ANCILLARY STATISTICS, PIVOTAL QUANTITIES
AND CONFIDENCE STATEMENTS

by

D. Basu

April 1980
FSU Statistics Report M 577

The Florida State University
Department of Statistics
Tallahassee, FL. 32306

Abstract

Several basic concepts of statistics that originated in the writings of R.A. Fisher are critically reviewed. Some deeprooted misconceptions about the meaning of ancillary information and confidence statements are carefully analysed.

Key Words: Ancillary information, conditionality argument, pivotal quantities and quantals, fiducial distribution, confidence statement, label set.

Research partially supported by NSF Grant No. 79-046903
[Presented at the Canadian Conference of Applied Statistics, Concordia University, Montreal, April 29 - May 1, 1981].
ON ANCILLARY STATISTICS, PIVOTAL QUANTITIES
AND CONFIDENCE STATEMENTS: A REVIEW

by

D. Basu
(The Florida State University)

1. INTRODUCTION

Information is perhaps the most commonly used expression in the
statistical literature, yet it is not possible to reach a consensus on the
definition and usage of the term. However, in the particular situation
where the problem is to predict the value of a random variable $X$ with
supposedly known distribution $p(\cdot)$, we can generally agree that the infor-
mation on $X$ may be characterized by $p(\cdot)$ itself. And if $Y$ is another random
variable such that the conditional distribution $p(\cdot|Y)$ of $X$, given $Y$,
can be figured out then, having observed $Y$ we can claim that the information
on $X$ has been altered from $p(\cdot)$ to $p(\cdot|Y)$. [To avoid multiplicity of notations
we are not distinguishing between a random variable and a typical
value of the variable.] If $p(\cdot|Y)$ is the same for all values of $Y$, then
$X$ is stochastically independent of $Y$ and this may be characterized by saying
that $Y$ has no information on $X$. And we know how to prove then that $X$ has
no information on $Y$. 
A parametric problem of statistical inference is similar to the above situation but not quite so to a non-Bayesian. A problem begins with a parameter of interest $\theta \in \Theta$ and an observable random variable $X \in X$. The argument begins with the choice of a model $\{p(\cdot \mid \theta) : \theta \in \Theta\}$ for $X$, where $p(\cdot \mid \theta)$ is interpreted as the conditional probability distribution of $X$, given $\theta$. [In this article we do not consider the case of nuisance parameters.] Typically, the distributions $\{p(\cdot \mid \theta) : \theta \in \Theta\}$ are different for different values of $\theta$. So $\theta$ has information on $X$. The reverse proposition that $X$ has information on $\theta$, therefore, seems reasonable enough. However, this cannot be proved in probabilistic terms unless we go Bayesian and introduce a prior probability distribution $q(\cdot)$ for $\theta$. For five decades R.A. Fisher tried very hard to set up a non-Bayesian theory of information in the sample. That resulted in an astonishing sequence of novel ideas like sufficiency, likelihood, information function, ancillarity, reference set, conditionality argument, pivotal quantity, and fiducial argument. Many learned discourses by contemporary statisticians and philosophers on Fisher's theory have illuminated as well as clouded the statistical literature of our time. In the hope of dispelling some of the still lingering shadows, I propose to take yet another look at the twin concepts of ancillary statistic and pivotal quantity and the related issue of confidence statements.

2. ANCILLARY STATISTICS

A statistic $T$ is ancillary if the conditional distribution $p_T(\cdot \mid \theta)$ of $T$, given $\theta$, is the same for all $\theta$. If the problem is to predict $T$ then surely $\theta$ by itself has no information on $T$. Is it reasonable to say then that $T$ by itself has no information on $\theta$? Can we generate any information on a parameter $\theta$ by, say, rolling a symmetric die a number of times?
When we are arguing in the context of a specified model of an observable \( X \) and when \( T = T(X) \) is ancillary, then the act of recording only the T-value of \( X \) seems to be quite as useless a statistical exercise as that of rolling the die. R.A. Fisher may have been guided by a consideration of the above kind to arrive at the conclusion that an ancillary \( T \) by itself cannot possibly have any information on \( \theta \). A Bayesian would regard this as self evident, because, in the context of his prior opinion \( q(\cdot) \) on \( \theta \), whatever \( q \) might be, the posterior distribution of \( \theta \), given \( T = t \), will work out as

\[
q(\theta|t) = q(\theta)P_T(t|\theta) / \sum_{\theta} q(\theta)P_T(t|\theta)
\]

\[
= q(\theta)P_T(t|\theta) / P_T(t|\theta)
\]

\[
= q(\theta)
\]

for all \( \theta \) and \( t \). Since observing \( T \) in isolation cannot change the distribution of \( \theta \), the Bayesian must regard \( T \) as uninformative in itself. Recently, V.P. Godambe (1979,1980) has clouded the issue with his repeated assertions that there are situations where an ancillary statistic by itself can yield a quantum of information on the parameter. We shall take a good look at the Godambe intuition on ancillarity later.

The notion of sufficiency is closely related to the notion of ancillarity. The statistic \( S = S(X) \) is sufficient if the sample \( X \) becomes ancillary when it is conditioned by \( S \). In other words, \( S \) is sufficient if the conditional distribution \( p(\cdot|S,\theta) \) of \( X \), given \( S \) and \( \theta \), depends on \( (S,\theta) \) only through \( S \). Thus, once we know the \( S \)-value of \( X \), the parameter \( \theta \) has no further information for us on \( X \). Fisher's intuition that
"a sufficient statistic exhausts (summarizes) all the relevant information on \( \theta \) in \( X \)" was apparently based on an unconscious reversal of the roles of \( \theta \) and \( X \) in the previous sentence. Kolmogorov (1942) presented us with the correct Bayesian perspective on sufficiency with the following Definition: The statistic \( S \) is sufficient if, for every prior \( q(\cdot) \) for \( \theta \), the posterior distribution \( q(\cdot | X) \) of \( \theta \), given \( X \), depends on \( X \) only through \( S = S(X) \). In other words, \( S \) is sufficient if, for every prior \( q \), the variables \( X \) and \( \theta \) are conditionally independent, given \( S \). Likewise, \( T \) is ancillary if, for every prior \( q \), the variables \( T \) and \( \theta \) are independent of each other.

