SOME INEQUALITIES RELATING TO THE PARTIAL SUM OF BINOMIAL PROBABILITIES

By MASASHI OKAMOTO

(Receved June 2, 1958)

1. Introduction.

Uspensky [1, p. 102] gives an inequality relating to the partial sum of binomial probabilities: Let X be a random variable following a binomial distribution $B(n, p)$, arising from n repetitions of an event with probability p. Then it holds that

$$P(|X/n-p|\geqq c)<2e^{-nc^2/2}$$

for any constant $c>0$ and any p with $0<p<1$. Its proof, however, is too tedious, although elementary. In the following we shall give a simplified proof for a somewhat strengthened result (Theorem 1). By the same method we can also obtain some other inequalities which prove to be useful in Matusita's theory [2, 3] of test of fit, two-sample problem, test of independence, etc.

2. Two lemmas

We shall state two lemmas the first of which is a corollary of a theorem given by Chernoff (Theorem 1 in [4]).

LEMMA 1. *Let X be a random variable following $B(n, p)$ and x a constant, $0\leqq x\leqq 1$, which may depend on n or p. It holds then that*

(i) $P(X/n\geqq x)\leqq e^{-n\varphi(x)}$ *if* $x\geqq p$, *and*

(ii) $P(X/n\leqq x)\leqq e^{-n\varphi(x)}$ *if* $x\leqq p$,

where

$$\varphi(x)=x\log\frac{x}{p}+(1-x)\log\frac{1-x}{q}$$

and $q=1-p$.

LEMMA 2. *The function $\varphi(x)$ defined in Lemma 1 satisfies the following inequalities:*

(a) $\varphi(x)\geqq 2(x-p)^2$ *if* $0\leqq x\leqq 1$,

(b) $\varphi(x) \geqq \frac{(x-p)^2}{2pq}$ *if* $p \leqq x \leqq 1,\ p \geqq \frac{1}{2}$

(b′) $\varphi(x) \geqq \frac{(p-x)^2}{2pq}$ *if* $0 \leqq x \leqq p,\ p \leqq \frac{1}{2}$

(c) $\varphi(x) \geqq 2(\sqrt{x}-\sqrt{p})^2$ *if* $p \leqq x \leqq 1$,

(d) $\varphi(x) \geqq (\sqrt{p}-\sqrt{x})^2$ *if* $0 \leqq x \leqq p$,

where the equality sign holds in each case if and only if $x=p$.

PROOF. First we have

$$\varphi'(x)=\log\frac{x}{p}-\log\frac{1-x}{q}\ , \qquad \varphi''(x)=\frac{1}{x(1-x)} \geqq 0\ , \tag{1}$$

consequently

$$\varphi(p)=\varphi'(p)=0\ . \tag{2}$$

For the proof of (a), put $\varphi_1(x)=2(x-p)^2$. Then

$$\varphi_1(p)=\varphi_1'(p)=0 \tag{3}$$

and

$$\varphi_1''(x)=4 \leqq \varphi''(x) \quad \text{if } 0 \leqq x \leqq 1\ , \tag{4}$$

where the equality holds at a single point $x=1/2$. From (2), (3) and (4) we obtain

$$\varphi_1(x) \leqq \varphi(x) \quad \text{if } 0 \leqq x \leqq 1\ ,$$

with the equality sign only for $x=p$.

Concerning (b) and (b′), putting $\varphi_2(x)=\frac{(x-p)^2}{2pq}$, we can prove them similarly.

Re (c). Put $\varphi_3(x)=2(\sqrt{x}-\sqrt{p})^2$.

Re (d). The proof of this case is most lengthy. We shall first prove

$$\varphi(p-\sqrt{p}\,c) \geqq c^2/2 \quad \text{if } 0 \leqq c \leqq \sqrt{p}\ . \tag{5}$$

If $p \leqq 1/2$, then (b′) implies

$$\varphi(p-\sqrt{p}\,c) \geqq \frac{c^2}{2q} \geqq \frac{c^2}{2}$$

and if $p \geqq 1/2$, then (a) implies

$$\varphi(p-\sqrt{p}\,c) \geqq 2pc^2 \geqq c^2 \geqq c^2/2 .$$

Then we have (5).

Now we consider two cases :

Case (i) where x satisfies

$$(\sqrt{2}-1)^2 p \leqq x \leqq p . \tag{6}$$

Put $c=\sqrt{p}-\sqrt{x}$. Then (6) is equivalent to $0 \leqq c \leqq (2-\sqrt{2})\sqrt{p}$, which implies

$$x=(\sqrt{p}-c)^2 \leqq p-\sqrt{2p}\,c . \tag{7}$$

By (1) and (2) $\varphi(x)$ decreases monotonically in the interval $0 \leqq x \leqq p$. Therefore (5) and (7) give

$$\varphi(x) \geqq \varphi(p-\sqrt{2p}\,c) \geqq c^2 ,$$

which is (d) for the case (i).

