwise.[2] Thus we came down

[2] This is perhaps the central problem over which opinions differ. In setting down my thoughts on some of the difficulties to be faced my purpose is not to nail a flag to any mast, but to encourage discussion which may in the end lead to a clearing up of certain dusty corners of our minds.

on the side of using only probability measures which could be related to relative frequency. Of necessity, as it seemed to us, we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters—to use our terminology—as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities.

We also considered how far inferences and decisions could be based on the values of likelihood ratios and we first obtained for the critical or rejection regions, those bounded by contours in the sample space on which the appropriate likelihood ratio was constant. But looking back I think it is clear why we regarded the integral of probability density within (or beyond) a contour as more meaningful than the likelihood ratio—more readily brought into gear with the particular process of reasoning we followed.

The reason was this. We were regarding the ideal statistical procedure as one in which preliminary planning and subsequent interpretation were closely linked together—formed part of a single whole. It was in this connection that integrals over regions of the sample space were required. Certainly, we were much less interested in dealing with situations where the data are thrown at the statistician and he is asked to draw a conclusion. I have the impression that there is here a point which is often overlooked; I will come back to this in the example which I propose to discuss shortly.

3. The subjectivist approach. As I have said, these choices of Neyman and myself were deliberate, although at that time the issues may not have been as clearly before us as they are presented today. The up to date subjectivist or Bayesian considers that this was the wrong choice. He believes that unless the statistician attempts to express his notions of prior probability and his utility functions in a form which can be inserted into a mathematical mechanism, geared with his way of thought, he is falling down on his job. The ideas of the Bayesian are not of course new; what is new I think is the more precise formulation of the theory in mathematical terms and its application to a much wider range of situations than the 19th century users of inverse probability methods could have dreamed of.

If I am asked how I regard the views of writers on subjective probability, my answer is this: the approach of Professor Savage and others strikes me as extremely illuminating in a variety of ways and I certainly welcome further exploration along these lines. At the same time I must admit that there are some fundamental parts of the mechanism of subjective probability theory which simply will not at present get into gear with the way *I* think any more than they did 30 years ago. May be this is because I am getting old and have settled into a certain routine of thought—or may be I have some justification for an instinctive hunch that some things cannot always work. I do not pretend to know the answer.

Let me however illustrate some of my difficulties very briefly.

(a) We are told that "if one is being consistent, there is a prior distribution".

"A subjectivist feels that the prior distribution means something about the state of his mind and that he can discover it by introspection". But does this mean that if introspection fails to produce for me a stable and meaningful prior distribution which can be expressed in terms of numbers, I must give up the use of statistical method?

(b) Again, it is an attractive hypothesis that Bayesian probabilities "only differ between individuals because individuals are differently informed; but with common knowledge we have common Bayesian probabilities". Of course it is possible to define conceptual Bayesian probabilities and the "rational man" in this way, but how to establish that all this bears a close relation to reality?

It seems to me that in many situations, if I received no more relevant knowledge in the interval and could forget the figures I had produced before, I might quote at intervals widely different Bayesian probabilities for the same set of states, simply because I should be attempting what would be for me impossible and resorting to guesswork. It is difficult to see how the matter could be put to experimental test. Of course the range of problems is very great. At one end we have the case where a prior distribution can be closely related to past observation; at the other, it has to be determined almost entirely by introspection or (because we do not trust our introspection) by the introduction of some formal mathematical function, in Jeffreys' manner, to get the model started. In the same way utility and loss functions have sometimes a clear objective foundation, but must sometimes be formulated on a purely subjectivist basis.

To have a unified mathematical model of the mind's way of working in all these varied situations is certainly intellectually attractive. But is it always meaningful? I think that there is always this question at the back of my mind: can it really lead to my own clear thinking to put at the very foundation of the mathematical structure used in acquiring knowledge, functions about whose form I have often such imprecise ideas?