Following Neyman-Pearson (1936) we call an \( X \)-event \( A \) (a measurable subset of the sample space) similar (a similar region) if \( p(A|\theta) \) is a constant, say \( \alpha \), in \( \theta \). Despite the fact that the sample space is endowed (by the model) with a multiplicity of probability distributions, a similar event \( A \) with \( P(A|\theta) = \alpha \), like the impossible and the sure event, seems to be endowed with an absolute (unconditional, that is) probability \( \alpha \). As before, a Bayesian characterization of a similar event is that \( A \) is similar if it is independent of \( \theta \) for every prior \( q \). [It should be understood that we are always arguing in the context of a fixed model.]

The class \( A \) of similar regions  

i) contains the whole space \( \mathcal{A} \),  
ii) is closed for differences (if \( A \Subset B \) and both belong to \( A \) then so also does \( B - A \)) and \( \mathcal{A} \) is a monotone class (closed for monotone limits).  
It is what Dynkin (1965) calls a \( \lambda \)-system. A \( \lambda \)-system that is closed for intersection is a Borel field. The class \( A \) is typically not a Borel field. A statistic \( T \) is ancillary if and only if every \( T \)-event (a subset
of $X$ defined in terms of $T(X)$) belongs to $\mathcal{A}$. Whenever $\mathcal{A}$ is not a Borel field we can find two ancillary statistics $T_1$ and $T_2$ such that the pair $(T_1, T_2)$ is not ancillary. In such a situation we cannot have an ancillary statistic $T_0$ that is a maximum ancillary in the sense that every other ancillary statistic is a function of $T_0$.

**Example 1:** Let $X = (X_1, X_2)$ be a pair of i.i.d. random variables whose common distribution on the real line is known excepting for a location parameter $\theta$. Clearly, $T = X_1 - X_2$ is ancillary and, therefore, so also is every function $(T)$ of $T$. Contrary to popular impression $T$ is not the maximum ancillary. Indeed, a maximum ancillary can never exist in a situation like this. That the family $\mathcal{A}$ of similar regions is not a Borel field is seen as follows. For any $X$-event $A$ let $A^C$ denote its complement and $A^*$ the event obtained from $A$ by interchanging $X_1, X_2$ in its definition. We call $A$ symmetric in the co-ordinates if $A^* = A$. Clearly $(A^*)^* = A$ and $(AB)^* = A^*B^*$ for all $A, B$. Since the distribution of $X = (X_1, X_2)$ is symmetric in the co-ordinates for each $\theta$, we have $p(A^*|\theta) = p(A|\theta)$ for all $\theta$ and $A$. Now, let $A$ be an arbitrary $T$-event that is not symmetric, e.g., $A = \{(x_1, x_2): x_1 - x_2 > 1\}$ and let $B$ be an arbitrary symmetric event that is not a $T$-event, e.g., $B = \{(x_1, x_2): x_1^2 + x_2^2 < 1\}$. By definition $A$ and $A^* = \{(x_1, x_2): x_2 - x_1 > 1\}$ are similar regions. Consider $E = AB \cup A^*B^C$. Since $(A^*B^C)^* = A^{**}B^C = AB^C$, we have $p(A^{**}B^C|\theta) = p(AB^C|\theta)$ and so

$$p(E|\theta) = p(AB|\theta) + p(A^*B^C|\theta)$$

$$= p(AB|\theta) + p(AB^C|\theta)$$

$$= p(A|\theta)$$

for all $\theta$. That is, $E$ is a similar region even though it is not a $T$-event.
Observe that in this particular instance

\[ AE = AB = \{(x_1, x_2): x_1^2 + x_2^2 < 1, x_1 - x_2 > 1\} \]

and that this, being a bounded set, cannot be a similar region for a location parameter family. That is, it is not closed for intersection. The family of ancillary statistics is very large.

Example 2: Let \( X \) be uniformly distributed over the interval \((\theta, \theta + 1)\).

With a single observation on \( X \), can we have a nontrivial ancillary statistic?

Let \([X]\) be the integer part of \( X \) and \( \psi(X) = X - [X] \) the fractional part.

It is then easy to verify that \( \psi(X) \) is uniformly distributed over the interval \((0, 1)\) for all \( \theta \) and so it is an ancillary statistic. In this case \( \psi(X) \) is the maximum ancillary. If \( X_1, X_2, \ldots, X_n \) are \( n \) independent observations on \( X \), then the statistic \( T = (\psi(X_1), \psi(X_2), \ldots, \psi(X_n)) \) is ancillary and so also is the difference statistic \( D = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1) \).

