NOTE ON ORDERED RANDOM INTERVALS AND ITS APPLICATION

BY HIROJIRO AOYAMA

(Received Dec. 10, 1961)

1. Introduction

In the polls we are faced to a problem how to decide which candidate is approved to be elected. Further what percentage gain is a certain level is a problem. In practical use we have a criterion as follows:

Let n be the number of candidates, and k the fixed number. Then the reliable level to be elected will be $1/(k+n/k)$.

But this level is apt to be higher than that of actual data especially for larger k. Therefore we have to make a better level by considering the distribution of percentage gain of the candidate.

To solve this problem we may consider the distribution of ordered random intervals:

Let us assume $(n-1)$ points are dropped randomly on a fixed line of unit length. The lengths of the n intervals between points are arranged in order of ascending magnitude

$$y_1, y_2, \cdots, y_n ,$$

y_i being the i^{th} smallest interval. This distribution of y_i has been discussed by many authors [1], [2], [3], [4], [5], [6], but we would like to derive the distribution of y_i by another method, and consider the application to the above-mentioned problem.

2. Derivation of the distribution function of the ordered statistic y_r

Let $(n-1)$ points be dropped on the interval $[0, 1]$ and the lengths of the n intervals between these points be arranged in order of ascending magnitude $y_1^{(n)}, y_2^{(n)}, \cdots, y_n^{(n)}$. Then it is well known that the probability density function (p.d.f.) of any interval y is given by

$$\Pr(y, n)=(n-1)(1-y)^{n-2} . \qquad (1)$$

By the mathematical induction we would like to prove the well known formula of the p.d.f. of length of the r^{th} smallest interval $y_r^{(n)}$

$$f(y_r^{(n)})=n(n-1)\binom{n-1}{r-1}\sum_{i=0}^{r-1}(-1)^i\binom{r-1}{i}(1-y_r^{(n)}(n-r+i+1))^{n-2} \qquad (2)$$

where the summation is carried over as long as the terms $(1-y_r^{(n)}(n-r+i+1))$ are positive.

For $n=2$ it is evident that the equation (2) holds because

$$f(y_1^{(2)})=\begin{cases}2\,, & \text{for } \frac{1}{2}\geqq y_1^{(2)}\\ 0\,, & \text{otherwise}\end{cases}$$

$$f(y_2^{(2)})=\begin{cases}0\,, & \text{for } \frac{1}{2}\geqq y_2^{(2)}\\ 2\,, & \text{otherwise}\,.\end{cases}$$

Therefore, if we assume that (2) holds for $n-1$, we can prove in the following way that it is true for n:

If we get a sample value y, the conditional probability $\Pr(y_r^{(n)}=y\mid y, n)$ that y is the r^{th} smallest order statistic, multiplied by $nP_r(y, n)=n(n-1)(1-y)^{n-2}$ makes the p.d.f. $f(y_r^{(n)}=y)$.

On the other hand, the sum of other $(n-1)$ values equals $1-y$.

Putting $t=y/(1-y)$ and $t'=y'/(1-y)$, we can therefore apply (2) for $(n-1)$ which is the induction assumption for these $(n-1)$ intervals which make up the unit length in regard to t' scale.

In other words, the r^{th} smallest interval $t_r^{(n-1)}$ satisfies

$$f(t_r^{(n-1)})=(n-1)(n-2)\binom{n-2}{r-1}\sum_{i=0}^{r-1}(-1)^i\binom{r-1}{i}(1-t_r^{(n-1)}(n-r+i))^{n-3}\,.$$

Hence the conditional probability $\Pr(y_r^{(n)}=y\mid y, n)$ for $t<1$ is given as follows:

$$\Pr(y_r^{(n)}=y|y, n)=\Pr(t_{r-1}^{(n-1)}<t,\ t_r^{(n-1)}>t)=\int_0^t f(t_{r-1}^{(n-1)})dt_{r-1}^{(n-1)}-\int_0^t f(t_r^{(n-1)})dt_r^{(n-1)}$$

$$=\binom{n-1}{r-1}\sum_{r=0}^{r-1}(-1)^i\binom{r-1}{i}(1-(n-r+i)t)^{n-2}$$

$$=\frac{1}{(1-y)^{n-2}}\binom{n-1}{r-1}\sum_{i=0}^{r-1}(-1)^i\binom{r-1}{i}(1-(n-r+i+1)y)^{n-2}\,.$$

In the case of $t\geqq1$, we may put $f(t_r^{(n-1)})=0$.