4. The problem of King Hiero's crown. To make these reflections more concrete I will try to illustrate both the illumination and some of the difficulties of the subjectivist approach as they strike me, on an example, the broad lines of which originate from Professor L. J. Savage, who introduced it during a two-day discussion meeting at Birkbeck College, London, nearly two years ago.[3] The example, whose scope I have somewhat enlarged, though no doubt expressed in rather simplified terms seems to me to represent a type of situation which is not altogether unusual.

Savage has called it the problem of *King Hiero's Crown*. Briefly, the legend as brought up to date is this:

(a) King Hiero has ordered a new crown and he believes that the goldsmiths may have adulterated the gold, either with lead or with silver.

(b) Archimedes has hit on the idea (presumably unknown to the goldsmiths) of

[3] Professor Savage has been kind enough to welcome the use of this example here before the publication of his own talk in London.

determining the density of the crown by weighing it and a specimen of pure gold in air and in water.

(c) By this test, Archimedes is estimating a quantity λ by means of a measure x (which may be the mean of n independent test results, X_i).

(d) For pure gold $\lambda = 0$, for lead $\lambda > 0$, for silver $\lambda < 0$.

(e) Archimedes has found by earlier experiment that from weighing to weighing x will vary normally about λ with known standard error σ. (σ may equal $\Sigma/\sqrt{n}$ where Σ is the standard error of a single observation).

The King attaches some credence to the possibility that there is no cheating ($\lambda = 0$), and associates this with a prior probability I. $\bar{I} = 1 - I$ is the prior probability of cheating, and the prior distribution of λ, conditional on cheating, is $\pi(\lambda)$. If I' and $\bar{I}'$ are the posterior probabilities of no cheating and cheating, respectively, then it may be shown that

$$\frac{\bar{I}'}{I'} = \frac{\bar{I}}{I}\frac{\sigma}{\phi(x/\sigma)} \cdot \int_{-\infty}^{\infty} \pi(\lambda)\phi\left(\frac{x-\lambda}{\sigma}\right)\frac{d\lambda}{\sigma}, \tag{1}$$

where ϕ is the standardised Normal probability distribution function.

If $\pi(\lambda)$ is nearly uniform over a sufficiently long range having regard to σ, then (1) becomes approximately

$$\frac{\bar{I}'}{I'} = \frac{\bar{I}}{I}\frac{\sigma}{\phi(x/\sigma)}\pi(x). \tag{2}$$

Notice that in this model no attempt is made to introduce the degree of guilt nor to balance the utility of hanging innocent goldsmiths against allowing guilty ones to go free. As Savage has pointed out to me, it is quite possible and elegant mathematically to introduce a utility function into the problem. But I think that the example, without this complication, represents a common type of problem in which the consequences of the two different kinds of mistaken conclusions are incommensurable in terms of any readily acceptable numbers. I had thought it likely that Hiero would decide not to execute unless the odds against innocence were high, perhaps 25 to 1, or 10 to 1. But Professor Savage points out to me that a likely royal view in Hiero's days would be that the goldsmiths should be hanged unless the odds on their innocence were very high!

5. Numerical illustration of the theory. In the calculations illustrated by Figures 1 and 2, I have first taken six different prior distributions. For Cases 1, 2 and 3, $\bar{I}/I = 4{:}1$ and for Cases 4, 5 and 6, $\bar{I}/I = 1{:}1$. Two values for σ have been taken:

$\sigma = 0.25$, which might correspond, say, to $n = 4$, $\Sigma = 0.5$

$\sigma = 0.10$, corresponding to $n = 25$, $\Sigma = 0.5$.

For these values of σ if we explore only the case where $\lambda \geqq 0$ it is not necessary to specify the form of $\pi(\lambda)$ outside a certain section of the λ-scale, but we can