3. ANCILLARY INFORMATION

An ancillary statistic \( T \) by itself carries no information on the parameter \( \theta \), but in conjunction with another statistic \( Y \) (which, as we shall see later, may even be ancillary itself) may become fully informative (sufficient). Fisher's controversial theory of recovery of ancillary information (Basu, 1964) is based on a recognition of the fact that an ancillary statistic can carry a lot of potent information. A simple example is given to elucidate the Fisher method. Example 3: Consider a threefold multinomial model with cell probabilities \( \theta/2 \), \((1 - \theta)/2 \) and \( 1/2 \). Let the observed cell frequencies be \( n_1 \), \( n_2 \) and \( n - n_1 - n_2 \). Now the statistic \( S = (n_1, n_2 \text{ is sufficient}) \) and \( T = n_1 + n_2 \) being distributed as Bin \((n, 1/2)\),
is ancillary. The statistic $T$ by itself does not tell us anything about $\theta$ that we did not know already, but in a sort of way it tells us something about how informative the sample is. For instance, if $T$ is zero then we know that the likelihood function generated by the data is the constant $1/2^N$ and so the sample is void of any information on $\theta$. In a sense it seems clear that larger the observed value of $T$ the more informative the sample is. The maximum likelihood (ML) estimate $\hat{\theta}$ of $\theta$ is $n_1/T$ if $T \neq 0$ and is undefined if $T = 0$. The ML estimator $\hat{\theta}$, assuming that we have suitably defined it for the case $T = 0$, is not a sufficient statistic and so, according to Fisher, there will be a loss of information if we evaluate $\hat{\theta}$ in terms of its (marginal) distribution $p_{\hat{\theta}}(\cdot|\theta)$. The statistic $T$ is an ancillary complement to $\hat{\theta}$ in the sense that the pair $(\hat{\theta}, T)$, being equivalent to $S = (n_1, n_2)$, is fully informative. This is the kind of situation where Fisher (1935) wants us to evaluate the ML estimator $\hat{\theta}$ by conditioning it by the ancillary statistic $T$. The conditional distribution of $n_1$, given $T, \theta$, is $\text{Bin}(T, \theta)$, when $T \neq 0$. Thus $E(\hat{\theta}|T, \theta) = \theta$, $V(\hat{\theta}|T, \theta) = \theta(1-\theta)/T$ and the Fisher informative content of $\hat{\theta}$, conditionally on $T$, is $T/\theta(1-\theta)$. Fisher would regard $\hat{\theta}$ to be an unbiased estimate of $\theta$ with small variance and large information if $T$ is large. The smaller the observed value of $T$, the less informative $\hat{\theta}$ will be as an estimate of $\theta$. In the extreme case when $T = 0$, there is no information at all.

This example brings out the Fisher dilemma. He recognized that as a rule not all samples are equally informative and so a sample space analysis of the data may not be quite appropriate. He realized that the information in the data obtained is fully summarized in the corresponding likelihood
function but did not quite know how to analyze the likelihood function in a non-Bayesian fashion. [In Fisher (1956) we do find some half-hearted advice on how to analyze the likelihood function. This has been fully scrutinized (and rejected) by the author in Basu (1975).] So he had to go for a compromise solution: Analyze the data in sample space terms, but suitably restrict the sample space by clustering together all sample points that are, in some qualitative sense, as informative as the sample actually observed.

In the present case the Fisher solution seems pretty attractive. With \( n = 15 \), the two samples \((1, 2, 12)\) and \((12, 2, 1)\) are qualitatively quite different. It seems reasonable enough to say that the data \((1, 2, 12)\) should be interpreted as 1 success in a succession of 3 Bernoulli trials with \( \theta \) as the probability of success. Similarly the data \((12, 2, 1)\) should be looked upon as 12 successes in 14 Bernoulli trials. In the first case \(1/3\) is an unbiased estimate of \( \theta \) with variance \( \theta (1 - \theta)/3 \), whereas in the second case \(12/14\) is an unbiased estimate of \( \theta \) with variance \( \theta (1 - \theta)/14 \).

Admittedly, it is not easy to make an unconditional sample space analysis of the data \( X = (n_1, n_2, n_3) \). Our heart reaches out for the ML estimator \( n_1 | (n_1 + n_2) \) but we worry about it being undefined when \( T = n_1 + n_2 = 0 \). We do not relish the idea of figuring out the bias, variance and the information content of the estimator. We recognize many unbiased estimators like \( 2n_1/n, 1 - 2n_2/n, \) and \( 1/2 + (n_1 - n_2)/n \) but we like none of them as we recognize that they can take values outside the parameter space \((0, 1)\). The minimum sufficient statistic \( S = (n_1, n_2) \) is not complete, there does not exist a minimum variance unbiased estimator of \( \theta \), no unbiased estimator
can be admissible and so on. All the difficulties seem to be stemming from the fact that $T$ is ancillary. So why not cut the Gordian knot by holding fixed the observed value of $T = n_1 + n_2$ just as we do for the sample size $n = n_1 + n_2 + n_3$?

The sample size analogy for an ancillary statistic comes from Fisher (1935) who wrote: "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 the sample is always expected to play, in telling us what reliance to place on the result".

Example 4: An experiment consists of $n$ Bernoulli trials where $n = 5$ or 100 depending on the flip of a fair coin. The sample is $X = (Y, n)$, where $Y$ is the number of successes in $n$ trials. The ML estimate of $\theta$ is $\hat{\theta} = Y/n$. The sample size $n$ is an ancillary statistic now. How to analyze the data $X = (3, 5)$? Clearly, the data is qualitatively very different from the possible data $(60, 100)$ although the ML estimate is the same in either case. As Fisher would have put it, the ML estimate is insufficient (in this case) to identify the full information content of the data, namely, the likelihood function. In terms of the conditional model $p(\cdot | n, \theta)$ for the data, the ML estimator $\hat{\theta}$ is fully sufficient.

