ORDER STATISTICS FOR DISCRETE CASE WITH A NUMERICAL APPLICATION TO THE BINOMIAL DISTRIBUTION

By MINORU SIOTANI

(Received Oct. 15, 1956)

1. Introduction

In taste-testing, we frequently meet with the necessity of testing the homogeneity of k mutually independent frequencies of occurrences of a certain event and also testing the significance of the largest or the smallest of a set of the observed frequencies. If the number of trials is sufficiently large, we can easily carry out the test by using the normal approximation. But we have frequently experienced the situation that the number of trials is not so large enough to apply the normal application, in the routine work, for example, in tasting-wine.

In this paper, we consider the probability distributions of the largest value, the smallest value and the range of k observations taken from a discrete type population. An application to the comparison of k independent observed numbers each of which is subject to the binomial law will also be given.

2. Distributions of the largest value and the smallest value

Let x be the discrete random variable which may take the values $0, 1, 2, \cdots$ with probabilities $p(0)$, $p(1)$, $p(2)$, $\cdots$, respectively, where $p(\alpha)\geqq 0$ and $\sum_{\alpha=0}^{\infty} p(\alpha)=1$. When x takes only the finite number of values, $0, 1, 2, \cdots, M$, we interpret that $p(M+\alpha)\equiv 0$ for $\alpha=1, 2, \cdots$. Suppose that we make k independent observations on x and have the values, $x_{(1)}, x_{(2)}, \cdots, x_{(k)}$ $(k\geqq 2)$. Let x_1 be the largest value among them, i.e..

$$(1) \qquad x_1=\max\{x_{(1)}, x_{(2)}, \cdots, x_{(k)}\} ,$$

and let x_2 be the smallest

$$(2) \qquad x_2=\min\{x_{(1)}, x_{(2)}, \cdots, x_{(k)}\} .$$

The probability distributions of x_1 and x_2 are easily obtained. Putting

$$\sum_{\alpha=0}^{s} p(\alpha) \equiv P(s)\,,$$

we have

$$p^*(x_1) = \sum_{i=1}^{k} \binom{k}{i} \{p(x_1)\}^i \{P(x_1-1)\}^{k-i} \tag{3}$$

$$= \{P(x_1)\}^k - \{P(x_1-1)\}^k \tag{4}$$

where $x_1=0, 1, 2, \cdots$ and $P(x_1-1)=0$ when $x_1=0$. If $p(\alpha)=1/N$ for $\alpha=1, 2, \cdots, N$ and $p(\alpha)=0$ for $\alpha=0, N+1, N+2, \cdots$, then we have

$$p^*(x_1) = \{x_1^k - (x_1-1)^k\} N^{-k}\,,$$

which is the probability distribution of the largest number in k drawings from a bowl containing balls numbered 1 to N when random sampling with replacement is used (Feller, 1950, pape 176).

For the binomial distribution with parameters p and N,

$$p(\alpha) \equiv b(\alpha;\, p,\, N) = \binom{N}{\alpha} p^{\alpha}(1-p)^{N-\alpha}\,,$$

$$P(x_1) \equiv B(x_1;\, p,\, N) = \sum_{\alpha=0}^{x_1} b(\alpha;\, p,\, N)$$
$$= 1 - I_p(x_1+1,\ N-x_1)\,,$$

where $I_p(a, b)$ is the incomplete beta function. Therefore, we have

$$p^*(x_1;\, p,\, N) = \{B(x_1;\, p,\, N)\}^k - \{B(x_1-1;\, p,\, N)\}^k \tag{5·a}$$

$$= \{1 - I_p(x_1+1,\ N-x_1)\}^k - \{1 - I_p(x_1,\ N-x_1+1)\}^k \tag{5·b}$$

where $x_1=0, 1, 2, \cdots, N$. The numerical values of $p^*(x_1; p, N)$ are easily obtained from (5·a) by using the "Tables of the Binomial Probability Distribution" (U.S.D.E. 1950) or from (5·b) by using the "Tables of the Incomplete Beta-Function" (Pearson, 1948) for fixed values of k, p and N.