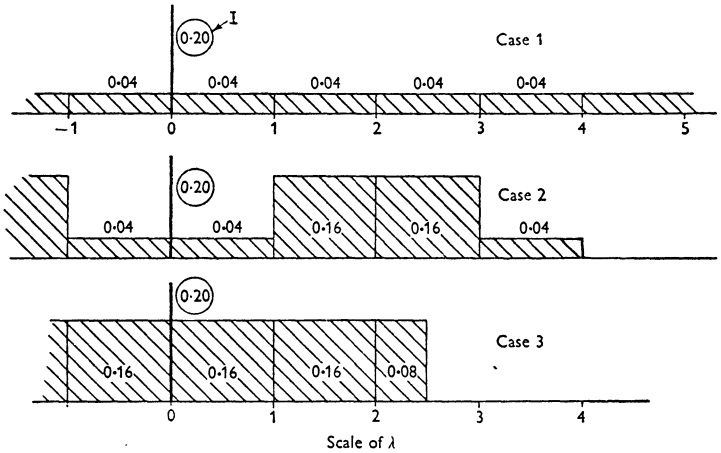

FIG. 1. Prior distribution of λ for Cases 1, 2 and 3.

NOTE: the numbers associated with the shaded blocks are the integrals of $\bar{I}\pi(\lambda)$ for unit (or in one case half-unit) intervals of λ.

if we like suppose $\pi(\lambda)$ to be symmetrical about $\lambda = 0$. *Within the range shown,*[4] it was supposed that:

$$\begin{aligned}
&\text{for Cases 1 and 4,} && \pi(\lambda) = 0.05 && \text{for} && -1 \leqq \lambda \leqq 4,\\
&\text{2 and 5,} && \pi(\lambda) = 0.05 && \text{for} && -1 \leqq \lambda \leqq 1\\
& && = 0.20 && \text{for} && 1 < \lambda \leqq 3\\
& && = 0.05 && \text{for} && 3 < \lambda \leqq 4,\\
&\text{3 and 6,} && \pi(\lambda) = 0.20 && \text{for} && -1 \leqq \lambda \leqq 2.5\\
& && = 0 && \text{for} && \lambda > 2.5.
\end{aligned}$$

Cases 1 and 4 correspond to a situation in which the King's opinion about the extent of cheating, if it has occurred, is very "diffuse". Thus it could be that $\pi(\lambda) = 0.05$ for $-10 \leqq \lambda \leqq 10$.

Cases 2 and 5, might represent the position if the King argued as follows: On the one hand the goldsmiths will not risk including so much base metal that it might be obvious; on the other it will hardly seem to them worthwhile adulterating the gold to only a small extent.

Cases 3 and 6 use an intermediate form for $\pi(\lambda)$.

The crude step functions were of course introduced to simplify the calculations. Figure 1 illustrates the prior distributions for Case 1–3 only: the value of $I = 0.2$ is given and also the integral of $\bar{I}\pi(\lambda) = 0.8\,\pi(\lambda)$ over unit (or $\frac{1}{2}$ unit) intervals of λ.

[4] It will be noticed that I have not committed myself to the form of $\pi(\lambda)$ outside the range of λ needed in my discussion. This explains, e.g., why I have not defined $\pi(\lambda)$ for $\lambda < -1$.

