On Sufficiency and the Exponential Family

By D. A. S. FRASER
University of Toronto

[Received July 1961. Revised January 1962]

SUMMARY

The use of the likelihood function as the minimal sufficient statistic provides a simple and direct proof of results by Barankin and Katz (1959) and Barankin (1961) concerning minimal sufficient statistics. In Section 2 the likelihood function is used to analyse the effect of sampling on the dimensionality of the sufficient statistic and it provides a simple and transparent proof that fixed dimension for the sufficient statistic is restricted to the exponential family. In Section 3 a class of selected exponential distributions is shown to be complete.

1. INTRODUCTION

THE notion of a sufficient statistic was introduced by R. A. Fisher (1920–22). The general theory surrounding such statistics has been investigated by Fisher† and others including Neyman (1935), Darmois (1935), Pitman (1936), Koopman (1936), Halmos and Savage (1949), Lehmann and Scheffé (1950), Dynken (1951), Bahadur (1954), Barankin and Katz (1959), and Barankin (1961). These papers develop criteria for sufficiency, methods of construction for sufficient statistics, and a characterization of those forms of distribution which, for any sample size, admit a sufficient statistic of fixed dimensionality.

The likelihood function appears prominently in the writings of R. A. Fisher both as a means of summarizing data and as a basic entity in mathematical analyses of statistical problems. An important property of the likelihood function, which has only recently received explicit recognition, is that it is a minimal sufficient statistic in its dependence on the observable variable. Interestingly, this property is in part implicit in an early paper by Fisher (1925).

The likelihood function provides one mode of expression for the minimal sufficient statistic. Several of the recent papers on sufficiency, Barankin and Katz (1959) and Barankin (1961), have not explicitly used this mode of expression. In this section some notation and properties of the likelihood function will be recorded for subsequent reference, and some examples will be given of the simplicity of interpretation that results from the use of this mode of expression for the sufficient statistic.

Let x be a variable taking values in an observation space X and let θ be a parameter that indexes the distributions for x and takes values in a parameter space S. Suppose that these distributions are specified by a function $f(x|\theta)$ taken to be the probability density function in the continuous case, the probability function in the discrete case and a combination of these in the mixed case.

† Referee's Note: see especially the 1934 paper "Two new properties of mathematical likelihood" (*Contributions to Mathematical Statistics*, Paper 24). In Section 2.5 the Darmois–Koopman–Pitman expression is incidentally derived, while the first paragraph on p. 289 and the second paragraph on p. 300 are also particularly relevant to the present paper.

The likelihood function $L(\theta|x)$ is a function of θ determined by an outcome x in the following way,

$$L(\theta|x) = cf(x|\theta),$$

and it is left indeterminate to an arbitrary constant factor c that does not depend on θ. Other definitions are possible such as the equivalence class of functions of θ proportional to $f(x|\theta)$, but the definition just given is closest to Fisher's usage and it does lead to suggestive manipulation in concrete problems. In some contexts Fisher uses a different mode of expression for this function and refers to

$$l(\theta|x) = k + \log L(\theta|x)$$

as the likelihood function; in this logarithmic form, the function of θ is left indeterminate to an additive constant and the value $-\infty$ must be permitted for those θ values having $f(x|\theta) = 0$. For a sample of n, the individual likelihood functions combine in the simple way

$$l(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} l(\theta|x_i);$$

and, in fact, for any succession of observations the likelihood functions add:

$$l(\theta,|x, y) = l(\theta|x) + l(\theta|x; y),$$

where $l(\theta|x; y)$ is the likelihood for θ from the conditional distribution of y, given x.

For an outcome x, let S_x, the *selection statistic*, designate the *possible* θ values for that x:

$$S_x = \{\theta|L(\theta|x) > 0\}.$$

If $S_x = S$ for all x then the region of positive density does not depend on the parameter and the problem is sometimes called regular. For a sample of n, the statistics of selection combine in the simple way

$$S_{x_1,\ldots,x_n} = \bigcap_{i=1}^{n} S_{x_i}.$$

The range of the statistic of selection is a class of subsets of the parameter space: the sets S_x and the sets obtained by finite intersections.