Thus we obtain, from the above argument

$$f(y_r^{(n)}=y)=n\Pr(y, n)\Pr(y_r^{(n)}=y|y, n)$$
$$=n(n-1)\binom{n-1}{r-1}\sum_{i=0}^{r-1}(-1)^i\binom{r-1}{i}(1-(n-r+i+1)y)^{n-2}$$

and (2) for n.

Thus (2) holds for all n by the mathematical induction.

3. Derivation of the simultaneous distribution function of the ordered statistics y_r and y_s

The method to derive the simultaneous probability density function of y_r and y_s with $r<s$

$$f(y_r^{(n)}, y_s^{(n)})=(n-1)(n-2)\binom{n}{s}\binom{s}{r+1}r(r+1)\sum_{i=0}^{r-1}\sum_{j=0}^{s-r-1}(-1)^{i+j}$$
$$\times\binom{r-1}{i}\binom{s-r-1}{i}\{1-(s-r+i-j)y_r^{(n)}-(n-s+j+1)y_s^{(n)}\}^{n-3} \qquad (3)$$

is quite similar to that for (2).

For $n=3$ it is evident that for any r and s

$$f(y_r^{(3)}, y_s^{(3)})=12$$

because $\Pr(y_1, y_2, 3)=2$, and $\Pr(y_r^{(3)}=y_1, y_s^{(3)}=y_2|y_1, y_2, 3)=1$ when y_1 and y_2 lie in the existence region. On the other hand the right hand side of (3) for $n=3$ equals 12 for any r and s.

At first the simultaneous p.d.f. $\Pr(y_1, y_2\, n)$ of any y_1 and y_2 is

$$(n-1)(n-2)(1-y_1-y_2)^{n-3} \qquad (4)$$

which is easily shown. We put $t_1=y_1/(1-y_1-y_2)$ and $t_2=y_2/(1-y_1-y_2)$, and consider the t-scale: $t=y/(1-y_1-y_2)$, the sum of $(n-2)$ intervals being equal to $1-y_1-y_2$.

If we assume the equation (3) holds for $n-2$, we can prove that it is true for n. For this purpose we may derive the equation

$$n(n-1)\Pr(y_r^{(n)}=y_1, y_s^{(n)}=y_2|y_1, y_2, n)\Pr(y_1, y_2, n)$$
$$=f(y_r^{(n)}=y_1, y_s^{(n)}=y_2)\ . \qquad (5)$$

Let the conditions A_1, A_2, B_1 and B_2 be as follows:

$$A_1\colon t_{r-1}^{(n-2)}<t_1\ , \qquad B_1\colon t_{s-2}^{(n-2)}<t_2$$
$$A_2\colon t_r^{(n-2)}>t_1\ , \qquad B_2\colon t_{s-1}^{(n-2)}>t_2\ .$$

As was shown in section 2, $\Pr(A_1\times A_2)=\Pr(A_1)-\Pr(\bar{A}_2)$, where $\bar{A}_2$ is the complement of A_2, so that the simultaneous p.d.f. in this special case for A_1 and A_2 is given by the difference of simple p.d.f.. Hence, under given t_1 and t_2,

$$
\begin{aligned}
\Pr(y_r^{(n)}=y_1,\ y_s^{(n)}=y_2|y_1,\ y_2,\ n)&=\Pr(A_1\times A_2,\ B_1\times B_2)\\
&=\Pr(A_1-\bar{A}_2,\ B_1-\bar{B}_2)=\Pr(A_1,\ B_1)-\Pr(\bar{A}_2,\ B_1)\\
&\quad-\Pr(A_1,\ \bar{B}_2)+\Pr(\bar{A}_2,\ \bar{B}_2)\ . \qquad (6)
\end{aligned}
$$

