RECOVERY OF ANCILLARY INFORMATION*

By D. BASU

Indian Statistical Institute

1. Introduction

The main upsurge of late Professor R. A. Fisher's theory of Statistical Inference took place within a brief span of about 10 years (1920-30) after the first world war. It was during this period that Fisher came out with the brilliant and now famous notions of (a) likelihood, (b) fiducial probability, (c) information and intrinsic accuracy, (d) sufficiency and (e) ancillary statistics and recovery of information — concepts around which the superstructure of the theory is built.

Many eminent statisticians and mathematicians have made detailed localised studies of some particular aspect of Fisher's theory and some of these studies gave rise to important streams of fundamental research in statistical theory. The author (Basu, 1959) made a very much localised study of the notion of ancillary statistics from the purely mathematical point of view. This note is a follow up study of the earlier paper (Basu, 1959) from the statistical angle. Here we discuss the very controversial subject matter of 'recovery of ancillary information' through proper choice of 'reference sets.' For the purpose of pinpointing our attention to the basic issues raised, we restrict ourselves to the one-parameter set-up only.

In the one parameter set-up Fisher defines an ancillary statistic as one whose probability (sampling) distribution is free of the parameter and which, in conjunction with the maximum likelihood estimator of θ (the parameter), is sufficient. The use of ancillary statistics has been recommended in two different inference situations namely the point estimation problems and the testing of hypotheses problems.

In point estimation problems the use of a suitably chosen ancillary statistic is recommended in the situations where the maximum likelihood estimator T of θ is not a sufficient statistic. The use of T as an estimator of θ will then entail a certain loss of information which, according to Fisher, may be meaningfully (at least in the large sample case) measured, in some situations, as follows. The information contained in the whole sample X is defined as

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta) \,\middle|\, \theta \right]$$

where $f(x \mid \theta)$ is the frequency or density function for X. Similarly the information contained in a statistic T (which may be vector-valued) is measured by the function

$$J(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log g(T \mid \theta) \,\middle|\, \theta \right]$$

* The paper has been included in *Contributions to Statistics*, Presented to Professor P. C. Mahalanobis on the occasion of his 70th birthday.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_19,

where $g(t \mid \theta)$ is the frequency or density function for the estimator T. The difference

$$\lambda(\theta) = I(\theta) - J(\theta)$$

may then be taken as a measure of the information lost. Under certain regularity conditions the following results may be proved :

(1) $\lambda(\theta) \geqslant 0$ for all values of θ;

(2) $\lambda(\theta) \equiv 0$ if and only if T is a sufficient statistic;

(3) if T together with the statistic Y be sufficient for θ then the information contained in the pair (T, Y) is $I(\theta)$;

(4) if Y be ancillary and the pair (T, Y) is sufficient then

$$I(\theta) = E[J(\theta \mid Y) \mid \theta]$$

where $J(\theta \mid y)$ is the conditional amount of information contained in T under the condition that Y takes the value y, i.e.

$$J(\theta \mid y) = E\Big[-\frac{\partial^2}{\partial \theta^2} \log f(T \mid y, \theta) \mid Y = y, \theta \Big]$$

where $f(t \mid y, \theta)$ is the conditional frequency (density) function for T under the condition $Y = y$. The relation (4) follows directly from the observation that the joint frequency (density) function $h(t, y \mid \theta)$ of (T, Y) may be factorized as

$$h(t, y \mid \theta) = g(y) f(t \mid y, \theta)$$

where $g(y)$ is the θ-free (Y being ancillary) marginal frequency (density) function for Y.

Now, consider a situation where the maximum likelihood estimator T is not sufficient but where we are able to find another statistic Y whose marginal distribution is θ-free and which complements T in the sense that the pair (T, Y) is jointly sufficient. The statistic Y by itself contains no information about θ. But in a sense it 'summarises in itself' the quantum of information $\lambda(\theta)$ that is lost in the use of T as an estimator of θ. The problem is how to recover this apparent loss of information. According to Fisher it is a mistake to calculate the information content of T with reference to the whole sample space, i.e. with reference to the marginal distribution of T. The appropriate measure of the information content of T is $J(\theta \mid Y)$ not $J(\theta)$. Having observed T and Y we should consider the conditional distribution of T with the observed value of Y as the condition. We take as our 'reference set' not the whole sample space but the sub-set of those sample points that could give rise to the observed value of the ancillary statistic Y.

