# A NOTE ON A THEOREM OF DYNKIN ON NECESSARY AND SUFFICIENT STATISTICS

Peter Tan

1. Introduction. In his paper "Necessary and sufficient statistics for a family of probability distributions", Dynkin (1951) establishes the important concept of rank for such a family with this conclusion: "If the rank is infinite, then the family has no non-trivial sufficient statistic in any size of sample." His concept of rank is based on a theorem, Theorem 2 described below, which has been pointed out by Brown (1964) to be invalid under its hypotheses. This note shows that Dynkin's Theorem 2 remains valid under its original hypotheses provided that the set (in Dynkin's notation) $\Delta - S$ is countable.

2. Main result. The concept of sufficient statistics introduced first by the late Sir R.A. Fisher in 1921 has until most recently been generally accepted as occupying a very important position in the theory of statistical inference. E.B. Dynkin in 1951 defined the term 'necessary statistic' as a statistic which is a function, or which depends on, every sufficient statistic. A necessary and sufficient statistic is also called a minimal sufficient statistic. Further, as a method of reduction, a statistic $t(x)$ is said to be trivial if it is equivalent to the identity mapping $\epsilon(x) = x$, i.e., if they depend on each other in a subregion of the domain of definition of these functions.

Let $p(x, \theta)$ be the probability density of a family $\Gamma$ of probability distributions defined in a region $X$ of m-dimensional space $R^m$ for an arbitrary parameter $\theta \in \Omega$. $\Gamma$ will be said to be regular in $X$ if $p(x, \theta)$ is a positive continuous function for all $\theta \in \Omega$. Further, $\Gamma$ is said to be piecewise smooth in $X$ if for all $\theta \in \Omega$ there exists an open set $A \subset X$ such that its closure $\overline{A} = \overline{X}$ and if the partial derivatives $\frac{\partial p(x, \theta)}{\partial x_j}$ exist and are continuous in $A$.

The function

$$g_x(\theta) = \ln p(x, \theta) - \ln p(x, \theta_0), \qquad \theta_0 \in \Omega,$$

is shown by Dynkin to be a necessary and sufficient statistic for $\Gamma$ in the regular case. D.A.S. Fraser (1963) has further shown that, considered as the likelihood function, $g_x(\theta)$ is minimal sufficient in both regular or nonregular cases.

The following theorem is due to Dynkin.



https://doi.org/10.4153/CMB-1969-046-5 Published online by Cambridge University Press

THEOREM 2 (Dynkin, 1951). Let the system $\Gamma$ of one-dimensional probability distributions be regular and piecewise smooth in the interval $\Delta$. We shall denote by $L(\Gamma, \Delta)$ the minimal linear space of functions, defined in $\Delta$, consisting of constants and functions $g_x(\theta)$ for any $\theta \in \Omega$. Let the dimension of $L(\Gamma, \Delta)$ be $r+1$ (not excluding $r = \infty$). Then

(a) for every finite $n \leq r$ any sufficient statistic for a sample of size $n$ is trivial;

(b) if the functions $1, \phi_1(x), \ldots, \phi_r(x)$ are bases in $L(\Gamma, \Delta)$ then for any $n \geq r$ the system of functions

$$\Upsilon_i(x_1, \ldots, x_n) = \phi_i(x_1) + \phi_i(x_2) + \ldots + \phi_i(x_n), \quad (i = 1, 2, \ldots, r)$$

is functionally independent and forms a necessary and sufficient statistic for the sample of size $n$.

L. Brown (1964) points out an error in the proof of (a) above and suggests that to insure the validity of the theorem, the density $p(x, \theta)$ be required to be continuously differentiable in $\Delta$.

This theorem (and other results) obtained by Dynkin in its original form is contained in a French book L'Estimation Statistique by D.D. De Rauly (1966). Its corrected form due to Dynkin and Brown is found in Statistical Problems with Nuisance Parameters by J.V. Linnik (In Russian, 1966; English translation, 1968).

O.G. Zhuravlev (1966) has extended this Dynkin's theorem from the case of a random sample of $n$ observations to the case of a sequence of $n$ independent random variables.

We note that there is an additional error in Dynkin's proof of part (b), which is also employed in Zhuravlev's proof of Lemma 3. The incorrect argument in Dynkin's paper is to prove the statement: "If $1, \phi_1(x), \ldots, \phi_s(x)$ are linearly independent functions in $L$, then for $n \geq s$ the system of functions $\Upsilon_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} \phi_i(x_j)$ $(i = 1, 2, \ldots, s)$ is functionally independent." The assertion which is invalid under the hypotheses of Theorem 2 is as follows.

If

$$\phi'(x) = a_1\phi_1'(x) + a_2\phi_2'(x) + \ldots + a_s\phi_s'(x) = 0$$

for all $x$ in $A$, where the $a_i$ do not depend on $x$ and $a_s \neq 0$, and the prime denotes the derivative of a function, then the function

$$\phi(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_s\phi_s(x)$$



https://doi.org/10.4153/CMB-1969-046-5 Published online by Cambridge University Press

is constant in any connected region of the open set $A \subset X$, and therefore takes no more than an enumerable number of different values in $A$. From the continuity of $\phi(x)$ in $\Delta$ and the condition $\overline{A} = \overline{\Delta}$ it follows that $\phi(x)$ is constant in $\Delta$ (Dynkin, p. 25).