For the Poisson distribution with parameter λ,

$$p(\alpha) \equiv p(\alpha;\, \lambda) = \frac{\lambda^{\alpha}}{\alpha!} e^{-\lambda}$$

$$P(x_1) \equiv P(x_1;\, \lambda) = \sum_{\alpha=0}^{x_1} p(\alpha\,;\, \lambda)$$
$$= 1 - I_{\lambda}(x_1+1)\,,$$

where $I_{\lambda}(a)$ is the incomplete gamma function. Then we have

$$p^*(x_1;\ \lambda)=\{P(x_1;\ \lambda)\}^k-\{P(x_1-1;\ \lambda)\}^k, \tag{6·a}$$

$$=\{1-\boldsymbol{I}_\lambda(x_1+1)\}^k-\{1-\boldsymbol{I}_\lambda(x_1)\}^k, \tag{6·b}$$

where $x_1=0,\ 1,\ 2,\ \cdots$. Formula (6·a) is for using the "Tables of the Poisson Distribution" (Kitagawa, 1951) and (6·b) is for using the "Tables of the Incomplete Γ-Function" (Pearson, 1951).

Similarly we can obtain the distribution of x_2 as follows:

$$p_*(x_2)=\sum_{i=1}^{k}\binom{k}{i}\{p(x_2)\}^i\{1-P(x_2)\}^{k-i} \tag{7}$$

$$=\{1-P(x_2-1)\}^k-\{1-P(x_2)\}^k \tag{8}$$

where $x_2=0,\ 1,\ 2,\ \cdots$ and $P(x_2-1)=0$ when $x_2=0$. The expressions for the binomial case and the poisson case are obvious.

3. Joint distribution of x_1 and x_2 and distribution of the range.

In this section, we shall consider the joint probability distribution of the largest value x_1 and the smallest value x_2 and also the distribution of the range, $R=x_1-x_2$. If we take into consideration, only the case when $x_1\neq x_2$ the required distributions are the conditional ones under the restriction that k sampled values $x_{(1)},\ x_{(2)},\ \cdots,\ x_{(k)}$ are not of the same value. But the unconditional probability distributions are easily obtained. First we shall derive the joint distribution of x_1 and x_2.

For the case when $x_1=x_2$,

$$P_r\{x_1=x_2=\alpha\}=\{p(\alpha)\}^k, \qquad \alpha=0,\ 1,\ 2,\ \cdots, \tag{9}$$

and for $x_1>x_2$,

$$p(x_1,\ x_2)=\sum_{i,j}{}'\frac{k!}{i!j!(k-i-j)!}\{p(x_2)\}^i\{P(x_1-1)-P(x_2)\}^{k-i-j}\{p(x_1)\}^j, \tag{10}$$

where $\sum_{i,j}'$ denotes the summation that i and j run over $1,\ 2,\ \cdots,\ k-1$, while being subject to the restriction $k\geqq i+j$. Then,

$$\begin{aligned} p(x_1,\ x_2)=&\{p(x_1)+p(x_2)-P(x_1-1)-P(x_2)\}^k\\ &-\sum_{i=0}^{k}\binom{k}{i}\{p(x_2)\}^i\{P(x_1-1)-P(x_2)\}^{k-i}\\ &-\sum_{j=0}^{k}\binom{k}{j}\{p(x_1)\}^j\{P(x_1-1)-P(x_2)\}^{k-j}\\ &+\{P(x_1-1)-P(x_2)\}^k \end{aligned} \tag{11}$$

$$=\{P(x_1)-P(x_2-1)\}^k-\{P(x_1)-P(x_2)\}^k$$
$$-\{P(x_1-1)-P(x_2-1)\}^k+\{P(x_1-1)-P(x_2)\}^k,$$

where $x_1>x_2$ and $P(x_2-1)=0$ when $x_2=0$.