RECOVERY OF ANCILLARY INFORMATION

The following two quotations from Fisher's writings emphasise the analogy that he has repeatedly drawn between the sample size and the ancillary statistic.

"Having obtained a criterion for judging the merits of an estimate in the real case of finite samples, the important fact emerges that, though sometimes the best estimate we can make exhausts the information in the sample, and is equivalent for all future purposes to the original data, yet sometimes it fails to do so, but leaves a measurable amount of the information unutilized. How can we supplement our estimate so as to utilize these too? It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it. Their function is, in fact, analogous to the part which the size of our sample is always expected to play, in telling us what reliance to place on the result." (Fisher, 1935).

"When sufficient estimation is possible, there is no problem, but the exhaustive treatment of the cases in which no sufficient estimate exists is now seen to be an urgent requirement. This at present is in the interesting stage of being possible sometimes, though, so far as we know, not always. I have spoken of the sufficient estimates as containing in themselves the whole of the information provided by the data. This is not strictly accurate. There is always one piece of additional, or ancillary, information which we require, in conjunction with even a sufficient estimate, before this can be utilized. That piece of information is the size of the sample or, in general, the extent of the observational record. We always need to know this in order to know how reliable our estimate is. Instead of taking the size of the sample for granted, and saying that the peculiarity of the cases where sufficient estimation is possible lies in the fact that the estimate then contains all the further informations required, we might equally well have inverted our statement, and, taking the estimate of maximum likelihood for granted, have said that the peculiarity of these cases was that, in addition, nothing more than the size of the sample was needed for its complete interpretation. This reversed aspect of the problem is the more fruitful of the two, once we have satisfied ourselves that, when information is lost, this loss is minimised by using the estimate of maximum likelihood. The cases in which sufficient estimation is impossible are those in which, in utilizing this estimate, other ancillary information is required from the sample beyond the mere number of observations which compose it. The function which this ancillary information is required to perform is to distinguish among samples of the same size those from which more or less accurate estimates can be made; or, in general, to distinguish among samples having different likelihood functions, even though they may be maximised at the same value. Ancillary information never modifies the value of our estimate; it determines its precision." (Fisher, 1936).

Often the 'extent of observational record' is planned in advance and is taken for granted in the subsequent analysis of the data. If we take n independent observations on a normal variable with unknown mean θ and known standard deviation,

we see no need to bother about any characteristic of the sample other than the sample mean $\bar{x}$; but yet the fact remains that without some knowledge about n the maximum likelihood estimator $\bar{x}$ of θ will be hardly of any use to any statistician. Along with the information that the sample is drawn from a normal population and the observed value of $\bar{x}$, we need to know the value of the sample size n. The 'reliability' of the estimator $\bar{x}$ is interpreted in terms of its average performance in repeated sampling with the fixed sample size n.

What happens if 'chance' plays (or is allowed to) a part in the determination of n? Suppose we toss a true coin and, depending on whether the outcome is a 'head' or a 'tail', we draw a sample of size 10 or 100. It is easily verified that the sample mean $\bar{x}$ is still the maximum likelihood estimator of θ but that it no longer is a sufficient statistic. It is the pair $(\bar{x}, n)$, where n is the (variable) sample size, that is sufficient for θ. Here n is an ancillary statistic taking the two values 10 and 100 with equal probabilities. Now, having drawn a sample of size n (which is either 10 or 100) and having estimated θ by the sample mean $\bar{x}$, how does the statistician report the 'reliability', the precision, the information content of the estimate? There is no gainsaying the fact that a sample of size 10 will lead to a less reliable estimate than a sample of size 100. Having drawn a sample of size 10 should the statistician turn a blind eye to the actual smallness of the sample size and try to figure out the long run performance of his estimation procedure in a hypothetical series of experimentations in which 50% of the cases he draws sample of size 10 and the other 50% of the cases the sample size is 100? What should be the reference set for judging the performance characteristic of the estimator—the 10 dimensional Euclidean space R_{10} or the union of R_{10} and R_{100}? The author agrees with Fisher that, having drawn a sample with the ancillary statistic $n = 10$, the statistician should judge (if at all he must) the performance of the maximum likelihood estimator $\bar{x}$ in the conditional sample space (restricted reference set) R_{10}. However, the author feels that Fisher, in his writings on ancillary statistics and choice of reference sets, has pushed the above analogy with the sample size a little too far, thereby giving rise to some logical difficulties the real nature of which will be discussed later.

