MINIMUM VARIANCE UNBIASED ESTIMATION OF THE DISTRIBUTION FUNCTION ADMITTING A SUFFICIENT STATISTIC

G. P. PATIL AND J. K. WANI

(Received Dec. 2, 1964)

1. Introduction and the main result

Let a real-valued random variable X have the distribution function (df) $F(x; \theta)$ where θ is a scalar or a vector parameter. The situations where one wants to estimate $g(\theta)=F(a; \theta)$ with "a" known on the basis of a random sample of size n, arise in fields of application and have attracted the attention of the statisticians from time to time. Using the theory of transforms, Kolmogorov [2] has investigated some problems of this nature for the normal distribution. Laurent [3] has treated the problem for the two-parameter exponential distribution by using conditional distributions. Patil [4] has provided the results for the generalized power series distribution using power series expansions whereas Tate [6] has solved the problem in generality for distributions with scale and location parameters, where he writes the unbiased estimates as unknown functions in integral equations of the convolution type and recovers them by integral transform methods and further applies the results obtained to a few specific distributions. In this paper, we solve the problem of estimating $g(\theta)=F(a; \theta)$ for several of the important df's $F(x; \theta)$ which admit sufficient statistics by using a uniform technique made available by the existence of the sufficient statistics. We have in particular the following

THEOREM. *Let $x_1, x_2, \cdots, x_n$ be a random sample of size n on the distribution function $F(x; \theta)=P_\theta(X \leqq x)$ where x is real-valued and θ is either a scalar or a vector. Let $t(x_1, \cdots, x_n)$ be a scalar or vector statistic sufficient and complete for θ. Then the conditional distribution function of x_1 given t, $\phi(a, t)=P(x_1, \leqq a \mid t)$, is the minimum variance unbiased estimate of $F(a; \theta)$ where "a" is a known constant.*

PROOF. It is clear that because of the sufficiency of t,

$$\phi(a, t)=P(x_1 \leqq a \mid t)$$

is independent of θ and is therefore a statistic. Also

$$\begin{aligned} E_\theta[\phi(a, t)] &= E_\theta[P(x_1 \leqq a \mid t)] \\ &= P_\theta(x_1 \leqq a) \\ &= F(a; \theta) . \end{aligned}$$

Because of the completeness of t, $\phi(a, t)$ is therefore the minimum variance unbiased estimate of $F(a; \theta)$ as the Rao-Blackwell-Lehmann-Scheffé theorem now applies.

2. The normal distribution function

The probability density function (pdf) of the normal distribution with the vector parameter $\theta=(\mu, \sigma^2)$ is given by

$$f(x; \theta)=\frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left[-(x-\mu)^2/2\sigma^2\right] \qquad -\infty<x<\infty$$

where $-\infty<\mu<\infty$, and $0<\sigma^2<\infty$. Three cases arise:

2.1 *The mean μ unknown, σ known.* Without loss of generality we may take $\sigma=1$. Now the joint pdf of $(x_1, \cdots, x_n)$ is

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= \prod_{i=1}^{n} f(x_i; \theta) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\sum_{i=1}^{n}(x_i-\mu)^2/2\right] \end{aligned}$$

and we know that $t(x_1, \cdots, x_n)=\sum_{i=1}^{n} x_i/n=\bar{x}$ is sufficient and complete for μ. In order to find the conditional distribution of x_1 given t, we note that the joint pdf of $(x_1, \cdots, x_{n-1}, t)$ can be obtained as

$$\begin{aligned} h(x_1, \cdots, x_{n-1}, t)=&\frac{n}{(\sqrt{2\pi})^n} \exp\left[-\sum_{1}^{n-1}(x_i-t)^2/2\right] \\ &\times \exp\left[-\left\{\sum_{1}^{n-1}(x_i-t)\right\}^2/2\right] \exp\left[-n(t-\mu)^2/2\right] \end{aligned}$$

since the Jacobian of the transformation is n.