The probability distribution of the range

$$R_k=x_1-x_2 \tag{13}$$

of k observed values, $x_{(1)}, x_{(2)}, \cdots, x_{(k)}$ is obtained from (9) and (11).

Since R_k is zero if and only if all values are same,

$$P_r\{R_k=0\}=\sum_{\alpha=0}^{\infty}\{x(\alpha)\}^k. \tag{14}$$

The remaining part of the distribution of R_k is obtained by transforming the formula (11) into the expression of R_n and x_2 and summing up the resultant with respect to x_2, that is,

$$p(x_2, R_k)=\{P(R_k+x_2)-P(x_2-1)\}^k-\{P(R_k+x_2)-P(x_2)\}^k \tag{15}$$
$$-\{P(R_k+x_2-1)-P(x_2-1)\}^k+\{P(R_k+x_2-1)-P(x_2)\}^k$$

where $R_k\geqq 1$. If the original random variable x can take only the finite number of values, 0, 1, 2, $\cdots$, M, the summation with respect to x_2 is taken over 0, $\cdots$, $M-R_k$, and when there is no such a finite value M, we must sum up over the whole set (0, 1, 2, $\cdots$).

4. Mean and variance of x_1, x_2 and R_k

First we consider the case when there is a finite number M which is the largest value x takes.

Since $P(M)=1$,

$$E(x_1)=\sum_{x_1=0}^{M}x_1p^*(x_1)=\sum_{x_1=0}^{M}x_1[\{P(x_1)\}^k-\{P(x_1-1)\}^k]$$
$$=M-\sum_{x_1=0}^{M-1}\{P(x_1)\}^k.$$

Hence we have

$$E(x_1)=\sum_{x_1=0}^{M-1}[1-\{P(x_1)\}^k]. \tag{16}$$

For the variance of x_1, we have

$$E(x_1^2)=\sum_{x_1=0}^{M}x_1^2p^*(x_1)=\sum_{x_1=0}^{M}x_1^2[\{P(x_1-1)\}^k-\{P'x_1-1)\}^k]$$
$$=M^2-\sum_{x_1=0}^{M-1}(2x_1+1)\{P(x_1)\}^k$$

$$=\sum_{x_1=0}^{M-1}(2x_1+1)[1-\{P(x_1)\}^k]\,,$$

$$=2\sum_{x_1=0}^{M-1}x_1[1-\{P(x_1)\}^k]+E(x_1)\,,$$

therefore,

$$D^2(x_1)=2\sum_{x_1=0}^{M-1}x_1[1-\{P(x_1)\}^k]+E(x_1)\{1-E(x_1)\}\;. \tag{17}$$

Similarly, we can evaluate the mean and variance of x_2, that is,

$$E(x_2)=\sum_{x_2=0}^{M-1}\{1-P(x_2)\}^k, \tag{18}$$

$$D^2(x_2)=2\sum_{x_2=0}^{M-1}x_2\{1-P(x_2)\}^k+E(x_2)\{1-E(x_2)\}\;. \tag{19}$$

The mean of the range, $R_k=x_1-x_2$, can easily be obtained from (16) and (17);

$$E(R_k)=E(x_1)-E(x_2)$$

$$=\sum_{x_1=0}^{M-1}[1-\{P(x_1)^k]-\sum_{x_2=0}^{M-1}\{1-P(x_2)\}^k.$$

This can be rewritten as

$$E(R_k)=\sum_{x_1=0}^{M-1}[1-\{P(x_1)\}^k-\{1-P(x_1)\}^k]\,, \tag{20}$$

which is the corresponding form to Tippett's formula for the continuous case (Tippett, 1925).

