Sankhyā : The Indian Journal of Statistics
1989, Volume 51, Series A, Pt. 2, pp. 205-211.

A STOCHASTIC REPRESENTATION OF THE LOGARITHM OF *P*-VALUES AND RELATED RESULTS

By TAPAS K. CHANDRA

Indian Statistical Institute

SUMMARY. A crucial lemma of Ghosh (1971) is used to give a novel stochastic representation of *P*-values. Further examples where the lemma can be fruitfully used are also given. The results obtained here extend the earlier results of other authors.

1. Lemmas

It is assumed that the reader is familiar with the papers Lambert and Hall (1982) and Bahadur, Chandra and Lambert (1983); these papers will be referred to as [LH] and [BCL] respectively. *Throughout the paper, t_1 and t_2 will denote two real numbers such that $t_1 > t_2$.*

Lemma 1 : *If $\{X_n\}$ and $\{Y_n\}$ are two sequences of random variables defined on the same probability space satisfying the conditions*

(a) *$\{Y_n\}$ is stochastically bounded ;*

(b) *for each t_1 and t_2,*

$$P(X_n > t_1, Y_n < t_2) \to 0 \text{ as } n \to \infty,$$

Then

$$P(X_n - Y_n \geq \epsilon) \to 0 \text{ as } n \to \infty \text{ for each } \epsilon > 0.$$

Proof : Let $\epsilon > 0$ and $\delta > 0$. It suffices to show that

$$\limsup_{n \to \infty} P(X_n - Y_n \geq \epsilon) < \delta. \qquad \ldots \quad (1)$$

By Condition (a), there exists a $\lambda > 0$ such that

$$P(|Y_n| \geq \lambda) < \delta \quad \text{for each} \quad n \geq 1.$$

Divide the interval $(-\lambda, +\lambda]$ into m subintervals $(a_1, a_2], (a_2, a_3], \ldots, (a_m, a_{m+1}]$ such that $a_1 = -\lambda$, $a_{m+1} = +\lambda$ and $a_{i+1} - a_i < \epsilon/2$ for each $i = 1, \ldots, m$; note that m does not depend on n.

Then for each n and i,

$$\{X_n - Y_n \geq \epsilon, a_i < Y_n \leq a_{i+1}\} \subset \{X_n > a_{i+1} + \epsilon/2, Y_n < a_{i+1} + \epsilon/4\}. \quad \ldots \quad (2)$$

AMS (1970) *subject classification :* Primary 62F20 ; Secondary 62G20.

Key words and phrases : *P*-value ; Bahadur efficiency ; Asymptotic normality ; Exponential family.

Hence

$$P(X_n - Y_n \geqslant \epsilon)$$
$$\leqslant \delta + P(X_n - Y_n \geqslant \epsilon, |Y_n| < \lambda)$$
$$\leqslant \delta + \sum_{i=1}^{m} P(X_n - Y_n \geqslant \epsilon, a_i < Y_n \leqslant a_{i+1})$$

Letting $n \to \infty$, using (2) and Condition (b), we get (1).

The following lemma is due to Ghosh (1971).

Lemma 2 : *Let $\{X_n\}$ and $\{Y_n\}$ be as in Lemma* 1. *Assume furthermore that for each t_1, t_2*

$$P(X_n < t_2, Y_n > t_1) \to 0 \text{ as } n \to \infty.$$

Then $X_n - Y_n$ converges to 0 *in probability.*

Lemma 3 : *Let $\{d_{1n}\}$ and $\{d_{2n}\}$ be two sequences of positive reals such that $d_{1n} \to \infty$ and $\{d_{2n}\}$ any sequence of reals. Let $\{d_{1n}(X_n - b)\}$ be stochastically bounded and $H_n : R^1 \to R^1$ be an increasing function satisfying the condition that there exist $a_1 \in R^1$, $a_2 > 0$ and an open interval I containing b such that for each real k,*

$$H_n(b + k/d_{1n}) = d_{2n}\, a_1 + d_{3n}\, k a_2 + o(d_{3n}).$$

Then

$$H_n(X_n) = d_{2n}\, a_1 + d_{1n}\, d_{3n}\, a_2 (X_n - n) + o_p(d_{3n}).$$

Proof : Put

$$M_n = (H_n(X_n) - d_{2n}\, a_1)/(d_{3n}\, a_2).$$

We shall show, using Lemma 2, that

$$M_n - d_{1n}(X_n - b) \xrightarrow{P} 0.$$

To this end, fix t_1 and t_2 and let $m \geqslant 1$ be such that for each $n \geqslant m$ the interval I contains $b + t_1/d_{1n}$. Note that for $n \geqslant m$,