Case (ii) where x satisfies

$$0 \leqq x \leqq (\sqrt{2}-1)^2 p . \tag{8}$$

If we define the function $\psi(x)$ as

$$\psi(x)=\varphi(x)-(\sqrt{p}-\sqrt{x})^2 , \tag{9}$$

then its first two derivatives are

$$\psi'(x)=\log\frac{x}{p}-\log\frac{1-x}{q}-\left(1-\sqrt{\frac{p}{x}}\right) , \tag{10}$$

$$\psi''(x)=\frac{1}{2\sqrt{x^3}\,(1-x)}\{2\sqrt{x}-\sqrt{p}\,(1-x)\} . \tag{11}$$

Since the formula in the braces of (11) increases monotonically for $x \geqq 0$ and its value at $x=(\sqrt{2}-1)^2 p$ is easily seen to be non-positive, we obtain for any value of x in the interval (8)

$$\psi''(x) \leqq 0 .$$

Since we have from (9) and (10)

$$\psi(0)=-\log q-p\geqq 0 \text{ and } \psi'(0)=\infty ,$$

we have only to show

$$\psi((\sqrt{2}-1)^2p)\geqq 0 \qquad \text{if } 0\leqq p\leqq 1 \tag{12}$$

in order to prove $\psi(x)\geqq 0$ for any x in (8). Now, from (9) we have

$$\psi((\sqrt{2}-1)^2p)=[1-(\sqrt{2}-1)^2p]\log\frac{1-(\sqrt{2}-1)^2p}{q}$$
$$+2(\sqrt{2}-1)^2[\log(\sqrt{2}-1)-1]p=\zeta(p)\ (say) .$$

The function $\zeta(p)$ is defined in $0\leqq p\leqq 1$. Since $\zeta(0)=0$, in order to prove (12) or $\zeta(p)\geqq 0$ it suffices to verify

$$\zeta'(p)\geqq 0 \qquad \text{for } 0\leqq p\leqq 1 . \tag{13}$$

The derivative of $\zeta(p)$ can be expressed as

$$\zeta'(p)=(\sqrt{2}-1)^2[2\log(\sqrt{2}-1)-3]-(\sqrt{2}-1)^2\log\xi(p)+\xi(p) , \tag{14}$$

where

$$\xi(p)=\frac{1-(\sqrt{2}-1)^2p}{q} .$$

The function $\xi(p)$ defined in $0\leqq p\leqq 1$ is clearly monotone-increasing and therefore

$$\xi(p)\geqq\xi(0)=1, \qquad 0\leqq p\leqq 1 , \tag{15}$$

which implies

$$\log\xi(p)\leqq\xi(p)-1, \qquad 0\leqq p\leqq 1 . \tag{16}$$

Finally (14), (15) and (16) together imply (13), for

$$\zeta'(p)\geqq(\sqrt{2}-1)^2[2\log(\sqrt{2}-1)-3]-(\sqrt{2}-1)^2[\xi(p)-1]+\xi(p)$$
$$\geqq(\sqrt{2}-1)^2[2\log(\sqrt{2}-1)-3]+1>0 .$$

It will readily be seen that the equality condition in (d) is $x=p$. This completes the proof of Lemma 2.

3. Theorems

Let X be a binomial variate with $B(n,p)$, $0<p<1$, and c a non-negative constant depending possibly on n or p. From Lemmas 1 and 2 in the preceding section we have readily the following theorems.

THEOREM 1

(i)
$$P\left(\frac{x}{n}-p\geqq c\right)<e^{-2nc^2},$$

(ii)
$$P\left(\frac{x}{n}-p\leqq -c\right)<e^{-2nc^2}.$$

THEOREM 2

(i)
$$P\left(\frac{x}{n}-p\geqq c\right)<\exp\left(-\frac{nc^2}{2pq}\right) \qquad \textit{for } p\geqq\frac{1}{2},$$

(ii)
$$P\left(\frac{x}{n}-p\leqq -c\right)<\exp\left(-\frac{nc^2}{2pq}\right) \qquad \textit{for } p\leqq\frac{1}{2}.$$

THEOREM 3

$$P\left(\sqrt{\frac{x}{n}}-\sqrt{p}\geqq c\right)<e^{-2nc^2}.$$

THEOREM 4

$$P\left(\sqrt{\frac{x}{n}}-\sqrt{p}\leqq -c\right)<e^{-nc^2}.$$

We note that the equality signs for $c=0$ which are to be present in these formulas in applying Lemmas 1 and 2 are absent there. This is justified by the direct consideration of properties of the binomial distribution, where we restrict p in the open interval $0<p<1$.