Any function of the minimal sufficient statistic is called a necessary statistic by Dynken (1951). Any characteristic of the likelihood function is thus a necessary statistic, and an example is the statistic of selection S_x.

Dynken (1951, Theorem 7) shows that the combination of S_x and an expression which amounts to the likelihood function is a minimal sufficient statistic. But since the statistic S_x is itself a function of the likelihood and is thus redundant in the combination, it follows that Dynken's proof establishes that the likelihood function is a minimal sufficient statistic.

In many contexts it is convenient to have the minimal sufficient statistic expressed in real or vector-valued form. Any coordinate of such a form is a real-valued necessary statistic. Thus the problem of finding a vector-valued form for the sufficient statistic is the problem of finding *the simplest set of real-valued characteristics of the likelihood function that together characterize the likelihood function.*

Consider the case in which the range of positive density does not depend on the parameter: $S_x = S$. The likelihood function $l(\theta|x)$ is then real valued, and of course indeterminate to an additive constant; it follows that it is completely characterized by

all differences $l(\theta_1|x)-l(\theta_0|x)$ with respect to a reference point θ_0. This functional form was used by Pitman (1936) in his analysis of sufficiency. And it was identified as the minimal sufficient statistic by Dynken (1951). Real-valued necessary statistics can be generated by forming any specific difference

$$l(\theta_1|x)-l(\theta_0|x),$$

or any specific derivative

$$\frac{d^r}{d\theta_1^r}l(\theta|x),$$

or if the parameter is vector-valued any mixed partial derivative. One approach then to finding a *simple minimal sufficient statistic is to form collections of differences and derivatives of* $l(\theta|x)$ *in search of a collection that completely characterizes the likelihood function.* For sampling from a normal distribution the derivatives ($\partial l/\partial\mu = \Sigma x_i$, $\partial l/\partial\sigma^2 = \frac{1}{2}\Sigma x_i^2$) at $(\mu, \sigma^2) = (0, 1)$ characterize the likelihood function and thus form the minimal sufficient statistic.

Barankin and Katz (1959) give a detailed study of minimum dimensionality of sufficient statistics. They consider collections of first derivatives of the likelihood function $l(\theta|x)$:

$$\frac{d}{d\theta_1}l(\theta|x),\ldots,\frac{d}{d\theta_r}l(\theta|x),$$

and if θ is vector-valued these derivatives can be with respect to varying coordinates at the points $\theta_1, \ldots, \theta_r$. These are necessary statistics. For a neighbourhood of a point x they examine the rank of the Jacobian of these statistics with respect to the coordinates of the variable x; in doing this they are checking for functional dependence among the statistics, and hence for redundancy. In essence, they then consider forming a sequence of functionally independent statistics of the above form; if at some stage no further independent statistics can be found it follows that the sequence so obtained characterizes the likelihood function and is thus a mode of expression for the minimal sufficient statistic in that neighbourhood. It is of interest that a large amount of their analysis is essentially not directed at sufficiency *per se* since the sufficiency is contained in the likelihood function, but is rather directed at the primarily non-statistical problem of removing redundancy within a large class of necessary statistics. Thus the use of the likelihood function and the discussion on necessary statistics in the preceding paragraph give a simple proof of the main results in the Barankin and Katz paper and in addition separate the statistical aspects from the essentially analytical aspects.

Consider again the case where the range of positive probability density depends on the parameter. Certainly a labelling of the sets S_x is needed as part of any simple form for the minimal sufficient statistic. The discussion in the preceding paragraph then suggests that further necessary statistics be obtained by taking derivatives and differences of $l(\theta|x)$ on S_x in search of a collection that will characterize the likelihood functions that have that value S_x for the selection statistic.

2. The Dimensionality of Sufficient Statistics Under Sampling

The problem of characterizing the distribution forms that yield a sufficient statistic of fixed dimensionality regardless of the sample size was investigated in three almost-contemporary papers by Darmois (1935), Koopman (1936), and Pitman (1936). A fairly compact and general analysis using results concerning linear spaces may be

found in Dynken (1951). In this section a somewhat more general problem will be considered, the problem of relating the behaviour of the sufficient statistic under increasing sample size to the functional form of the distribution being sampled.