In problems of testing Fisher uses ancillary statistics for the determination of the 'true' level of significance. Having selected the test criterion—a measure of the extent to which the observed sample departs from the expected one under the null hypothesis—the level of significance is the probability (under the null hypothesis) of getting a sample with a larger criterion score than the one actually obtained. In the presence of a suitable ancillary statistic Y, Fisher recommends that the level of significance of a test should be computed by referring to the conditional sample space determined by the set of all possible samples for which the value of Y is the one presently observed.

The following example worked out in Fisher (1956, pp. 163-69) is reproduced here with quotations with the idea of bringing out the essential features of the method

of 'Recovery of Information' (as envisaged by Fisher in the context of point estimation) through proper choice of 'reference sets'.

Example: Let us suppose that we have N pairs of independent observations on the pair (X, Y) of positive random variables with joint probability density function

$$p(x, y \mid \theta) = e^{-\left(\theta x + \frac{y}{\theta}\right)}, \quad x > 0, y > 0, \theta > 0.$$

Let (X_i, Y_i), $i = 1, 2, \ldots, N$, be the N pairs of observations and

let $$T = \sqrt{\Sigma Y_i / \Sigma X_i} \text{ and } U = \sqrt{(\Sigma X_i)(\Sigma Y_i)}.$$

It is easy to check that

(i) T is the maximum likelihood estimator of θ,

(ii) T is not sufficient for θ,

(iii) the pair (T, U) is jointly sufficient for θ,

(iv) U is an ancillary statistic, i.e. the marginal distribution of U is θ-free.

"Since the likelihood cannot be expressed in terms of only θ and T, there will be no sufficient estimate, and some information will be lost if the sample is replaced by the estimate T only."

The amount of information supplied by the whole sample (of N pairs of observations) is

$$I(\theta) = 2N/\theta^2$$

while the amount of information contained in the statistic T is

$$J(\theta) = \frac{2N}{\theta^2} \frac{2N}{2N+1}.$$

"The loss of information is less than half the value of a single pair of observations, and never exceeds one third of the total available. Nevertheless its recovery does exemplify very well the mathematical processes required to complete the logical inference."

Here, U is the ancillary statistic and so we have to consider the conditional distribution of T given U. From this conditional distribution the conditional information content of T works out as

$$J(\theta \mid U) = \frac{2N}{\theta^2} \frac{K_1(2U)}{K_0(2U)}$$

where K_0 and K_1 are Bessel functions.

Let us note that the information content $J(\theta \mid U)$ "depends upon the value of U actually available, but has an average value, when variations of U are taken into account, of

$$2N/\theta^2,$$

the total amount expected on the average from N observations, none is now lost. The information is recovered and inference completed by replacing the distribution

of T for given size of sample N, by the distribution of T given U, which indeed happens not to involve N at all. In fact, U has completely replaced N as a means of specifying the precision to be ascribed to the estimate. In both cases the estimate T is the same, the calculation of U enables us to see exactly how precise it is, not on the average, but for the particular value of U supplied by the sample."

2. THE SAMPLE SIZE ANALOGY

This section is devoted to a detailed discussion on the oft-drawn analogy between the sample size and an ancillary statistic. While drawing the analogy Fisher seems to be always thinking of the sample size n as the only determinant of the sampling experiment. The reliability of an estimate is assessed in terms of the average performance of the experimental procedure in a hypothetical series of experimentations with the sample size n fixed at the level actually obtained in the sample at hand. Fisher always interpreted the reliability (information content, variance etc.) of an estimate in terms of the average performance of some well-defined experimental (estimation) procedure in some hypothetical sequence of experimentations. When Fisher talks of the reliability of an estimate the adjective 'reliability' is used only as a transferred epithet and is actually meant to be attached to the estimation procedure that has given rise to the estimate.

In the example on page 56 the statistician (for some unknown reasons) decided to choose between a sample of size 10 or one of size 100 on the basis of the flip of a coin. Here a random choice is being made between two sampling experiments $\mathcal{E}_{10}$ and $\mathcal{E}_{100}$ with the associated sample spaces R_{10} and R_{100} and the corresponding probability distributions. More generally, suppose $\mathcal{E}_\nu$ is the sampling experiment corresponding to a sample of size ν and suppose the observed sample size n is determined by a θ-free (parameter-free) chance mechanism. Once we recognize that the estimate T of θ is generated by the random choice $\mathcal{E}_n$ from the family $\{\mathcal{E}_\nu\}$ of available experimental procedures we ought to transfer the reliability index of the chosen experiment $\mathcal{E}_n$ to the estimate T. When the statistician is forced to make a selection from a family of available experimental procedures he should report the reliability of the procedure actually selected by him. The following example is somewhat more realistic than the one considered on page 56.