To get the formula of the variance of R_k, we shall evaluate $E(R_k^2)$.

$$E(R_k^2)=\sum_{x_1=1}^{M}\sum_{x_2=0}^{x_1-1}(x_1-x_2)^2[\{P(x_1)-P(x_2-1)\}^k-\{P(x_1)-P(x_2)\}^k]$$

$$-\sum_{x_1=1}^{M}\sum_{x_2=0}^{x_1-1}(x_1-x_2)^2[\{P(x_1-1)-P(x_2-1)\}^k-\{P(x_1-1)-P(x_2)\}^k]$$

$$=\sum_{x_1=1}^{M}x_1^2[\{P(x_1)\}^k-\{P(x_1-1)\}^k]$$

$$-\sum_{x_1=1}^{M}\sum_{x_2=0}^{x_1-1}\{2(x_1-x_2)-1\}[\{P(x_1)-P(x_2)\}^k-\{P(x_1-1)-P(x_2)\}^k]$$

$$=\sum_{x_1=1}^{M}(2x_1-1)[1-\{P(x_1-1)\}^k]$$

$$-\sum_{x_2=0}^{M-1}\{2(M-x_2)-1\}\{1-P(x_2)\}^k$$

$$+2\sum_{x_1=1}^{M}\sum_{x_2=0}^{x_1-1}\{P(x_1-1)-P(x_2)\}^k$$

$$=2\sum_{x_1=1}^{M}\sum_{x_2=0}^{x_1-1}[1-\{P(x_1-1)\}^k-\{1-P(x_2)\}^k+\{P(x_1-1)-P(x_2)\}^k]$$

$$-\sum_{x_1=1}^{M}[1-\{P(x_1-1)\}^k-\{1-P(x_1-1)\}^k]\,.$$

Hence the variance of R_k is expressible in the form

$$D^2(R_k)=2\sum_{x_1=0}^{M-1}\sum_{x_2=0}^{x_1}[1-\{P(x_1)\}^k-\{1-P(x_2)\}^k+\{P(x_1)-P(x_2)\}^k]$$
$$-E(R_k)\{1+E(R_k)\}\ , \tag{21}$$

which is the corresponding form to the continuous case (Tippett, 1925).

Now, we consider the case when the original variable x assumes 0 and all positive integral values. Let us consider, for example, the mean and variance of x_1 in this case. For any finite M and fixed k, we have

$$\sum_{x_1=0}^{M}x_1p^*(x_1)=\sum_{x_1=0}^{M-1}[\{P(M)\}^k-\{P(x_1)\}^k] \tag{22}$$

and

$$\sum_{x_1=0}^{M}x_1^2p^*(x_1)=\sum_{x_1=0}^{M-1}(2x_1+1)[\{P(M)\}^k-\{P(x_1)\}^k]\,. \tag{23}$$

Here $P(M)$ is not equal to 1. If the original distribution has finite mean and variance, then $E(x_1)$ and $E(x_1^2)$ exist, for

$$\sum_{x_1=0}^{M}x_1^{\nu}p^*(x_1)=\sum_{x_1=0}^{M}x_1^{\nu}\left(\sum_{i=1}^{k}\binom{k}{i}\{p(x_1)\}^i\{P(x_1-1)\}^{k-i}\right),$$
$$<(2^k-1)\left(\sum_{x_1=0}^{M}x_1^{\nu}p(x_1)\right),$$
$$<(2^k-1)\left(\sum_{\alpha=0}^{\infty}\alpha^{\nu}p(\alpha)\right).\qquad(\nu=1,\ 2)$$

Consequently, we have, from (22) and (23),

$$E(x_1)=\lim_{M\to\infty}\sum_{x_1=0}^{M}x_1p^*(x_1)=\sum_{x_1=0}^{\infty}[1-\{P(x_1)\}^k] \tag{24}$$

and

$$E(x_1^2)=\lim_{M\to\infty}\sum_{x_1=0}^{M}x_1^2p^*(x_1)=\sum_{x_1=0}^{\infty}(2x_1+1)[1-\{P(x_1)\}^k] \tag{25}$$
$$=2\sum_{x_1=0}^{\infty}x_1[1-\{P(x_1)\}^k]+E(x_1)\,.$$