Each term of the right hand side of (6) is calculated by the assumption. For example,

$$
\begin{aligned}
\Pr(A_1,\ B_1)&=\Pr(t_{r-1}^{(n-2)}<t_1,\ t_{s-2}^{(n-2)}<t_2)\\
&=(n-3)(n-4)\binom{n-2}{s-2}\binom{s-2}{r}(r-1)r\sum_{i=0}^{r-2}\sum_{j=0}^{s-r-2}(-1)^{i+j}\binom{r-2}{i}\binom{s-r-2}{j}\\
&\quad\times\int_0^{t_1}\int_{t_{r-1}^{(n-2)}}^{t_2}\{1-(s-r-1+i-j)t_{r-1}^{(n-2)}-(n-s+j+1)t_{s-2}^{(n-2)}\}^{n-5}dt_{s-2}^{(n-2)}dt_{r-1}^{(n-2)}\\
&=\binom{n}{s}\binom{s}{r+1}\frac{r(r+1)}{n(n-1)}\sum_{i=0}^{r-1}\sum_{j=0}^{s-r-1}(-1)^{i+j}\binom{r-1}{i}\binom{s-r-1}{j}\\
&\quad\cdot\Bigg[\frac{ij}{(n-s+j)(s-r-1+i-j)}\{(1-(s-r-1+i-j)t_1-(n-s+j)t_2)^{n-3}\\
&\quad-(1-(n-s+j)t_2)^{n-3}\}+\frac{ij}{(n-s+j)(n-r+i)}\{1-(1-(n-r+i)t_1)^{n-3}\}\Bigg]\ .
\end{aligned}
$$

In a similar way we obtain

$$
\begin{aligned}
\Pr(y_r^{(n)}=y_1,\ y_s^{(n)}=y_2|y_1,\ y_2,\ n)&=\binom{n}{s}\binom{s}{r+1}\frac{r(r+1)}{n(n-1)}\sum_{i=0}^{r-1}\sum_{j=0}^{s-r-1}(-1)^{i+j}\binom{r-1}{i}\\
&\quad\times\binom{s-r-1}{j}\{1-(s-r-1+i-j)t_1-(n-s+j)t_2\}^{n-3}\\
&=\frac{1}{n(n-1)(1-y_1-y_2)^{n-3}}\binom{n}{s}\binom{s}{r+1}r(r+1)\sum_{i=0}^{r-1}\sum_{l=0}^{s-r-1}(-1)^{i+j}\binom{r-1}{i}\binom{s-r-1}{j}\\
&\quad\times\{1-(s-r+i-j)y_1-(n-s+j+1)y_2\}^{n-3} \qquad (7)
\end{aligned}
$$

because the terms vanish which do not include i and j at the same time.

Hence from (4) and (5) $f(y_r^{(n)},\ y_s^{(n)})$ is given by (3). Thus (3) holds in general by the mathematical induction.

4. Derivation of the simultaneous distribution function of the ordered statistics y_r, y_s and y_u

The procedure for this problem is quite the same as the above mentioned method. The simultaneous p.d.f. of $y_r^{(n)}$, $y_s^{(n)}$ and $y_u^{(n)}$ with $r<s<u$ is given by

$$
\begin{aligned}
f(y_r^{(n)},\ y_s^{(n)},\ y_u^{(n)})&=(n-1)(n-2)(n-3)r(r+1)(r+1)\binom{n}{u}\binom{u}{s+1}\binom{s+1}{r+2}\\
&\quad\times\sum_{i=0}^{r-1}\sum_{j=0}^{s-r-1}\sum_{k=0}^{u-s-1}(-1)^{i+j+k}\binom{r-1}{i}\binom{s-r-1}{j}\binom{u-s-1}{k}\{1-(s-r+i-j)y_r^{(n)}\\
&\quad-(u-s+j-k)y_s^{(n)}-(n-u+k+1)y_u^{(n)}\}^{n-4}\ . \qquad (8)
\end{aligned}
$$

In this case we put the conditions

$$\begin{array}{lll} A_1\colon t_{r-1}^{(n-3)}<t_1\ , & B_1\colon t_{s-2}^{(n-3)}<t_2\ , & C_1\colon t_{u-3}^{(n-3)}<t_3 \\ A_2\colon t_{r}^{(n-3)}>t_1\ , & B_2\colon t_{s-1}^{(n-3)}>t_2\ , & C_2\colon t_{u-2}^{(n-3)}>t_3 \end{array}$$

with $t_i=y_i/(1-y_1-y_2-y_3)(i=1, 2, 3)$.