Example : Suppose, from a finite population of N units, we draw a sample of size s with replacements. Let $X_1, X_2, \ldots, X_N$ be the population values and $x_1, x_2, \ldots, x_s$ the s sample values (arranged in order of their appearances). The problem is to estimate the population mean $\overline{X} = (X_1+\cdots+X_N)/N$ and the obvious estimate is the sample mean $\bar{x} = (x_1+\ldots+x_s)/s$ which has a standard deviation of $\sigma/\sqrt{s}$ where σ^2 is the population variance. But will it be correct to recognize s as the true sample size? Since the sample was drawn with replacements, it is plausible that some of the population units came repeatedly in the sample. In realistic sample survey situations, we can usually distinguish between the population units. Consider the extreme situation where all the s sample units happen to be the same (population unit).

Confronted with a situation like this the statistician would surely hesitate to report the reliability of his estimate as $\sigma/\sqrt{s}$. In this extreme case the honest statistician will have to admit that he had drawn an unlucky sample whose effective size is only 1 (and not s) and report the reliability of his estimate as σ (and not $\sigma/\sqrt{s}$). More generally, consider the situation where n is the number of distinct units in the sample and $x_1^*, x_2^*, \ldots, x_n^*$ are the corresponding sample values (arranged say in an increasing order of their population unit-indices). It is easy to see that the probability distribution of n involves only N and s, and since they are known constants, the statistic n is ancillary. If we define $\bar{x}^*$ as

$$\bar{x}^* = (x_1^* + \ldots + x_n^*)/n$$

then it is well recognized [see Basu (1958) for a detailed discussion on this] that $(x_1^*, x_2^*, \ldots, x_n^*)$ is a sufficient statistic and that $\bar{x}^* = E(\bar{x} \mid x_1^*, \ldots, x_n^*)$. Hence, by the Rao-Blackwell theorem, $\bar{x}^*$ is better than $\bar{x}$ as an unbiased estimator of $\bar{X}$. All right, we agree to estimate $\bar{X}$ by $\bar{x}^*$ and forget all about $\bar{x}$. But our troubles are not yet over. What is the standard deviation of $\bar{x}^*$? For a fixed n, the conditional distribution of $(x_1^*, \ldots x_n^*)$ is the same as that of a simple random sample of size n from the population of N values. We see then that the ancillary statistic n is really a sample size. When we draw a sample of size s with replacements and agree to take note of only the sufficient statistic $(x_1^*, x_2^*, \ldots, x_n^*)$ we are effectively drawing a simple random sample of variable size n. If we denote by $\mathcal{E}_i$ $(i = 1, 2, \ldots, s)$ the experimental procedure of drawing a simple random sample of size i from the population of N units and by p_i the parameter-free probability that $n = i$, then the above experimental procedure is the same as that of selecting one of the experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_s$ with probabilities $p_1, p_2, \ldots, p_s$ (and then carrying out the experiment $\mathcal{E}_n$ so selected). From what we have said earlier, it then follows that we should assess the reliability of $\bar{x}^*$ in terms of that of the experiment $\mathcal{E}_n$ actually selected, i.e. $V(\bar{x}^*)$ should be reported as

$$\frac{N-n}{N-1}\,\frac{\sigma^2}{n} \qquad \ldots \quad (2.1)$$

and not as

$$E\left(\frac{N-n}{N-1}\,\frac{\sigma^2}{n}\right). \qquad \ldots \quad (2.2)$$

Let us repeat once again that when reporting the reliability of an estimate we are actually saying something about the long term average performance of some well-defined estimation procedure. Both (2.1) and (2.2) are reliability indices—(2.1) for the experimental procedure $\mathcal{E}_n$ with a fixed n and (2.2) for the experimental procedure where n is allowed to vary (in the parameter-free manner described earlier) from trial to trial. We may briefly summarise the basic Fisherian point of view (in the present context) as follows :

(a) Suppose a whole family $\{\mathcal{E}_y\}$, where y runs over an index set $\mathcal{Y}$, of statistical experiments is available any one of which may be meaningfully performed for the purpose of making a scientific inference about some physical quantity θ.