Suppose we slightly alter the experiment in Example 4 by determining the sample size $n$, not by the flip of a fair coin, but by the outcome of the first trial - the sample size is 5 in 100 according as the first trial yields a success or a failure. Now, $n$ is no longer an ancillary
statistic even though the quality of the sample depends very much on $n$. The likelihood function generated by an observed sequence of $n$ successes and failures is $\theta^y (1 - \theta)^{n-y}$, where $y$ is the number of successes in the sequence. Therefore, $(Y, n)$ is the minimum sufficient statistic and $\hat{\theta} = y/n$ is still the ML estimate of $\theta$. To evaluate $\hat{\theta}$ in terms of its full (unconditional) model $p_\theta(X \mid \theta)$ will clearly entail a substantial loss of information. The recovery of information argument does not work here in view of the fact that we do not have an ancillary complement to $\hat{\theta}$.

In Basu (1964) there are other illustrations of how the argument can fail. For instance, there can be a multiplicity of ancillary statistics with no clear cut choice for one to condition the data with. There can also be situations where conditioning $X$ with an ancillary statistic $T$ reduces it to a degenerate random variable with the point of degeneracy depending on $T$ and $\theta$. In Example 2 we have such a situation because the conditional distribution of $X$, given $\psi(X)$ and $\theta$, is degenerate. In Example 7 we shall have a similar situation. A traditional sample space analysis of data with a degenerate sample space is unheard of.

With his recovery of ancillary information argument, Fisher was seeking for a via media between the Bayesian and what we now regard as the Neyman-Pearson way. As a compromise solution it failed as most such solutions do. From the point of view of history, it is more important to recognize what Fisher was attempting to do than the fact of his failure to do so. Clearly, he was trying to cut down the sample space to size. Why? Because he realized that not all sample points in the same sample space are equally informative. With a data of poor quality in hand, e.g., $X = (3, 5)$ in Example 4, it makes little sense to derive some comfort from the thought of a might have been excellent data, e.g., $X = (60, 100)$
or any other sample with \( n = 100 \). In this situation, why not evaluate the data \((3, 5)\) in the context of the reduced (conditional) sample space (reference set) consisting of only the six points \((0, 5), (1, 5), \ldots, (5, 5)\)?

Fisher (1936) wrote: "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 maximized at the same value". (Italics ours.) Clearly, Fisher thought that two samples, even though they may be of the same size, may be different in their information content in that a more (or less) accurate estimate of the parameter can be made from the one than from the other. It is not at all clear why he thought that his ancillary statistic can "distinguish among samples with different likelihood functions".

The minimum sufficient statistic \( S \) is the one that distinguishes between samples with different likelihood functions. [Two likelihood functions are equivalent if they differ only by a multiplicative constant. The minimum sufficient \( S \) is the statistic that partitions the sample size into points with equivalent likelihood functions.] The M.L. estimator \( \hat{\theta} \) is a function of \( S \). The ancillary statistic \( T \) complements \( \hat{\theta} \), that is, \((\hat{\theta}, T)\) is a one-one function of \( S \). Thus, \((\hat{\theta}, T)\) can distinguish among samples with different likelihood functions. If \( \hat{\theta}(x_1) = \hat{\theta}(x_2) \), then \( x_1 \) and \( x_2 \) generate different likelihood functions if and only if \( T(x_1) \neq T(x_2) \). Of course, \( T \) cannot distinguish among samples with different likelihood functions, for, if it could, it would have been the minimum sufficient statistic.

In the thirties, when Fisher came up with the recovery of information
argument, he was still struggling to justify his pet ML estimation method in some sample space terms. Later on (1956, pp. 66-73) we find Fisher willing to condition all the way and focus his undivided attention on the information generated by the data, namely, the likelihood function. His ancillarity argument was a mistake because it discarded the sample space only halfway.

Apart from recovering ancillary information, Fisher found two other uses for the conditionality argument. In the next section we briefly discuss how he extended the scope of his fiducial argument by conditioning pivotal quantities that were not based on what he called "sufficient estimates". In Basu (1977) Fisher's conditioning method for elimination of nuisance parameters has been discussed at some length.

4. PIVOTAL QUANTITIES

The notion of a pivotal quantity, perhaps the most innovative idea that came from Sir Ronald, is an extension of the concept of an ancillary statistic. A quantity $Q = Q(\theta, X)$ is a measurable function of the parameter $\theta \in \Theta$ and the sample $X \in \mathcal{X}$, that is, $Q$ is a map of the product space $\Omega = \Theta \times \mathcal{X}$ into a range space $\mathcal{R}$ that is usually taken to be the real line. A statistic is a quantity that is constant in $\theta$.

**Definition:** $Q: \Omega \rightarrow \mathcal{R}$ is a pivotal quantity, in short, a p-quantity, if the conditional distribution of $Q(\theta, X)$, given $\theta$, is the same for all $\theta$.

Thus, an ancillary statistic is the extreme case of a p-quantity that is constant in $\theta$. If $Q: \Omega \rightarrow \mathcal{R}$ is a p-quantity then so also is $h(Q)$ for every measurable function $h$ on $\mathcal{R}$. In the location parameter models of Examples 1 and 2, typical examples of p-quantities are $X_1 - \theta$, $\bar{X} - \theta$, $\tilde{X} - \theta$, etc., where $\tilde{X}$ is an equivariant statistic like the mean, median
is the mid-range. A more complex example of a p-quantity, a sheer joy to the nonparametricians, involving the location parameter or as follows.

**Example 5:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common c.d.f. $F(x - \mu)$, where $F$ also is unknown save for the facts that it is continuous at the origin and that $F(0) = 1/2$. With the location parameter $\mu$ as the parameter of interest define the quantity $Q(\mu, X)$ as the number of i's for which $X_i - \mu > 0$, $i = 1, 2, \ldots n$. Then $Q$ is a p-quantity since the conditional distribution of $Q$, given $\theta = (\mu, F)$, is $\text{Bin}(n, 1/2)$.