This last argument is easily seen to be invalid if we take $s = 1$ and $\phi_1(x)$ to be the Cantor function $f(x)$ defined below over the closed interval $[0, 1]$.

Let $\{E_n^k : k = 1, 2, \ldots, 2^n$ and $n = 0, 1, 2, \ldots\}$ be a countable number of disjoint open intervals in $[0, 1]$, whose union

$$A = \bigcup_{n=0}^{\infty} \bigcup_{k=1}^{2^n} E_n^k$$

is the complement of the Cantor ternary set. Then $A$ is open and $\overline{A} = [0, 1]$.

On $A$ we construct a function $g$ which is a constant over each open interval $E_n^k$:

$$g(x) = \frac{2k - 1}{2^{n+1}} \quad \text{for } x \in E_n^k .$$

Then the function

$$f(x) = \lim_{t \to x} g(t)$$

is a continuous function on the interval $\Delta = [0, 1]$ and its derivative $f'(x)$ is identically zero in the open set $A$. $f$ takes a constant value in each open interval $E_n^k$ but is a monotonically increasing function on $\Delta$ (Monroe, p. 193).

Dynkin's (and Zhuravlev's) error is avoided if we take Brown's suggestion of requiring the density function $p(x, \theta)$ to be continuously differentiable in $\Delta$, as is done by Linnik. The invalid argument of Dynkin's may also become valid if, in addition to the hypotheses of Theorem 2, we require merely that the set $\Delta - A$ be countable. Part (b) of Theorem 2 will then be valid.

With this additional requirement, but without requiring $p(x, \theta)$ to be continuously differentiable in the entire interval $\Delta$, the validity of part (a) of Theorem 2 follows from part (b).



https://doi.org/10.4153/CMB-1969-046-5 Published online by Cambridge University Press

If $1, \phi_1(x), \ldots, \phi_n(x)$ are linearly independent functions in $L$, then it follows from (b) that the functions $\gamma_i(x_1, x_2, \ldots, x_n)$, $(i = 1, 2, \ldots, n)$, are functionally independent in $A^n$. Hence there exists $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, $x_i^0 \in A$, where the Jacobian $\dfrac{\partial(\gamma_1, \ldots, \gamma_n)}{\partial(x_1^0, \ldots, x_n^0)} \neq 0$. In a neighborhood of such a point, $x_i$ are functions of the necessary statistic $(\gamma_1, \ldots, \gamma_n)$ and hence the identity statistic $\varepsilon(x) = (x) = (x_1, x_2, \ldots x_n)$ is also necessary. It follows that every sufficient statistic is trivial.

The error in Dynkin's proof of (a) pointed out by Brown is not a significant one.

3. Remarks. Based on Theorem 2, Dynkin defines the rank of a family $\Gamma$ of distributions in the domain $X$ as the greatest integer $r$, such that for any finite $n \leq r$ the family $\Gamma$ has no non-trivial sufficient statistic for the sample of size $n$ in the domain $X$. He further shows that, under the regularity conditions set forth in Theorem 2, if the family $\Gamma$ has finite rank $r$ in the interval $\Delta$ then $\Gamma$ is an $r$-parameter exponential family.

Fraser (1966) has recently generalized Dynkin's result by considering the structural dimension of a family under weaker assumptions.

4. Acknowledgement. Most of this note is contained in the author's Ph.D. thesis done under the supervision of Prof. D.A.S. Fraser at the University of Toronto with partial support from the National Research Council of Canada.

## REFERENCES

L. Brown, (1964), Sufficient statistics in the case of independent random variables. Annals Math. Statist. 35, 1456-1474.

Daniel Dumas De Rauly, (1966), L'Estimation Statistique. (Gauthier - Villars, Paris).

E.B. Dynkin, (1951), Necessary and sufficient statistics for a family of probability distributions. Selected Transl. Math. Statist. and Prob. 1, 17-40.

D.A.S. Fraser, (1963), On sufficiency and exponential family. J. Roy. Stat. Soc. B, 25, 115-123.

D.A.S. Fraser, (1966), Sufficiency for regular models. Sankhya A, 28, 137-144.

Ju. V. Linnik, (1968), Statistical problems with nuisance parameters. (Amer. Math. Soc., Providence, R.I.)

M.E. Monroe, (1953), Introduction to measure and integration. (Addison-Wesley, Reading, Massachusetts).

P.C. Tan, (1968), R-regular statistical models, sufficiency and conditional sufficiency. (Ph.D. thesis, University of Toronto).

O.G. Zhuravlev, (1966), Minimal sufficient statistics for a sequence of independent random variables. Theor. Probability Appl. XI, 282-291.

University of Toronto

Stanford University



https://doi.org/10.4153/CMB-1969-046-5 Published online by Cambridge University Press