(b) And suppose that the experiment that the statistician actually performs is recognized to be equivalent to the two-stage experiment of first selecting at random a point Y in $\mathcal{Y}$ and then performing the experiment $\mathcal{E}_Y$.

(c) Suppose, further, that the probability distribution of Y in $\mathcal{Y}$ is θ-free.

Under the above conditions, the true 'reference set' for the statistician is the experiment $\mathcal{E}_Y$ (with its associated sample space and probability distributions) and not the two-stage experiment described in (b) above. The reliability (information content, standard deviation, significance level etc.) of the inference made about θ should be assessed in terms of the average performance characteristic of the inference procedure in a long hypothetical sequence of independent repetitions under identical conditions of the experiment $\mathcal{E}_Y$ (where Y is supposed to be held fixed at the particular point that obtains in the present instance).

Under the above circumstances Fisher would like the statistician to say something to the following effect : "It was rather silly of me to let chance have a hand in the determination of the experiment $\mathcal{E}_Y$ for me. But I now recognize that I have performed the experiment $\mathcal{E}_Y$ and see no reason whatsoever to fuss about the other experiments in the family $\{\mathcal{E}_y\}$ that might have been handed down to me by chance. The inference that I make about θ is the most appropriate one for the experiment $\mathcal{E}_Y$, and the assessment of the reliability of my inference is made with reference to the experiment $\mathcal{E}_Y$ alone."

Now, let us consider a general inference situation and see whether the above arguments hold in the presence of an ancillary statistic. Let $\mathcal{E}$ be an arbitrary statistical experiment performed with a view to elicit some information about a physical quantity θ. From the mathematical standpoint we are then concerned with the trio $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ where $\mathcal{X} = \{x\}$ is the sample space and $\mathcal{A} = \{A\}$ is the σ-field of events on which the family $\mathcal{P} = \{P_\theta\}$ of probability measures is defined. For the sake of simplicity, we ignore the possibility of nuisance parameters and we assume that different possible values of θ are associated with different probability distributions. Now, let Y be an ancillary statistic taking values in the space $\mathcal{Y}$. Corresponding to each point y in $\mathcal{Y}$ we then have (under some regularity conditions) a trio $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$, where $\mathcal{X}_y$ is the sub-set of points of $\mathcal{X}$ for which $Y = y$ and $\mathcal{P}_y = \{P_\theta^y\}$ is the family of conditional probability distributions on a σ- field $\mathcal{A}_y$ of sub-sets of $\mathcal{X}_y$.

We have only to imagine a conceptual experiment $\mathcal{E}_y$ that gives rise to the trio $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$ and the analogy that we have been trying to drive home is complete. The statistical experiment $\mathcal{E}$ is then equivalent to the two-stage experiment of first observing the random variable Y (whose distribution is θ-free) and then performing the conceptual experiment $\mathcal{E}_Y$ leading ultimately to a point X in $\mathcal{X}$.

Why the insistence on Y being an ancillary statistic? The sample X that we arrive at through the experiment $\mathcal{E}$ (or the equivalent two-stage breakdown $Y-\mathcal{E}_Y$), is the only source of our information about θ. Now, according to Fisher, the likelihood function $L(\theta)$ for the sample X is the sole basis for making any judgement about θ.

Nothing else need be taken cognizance of. Let us observe that the likelihood function $L(\theta)$ is the same (excepting for a θ-free multiplicative constant) whether we consider X to be generated by the experiment $\mathcal{E}$ or the conceptual experiment $\mathcal{E}_Y$. This, according to the author, is the main explanation as to why in the above Y-decomposition of the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ into the family $\{\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y\}$ the statistic Y has to have a θ-free distribution. If Y be ancillary then the choice of the 'reference set'$(\mathcal{X}_Y, \mathcal{A}_Y, \mathcal{P}_Y)$ does not affect the likelihood scale.

3. A LOGICAL DIFFICULTY

The decomposition of the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ into the family of probability structures $\{\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y\}$ depends on the ancillary statistic Y. Which ancillary statistic Y to work with ? The author made a rather comprehensive study (Basu, 1959) of the family of ancillary statistics. It was noted that each of the two statistics Y_1 and Y_2 may individually be ancillary but jointly not so. Thus, in case of a controversy as to which one of the two ancillaries Y_1 and Y_2 should determine the 'reference set' one cannot solve the dilemma by referring to the conditional probability structure conditioned by the observed values of both Y_1 and Y_2. Consider the following example :

Example : Let (X, Y) have a bivariate normal distribution with zero means, unit standard deviations, and unknown correlation coefficient θ. If (X_i, Y_i), $i = 1, 2, \ldots, n$, be n pairs of independent observations on (X, Y) then we see at once that the set of n observations $(X_1, X_2, \ldots, X_n)$ on X is an ancillary statistic. Similarly $(Y_1, Y_2, \ldots, Y_n)$ also is ancillary. But the two ancillary statistics together is the whole data and is obviously sufficient. In regression studies the statistician often ignores the sampling variations in the observed X-values and treats them as preselected experimental constants. If, in the above situation, he be justified in similarly treating the observed Y-values also, then how do we define the sampling error for the estimate of θ?