Then we obtain

$$\begin{aligned} &\Pr(y_r^{(n)}=y_1,\ y_s^{(n)}=y_2,\ y_u^{(n)}=y_3|y_1,\ y_2,\ y_3,\ n) \\ &\quad=\Pr(A_1\times A_2,\ B_1\times B_2,\ C_1\times C_2) \\ &\quad=\Pr(A_1-\bar{A}_2,\ B_1-\bar{B}_2,\ C_1-\bar{C}_2) \\ &\quad=\Pr(A_1,\ B_1,\ C_1)-\Pr(\bar{A}_2,\ B_1,\ C_1)-\Pr(A_1,\ \bar{B}_2,\ C_1)+\Pr(\bar{A}_2,\ \bar{B}_2,\ C_1) \\ &\qquad-\Pr(A_1,\ B_1,\ \bar{C}_2)+\Pr(\bar{A}_2,\ B_1,\ \bar{C}_2)+\Pr(A_1,\ \bar{B}_2,\ \bar{C}_2)-\Pr(\bar{A}_2,\ \bar{B}_2,\ \bar{C}_2) \\ &=\frac{r(r+1)(r+2)}{n(n-1)(n-2)(1-y_1-y_2-y_3)^{n-4}}\binom{n}{u}\binom{u}{s+1}\binom{s+1}{r+2}\sum_{i=0}^{r-1}\sum_{j=0}^{s-r-1}\sum_{k=0}^{u-s-1}(-1)^{i+j+k} \\ &\qquad\times\binom{r-1}{i}\binom{s-r-1}{j}\binom{u-s-1}{k}\{1-(s-r+i-j)y_1-(u-s+j-k)y_2 \\ &\qquad-(n-u+k+1)y_3\}^{n-4} \end{aligned} \tag{9}$$

and

$$\Pr(y_1,\ y_2,\ y_3,\ n)=(n-1)(n-2)(n-3)(1-y_1-y_2-y_3)^{n-4} \tag{10}$$

so that

$$\begin{aligned} f(y_r^{(n)},\ y_s^{(n)},\ y_u^{(n)})&=n(n-1)(n-2)\Pr(y_r^{(n)}=y_1,\ y_s^{(n)}=y_2,\ y_u^{(n)}=y_3|y_1,\ y_2,\ y_3,\ n) \\ &\quad\cdot\Pr(y_1,\ y_2,\ y_3,\ n) \end{aligned} \tag{11}$$

from which it follows that (8) holds.

5. Application to polls

As was mentioned in the introduction we would like to consider the application of the problem to polls.

As before, let n and k be the number of candidates and the fixed number, respectively. If we may take the reliable level to be elected as the cutting point y^* of $f(y_{n-k+1}^{(n)})$ with $f(y_{n-k}^{(n)})$, we can easily obtain y^* by solving the algebraic equation $f(y_{n-k+1}^{(n)})=f(y_{n-k}^{(n)})$. In this case we have to solve the algebraic equation of $n-2$ degrees. In the following table we shall show the cutting points y^* in case of $n=3, 4, 5$ and compare them with the results from empirical formula $y_e=1/(k+n/k)$.

n \ k	3			4			5		
	y^*	y_e	y^{**}	y^*	y_e	y^{**}	y^*	y_e	y^{**}
1	$\frac{3}{7}\fallingdotseq 0.429$	$\frac{1}{4}$	$\frac{7}{15}\fallingdotseq 0.467$	$\frac{11-\sqrt{6}}{23}\fallingdotseq 0.372$	$\frac{1}{5}$	$\frac{23-2\sqrt{3}}{47}\fallingdotseq 0.416$	0.333	$\frac{1}{6}$	0.376
2	$\frac{1}{5}=0.2$	$\frac{2}{7}$	$\frac{3}{11}\fallingdotseq 0.273$	$\frac{2}{9}\fallingdotseq 0.222$	$\frac{1}{4}$	$\frac{13-\sqrt{5}}{41}\fallingdotseq 0.263$	0.214	$\frac{2}{9}$	0.247
3				$\frac{7-\sqrt{8}}{37}\fallingdotseq 0.113$	$\frac{3}{13}$	$\frac{5-\sqrt{2}}{23}\fallingdotseq 0.156$	0.126	$\frac{3}{14}$	0.175
4							0.056	$\frac{4}{21}$	0.101