The means and variances of x_2 and R_k are also obtained in a similar way and have the expressions (18), (19), (20) and (21) with the symbol ∞ replacing M in the upper limits of the summation, provided that x has the finite mean and variance. Thus, for the range R_k, we have

$$(26) \qquad E(R_k)=\sum_{x_1=0}^{\infty}[1-\{P(x_1)\}^k-\{1-P(x_1)\}^k]$$

and

$$(27) \qquad D^2(R_k)=2\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{x_1}[1-\{P(x_1)\}^k-\{1-P(x_2)\}^k+\{P(x_1)-P(x_2)\}^k]$$
$$-E(R_k)\{1+E(R_k)\}$$

5. Application to the binomial distribution

In this section we shall illustrate an application of the order statistics considered in the preceeding sections for binomial case. Let $\nu_{(i)}$ $(i=1, 2, \cdots, k)$ be the observed number of occurrences of a certain event in the ith succession of N trials, where N is fixed for each succession, and let p_i $(i=1, 2, \cdots, k)$ be the probability of the event in the ith experiment. In order to test the homogeneity of k experiments, that is, to test the hypothesis that

$$(28) \qquad H;\ p_{(1)}=p_{(2)}=\cdots=p_{(k)} \ (\equiv p, \text{ say}),$$

we can use the range defined by

$$(29) \qquad R_k(N, p)\equiv \nu_1-\nu_2$$

where ν_1 and ν_2 are the largest and the smallest values of k observations, respectively. Suppose that we are in the situation that the conditions for applying the normal approximation formulas are not fulfilled but Nk is large enough for the estimate of p, to be taken for the true value under the hypothesis H, that is,

$$(30) \qquad \hat{p}=\frac{\nu_{(1)}+\nu_{(2)}+\cdots+\nu_{(k)}}{Nk}\approx p\ .$$

Then, from (15), we can evaluate the probability

$$(31) \qquad P_r\{R_k(N, p)\geqq r_k\}\equiv U(r_k;\ k, N, p)$$

where r_k is the observed value of the range in a specified comparison. This value of probability is compared with the significance level α:

if $U(r_k;\ k,\ N,\ p)<\alpha$, the hypothesis H is rejected on the $100\alpha\%$ level and otherwise accepted.

For numerical example, we have the following data, which are obtained in tasting wine. Each of $k=8$ brands is tasted by $N=17$ judges and each judge makes decision whether or not the wine tasted

i	1	2	3	4	5	6	7	8
$\nu_{(i)}$	14	12	6	13	13	15	8	13

is classified in the 1st class of quality. $\nu_{(i)}$ is the number of the judgement of the 1st class of quality. Then we have

$$r_8=15-6=9\,,$$

$$p\approx\frac{14+12+\cdots+13}{8\times 17}=0.69\,.$$

$$U(9;\ 8,\ 17,\ 0.69)=U(9;\ 8,\ 17,\ 0.31)=0.02793\,,$$

hence the hypothesis of homogeneity is rejected on the 5% level of significance. The mean and variance of R_8 (17, 0.69) are calculated, from the formulas (20) and (21), as

$$E\{R_8(17,\ 0.69)\}=E\{R_8(17,\ 0.31)\}=5.3531\,,$$

$$D^2\{R_8(17,\ 0.69)\}=D^2\{R_8(17,\ 0.31)\}=2.4291\,.$$

This means that our observed value of the range, 9, is deviated from the mean value, 5.3531, by 2.98 times the standard deviation, 1.5586.

By a simple inspection of our data, it can be seen that the heterogeneity is caused by the difference between two groups.

$$G_1:\ \ \nu_{(1)}=14,\ \nu_{(2)}=12,\ \nu_{(4)}=13,\ \nu_{(5)}=13,\ \nu_{(6)}=15,\ \nu_{(8)}=13\,,$$

$$G_2:\ \ \nu_{(3)}=6,\ \nu_{(7)}=8\,.$$

In fact, if we carry out the same procedure as before as to each group, the hypothesis of homogeneity is accepted. Moreover, since G_1 can now be grouped into a single sample from a population with the parameter p_1, we can test the null hypothesis that $p_1=\frac{1}{2}$ against the alternative hypothesis that $p_1>\frac{1}{2}$ (assuming that the brand with $p>\frac{1}{2}$ is classified into the 1st class of quality) by the usual test procedure based on the normal approximation. The similar argument is made for the group G_2.