Let x be a variable with probability or probability density function $f(x|\theta)$. The log-likelihood function provides an expression for the minimal sufficient statistic:

$$l(\theta|x) = k + \log f(x|\theta).$$

Suppose first that the region of positive density does not depend on the parameter. The indeterminate constant k can then be removed simply by relating to a reference value, say θ_0, on the parameter space:

$$l(\theta|x) = \log f(x|\theta) - \log f(x|\theta_0).$$

For a single observation x the maximal reduction that can be made without losing sufficiency is provided by the transformation from x to $l(\theta|x)$.

Now consider n observations: $(x_1, \ldots, x_n)$. From this, n observations are obtained on the reduced statistic $l(\theta|x)$:

$$(l(\theta|x_1), \ldots, l(\theta|x_n)).$$

The theory in Section 1 then establishes that the maximal reduction is provided by going to the likelihood for the n observations:

$$l(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} l(\theta|x_i).$$

The investigation of sufficiency under sampling thus becomes an investigation of the transformation from

$$(l(\theta|x_1), \ldots, l(\theta|x_n))$$

to

$$\sum_{i=1}^{n} l(\theta|x_i);$$

that is, of the transformation from $(l_1, \ldots, l_n)$ to

$$\sum_{i=1}^{n} l_i,$$

where each l_i can range over the class of real-valued functions $\{l(\theta|x)\,|\,x \in X\}$. Various cases can be distinguished.

First, the class of functions $\{l(\theta|x)\,|\,x \in X\}$ may be contained in a linear space of finite dimension. Let r be the minimum dimension, and let $(\phi_1(\theta), \ldots, \phi_r(\theta))$ be a basis for generating this space from a reference element $\phi_0(\theta)$. Then any element of the class of likelihood functions can be represented as

$$l(\theta|x) = \phi_0(\theta) + \sum_{j=1}^{r} a_j(x)\,\phi_j(\theta),$$

and the probability density function takes the form

$$f(x|\theta) = f(x|\theta_0) \exp\left\{\phi_0(\theta) + \sum_{j=1}^{r} a_j(x)\,\phi_j(\theta)\right\};$$

these densities are said to be of Koopman–Darmois–Pitman form or to be densities of an *exponential* family. For a sample of n the minimal sufficient statistic has the alternative mode of expression,

$$\left(\sum_{i=1}^{n} a_1(x_i), \ldots, \sum_{i=1}^{n} a_r(x_i)\right),$$

and is of dimension r regardless of the sample size (for very small sample sizes some of the coordinates may be redundant).

Second, the class of functions may contain only a countable number of linearly independent functions. As a special case of this the distinct elements in the class $\{l(\theta|x)|x \in X\}$ may be linearly independent; suppose they are arranged in a sequence, numbered and thereby indexed by a variable y. With this mode of expression, n observations $(x_1, \ldots, x_n)$ yield n observations on the likelihood function $(y_1, \ldots, y_n)$. The maximal reduction is obtained by going to the unordered set of likelihood functions $\{y_1, \ldots, y_n\}$. With this mode of expression the minimal sufficient statistic can take as its "value" any set of n positive integers. As n increases this statistic has a range that becomes progressively more complicated.

Third, the class of functions may contain a continuum of functions, no finite set of which is linearly dependent. As a special case these distinct elements in the class $\{l(\theta|x)|x \in X\}$ may be linearly independent. Let y be a continuous real variable that indexes this class. Then the minimal sufficient statistic for a sample of n is the set $\{y_1, \ldots, y_n\}$ of n real variables and the dimensionality of the statistic increases with each increase in the sample size.

More generally for the preceding two cases there will be an increase in the complexity of the minimal sufficient statistic as n increases unless the class

$$\{l(\theta|x) - l(\theta|x_0)|x \in X\}$$

is closed under addition; that is, unless

$$\sum_{i=1}^{n} [l(\theta|x_i) - l(\theta|x_0)]$$

belongs to the class for all $x_1, \ldots, x_n$ and all n.