A Bayesian definition of a p-quantity will run along the following lines.

**Definition:** $Q(\theta, X)$ is a p-quantity with respect to a model $\{p(\cdot|\theta): \theta \in \Theta\}$ for the sample $X$, if, irrespective of the prior distribution $q$ of $\theta$, $Q$ and $\theta$ are stochastically independent.

Equivalently, $Q$ is a p-quantity if its predictive distribution depends on the prior $q$ and the model $\{p(\cdot|\theta)\}$ only through the latter.

With $n$ i.i.d. observations $X_1, X_2, \ldots, X_n$ on a $N(\theta, 1)$, Fisher derived the fiducial distribution of $\theta$ as $N(\bar{X}, 1/n)$ by using the p-quantity $\bar{X} - \theta$ as a pivot. It is not the purpose of this article to examine the fiducial argument once again. During the past five decades the argument has been thoroughly examined many times and has been declared, with a few exceptions, as logically invalid. We propose to examine the confidence statement argument that Neyman-Pearson in the early thirties, synthesized from the fiducial argument. But before we get into that it will be useful to examine why Fisher chose the particular p-quantity $\bar{X} - \theta$ as his pivot.
Fisher regarded the p-quantity $\bar{X} - \theta$ as the correct pivot for the fiducial argument because i) $\bar{X}$ is a natural estimate of $\theta$, ii) $\bar{X}$ is a (minimum) sufficient statistic and iii) the range of variation of $\bar{X}$ is the same as that of the parameter $\theta$, namely, the whole real line. Fisher's fiducial logic would not permit us to use $X_1 - \theta$ as a pivot because $X_1$ is not a "sufficient estimate" of $\theta$.

The weak p-quantity $X_1 - \theta$ can, however, be made strong by proper conditioning! The difference statistic $D = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is the ancillary complement of $X_1$. Consider the conditional distribution of $X_1 - \theta$ for fixed $D$ and $\theta$. Since $X_1 - \theta = \bar{X} - \theta + (X_1 - \bar{X})$, $X_1 - \bar{X}$ is a function of $D$, and $\bar{X}$ is independent of $D$ for fixed $\theta$, it follows that $X_1 - \theta$, given $D$ and $\theta$, is distributed as $N(X_1 - \bar{X}, 1/n)$. Thus, $X_1 - \theta$ remains a p-quantity even when it is conditioned by $D$. If the fiducial argument is based on the conditioned p-quantity $X_1 - \theta | D$ then we arrive at the correct fiducial distribution $N(\bar{X}, 1/n)$ for $\theta$.

Fisher generalized the above conditional pivotal quantity argument to the case of a location parameter model as follows. Let $X = (X_1, X_2, \ldots, X_n)$ be the sample with density function $f(x_1 - \theta, x_2 - \theta, \ldots, x_n - \theta)$. There are many p-quantities like $X_1 - \theta, \bar{X} - \theta, \tilde{X} - \theta, \ldots$, etc., where $\tilde{X}$ is any equivariant estimator of $\theta$. On which one of these shall we base the fiducial argument? It does not matter as long as we condition the chosen p-quantity by the difference statistic! Suppose we choose the sickly p-quantity $X_1 - \theta$, condition it by $D$, thus arriving at a density function, say, $p(t)$. Chanting the fiducial argument we then arrive at the fiducial density function
p(X_1 - t) for the parameter \( \theta \). But had we started off with a stronger looking p-quantity \( \tilde{X} - \theta \), where \( \tilde{X} \) is an equivariant estimate of \( \theta \) like
the mean or the median, then, in view of the fact that \( \tilde{X} - X \), is a function of \( D \), the conditional density function for \( \tilde{X} - \theta \), given \( D \) and \( \theta \), would have been \( p(t - \tilde{X} + X_1) \). Therefore, the fiducial distribution of \( \theta \), with \( \tilde{X} - \theta \) held as the pivot, would also have worked out as \( p(X_1 - t) \).

Many of our contemporary statisticians -- Pitman, Barnard, Kempthorne, Fraser to name a few -- have been greatly (and very diversely) influenced by the above conditional pivot argument of Fisher. For one thing, the extensive study of invariance as a data reduction principle originated in the fiducial distribution \( p(X_1 - t) \) for the location parameter \( \theta \). Note that the mean

\[
\hat{\theta} = \int t p(x_1 - t) \, dt \\
= \int (x_1 - t) p(t) \, dt \\
= x_1 - \int t p(t) \, dt \\
= x_1 - \text{E}(x_1 - \theta | D, \theta) \\
= x_1 - \text{E}(x_1 | D, \theta = 0)
\]

of the fiducial distribution is the Pitman estimator (the best equivariant estimator with squared error loss function) of \( \theta \). Similarly, the median of the fiducial distribution is the best equivariant estimator when the absolute error is taken as the loss function. If the proof of the pudding is in the eating then one may argue that Fisher's fiducial argument is plausible at least in the case of a location parameter.
The author regards the following proposition as a well established statistical metatheorem: No inferential argument in statistics has anything going for it unless a sensible Bayesian interpretation can be found for it. It so happens that the fiducial distribution $p(X_1 - \theta)$ for the location parameter $\theta$ may be interpreted as the Bayes posterior with the (improper) uniform prior over the real line. So it is possible to condition the sample $\bar{X}$ all the way down to its observed value and derive a distribution for $\theta$ directly from the likelihood function. However, the uniform prior over the unbounded real line can hardly be regarded as a sensible representation of anyone's ignorance about the parameter $\theta$.