When the alternative to the homogeneity is that one of the numbers $\nu_{(i)}$ substantially exceeds the expected values of the others, which may be homogeneous or not, the test procedure based on the range, though applicable, seems not suitable. In this situation, we shall use the largest value ν_1. Since, for the binomial case,

$$\max\{N-\nu_{(1)},\ N-\nu_{(2)},\ \cdots,\ N-\nu_{(k)}\}=N-\min\{\nu_{(1)},\ \nu_{(2)},\ \cdots,\ \nu_{(k)}\}\ ,$$

there is no need of considering the test based on the smallest value, ν_2.

Let us consider the case when Nk is sufficiently large as in the previous case when the range is used. Then, under the hypothesis of the homogeneity to be tested, $H:\ p_{(1)}=p_{(2)}=\cdots=p_{(k)}\equiv p$, we have

$$p\simeq\hat{p}=\frac{\nu_{(1)}+\nu_{(2)}+\cdots+\nu_{(k)}}{Nk}=\frac{\bar{\nu}}{N}$$

$$\sigma^2=Np(1-p)\simeq N\hat{p}(1-\hat{p})=\hat{\sigma}^2$$

where σ^2 is the common variance of $\nu_{(i)}$ under H. In this case, the test based on the statistic $(\nu_1-N\hat{p})/\hat{\sigma}\simeq(\nu_1-Np)/\sigma$ is equivalent to the test based on the ν_1 only. Thus, if we denote the observed value of ν_1 in a specified comparison by ν_1^*, we can evaluate the probability

$$P_r\{\nu_1\geqq\nu_1^*\}=W(\nu_1^*;\ k,\ N,\ p) \tag{32}$$

from the formula (5·a) or (5·b) for fixed values of k, N and p.

If $W(\nu_1^*:\ k,\ N,\ p)<\alpha$, the hypothesis H is rejected on the $100\alpha\%$ level and otherwise accepted, where α is the prescribed level of significance.

For the data:

$$\nu_{(1)}=9,\ \nu_{(2)}=10,\ \nu_{(3)}=15,\ \nu_{(4)}=9$$
$$N=16$$

we have

$$\nu_1^*=15,\ \ p\simeq\frac{9+10+15+9}{16\times4}=0.67\ ,$$

$$W(15;\ 4,\ 16,\ 0.67)=0.05730$$

hence we accept the hypothesis H on the 5% level

In order to give the convenience to the practical testing for the binomial case, the author and Mr. T. Huziwara are now preparing the the tables of $U(r_k;k,N,p)$, $W(\nu_1^*;\ k,\ N,p)$, $E\{R_k(N,p)\}$ and $D^2\{R_k(N;p)\}$, and we shall publish them in the near future.

6. Acknowledgement

I wish to express my gratitude to Mr. T. Sakaguti and Mr. H. Hukumitu of TAKARA-SYUZO CO. who called my attention to this research.

REFERENCES

[1] Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. I. John Wiley and Sons., New York (1950).

[2] Kitagawa, T., *Tables of the Poisson Distribution.* Baihūkan, Tokyo (1951).

[3] Pearson, K., *Tables of the Imcomplete Beta-Function.* Biometrika Office, London (1948).

[4] Pearson, K., *Tables of the Incomplete Γ-Function.* Cambridge University, London (1951).

[5] Tippett, L. H. C., On the extreme individuals and the range of samples taken from normal population. *Biometrika,* **17**, 364. (1925).

[6] United States Department of Commerce, *Tables of the Binomical Probability Distribution. App. Math. Ser.*, No. 6; National Bureau of Standards. (1950).