$$d_{1n}(X_n - b) > t_1$$
$$\Rightarrow H_n(X_n) \geqslant H_n(b + t_1/d_{1n})$$
$$\Rightarrow M_n \geqslant (H_n(b + t_1/d_{1n}) - d_{2n}\, a_1)/(d_{3n}\, a_2) = t_1 + o(1)$$
$$\geqslant t_2 \text{ for all sufficiently large } n$$

(since $t_1 > t_2$) ; thus for all sufficiently large n,

$$\{M_n < t_2, d_{1n}(X_n - b) > t_1\} = \phi.$$

Similarly, for all sufficiently large n,

$$\{M_n > t_1, d_{1n}(X_n - b) < t_2\} = \phi.$$

Lemma 4 : *If $\{X_n\}$ and $\{Y_n\}$ are two sequences of random variables defined on the same probability space such that*

$$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{d} X$$

and for each t_1, t_2

$$P(X_n > t_1,\ Y_n < t_2) \to 0 \text{ as } n \to \infty,$$

then $(X_n, Y_n) \xrightarrow{d} (X, X)$.

Proof : By Lemma 1, $P(X_n - Y_n \geqslant \epsilon) \to 0$ as $n \to \infty$ for each $\epsilon > 0$. Now arguments used in the proof of Proposition 2.5, [BCL] complete the proof.

Lemma 5 : *Let F be the distribution function of a random variable X. Then*

$$u \leqslant P(F(X-) < u), \text{ for each } u \in [0, 1].$$

Lemma 6 : *Let $\{F_n\}$ be a sequence of distribution functions converging to a distribution function F weakly. If $x_n \to x$, then*

$$F(x-) \leqslant \liminf F_n(x_n) \leqslant \limsup F_n(x_n) \leqslant F(x).$$

Proofs of Lemmas 5 and 6 are well-known and are omitted.

2. Applications

We first obtain a stochastic representation (see Part (a) below) of the logarithms of P-values which may be regarded as a useful novel extension of Lemma 4.1 of [LH]. [LH] shows that the asymptotic distribution of the normalised logarithms of P-values is lognormal under certain condition ; our stochastic representation clearly points out why it is so, and further it also points out that other limiting distributions are also possible if the condition of the asymptotic normality is replaced by similar other conditions.

Let $\{X_n\}$ be a sequence of random variables defined on $(\Omega, \mathcal{A})$ and the P_θ be a probability measure on $(\Omega, \mathcal{A})$ for each θ in Θ ; let Θ_0 be a proper non-empty subset of Θ. Let

$$G_n(t) = \sup\{P_\theta(X_n \geqslant t) : \theta \text{ in } \Theta_0\},$$
$$L_n = G_n(X_n).$$

Fix a θ in $\Theta - \Theta_0$.

Theorem 1 : *Under the above set-up assume that*

(i) $\{n^{1/2}(X_n - b(\theta))\}$ *is stochastically bounded under θ,*

(ii) *there exist* $a_1(\theta) \in R^1$, $a_2(\theta) > 0$ *and an open interval* I *containing* $b(\theta)$ *such that for each* t *in* R^1,

$$-\log G_n(b(\theta)+t/n^{1/2}) = n\,a_1(\theta)+n^{1/2}\,t\,a_2(\theta)+o(n^{1/2}).$$

Then

(a) $-\log L_n = n\,a_1(\theta)+n\,a_2(\theta)\,(X_n-b(\theta))+o_p(n^{1/2})$.

(b) $\sqrt{n}(X_n-b(\theta))$ *converges weakly iff* $\dfrac{-\log L_n - n\,a_1(\theta)}{\sqrt{n}\,a_2(\theta)}$ *converges weakly, in which case the limits are same.*

Proof: Use Lemma 3 with $H_n = -\log G_n$.