Sir Ronald is no longer with us but his pivotal quantities are very much alive. As we shall see in the next section, all confidence statements are based on p-quantities. Fisher tried hard but failed to establish a coherent theory for the choice of an appropriate p-quantity. Currently there appears to be no law and order regarding the choice of the pivot. Often a sizeable part of the data is ignored in order to facilitate the choice of the pivot. Stein's (1945) classical work on fixed width confidence interval is an example of this kind. Sometimes post-randomization variables are deliberately introduced in the data so as to create a p-quantity in terms of the extended data. For instance, if the data generated by a sequence of n Bernoulli trials be enhanced by a randomization variable uniformly distributed over $(0, 1)$, then we can construct a 95% confidence interval for the probability parameter $p$. Finally, there is the curious (Fisher inspired) method of holding a substantial part of the data as fixed and then recognizing that a quantity $Q(\theta, X)$ becomes a
p-quantity in terms of the restricted reference set. Godambe and Thompson (1971) and Seheult (1980) are only two of many such licentious use of the fiducial argument that Fisher himself thought to be of rather limited use.
5. CONFIDENCE STATEMENTS

A quantity Q was defined earlier as a measurable function of \((\theta, X)\). A measurable subset of \(\Omega = \Theta \times X\) will be called a quantal.

Let \(Q: \Omega \to \mathbb{R}\) be an arbitrary pivotal quantity (p-quantity) and let \(B\) be a measurable subset of \(R\). Consider the quantal \(E = \{(\theta, X): Q(\theta, X) \in B\}\)

For each \(\theta \in \Theta\), let \(E^\theta = \{X: (\theta, X) \in E\}\) denote the \(\theta\)-section of \(E\).

For a given \(\theta\), the statements \(Q(\theta, X) \in B\) and \(X \in E^\theta\) are identical.

Since \(Q\) is a p-quantity, the conditional probability of the former statement, given \(\theta\), is a constant in \(\theta\), and therefore, so also is \(\text{Prob}(X \in E^\theta | \theta)\). This motivates the following

**Definition:** In the context of a model \(\{p(\cdot | \theta)\}\) for the sample \(X\), a subset \(E\) of \(\Omega = \Theta \times X\) is called a p-quantal if \(p(E^\theta | \theta)\) is a constant in \(\theta\). The constant value of \(p(E^\theta | \theta)\) is called the size of the p-quantal \(E\).

The empty set and the whole space \(\Omega\) are trivial examples of p-quantals of size 0 and 1 respectively. If \(A\) is a similar region of size \(\alpha\) then \(\Theta \times A\) is a p-quantal of size \(\alpha\). Like the family \(A\) of similar regions, the family \(\xi\) of p-quantals is a \(\lambda\)-system of sets. The indicator function of a p-quantal is a p-quantity. A function \(Q(\theta, X)\) is a p-quantity if and only if every subset of \(\Theta \times X\) defined in terms of \(Q\) is a p-quantal. Analogous to the Bayesian definition of a p-quantity on page 13, the notion of a p-quantal may be redefined as

**Definition:** The set \(E \subset \Omega\) is a \(\alpha\)-p-quantal if, irrespective of the scientist's prior distribution (of belief) on \(\Theta\), the event \(E\) is independent of the random variable \(\theta\). Equivalently, \(E\) is a p-quantal if its probability is well defined in terms of the model \(\{p(\cdot | \theta)\}\) alone.
R.A. Fisher's fiducial argument was severely restrictive in that only a few simplistic statistical models could cope with his stringent requirements for the right pivotal quantity. Jerzy Neyman and E.S. Pearson rejected the fiducial logic but they nevertheless accepted a part of it and generalized it to the limit. Neyman-Pearson's confidence statement argument is based not on the limited stock of the Fisherian p-quantities but on the plentiful supply of p-quantals. Any p-quantal E can be the basis of a confidence statement as described below.

Let $E_X = \{\theta : (\theta, X) \in E\}$ denote the X-section of the p-quantal E. Clearly, the three statements $(\theta, X) \in E$, $X \in E^\theta$ and $\theta \in E_X$ are logically equivalent. If we regard $E_X^\theta$ as a random set determined by the random variable X, then, prior to the observation of X, the probability of $\theta \in E_X^\theta$ is well defined and is independent of any prior opinion (or lack of it) that the scientist may have on the parameter $\theta$. Suppose the p-quantal E is of size 0.95. Then, prior to the observation of X, the scientist is 95% sure (ordinary probability) that the true $\theta$ will belong to the set $E_X^\theta$ that is going to be determined by the observance of X. Once X is observed and the particular $E_X^\theta$ determined then is it still reasonable for the scientist to assert that he or she is still 95% sure that $\theta \in E_X^\theta$? Fisher would have answered the question with a firm negative if he found that the p-quantal E has not been defined in the right manner in terms of the right pivotal quantity. Neyman-Pearson's theory of confidence statements suffers from no such inhibition. Any
p-quantal of size 0.95 is regarded (by a Neyman-Pearsonian) as a generator of a 95% confidence set estimator \( E_\theta \), where \( \theta \in \{X_1, X_2, \ldots \} \). If we consider the hypothetical population of a sequence of observations \( X_1, X_2, \ldots \) on \( X \) with \( \theta \) fixed at its true value, then is it not correct to say that 95% of the sets \( E_{X_1}, E_{X_2}, \ldots \) will cover \( \theta \) and only 5% will not?