Since the pdf of t is known to be

$$\frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-n(t-\mu)^2/2\right],$$

we get after some simplification

$$(1) \qquad h(x_1, \cdots, x_{n-1}|t)=\frac{\sqrt{\det(\sigma^{ij})}}{(\sqrt{2\pi})^{n-1}}\exp\left[-\sum_1^{n-1}\sum_1^{n-1}\sigma^{ij}(x_i-t)(x_j-t)/2\right]$$

where $\sigma^{ij}-1=\delta_{ij}$, the Kronecker delta. It is clear from (1) that the conditional distribution of $(x_1, x_2, \cdots, x_{n-1})$ given t is the multivariate normal distribution with the mean vector $(t, t, \cdots, t)$ and the variance-covariance elements σ_{ij} given by $\sigma_{ij}+1/n=\delta_{ij}$, the Kronecker delta, from which follows that the conditional distribution of x_1 given t is normal with mean t and variance $(n-1)/n$. Thus, the minimum variance unbiased estimate of

$$F(a;\mu)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}\exp[-(x-\mu)^2/2]dx$$

is given by

$$(2) \qquad \phi(a, t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{A}\exp[-x^2/2]dx$$

where

$$A=(a-t)/\sqrt{(n-1)/n}\,,$$

a result obtained by Kolmogorov [2] by using a completely different method which involves inversion of the heat equation

$$\frac{\partial\theta}{\partial t}=\frac{\partial^2\theta}{\partial z^2}\,.$$

2.2 *The variance σ^2 unknown, μ known.* Without loss of generality we may take $\mu=0$. We know that $t=\sum_{i=1}^{n}x_i^2$ is a complete sufficient statistic for σ^2. Further, we note that the transformation of $(x_1, \cdots, x_n)$ to $(x_1, \cdots, x_{n-1}, t)$ is two to one with $1/2x_n$ as the Jacobian of transformation. Writing

$$y_1^2=t$$

and

$$y_k^2=y_{k-1}^2-x_{k-1}^2 \qquad k=2, 3, \cdots, n-1$$

we obtain the pdf of $(x_1, \cdots, x_{n-1}, t)$ as

$$h(x_1, \cdots, x_{n-1}, t)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n(t-\sum_1^{n-1}x_i^2)^{-1/2}\exp[-t/2\sigma^2]$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n(y_{n-1}^2-x_{n-1}^2)^{-1/2}\exp[-t/2\sigma^2]\,.$$

The repeated application of the result that

$$\int_{-y_m}^{y_m}(y_m^2-x_m^2)^{k/2}dx_m=y_m^{k+1}B\left(\frac{k+2}{2},\ \frac{1}{2}\right)$$

where $B(m, n)$ is the beta integral with parameters m and n gives

$$h(x_1, t)=\frac{[\Gamma(1/2)]^{(n-1)/2}}{(\sqrt{2\pi}\sigma)^n\Gamma\left(\frac{n-1}{2}\right)}(t-x_1^2)^{(n-3)/2}\exp[-t/2\sigma^2]$$

after integrating out $x_{n-1}, \cdots, x_2$ between their respective ranges. Using the well-known form of the distribution of t and dividing $h(x_1, t)$ by it, we obtain the conditional pdf of x_1 given t as

$$h(x_1\mid t)=\frac{1}{B\left(\frac{n-1}{2},\ \frac{1}{2}\right)}t^{1/2-1}\left(1-\frac{x_1^2}{t}\right)^{(n-1)/2-1}.$$

Noting that the distribution of $x_1^2\mid t$ for a given t is the beta distribution with parameters $(n-1)/2$ and $1/2$, we expect now $\phi(a, t)$ in the form of incomplete beta function. To be specific

$$(3)\qquad \phi(a, t)=\begin{cases}\frac{1}{2}I_A\left(\frac{n-1}{2},\ \frac{1}{2}\right) & \text{if } a<0\\ \frac{1}{2} & \text{if } a=0\\ 1-\frac{1}{2}I_A\left(\frac{n-1}{2},\ \frac{1}{2}\right) & \text{if } a>0\end{cases}$$

where $A=a^2/t$ and $I_A(m, n)=\frac{1}{B(m, n)}\int_0^A(1-x)^{m-1}x^{n-1}dx$ is the incomplete beta function tabulated by Karl Pearson [5].