Again the use of the likelihood function to represent the minimal sufficient statistic has separated out the essential statistical component of the problem thereby reducing it to one on linear spaces. In Dynken the linear space consists of functions of the variable x while here the functions are of the parameter θ. The two approaches relate to one another in the same way as a row vector analysis relates to a column vector analysis in the determination of the rank of a matrix. The analysis here, however, is not restricted to continuous densities having a piece-wise smooth derivative.

Barankin (1961) investigates the minimal sufficient statistic for a sample from an exponential family:

$$f(x|\theta) = f(x|\theta_0) \exp\left\{\phi_0(\theta) + \sum_{j=1}^{r} a_j(x)\phi_j(\theta)\right\}.$$

He gives conditions in his Theorem 4 under which

$$\left(\sum_{i=1}^{n} a_1(x_i), \ldots, \sum_{i=1}^{n} a_r(x_i)\right)$$

represents the minimal sufficient statistic for a sample of size n. The use of the likelihood function and the discussion in Section 1 provide immediate access to Barankin's condition. For consider the likelihood function for a single observation

$$l(\theta|x) = k1 + \sum_{j=1}^{r} a_j(x)\,\phi_j(\theta).$$

The functions $1, \phi_1(\theta), \ldots, \phi_r(\theta)$ generate a linear space containing the functions $l(\theta|x)$. If the functions are linearly dependent, they can be expressed linearly in terms of fewer functions and the expression reduced so that the new functions are linearly independent. In reduced form the functions $1, \phi_1(\theta), \ldots, \phi_r(\theta)$ provide a basis for coordinates of the linear space containing the likelihood functions. Then for a single observation the statistic

$$(a_1(x), \ldots, a_r(x))$$

indexes the likelihood functions and is thereby minimal sufficient, and for a sample of n the statistic

$$\left(\sum_{i=1}^{n} a_1(x_i), \ldots, \sum_{i=1}^{n} a_r(x_i)\right)$$

indexes the likelihood functions and is thereby minimal sufficient.

Barankin supplies the further condition that the functions $a_1(x), \ldots, a_r(x)$ be linearly independent. But again this is concerned with the secondary problem of removing redundancy among coordinates of the statistic and it can be treated along with functional dependence in each problem individually.

To complete this section consider briefly the case having the region of positive density dependent on the parameter and in particular the problem of finite dimension for the minimal sufficient statistic under increasing sample size.

The selection statistic $S_{x_1 \ldots x_n}$ is a necessary statistic. If the sufficient statistic is to have fixed dimensionality then the class

$$\left\{S_{x_1 \ldots x_n} = \bigcap_{i=1}^{n} S_{x_i} \,\middle|\, x_i \in X, \quad n = 1, 2, \ldots\right\}$$

of sets on the parameter space must admit indexing by a finite number of real variables $u_1, \ldots, u_r$: $S(u_1, \ldots, u_r)$. The statistic of selection is then expressible in terms of $u_1(x_1, \ldots, x_n), \ldots, u_r(x_1, \ldots, x_n)$, where

$$S_{x_1 \ldots x_n} = S(u_1(x_1, \ldots, x_n), \ldots, u_r(x_1, \ldots, x_n)).$$

Consider points x having $S_x \supset S(u_1, \ldots, u_r)$ for some value of the selection statistic $(u_1, \ldots, u_r)$. For such points the finite dimensionality of the sufficient statistic requires, by the argument earlier in this section, that

$$l(\theta|x) = b_1(x)\,\phi_1(\theta) + \ldots + b_s(s)\,\phi_s(\theta)$$

for all x having $S_x \supset S(u_1, \ldots, u_r)$ and for all θ in $S(u_1, \ldots, u_r)$. And for