Fisher recommended the choice of the 'reference set' with the help of an ancillary Y that complements the maximum likelihood estimator T in the sense that T is not sufficient but the pair (T, Y) is. This, however, is not a sufficient specification for Y. The two statistics Y_1 and Y_2 may each be ancillary complements to the maximum likelihood estimator T and lead to different 'reference sets' and different reliability indices for the estimator T. The following very simple example clearly brings out the above possibility.

Example : Suppose we have a biased die about which we have enough information to assume the following probability distribution :

scores :	1	2	3	4	5	6
probability distributions	$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$

where the parameter θ can take any value in the closed interval [0, 1]. Let $\mathcal{E}$ stand

for the experiment of rolling the die only once leading to the observed score X. It is easily seen that the maximum likelihood estimator T is defined as follows :

X :	1	2	3	4	5	6
$T(X)$:	0	0	0	1	1	1

and leads to the partition (1, 2, 3), (4, 5, 6) of the sample space. Here X is the minimal sufficient statistic and T is not sufficient. In this example there are six non-equivalent[1] ancillary complements to T. They may be listed as follows :

X	1	2	3	4	5	6
$Y_1(X)$	0	1	2	0	1	2
$Y_2(X)$	0	1	2	0	2	1
$Y_3(X)$	0	1	2	1	0	2
$Y_4(X)$	0	1	2	2	0	1
$Y_5(X)$	0	1	2	1	2	0
$Y_6(X)$	0	1	2	2	1	0

Each of the six statistics $Y_1, Y_2, \ldots, Y_6$ is a maximal[2] ancillary.

The statistic T induces the partition (1, 2, 3) and (4, 5, 6) whereas Y_1 induces the partition (1, 4), (2, 5) and (3, 6). Since each Y_1-partition intersects each T-partition in a one-point set, it follows that the pair (T, Y_1) is equivalent to X which is the minimal sufficient statistic. Thus, Y_1 is an ancillary complement to T. In like manner, we prove that each of the other Y_i's is an ancillary complement to T. The joint probability distribution for T and Y_1 is described in the following table :

T \ Y_1	0	1	2	total
0	$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{2-\theta}{4}$
1	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$	$\frac{2+\theta}{4}$
total	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	1

The Y_1-decomposition of the experiment of rolling the biased die once is as follows :

"Choose one of the three pairs (1, 4), (2, 5) and (3, 6) with probabilities 1/6, 1/3, and 1/2 respectively. Then select one from the chosen pair with probabilities

$\frac{1-\theta}{2}$ and $\frac{1+\theta}{2}$, if the chosen set is (1, 4),

or, $\frac{2-\theta}{4}$ and $\frac{2+\theta}{4}$, if the chosen set is (2, 5),

or, $\frac{3-\theta}{6}$ and $\frac{3+\theta}{6}$, if the chosen set is (3,6)".

[1]Two statistics are said to be equivalent if they lead to the same partition of the sample space.

[2]See Basu (1959) for the definition of a maximal ancillary.

The physical experiment of rolling the biased die once may then be imagined to be equivalent to the two-stage (conceptual) experiment of first choosing (in a θ-free manner) one of three biased coins and then tossing the selected coin once.

Let us observe that we have six different decompositions of $\mathcal{E}$ corresponding to the six different ancillary complements to T. How do we recover the information lost in T ?

Let us suppose that the observed value of X is 5. The corresponding values of T, Y_1, Y_2 and Y_3 are respectively 1, 1, 2 and 0. The conditional distributions of the maximum likelihood estimator T under the three conditions $Y_1 = 0$, $Y_2 = 2$, and $Y_3 = 0$ are as follows :

range of T :		0	1
conditional probability distribution of T under the condition	$Y_1 = 1$	$\frac{(2-\theta)}{4}$	$\frac{(2+\theta)}{4}$
	$Y_2 = 2$	$\frac{(3-\theta)}{5}$	$\frac{(2+\theta)}{5}$
	$Y_3 = 0$	$\frac{(1-\theta)}{3}$	$\frac{(2+\theta)}{3}$

Thus, in this situation we find that different choices of ancillary statistics lead to different 'reference sets' and different reliability indices for the estimator T. There exists no unique way of recovering the ancillary information.