4. Application to Matusita's multionomial distance.

Let F be a multinomial distribution with k classes and a set of probabilities $(p_1, \cdots, p_k)$, $p_i>0$, $\sum p_i=1$, and let S_n be an empirical distribution with relative frequencies $(n_1/n, \cdots, n_k/n)$, $(\sum n_i=n)$. Matusita [2], [3] defined the distance between S_n and F by the formula

(17)
$$\|S_n-F\|^2=\sum_{i=1}^{k}\left(\sqrt{\frac{n_i}{n}}-\sqrt{p_i}\right)^2.$$

which we shall refer to as Matusita's multinomial distance. He and M. Motoo [5] proved that

(18)
$$P(\|S_n-F\|^2\geqq\eta^2)\leqq\frac{k^2+k-1}{(n\eta^2)^2}$$

for any positive constant η. Now we obtain from Theorems 3 and 4 the following

THEOREM 5

$$P(\|S_n - F\|^2 \geqq \eta^2) < k\left\{\exp\left(-\frac{2n\eta^2}{k}\right) + \exp\left(-\frac{n\eta^2}{k}\right)\right\}. \tag{19}$$

PROOF. Clearly

$$P(\|S_n - F\|^2 \geqq \eta^2) \leqq \sum_{i=1}^{k} P\left(\left|\sqrt{\frac{n_i}{n}} - \sqrt{p_i}\right| \geqq \frac{\eta}{\sqrt{k}}\right).$$

Since for each i the random variable n_i is distributed according to $B(n, p_i)$, we have from Theorems 3 and 4

$$P\left(\sqrt{\frac{n_i}{n}} - \sqrt{p_i} \geqq \frac{\eta}{\sqrt{k}}\right) < \exp\left(-\frac{2n\eta^2}{k}\right),$$

$$P\left(\sqrt{\frac{n_i}{n}} - \sqrt{p_i} \leqq -\frac{\eta}{\sqrt{k}}\right) < \exp\left(-\frac{n\eta^2}{k}\right),$$

whence the required inequality follows.

We shall compare our result (19) with that of Matusita and Motoo (18), that is, we shall ask which of

$$A = \frac{k^2 + k - 1}{(n\eta^2)^2} \quad \text{and} \quad D = k(e^{-2n\eta^2/k} + e^{-n\eta^2/k})$$

is better (smaller in value). If we put $A' = k^2/(n\eta^2)^2$, which is better than A, it holds identically

$$D = k(e^{-2/\sqrt{A'}} + e^{-1/\sqrt{A'}}).$$

Now we mention two examples of the comparison of A and D: For $A' = 1/25$

$$D = k(e^{-10} + e^{-5}) \leqq 1/25 = A' \leqq A \qquad \text{if } k \leqq 5,$$

and for $A' = 1/100$

$$D = k(e^{-20} + e^{-10}) \leqq 1/10 = A' \leqq A \qquad \text{if } k \leqq 220.$$

Though the comparison depends on k, the number of classes, if A is around or below 0.01, then D is seen to be better than A in almost all practical cases.

OSAKA UNIVERSITY

REFERENCES

[1] J. V. Uspensky, *Introduction to Mathematical Probability*, New York, 1937.

[2] K. Matusita, Decision rules based on the distance for problems for fit, two samples and

estimation, *Ann. Math. Stat.*, Vol. 26 (1955), pp. 631-640.

[3] Matusita and H. Akaike, Decision rules, based on the distance, for the problems of independence, invariance and two samples, *Ann. Inst. Stat. Math.*, Vol. 7 (1956), pp. 67-80.

[4] H. Chernoff, A measure of asymoptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Stat.*, Vol. 23 (1952), pp. 493-507.

[5] K. Matusita and M. Motoo, On the fundamental theorem for the decision rule based on distance || ||, *Ann. Inst. Stat. Math.*, Vol. 7 (1956), pp. 137-142.

ERRATA

These Annals Vol. IX, No. 3

P. 204, in the determinant of the second member in formula (9): read "$1-\alpha_{nn}$" instead of "$-\alpha_{nn}$".

P. 207, 1st line: read "quantitative" instead of "quantitive".

P. 211, the last line: read "$\cdots+a_{nk}^{0}q_n+\frac{R_k'}{X_k'P_k^0}$" instead of "$\cdots+a_{nk}^{0}q_n\frac{R_k'}{X_k'P_k^0}$".

Vol. X, No. 1

P. 33, Theorems 1∼4: read "X" instead of "x".