Remark 1: Let $X_{n,i}$, $i = 1, \ldots, k$, be sequences of random variables defined on $(\Omega, \mathcal{A})$ such that under θ,

$$(n^{1/2}(X_{n,1}-b_1(\theta)), \ldots, n^{1/2}(X_{n,k}-b_k(\theta)))$$

converges in distribution to the distribution function F on R^k. Let $L_{n,i}$ be defined as above with X replaced by $X_{n,i}$. Then the random vector $(M_{n,1}, \ldots, M_{n,k})$ converges in distribution to F under θ, where

$$M_{n,i} = (-\log L_{n,i} - n\,a_1(\theta))/(n^{1/2}\,a_2(\theta)),\ i = 1, \ldots, k.$$

Lemma 4.1 of [LH] follows as a special case.

Remark 2: The above definitions of G_n and the $G_{n,i}$ have not been fully used; the fact that they are decreasing functions is only required.

We shall now consider some results of [BCL]. Let $\{\mathcal{B}_n\}$ be a sequence of sigma-fields such that $\mathcal{B}_n \subset \mathcal{A}$ for each $n \geqslant 1$. Let F be a probability distribution function on R^1. Consider the following assumptions.

Assumption (A): *There exist a real* $\nu := \nu(\theta)$, *a* $\gamma := \gamma(\theta)$ *in* Θ_0 *and a positive integer* $m := m(\theta)$ *such that* $P_\theta \ll P_\gamma$ *on* $\mathcal{B}_n$ *for each* $n \geqslant m$ *and*

$$R_n := n^{-1/2}\,(\log r_{n,\theta,\gamma} - n\,\nu) \qquad \ldots\ (3)$$

is stochastically bounded under θ, *where* $r_{n,\theta,\gamma}$ *is a finite non-negative* $\mathcal{B}_n$*-measurable function satisfying*

$$P_\theta(dw) = r_{n,\theta,\gamma}\,(w)\,P_\gamma\,(dw) \text{ on } \mathcal{B}_n.$$

We say that Assumption (B) *holds if Assumption* (A) *holds and under* θ, R_n *converges in distribution to* F.

It follows from Lemma 5 that $P_\theta(L_n \leqslant \alpha) \leqslant \alpha$ for each $n \geqslant 1$ and $\alpha \geqslant 0$; hence the following inequality holds as shown in [BCL]:

$$P_\theta(L_n \leqslant \alpha, r_{n,\theta,\tau} \leqslant k) \leqslant k\,\alpha \qquad \ldots \quad (4)$$

for each $k \geqslant 0$ and $\alpha \geqslant 0$.

The following theorems are useful extensions of Propositions 2.1, 2.2, 2.8 of [BCL].

Theorem 2(a): *Under Assumption* (A), *for each* $\epsilon > 0$

$$P_\theta(\log L_n + \log r_{n,\theta,\tau} \leqslant -n^{1/2}\,\epsilon) \to 0 \text{ as } n \to \infty.$$

(b) *Under Assumption* (B), *for each sequence* $z_n \to z$

$$\liminf_{n\to\infty} P_\theta\,(M_n < z_n) \geqslant F(z-)$$

where

$$M_n = (-\log L_n - n\nu)/n^{1/2}. \qquad \ldots \quad (5)$$

Proof: (a) We shall use Lemma 1. Then

$$\begin{aligned} &P_\theta(M_n > t_1,\ R_n < t_2) \\ &\quad = P_\theta(L_n < \exp(-n\nu - n^{1/2}\,t_1),\ r_{n,\theta,\tau} < \exp(n\nu + n^{1/2}\,t_2)) \\ &\quad \leqslant \exp(-n^{1/2}\,(t_1 - t_2)) \quad \text{by (4)} \\ &\quad \to 0 \text{ as } n \to \infty \text{ (since } t_1 > t_2). \end{aligned}$$

(b) Put $\alpha_n = \exp(-n\nu - n^{1/2}\,z_n)$ and $k_n = (n\,\alpha_n)^{-1}$. Then

$$P_\theta\,(M_n < z_n) \geqslant P_\theta(R_n \leqslant z_n - n^{1/2}\log n) - 1/n \qquad \text{by (4)}.$$

In view of Lemma 6 and Assumption (B), the desired inequality follows.