4. Conceptual statistical experiments

The author believes that the real trouble lies in our failure to recognize the difference between a real (performable) and a conceptual (non-performable) statistical experiment.[1] Every real experiment gives rise to a probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ but the converse is not true. On page 60 we saw how the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, generated by the experiment $\mathcal{E}$, may be decomposed into the family $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$, $y \in \mathcal{Y}$, of probability structures and there we conceived of an experiment $\mathcal{E}_y$ corresponding to $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$. In general, the experiments $\mathcal{E}_y$ are non-performable. If (as in page 56) the statistician selects (on the flip of a coin) between a sample of size 10 and one of size 100, he is making a random choice between two performable statistical experiments. But in the example considered on page 61 the statistician can only conceive (in six different ways) of a breakdown of the experiment of once rolling the biased die into a two-stage experiment of first making a (θ-free) random selection between three biased coins and then tossing the selected coin once. In this example the statistician has a die to experiment with; but where are the coins ?[2]

[1] The author does not think it necessary to enter into a lengthy discussion on the reality or performability of a statistical experiment.

[2] Of course, the experiment of rolling the die repeatedly until, say, either 2 or 5 appears (and then observing only the final score) is essentially equivalent to tossing once a biased coin with probabilities $(2-\theta)/4$ and $(2+\theta)/4$. But who is interested in such a wasteful experiment ? The author would classify such experiments under the conceptual (non-performable) category.

When the sample size n is determined in a θ-free manner the statistician may be justified in regarding n as a pre-fixed experimental constant and in contemplating the long term performance characteristic of the experimental procedure[1] with the sample size fixed at the level actually obtained. More generally, when the statistician is presented with a θ-free choice between a family $\{\mathcal{E}_y\}$ of performable experimental procedures then it would be correct to treat the Y actually obtained as a pre-determined experimental constant. His sole concern should be the experiment $\mathcal{E}_Y$ he is actually presented with and none of the other members of the family $\{\mathcal{E}_y\}$. Difficulty arises in the attempted generalization to non-performable meaningless experiments. The kind of situations where an experiment $\mathcal{E}$ may be decomposed into a family $\{\mathcal{E}_y\}$ of real experiments are very rare indeed. The author is not aware of any example where such a decomposition (into real experiments) may be effected in more than one way. The elementary example considered on page 61 establishes the possibility of a multiplicity of ancillary decompositions into conceptual experiments. The ancillary argument of Fisher cannot be extended to such cases. The sample size analogy for the ancillary statistic appears to be a false one. We end this discourse with a very elementary example where there exists an essentially unique maximal[2] ancillary decomposition of the experiment but yet the ancillary argument leads us to a rather curious and totally unacceptable 'reference set'.

Example : Let X be an observable random variable with uniform probability distribution over the interval $[\theta, \theta+1)$ where $0 \leqslant \theta < \infty$. For the sake of simplicity we consider the case of a single observation on X. The sample space $\mathcal{X}$ is the half line $[0, \infty)$. The likelihood function, for the observation X, is

$$L(\theta) = \begin{cases} 1 & \text{if } X-1 < \theta \leqslant X \\ 0 & \text{otherwise} \end{cases}$$

Thus, the integer part $[X]$ of X has as good a claim to be considered a maximum likelihood estimator for θ as any other point in the interval $(X-1, X]$. It is easy to check that $[X]$ is not a sufficient statistic. Do there exist an ancillary complement to $[X]$?

Consider the fractional part $\varphi(X) = X-[X]$ of X. It is not difficult to show that $\varphi(X)$ is an ancillary statistic with uniform probability distribution over the unit interval $[0, 1)$. Indeed, it is possible to show that $\varphi(X)$ is an essentially maximal ancillary in the sense that every ancillary statistic is essentially a function of $\varphi(X)$. As $X = [X]+\varphi(X)$, it follows at once that the pair $([X], \varphi(X))$ is equivalent to X and hence $\varphi(X)$ is the ancillary complement to $[X]$.