2.3 *Both μ and σ^2 unknown.* We have

$$t=(t_1, t_2)=(\bar{x}, \Sigma(x_i-\bar{x})^2)$$

as a complete sufficient statistic in this case. Transformation of $(x_1, \cdots, x_n)$ to $(x_1, \cdots, x_{n-2}, t_1, t_2)$ is two to one since the interchange of x_n and x_{n-1} does not alter the image point. The Jacobian of the transformation is available as $\frac{n}{2}(x_n-x_{n-1})^{-1/2}$. Writing $C_k=\sum_1^k(x_i-t_1)^2$ and $B_k=\sum_1^k(x_i-t_1)$ for $0\leqq k\leqq n-2$ and solving $t_1=\sum_1^n x_i/n$ and $t_2=\sum_1^n(x_i-t_1)^2$ for x_{n-1} and x_n, we find the Jacobian

$$(4) \qquad |J| = \left| \frac{n}{2}(2t_2 - 2C_{n-2} - B_{n-2}^2)^{-1/2} \right| .$$

Thus the pdf of $(x_1, \cdots, x_{n-2}, t_1, t_2)$ is available as

$$h(x_1, x_2, \cdots, x_{n-2}, t_1, t_2) = \frac{n}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{t_2}{2\sigma^2} - \frac{n(t_1-\mu)^2}{2\sigma^2} \right] \cdot |J| .$$

We may note that the ranges of $x_1, x_2, \cdots, x_{n-2}$ can be obtained from the fact that the discriminant obtained in solving the equations (4) must be non-negative since x_{n-1} and x_n are both real-valued.

Using successively the easily verifiable result that

$$\int (-ax^2+bx+c)^{k/2} dx = \left(\frac{b^2+4ac}{4a} \right)^{(k+1)/2} \frac{1}{\sqrt{a}} B\left(\frac{k+2}{2}, \frac{1}{2} \right) \qquad a>0$$

where integration is taken over the range

$$(-\sqrt{b^2+4ac}+b)/2a < x < (\sqrt{b^2+4ac}+b)/2a$$

and integrating out $x_{n-2}, \cdots, x_2$ and dividing the expression so obtained by the joint pdf of (t_1, t_2) which is well-known, we get ultimately the conditional pdf of x_1 given (t_1, t_2) as

$$(5) \qquad h(x_1|(t_1, t_2)) = \frac{\sqrt{n}}{B\left(\frac{n-2}{2}, \frac{1}{2} \right)} \frac{1}{\sqrt{(n-1)t_2}} \left[1 - \frac{n(x_1-t_1)^2}{(n-1)t_2} \right]^{(n-2)/2-1} .$$

Now, (5) suggests that the conditional distribution of $\frac{n(x_1-t_1)^2}{(n-1)t_2}$ for given $t=(t_1, t_2)$ is the beta distribution with parameters $(n-2)/2$ and $1/2$ and that the required minimum variance unbiased estimate is available from the equations (3), except that now $A = \frac{n(a-t_1)^2}{(n-1)t_2}$ and that the three cases that arise are according as $a<t_1$, $a=t_1$ and $a>t_1$ respectively. It may be mentioned here that Kolmogorov [2] has investigated a somewhat similar problem for this case through a different approach.