The frequency interpretation of the confidence statement \( \theta \in E_X \) that is implicit in the above hypothetical question is the cornerstone of the Neyman-Pearson argument. Fisher always maintained that the argument was a logical error. I vividly recall an occasion (Winter of 1955, Indian Statistical Institute, Calcutta) when Professor Fisher bluntly asserted that a confidence interval, unless it coincides with a fiducial interval, cannot be interpreted in frequency probability terms. At the end of the Fisher seminar, Professor Mahalanobis, who was chairing the session, invited me to comment on the proposition. So, shaking in my shoes, I rose to defend Professor Neyman and gave the standard frequency interpretation of confidence statements. Professor Fisher summarily dismissed my explanation as yet another example of the "acceptance test" type argument.

It is important for us to understand the Fisher point of view. According to Fisher (1956, p. 77) the population (an hypothetical sequence of repeated trials with certain elements held fixed) to which a certain inferential statement is referred to for probabilistic interpretation is a figment of the mind. Only in some "acceptance test" type situations, Fisher would concede that the population (reference set) is well defined. In problems of scientific inference it should
be recognized that the data can be differently interpreted in terms of different reference sets. Of course, Fisher alone knew how to artfully choose the reference set to suit an individual scientific problem!

Let us go back to a situation that we have already discussed in Section 3. Suppose with a sample \( X = (X_1, X_2, \ldots, X_n) \) of \( n \) i.i.d. \( N(\theta,1) \)'s, a misguided scientist chooses to make a confidence statement based on the first observation, \( X_1 \) alone. Since the \( p \)-quantity 
\[
X_1 - \theta \sim N(0,1),
\]
the 95% confidence interval is 
\[
I = (X_1 - 1.96, X_1 + 1.96).
\]
Following Neyman the scientist will claim that the random interval \( I \) covers the true \( \theta \) with 95% probability. But following Fisher the scientist may recognize quite a different probability for \( I \) covering \( \theta \). Since \( X_1 \) is not a "sufficient estimate" of \( \theta \), the scientist will have to look for an ancillary complement to \( X_1 \). Suppose \( d = \bar{X} - X_1 \) is chosen as the ancillary complement to \( X_1 \). [Note that \( \bar{X} - X_1 \) is an ancillary statistic and that \( (X_1, \bar{X} - X_1) \) is a sufficient statistic.]

The conditional probability
\[
\Pr(|X_1 - \theta| < 1.96|d, \theta) = \Pr(|\bar{X} - \theta - d| < 1.96|d, \theta) = \Pr(|N(d, 1/n)| < 1.96)
\]
and this is the same as Fisher's fiducial probability that \( \theta \) is in \( I \).

That the choice of the reference set (or the conditioning statistic) can dramatically affect the confidence co-efficient is evident from this example.

Most of my statistical colleagues can never cease to admire the
sheer elegance and simplicity of Neyman-Pearson's confidence statement argument. And then it is so wonderfully easy to construct a p-quantal E. Choose and fix a number \( \alpha \) in \((0,1)\); for each \( \theta \in \Theta \), choose and fix a subset \( E^\theta \) of \( X \) such that \( p(E^\theta | \theta) = \alpha \); and then define \( E \) as

\[
E = \{(\theta, x) : \theta \in \Theta, x \in E^\theta \}.
\]

By construction \( E \) is a p-quantal and so the family \( \{E_x : x \in X\} \) of \( X \)-sections of \( E \) is a 100\( \alpha \) % confidence set estimator of \( \theta \). With \( E_x \) as the confidence set corresponding to the observed sample \( X \), can any evidential meaning be attached to the assertion \( \theta \in E_x \)? Suppose on the basis of sample \( X \) one can construct a 95% confidence interval estimator for the parameter \( \theta \), then does it mean that (the random variable) \( X \) has information on \( \theta \) in some sense?

Anyone who fails to answer both the questions quickly and firmly in the negative is invited to take a look at the following simple example.

**Example 6:** The parameter \( \theta \) lies somewhere in the unit interval \((0,1)\) and the sample \( X \) is a pure randomization variable having the \( \theta \)-free uniform distribution over \((0,1)\). Surely \( X \) has no information on \( \theta \), also no evidential meaning can be attributed to any inference about \( \theta \) based on \( X \). That there is a plentiful supply of 95% confidence intervals for \( \theta \) is seen as follows. Choose and fix any subset \( B \) of \((0,1)\) and then define the quantal \( E \) as the union of the two sets \( B \times (0, .95) \) and \( B^c \times (0.05, 1) \). For each \( \theta \) the section \( E^\theta \) is either \((0, .95) \) or \((0.05, 1) \), and so \( E \) is a p-quantal of size 0.95. The 95% confidence intervals \( \{E_x \} \) based on the p-quantal \( E \) are
\begin{align*}
E_X = \begin{cases} 
B & \text{if } 0 < X \leq 0.05 \\
(0,1) & \text{if } 0.05 < X < 0.95 \\
B^c & \text{if } 0.95 \leq X < 1.
\end{cases}
\end{align*}

Of course it does not make any sense to say that 95% confidence can be placed on the statement \( \theta \in F_X \) irrespective of what \( X \) turns out to be.

As I said in the last paragraph of the previous section, the statistical literature is full of many singular kinds of confidence statements. But has anyone ever dared to base confidence statements on no data, that is, on an uninformative part of the data? Strange as it may appear, the answer is yes. In the next section we briefly consider the Godambe (1979, 1980) contention that the label part of the survey data, despite being an ancillary statistic, is informative by itself.
6. INFORMATION IN THE LABELS

In current survey literature a lot of confusion and controversy surround the notion of the label set, that is, the set of names or label identities of the surveyed units. Following usual practice we denote the survey population as a set \( \{1, 2, \ldots, N\} \) of unit labels and regard the vector \( \theta = (Y_1, Y_2, \ldots, Y_N) \) of unknown characteristics of the \( N \) population units as the universal parameter. The survey design (sampling plan) selects a label set \( s \subset \{1, 2, \ldots, N\} \) with a known selection probability \( p(s) \). The survey fieldwork determines the set \( \{Y_i : i \in s\} \) of the \( Y \)-values of the units in \( s \). The survey data \( X \) then consist of the label set \( s \) and the corresponding set of \( Y \)-values.