For a given $\theta = [\theta]+\varphi(\theta)$, the observation $X = [X]+\varphi(X)$ lies in the interval $[\theta, \theta+1)$, i.e., with probability one,

$$\theta = [\theta]+\varphi(\theta) \leqslant [X]+\varphi(X) < [\theta+1]+\varphi(\theta) = \theta+1.$$

[1]Throughout this paper we are regarding the inference procedure as a well-defined part of the experimental procedure.

[2]That is, a decomposition with respect to a maximal ancillary as defined in Basu (1959).

From the above, it follows that

$$[X] = \begin{cases} [\theta] & \text{if } \varphi(X) \geqslant \varphi(\theta) \\ [\theta+1] & \text{if } \varphi(X) < \varphi(\theta) \end{cases} \qquad \ldots \quad (4.1)$$

Since $\varphi(X)$ has a uniform distribution over $[0, 1)$, it follows that the marginal distribution of $[X]$ is concentrated at the two points $[\theta]$ and $[\theta+1] = [\theta]+1$ with probabilities $1-\varphi(\theta)$ and $\varphi(\theta)$ respectively. Hence

$$\begin{aligned} E([X]\,|\,\theta) &= [\theta](1-\varphi(\theta))+([\theta]+1)\,\varphi(\theta) \\ &= [\theta]+\varphi(\theta) \\ &= \theta \end{aligned}$$

i.e. $[X]$ is an unbiased estimator of θ. And

$$V([X]\,|\,\theta) = \varphi(\theta)\ (1-\varphi(\theta)).$$

Now, since $\varphi(X)$ is the ancillary complement to $[X]$, let us see what 'reference set' it leads us to. Given $\varphi(X)$, the sample $X = [X]+\varphi(X)$ can vary over the restricted set

$$\varphi(X),\ 1+\varphi(X),\ 2+\varphi(X), \ldots\ldots\ldots$$

From the relation (4.1) it is now clear that, for any fixed θ, the conditional distribution of $X = [X]+\varphi(X)$, given $\varphi(X)$, is degenerate at the point

$$[\theta]+\varphi(X), \qquad \text{if} \quad \varphi(\theta) \leqslant \varphi(X),$$

or, at the point $\qquad [\theta+1]+\varphi(X), \text{ if} \quad \varphi(\theta) > \varphi(X).$

Thus, the 'reference set', corresponding to an observed value of the ancillary statistic $\varphi(X)$, is a one-point, degenerate probability structure. The conditional distribution of the maximum likelihood estimator $[X]$, given $\varphi(X)$ and θ, is degenerate at the point $[\theta]$ or $[\theta+1]$ depending on whether $\varphi(\theta) \leqslant \varphi(X)$ or $\varphi(\theta) > \varphi(X)$. Acceptance of this 'reference set' will alter the status of $[X]$ from a statistical variable to an unknown constant.

Writing $Y = \varphi(X)$, the two-stage Y-decomposition of the experiment $\mathscr{E}$ of making a single observation on X will then be as follows :

(i) Select a number Y at random (with uniform probability distribution) in the unit interval $[0, 1)$

(ii) Determine the value of $[\theta]$ and whether

(a) $\varphi(\theta) \leqslant Y$ or

(b) $\varphi(\theta) > Y$

and then write $$X = \begin{cases} [\theta]+Y & \text{in case of (a)} \\ [\theta]+1+Y & \text{in case of (b).} \end{cases}$$

The second stage of the Y-decomposition is clearly non-performable.

REFERENCES

BASU, D. (1958) : On sampling with and without replacements. *Sankhyā*, **20**, 287.

——— (1959) : The family of ancillary statistics. *Sankhyā*, **21**, 247.

COX, D. R. (1958) : Some problems connected with statistical inference. *Ann. Math. Stat.*, **29**, 357.

FISHER, R. A. (1925) : Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700.

——— (1934) : Two new properties of mathematical likelihood. *Proc. Royal Soc.*, **144**A, 285.

——— (1935) : The logic of inductive inference. *J. Roy. Stat. Soc.*, **98**, 39.

——— (1936) : Uncertain inference. *Proc. American Academy of Arts and Sciences*, **71**, 245.

——— (1956) : *Statistical Methods and Scientific Inference*, Oliver and Boyd, London.

OWEN, A. R. G. (1948) : Ancillary statistics and fiducial distribution. *Sankhyā*, **9**, 1.

RAO, C. R. (1952) : Minimum variance estimation in distributions admitting ancillary statistics. *Sankhyā*, **12**, 53.