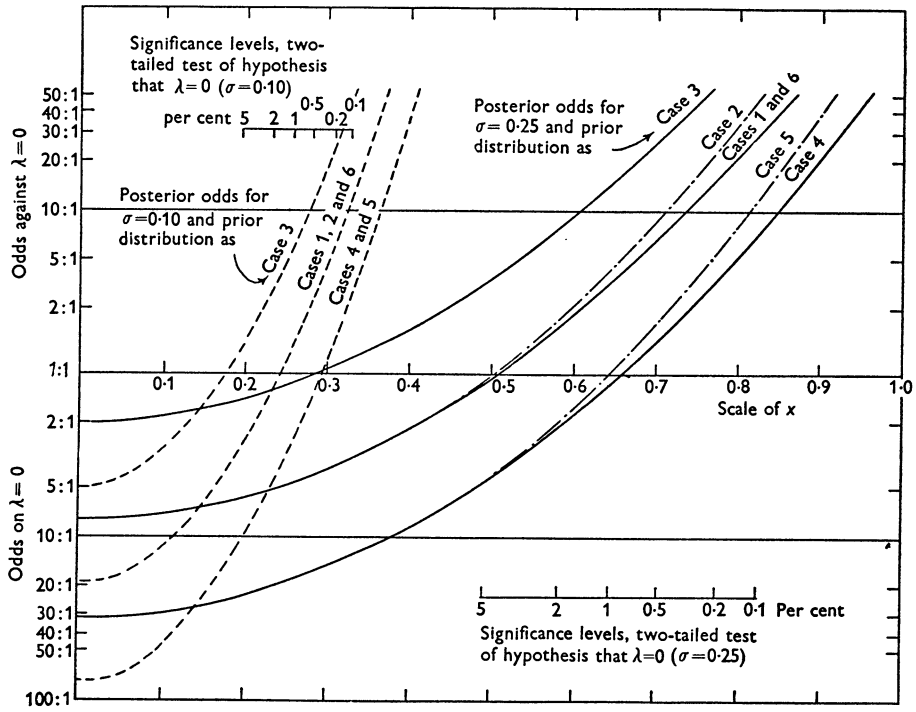

FIG. 2. Curves for posterior odds plotted against x, the estimator of λ.

In Figure 2 are shown the posterior odds derived from equation (1), or, where adequate, using equation (2). Also included are scales showing significance levels for a two-tailed test of the hypothesis that $\lambda = 0$; e.g., for $\sigma = 0.25$ the 5% levels of the two-tailed test fall at $x = \pm 0.25 \times 1.96 = \pm 0.49$. The values of $\bar{I}'/I'$ for Case 1 and 6 are the same, since within the range of x considered equation (2) is appropriate and $\bar{I}\pi(x)/I$ has the same value for both cases.

Accepting equations (1) and (2) as meaningful, consider a few of the points brought out by Figure 2:

(a) If $\sigma = 0.25$ and the King is of opinion that the posterior odds on guilt should be at least 10 to 1 before he hangs the goldsmiths, the critical value for x will fall at $x_c = 0.61$ for Case 3 (corresponding to the 1.5% significance level to the null hypothesis test) and at $x_c = 0.85$ (corresponding to the 0.07% significance level) for Case 4. In other words, if as in Case 4 the King believes that the goldsmiths are as likely to cheat as not ($I = \bar{I}$) and that if they cheat the odds are 9 to 1 that $|\lambda| > 1$, he can afford to put the critical x-level much further out than in Case 3.

(b) On the other hand, he may have an entirely different outlook on what is just and on the value of human life where a King's position is concerned. He may, therefore, decide to hang unless the posterior odds are at least 10 to 1 in favour of innocence and so will always hang in Cases 1, 2, 3 and 6 and will hang

unless $x < x_c = 0.38$ in Cases 4 and 5. Conventional significance levels have here no bearing on his decision.

(c) With $n = 25$, $\sigma = 0.10$ the relation between posterior odds and the scale of significance levels is clearly not the same as for $n = 4$, $\sigma = 0.25$.

Approached in this way there seems to me no doubt that the lessons to be drawn from results such as these, suggested by Figure 2, are illuminating because on certain assumptions they give precision to the way in which a rational man will react to the information he possesses and the objects he has in view. It is, however, clear that the critical level is very sensitive to the prior distribution adopted and also, of course to the King's opinion on the relative importance of punishing the guilty and hanging the innocent. Does this mean that a model which is clarifying in theory may in practice be impossible to use because it calls for the introduction of parameters whose values do not really exist, or is the lesson that this method of approach is of value just because it forces the King to face up to issues which he would otherwise have failed to appreciate fully?

6. The relation between preliminary planning and subsequent behavior. So far it has been supposed that the legend starts from the point where Hiero has views on I and $\pi(\lambda)$, is given x and has to decide whether to act as though the goldsmiths were innocent or guilty; this was the form in which Professor Savage originally stated the problem in the context in which the analysis aspect of statistics, rather than the unified problems of design and analysis, was the centre of the discussion. But in so far as the ideal statistical situation is one in which preliminary planning and subsequent interpretation of the results of an experiment are closely linked together, it seems useful (and with this Professor Savage fully agrees) to look at the legend from a slightly different point of view. If it is granted that Archimedes has a scientific approach, we may suppose that he and King Hiero will have thought round their method of testing the goldsmiths *before* x is known to them. They must, indeed, do some preliminary thinking, for if they have not pretty clear views on the values they will give to I and $\pi(\lambda)$ before x is known, they will find it hard to be unprejudiced in assigning values for these expressions afterwards. Thus it is likely that a survey of possibilities somewhat of the kind presented in my diagrams would have been carried out, perhaps while the goldsmiths were putting the finishing touches to the crown.