$$S(u'_1, \ldots, u'_r) \supset S(u_1, \ldots, u_r)$$

the likelihood must take the form

$$l(\theta|x) = b'_1(x)\,\phi'_1(\theta) + \ldots + b'_s(x)\,\phi'_s(\theta)$$

for all x in the smaller class having $S_x \supset S(u'_1, \ldots, u'_r)$ and for all θ in the larger class $S(u'_1, \ldots, u'_r)$. These two forms must of course agree on the region of overlap. This,

unfortunately, is not enough to require the family to be the natural generalization of an exponential family—a selected exponential family:

$$f(x|\theta) = f(x)\exp\left\{\phi_0(\theta) + \sum_{j=1}^{r} a_j(x)\,\phi_j(\theta)\right\}\psi(x,\theta),$$

where ψ takes only the values 0 and 1 and determines the region of positive density

$$\psi(x,\theta) = 1, \quad \theta \in S_x,$$
$$= 0, \quad \text{otherwise.}$$

In fact, it seems possible to have quite a complex interplay of the forms for the functions a and ϕ with the values for the statistic S_x. Such interplays would seem, however, to have pattern specifically designed to achieve finite dimension for the sufficient statistic at high cost in terms of pathology of form.

3. A Complete Class of Selected Exponential Densities

The minimal sufficient statistic for a sample of n from a Koopman–Darmois–Pitman model,

$$f(x|\theta) = f(x)\exp\left\{\sum_{j=1}^{r} a_j(x)\,\phi_j(\theta)\right\},$$

is known to be complete if the functions $\phi_1, \ldots, \phi_r$ have an adequate range; see, for example, Lehmann (1959). A natural generalization of this was mentioned briefly at the end of the preceding section; it involved an underlying distribution of exponential form which was restricted to a region of positive density that depended on the parameter. In this section such a *selected exponential* family is shown to have a complete sufficient statistic for any sample size, again provided the range of the natural parameters is large enough.

Let x be a variable with a probability or probability density function

$$f(x|\theta,\eta) = c(\theta,\eta)\exp\left[\Sigma a_j(x)\,\theta_j\right]\psi(x,\eta),$$

where $-\infty < \theta_j < \infty$ and $\psi(x,\eta)$ takes only the values 0 and 1. Let X_η designate the region of positive density

$$X_\eta = \{x|\psi(x,\eta) = 1\},$$

and suppose that the class of such regions of positive densities is closed under the operation of intersection; that is, for any regions $X_{\eta'}$ and $X_{\eta''}$ there is a value η for the parameter of selection such that

$$X_\eta = X_{\eta'} \cap X_{\eta''}$$

or the intersection is empty. For this model, the minimal sufficient statistic

$$\left(\sum_{i=1}^{n} a_1(x_i), \ldots, \sum_{i=1}^{n} a_r(x_i);\ S_{x_1 \ldots x_n}\right)$$

for a sample of n is complete; a proof of this completes this section.

First, it suffices to consider the case $n = 1$. For it is seen that the density function

$$f(x_1, \ldots, x_n|\theta,\eta) = c^n(\theta,\eta)\exp\left[\sum_{i=1}^{n} a_1(x_i)\,\theta_1 + \ldots + \sum_{i=1}^{n} a_r(x_i)\,\theta_r\right]\Pi\psi(x_i,\eta),$$

for a sample of n has the same exponential form as the sampled distribution, and that the selection function $\Pi\psi(x_i, \eta)$ corresponds to derived regions of positive density that must also satisfy the closure property.

Consider the set of sample-space points having a particular value for the selection statistic S_x:

$$\{x \mid S_x = S^*\},$$

where

$$S_x = \{\eta \mid x \in X_\eta\}.$$

Two points x and y belong to such a set if they have $S_x = S_y$; that is, if membership or nonmembership in X_η is the same for x as for y for all η. Thus, if the collection of sets X_η is used to partition the sample space, the ultimate sets of the partition will be precisely those having particular values for the statistic S_x.