For a typical survey design the probability distribution of the label set \( s \) does not involve the parameter \( \theta = (Y_1, Y_2, \ldots, Y_N) \) and so \( s \) is an ancillary statistic. Can we then discard \( s \) as uninformative and marginalize the data \( X \) to the sample set of \( Y \)-values? Since \( s \) is ancillary, should we not invoke Fisher's conditionality principle and condition the data by holding \( s \) fixed? If we factor the likelihood function \( L(\theta) \) as

\[
L(\theta) = Pr(X|\theta) = p(s)Pr(X|s, \theta)
\]

and then discard the factor \( p(s) \) as a \( \theta \)-free constant, then it becomes clear that the survey design is not a determinant of the likelihood scale for the parameter. Invoking the likelihood principle, should we not declare then that at the data analysis stage we need not concern ourselves with any details of the survey design? These are some of the hotly debated issues in survey theory. With his stunning claim
that there are situations where the label set \( s \) can by itself supply a quantum of information on the parameter \( \theta \), Godambe has added a new dimension to the controversy. Let us briefly discuss the Godambe paradox.

Writing \( \gamma \) for the sample of \( Y \)-values, we can denote the survey data as \( X = (s, \gamma) \). Since \( \gamma = \{Y_i: i \in s\} \) and \( \theta = (\gamma_1, \gamma_2, \ldots, \gamma_n) \), it is clear that any statistic \( T = T(s, \gamma) \) can also be represented as a quantity \( Q(\theta, s) \). In particular, any ancillary statistic \( T \) has a dual representation as a \( p \)-quantity \( Q(\theta, s) \) based on the label set \( s \).

**Example 7:** The population consists of 100 individuals labeled as 1, 2, ..., 100. It is known that only one of the 100 individuals is black, the rest being all white. It is also known that the black individual is either 1 or 2. Writing 1 for black and 0 for white, the universal parameter \( \theta \) then takes one or other of the two values \( \theta_1 = (1, 0, 0, 0 \ldots, 0) \) or \( \theta_2 = (0, 1, 0, 0, \ldots, 0) \). Suppose an individual is selected at random (equal probability) from the population and let \( s, y \) be, respectively, the label index and the color value of the sample individual. Note that in this case each of the two co-ordinates of the sample \( X = (s, y) \) is an ancillary statistic, an example of how two ancillary statistics can jointly be sufficient. Also observe that \( y \) gets fully determined once \( s \) and \( \theta \) are known. In other words, the ancillary statistic \( y \) has a dual representation as a \( p \)-quantity \( Q(\theta, s) \). Since \( \Pr(y = 0|\theta) = 0.99 \), the \( (\theta, s) \)-event \( E = \{(\theta, s): Q(\theta, s) = 0\} \) is a \( p \)-quantal of size 0.99. Defining \( E_s \) as the \( s \)-section of \( E \), we have a 99% confidence set estimator of \( \theta \).
that is based on the ancillary statistic \( s \) alone. Note that \( E_1 = \{ \theta_2 \} \). So if a sample individual \( s \) is selected from the population and \( s \) happens to be 1, then, even before \( y \) is observed, Godambe would place 99% confidence on the proposition \( \theta = \theta_2 \). Similarly, \( s = 2 \) will indicate a 99% confidence or plausibility index for the proposition \( \theta = \theta_1 \). For any other value of \( s \) the confidence set \( E_s \) is the whole parameter space \( \{ \theta_1, \theta_2 \} \). Are we supposed to attach a confidence co-efficient of 99% to the proposition \( \theta \in \{ \theta_1, \theta_2 \} \) when \( s = 3 \)?!

The above example is a simplified version of the following example due to Godambe.

**Example 8.** The population consists of four individuals \( \{1, 2, 3, 4\} \) the parameter \( \theta \) is either \( \theta_1 = (-1, 1, -1, 1) \) or \( \theta_2 = (-1, 1, 1, -1) \). A simple random sample of size 2 is drawn. Let \( s = (s_1, s_2) \) be the label set and \( y = (y_1, y_2) \) the sample \( Y \)-values. The statistic \( T = |(y_1 + y_2)/2| \) is ancillary taking the two values 0 and 1 with probabilities 4/6 and 2/6 respectively. Rewriting \( T \) as

\[
Q(\theta,s) = \frac{1}{2} \sum_{i \in s} Y_i,
\]

defining the \( p \)-quantal \( E \) as \( \{(\theta,s): Q(\theta,s) = 0\} \), and then observing that \( E_{1,3} = \{ \theta_2 \} \) Godambe asserts that the partial data \( s = \{1,3\} \), despite being ancillary in the sense of Fisher, is informative (in itself) in that it makes the parameter value \( \theta_2 \) more "plausible" than \( \theta_1 \).

Example 6 of the previous section and the two examples of this section are really the same. They all demonstrate the fact that it is possible to construct sizeable confidence statements based on meaningless data. No evidential meaning can be attached to the confidence statement
argument when it is stated in its pure p-quantal form.

Like a blind man in a dark room groping for a black cat that is probably not there, we, statisticians, are still seeking after the true meaning of statistical information. In this endeavor we are guided (misguided) by our individual intuitions and hearsay knowledge. I wrote this article in the hope of sharing my own intuition and knowledge on the matter of ancillary information with my colleagues. If in that process I have hurt the feelings of any of my friends I am sorry.
REFERENCES