Remark 3: In [BCL], it is assumed that there exists a unique γ such that the following infimum is attained at γ;

$$\nu = \inf\{K(\theta\,;\,\theta_0) : \theta_0 \text{ in } \Theta_0\}$$

where $K(\theta\,;\,\theta_0)$ is a Kullback-Liebler information number. Suppose that for each θ in Θ, P_θ is the linear exponential family with the density (with respect to some sigma-finite measure)

$$f(x\,;\,\theta) = h(x)\exp(\theta'\,x - c(\theta)),$$

$\Theta \subset R^k$ being the associated natural parameter space.

Let $\Theta_0 = \{\theta \in \Theta : \theta^{(1)} = 0\}$; here $\theta^{(1)}$ stands for the vector consisting of the first p components of θ, $1 \leqslant p < k$. Then

$$K(\theta\,;\,\theta_0) = c(\theta_0) - c(\theta) + \theta'\,\nabla\,c(\theta) - \theta_0'\,\nabla c(\theta_0)$$

where $\nabla c(\theta)$ is the gradient of $c(\theta)$. It, therefore, follows that if $\tilde{\theta}(x)$ is the maximum likelihood estimator of the last $(k-p)$ components of θ in Θ_0, then γ is given by

$$\gamma^{(1)} = 0,\ \gamma^{(2)} = \tilde{\theta}\,(E_\theta(\bar{X}^{(2)}))\ ;$$

here $\bar{X}^{(2)}$ is the vector consisting of the last $(k-p)$ components of the sample mean vector $\bar{X}$. For illustrations of this simple fact, one may consult the examples discussed in Koziol (1978).

Theorem 3 : *Suppose that Assumption* (B) *holds and* M_n *converges in distribution to* G. *If* F *and* G *are both symmetric about zero, then* $F = G$.

Proof : By Theorem 2(*b*),

$$G(z) \geqslant F(z-) \text{ for each real } z.$$

As F and G are symmetric about zero, F and G must be identical.

Remark 4 : The condition 'both F and G are symmetric about zero' in Theorem 3 can be replaced by other conditions ; consider, e.g., the condition 'there exists a positive real σ such that $G(z) = F(\sigma z)$ for each real z and F is strictly increasing', or the condition 'the means of F and G are finite and equal'.

Theorem 4 : *Suppose that Assumption* (*B*) *holds and that* $\{X_n\}$ *estisfies the following conditions* (*i*) *and* (*ii*).

(i) $P_\theta(R_n > t_1,\ \sqrt{n}(X_n - \nu) < t_2) \to 0$

(ii) *For all sufficiently large* n, *there exists a real-valued function* g_n *such that*

$$n^{-1} \log G_n(t) \leqslant -t + g_n(t) \text{ for each real } t$$

and

$$\limsup_{n \to \infty} h_n(k, \theta) \leqslant 0,$$

$$h_n(k\,;\theta) := \sup\{n^{1/2}\, g_n(t) : n^{1/2}\,|t-\nu| < k\}.$$

Then (a) $n^{1/2}\,(X_n - n^{-1} \log r_{n,\theta,\gamma}) \xrightarrow{P_\theta} 0$ *and* (*b*) *under* θ, M_n *converges to* F *in distribution.*

Proof : By Lemma 1, we have for each $\epsilon > 0$

$$P_\theta(n^{-1} \log r_{n,\theta,\gamma} - \bar{X}_n \geqslant n^{-1/2}\,\epsilon) \to 0.$$

Now use the arguments of the proof of Proposition 2.8 of [BCL].

Remark 4 : Suppose that M_n converges to F in distribution and Assumption (B) holds. Then

$$M_n - R_n \xrightarrow{P_\theta} 0 \text{ as } n \to \infty.$$

To prove this, note that by Theorem 2(a),

$$P_\theta(M_n - R_n \geqslant \epsilon) \to 0 \text{ as } n \to \infty$$

for each $\epsilon > 0$. An application of Lemma 4 completes the proof.

References

Bahadur, R. R., Chandra, T. K. and Lambert, D. (1983) : Some further properties of likelihood ratios on general sample spaces. *Statistics, Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* (Eds. Ghosh, J. K. and Roy, J.), 1-19.

Ghosh, J. K. (1971) : A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.*, **42**, 1957-1961.

Koziol, J. (1978) : Exact slopes of certain multivariate tests of significance. *Ann. Statist.*, **6**, 546-558.

Lambert, D. and Hall, W. J. (1982) : Asymptotic lognormality of *P*-values. *Ann. Statist.*, **10**, 44-64.