Now let $h(a; S)$ be a function of the sufficient statistic $(a_1, \ldots, a_r; S)$ and suppose that it has mean equal to zero for all values of the parameter:

$$\int h(a(x), S_x) f(x \mid \theta, \eta)\, dx = 0$$

for all θ, η. Completeness will be verified by showing that such a statistic $h(a, s)$ must be equal to zero almost everywhere. Let $H(a; \eta)$ be the conditional average of $h(a, S)$ given $a_1(x), \ldots, a_r(x)$. Since the ratio of density functions for two θ values, θ', θ'',

$$\frac{f(x \mid \theta', \eta)}{f(x \mid \theta'', \eta)} = \frac{c(\theta', \eta)}{c(\theta'', \eta)} \exp\left[\Sigma a_j(x)\,(\theta_j' - \theta_j'')\right]$$

does not depend on θ', θ'' for fixed $a_1, \ldots, a_r$, it follows that conditional average $H(a; \eta)$ does not depend on θ.

In terms of H the zero mean for h may be expressed as

$$\int_{X_\eta} H(a(x); \eta)\, c(\theta, \eta) \exp\left[\Sigma a_j(x)\, \theta_j\right] dx = 0$$

for each η; this holds for all θ. By the unicity theorem for multiple Laplace transforms it follows that, for each η,

$$H(a; \eta) = 0$$

for almost all a-values. And for any θ, and in particular for $\theta = 0$, the integral of $H(a; \eta)$ over any set A of a-values must be zero:

$$\int_A H(a(x); \eta)\, dx = 0$$

or, equivalently,

$$\int_A h(a(x), S_x)\, \psi(x, \eta)\, dx = 0$$

$$= \int_{A \cap X_\eta} h(a(x), S_x)\, dx.$$

But the algebra of sets X_η generates, as was shown at the beginning, the algebra of sets of the statistic S_x. Thus, the integral of h over an arbitrary set of a, S values is zero and h itself must be zero almost everywhere.

REFERENCES

BAHADUR, R. R. (1954), "Sufficiency and statistical decision functions", *Ann. math. Statist.*, **25**, 423–462.

BARANKIN, EDWARD W. (1961), "Application to exponential families of the minimum dimensionality problem for sufficient statistics", *B. int. statist. Inst.*, **38**, 141–150.

—— and KATZ, MELVIN (1959), "Sufficient statistics of minimal dimension", *Sankhyā*, **21**, 217–246.

DARMOIS, G. (1935), "Sur les lois de probabilités à estimation exhaustive", *C.R. Acad. sci. Paris*, **200**, 1265–1266.

DYNKEN, E. B. (1951), "Necessary and sufficient statistics for a family of probability distributions", *Uspehi Matem. Nauk.* (N.S.), **6** (1–41), 68–90.

FISHER, R. A. (1920), "A mathematical examination of the method of determining the accuracy of an observation by the mean error and by the mean square error", *M.N.R. astron. Soc.*, **80**, No. 8, 758–770.

—— (1922), "On the mathematical foundations of theoretical statistics", *Phil. Trans. Royal Society London*, A, **222**, 309–368.

—— (1925), "Theory of statistical estimation", *Proc. Camb. phil. Soc.*, **22**, 700–725.

FRASER, D. A. S. (1961), "Invariance and the fiducial method", *Biometrika*, **48**, 261–280.

HALMOS, P. R. and SAVAGE, L. J. (1949), "Application of the Radon–Nikodym theorem to the theory of sufficient statistics", *Ann. math. Statist.*, **20**, 225–241.

KOOPMAN, B. O. (1936), "On distribution admitting a sufficient statistic", *Trans. Amer. math. Soc.*, **39**, 399–409.

LEHMANN, E. L. (1959), *Theory of Hypothesis Testing*. New York: Wiley.

—— and SCHEFFÉ, HENRY (1950), "Completeness, similar regions, and unbiased estimation", *Sankhyā*, **10**, 305–340.

NEYMAN, J. (1935), "Su un teorema concernente le cosiddette statistiche sufficienti", *Inst. Ital. Atti Giorn.*, **6**, 320–334.

PITMAN, E. J. G. (1936), "Sufficient statistics and intrinsic accuracy", *Proc. Camb. phil. Soc.*, **32**, 567–579.