3. One parameter gamma distribution function

Let the random variable x have the gamma distribution with parameters θ and known m with the pdf

$$(6) \qquad f(x, \theta) = \frac{1}{\theta^m \Gamma(m)} x^{m-1} \exp(-x/\theta) \qquad x \geqq 0 .$$

We know that in this case $t=\sum_1^n x_i$ is a complete sufficient statistic for θ and the pdf of $(x_1, \cdots, x_{n-1}, t)$ may be easily obtained as

$$h(x_1, \cdots, x_{n-1}, t)=\frac{1}{\theta^{mn}[\Gamma(m)]^n}\exp(-t/\theta)\prod_1^{n-1} x_i^{m-1}(t-\sum_1^{n-1} x_i)^{m-1}$$

where $t>0$ and with $y_1=t$ and $y_k=y_{k-1}-x_{k-1}$, $0\leqq x_k\leqq y_k$ for $k=1, 2, \cdots, n-1$. Integrating out $x_{n-1}, x_{n-2}, \cdots, x_2$ and dividing the integral by the well-known form of the pdf of t, we obtain the conditional pdf of x_1 given t as

$$h(x_1|t)=\frac{1}{B(mn-m, m)}\left(\frac{x_1}{t}\right)^{m-1}\left(1-\frac{x_1}{t}\right)^{mn-m-1}\frac{1}{t}$$

which brings out that in this case

$$\phi(a, t)=I_A(mn-m, m) \tag{7}$$

where $A=a/t$.

4. Two parameter exponential distribution function

Let the random variable x have the exponential distribution with parameter $\theta=(\alpha, \sigma)$ with the pdf

$$f(x, \theta)=\frac{1}{\sigma}\exp[-(x-\alpha)/\sigma] \qquad x\geqq\alpha\,. \tag{8}$$

Three cases arise:

4.1 *Location parameter α known and scale parameter σ unknown.* Without loss of generality we can assume that $\alpha=0$ in which case (8) becomes a special case of (6).

4.2. *Location parameter α unknown and scale parameter σ known.* Without loss of generality we can assume that $\sigma=1$ and the pdf in (8) now reduces to $\exp[-(x-\alpha)]$. We know that $t=\min(x_1, \cdots, x_n)$ is complete sufficient for α. Further, we note that in this case the range of the distribution depends on its parameter α. It is proved by Huzurbazar [1] that if the range of the distribution with pdf $f(x)$ depends on its parameter α and $\alpha\leqq x\leqq b$ (b known), then the conditional distribution of any sample member x_i, $i=1, 2, \cdots, n$ given $t=\min(x_1, \cdots, x_n)$ is of mixed type and its pdf is given by

$$h(x_i|t)=\begin{cases}\dfrac{n-1}{n}\dfrac{f(x_i)}{\int_t^b f(x)dx} & x_i>t\\[2ex] \dfrac{1}{n} & x_i=t\,.\end{cases}$$

Thus for the case under consideration we have

$$h(x_1|t)=\begin{cases}\left(1-\dfrac{1}{n}\right)\exp\left[-(x_1-t)\right] & x_1>t\\[2ex] \dfrac{1}{n} & x_1=t\end{cases}$$

from which the desired minimum variance unbiased estimate comes out to be

$$(9)\qquad \phi(a,\,t)=\begin{cases}1-\left(1-\dfrac{1}{n}\right)\exp\left[-(a-t)\right] & a>t\\[2ex] \dfrac{1}{n} & a=t\\[2ex] 0 & a<t\,.\end{cases}$$

4.3. *Both α and σ unknown.* This case has been dealt with by Laurent [3] in detail following the approach similar to ours.

5. Two parameter rectangular distribution

Let the random variable x have the rectangular distribution with parameter $\theta=(\alpha,\,\beta)$ with the pdf

$$(10)\qquad f(x,\,\theta)=\frac{1}{\beta-\alpha}\qquad \alpha\leqq x\leqq\beta\,.$$

Two cases arise:

5.1. *Either α or β unknown.* Without loss of generality we can assume α known and known to be zero. Then (10) reduces to $f(x,\,\beta)=1/\beta$, $0\leqq x\leqq\beta$. We know that $t=\max(x_1,\,\cdots,\,x_n)$ is complete sufficient for β and following Huzurbazar [1] we get the conditional pdf of x_1 given t to be

$$h(x_1|t)=\begin{cases}\left(1-\dfrac{1}{n}\right)\dfrac{1}{t} & x_1<t\\[2ex] \dfrac{1}{n} & x_1=t\end{cases}$$

from which

$$\phi(a, t)=\begin{cases}\left(1-\frac{1}{n}\right)\frac{a}{t} & a<t \\ 1 & a\geqq t\,.\end{cases} \tag{11}$$

5.2. *Both α and β unknown.* In this case, $t=(t_1, t_2)$ where $t_1=\min(x_1, \cdots, x_n)$ and $t_2=\max(x_1, \cdots, x_n)$ is complete sufficient for $\theta=(\alpha, \beta)$. Following Huzurbazar [1] we get the conditional pdf of x_1 given t as

$$h(x_1|t)=\begin{cases}\frac{1}{n} & x_1=t_1 \\ \frac{n-2}{n}\cdot\frac{1}{t_2-t_1} & t_1<x_1<t_2 \\ \frac{1}{n} & x_1=t_2\end{cases}$$

which leads to the desired estimate

$$\phi(a, t)=\begin{cases}0 & a<t_1 \\ \frac{1}{n} & a=t_1 \\ \frac{1}{n}+\frac{(n-2)(a-t_1)}{n(t_2-t_1)} & t_1<a<t_2 \\ 1 & a\geqq t_2\,.\end{cases} \tag{12}$$

6. Generalized power series distribution

Let the discrete random variable x have the generalized power series distribution with the series function $f(\theta)=\sum a(x)\theta^x$, summation being taken over a set T of non-negative integers, and hence the probability function is given by

$$p(x, \theta)=\frac{a(x)\theta^x}{f(\theta)} \qquad x\in T\,. \tag{13}$$

It is shown by Patil [4] that $t=\sum_1^n x_i$ is complete and sufficient for θ and has the generalized power series distribution with series function $[f(\theta)]^n=\sum b(t, n)\theta^t$, say. It can be easily seen that the conditional probability function of x_1 given t is given by

$$h(x_1|t)=\frac{\left[\frac{a(x_1)\theta^{x_1}}{f(\theta)}\right]\left[\frac{b(t-x_1,\ n-1)\theta^{t-x_1}}{(f(\theta))^{n-1}}\right]}{\left[\frac{b(t,\ n)\theta^{t}}{(f(\theta))^{n}}\right]} \tag{14}$$

$$=\frac{a(x_1)b(t-x_1,\ n-1)}{b(t,\ n)}$$

from which it follows that for this case

$$\phi(a,\ t)=\sum_{x_1\leq a} h(x_1|t)\ . \tag{15}$$

The reductions of (15) to the special cases of the Binomial, Poisson, Negative Binomial and Logarithmic series distributions are available in Patil [4] where he obtains the results using the identities between different series expansions.

PENNSYLVANIA STATE UNIVERSITY AND MCGILL UNIVERSITY

REFERENCES

[1] V. S. Huzurbazar, "The general forms of distributions admitting sufficient statistics for parameters in nonregular cases," Typescript, 1964.

[2] A. N. Kolmogorov, "Unbiased estimates," *Translations Series 1*, Amer. Math. Soc., 11 (1950), 144-170.

[3] A. G. Laurent, "Conditional distribution of order statistics and distribution of the reduced *i*th order statistic of the exponential model," *Ann. Math. Statist.*, 34 (1963), 652-657.

[4] G. P. Patil, "Minimum variance unbiased estimation and certain problems of additive number theory," *Ann. Math. Statist.*, 34 (1963), 1050-1056.

[5] Karl Pearson, *Tables of Incomplete Beta Function*, Cambridge University Press, 1934.

[6] R. F. Tate, "Unbiased estimation: functions of location and scale parameters," *Ann. Math. Statist.*, 30 (1959), 341-366.