In this survey, we can imagine Hiero remarking somewhat as follows:

"You say that if we make four weighings of the crown ($\sigma = 0.25$) and if we agree on values for I and $\pi(\lambda)$ as, say, in Case 2, then the odds will be at least 10 to 1 against innocence if $x > 0.71$. So perhaps we might fix this as the critical value for hanging. But tell me, Archimedes, if we do take 0.71 what is the probability that (i) we shall hang innocent men, (ii) shall let off guilty men when they have actually adulterated the gold to the extent, say, of $\lambda = 1, 0.75, 0.5$?"

Archimedes of course might answer:

"Your Majesty, you have accepted a prior distribution $\pi(\lambda)$ which makes

a value of $|\lambda| < 1$ most improbable; you should not therefore ask this question."

To which the King might reply:

"Pray give me the answer which I asked for Archimedes. I am far too uncertain whether the particular prior distribution which we sketched out has any sure justification behind it."

Archimedes must therefore consider the operating characteristics or power function of the rule suggested. He will find in answer to Hiero's questions that if $\sigma = 0.25$,

(i) $$\Pr\{|x| > 0.71 \mid \lambda = 0\} = 0.0045,$$

(ii) $$\Pr\{x < 0.71 \mid \lambda = 0.5\} = 0.80$$

$$\Pr\{x < 0.71 \mid \lambda = 0.75\} = 0.44$$

$$\Pr\{x < 0.71 \mid \lambda = 1.0\} = 0.12.$$

This position the King may consider to be entirely unsatisfactory because of the large chance of failing to detect an amount of adulteration which he considers to be highly criminal.[5] He is neither prepared to rely on any prior distribution for λ nor does Archimedes' suggestion to formulate a value function appeal to him because he doubts whether this, too, would stand a critical scrutiny. So perhaps he will accept more readily a different specification of his wishes.

If Archimedes were to tell him that by making 25 rather than four measurements ($\sigma = 0.10$) and setting the critical value x_c at 0.25,

(a) the probability of hanging innocent men is about 0.01

(b) the probability of letting off men who have taken $\lambda = 0.5$ has the same value,

he might say:

"I realise, Archimedes, that the figures which you have put into this statement are somewhat arbitrary, but in my opinion they provide a solution which I can understand and accept as reasonable."

This reply of Hiero might be described as that of a man who after much thought finds one kind of arbitrary choice more meaningful than another. He would in short have come to the conclusion that it would be easier to specify his wishes in terms of an "indifference value" for λ at, say 0.25 and of two risks, (possibly but not necessarily equal) associated with wrong decisions when $\lambda = 0$ and $\lambda = 0.5$.

7. Conclusion. Having got this far I have reached a point where I am told by my Bayesian friends that we must try "to see as clearly as we can by reflexion and

[5] Presumably Hiero would regard *any* adulteration as criminal, but he might feel that by including only a small amount of silver or lead the men would have gained no profit adequate to outweigh their continual fear of later discovery.

introspection where the appeal of such a solution lies".[6] This is good counsel to which I have no objection; indeed my main purpose in venturing on this talk was to come to such a point where it was agreed that there were difficulties of many kinds which should be discussed jointly and dispassionately. But clearly I have had my say for today and must go no further now. Time might have counted for Hiero and Archmiedes too, and they, as well as working statisticians of today, might have had to call off further philosophical discussion and adopt a solution which was intelligible to them and in their judgment reasonable.

[6] In preparing this paper for publication, I quote the words of a very fair-minded referee.