As is seen from the table the empirical cutting point y_e is larger than y^* for larger k. But in the actual situation such as in the polls for the Members of the House of Representatives in Japan the empirical y_e is sometimes approved to be a pretty good level. To get a fine level from the above argument we can make use of the criterion

$$\begin{aligned}&\Pr(y^{(n)}_{n-k+1}=y|y,\, n)/\Pr(y^{(n)}_{n-k}=y|y,\, n)\\&=f(y^{(n)}_{n-k+1}=y)/f(y^{(n)}_{n-k}=y)=A\end{aligned} \tag{12}$$

that is, the critical point y^{**} is determined from this equation for a proper positive constant A.

The critical point y^{**} for $A=3$ are given in the above table.

In this manner we can get y^{**} by solving the above equation, but it is very complicated for large n and k. However, we can make use of approximate procedure.

If we would like to get the most probable k^{th} largest percentage gain y_k, we may solve the following equation, assuming other y_i except y_k mutually independent:

$$\frac{\partial}{\partial y_k}\log\left\{\int_0^{y_k}(n-1)(1-y)^{n-2}dy\right\}^{n-k}\cdot\left\{\int_{y_k}^1(n-1)(1-y)^{n-2}dy\right\}^{k-1}=0\ .$$

From this equation we obtain the approximate most probable value of y_k:

$$y_k=1-\sqrt[n-1]{\frac{k-1}{n-1}}\ . \tag{13}$$

In order to get a proper critical level y^{**} for k^{th} candidate, we may find a positive number B which satisfies the $y^{**}=(1+B)y_k$, but its magnitude depends on the actual situation in the polls.

On the other hand we can get the distribution of the difference D between $y_r^{(n)}$ and $y_{r+1}^{(n)}$ from the equation (3). That is, dropping the affix (n) we have

$$f(y_r, y_{r+1})=(n-1)(n-2)\binom{n}{r+1}r(r+1)\sum_{i=0}^{r-1}(-1)^i\binom{r-1}{i}$$
$$\times\{1-(1+i)y_r-(n-r)y_{r+1}\}^{n-3}$$

and

$$f(D)=(n-1)(n-r)\{1-(n-r)D\}^{n-2}\,. \tag{14}$$

Using this equation we obtain the mean of difference of percentage gains between the k^{th} candidate and the $(k+1)^{\text{st}}$ candidate($\bar{D}$) as is shown in the following table, where the mean percentage differences with parenthesis indicate the actual ones obtained from under 20 samples.

The difference between theoretical $\bar{D}$ and actual ones comes from the limit of the actual percentage gain in the polls. For example, the first successful candidate does not get over 50 percent in general. Therefore we must take into consideration the limit of region of D.

n \ k	3	4	5
6	(3.8) 5.6	(3.6) 4.2	3.3
7	(3.4) 4.8	(2.2) 3.6	2.9
8	(2.8) 4.2	(2.1) 3.1	(2.1) 2.5
9	3.7	2.8	(1.6) 2.2
10	3.3	2.5	(1.1) 2.0

For $n=6, k=4(r=2)$, for example, y_3 is not greater than 0.2 in the actual problem, so that y_2 is not smaller than $(1-0.2\times 4)/2=0.1$ and D is not greater than 0.1.

Hence we have

$$\bar{D}=\frac{\int_0^{0.1}5\cdot 4(1-4D)^4 D\,dD}{\int_0^{0.1}5\cdot 4(1-4D)^4 dD}=0.0346$$

which is very near to 3.6% (cf. the above table). Of course we can also derive this fact using the simultaneous p.d.f. $f(y_r, y_{r+1}, y_n)$, taking $y_n\leqq 0.4$.

These results can be used to discriminate the successful candidate from the unsuccessful one in the course of opening of the ballot.

6. Acknowledgments

The author is thankful to Mr. T. Komazawa, Mrs. Y. Taga and Mrs. T. Tino for their help in preparation of this paper. He also wishes to thank Mr. M. Siotani for his suggestion to bibliographies.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] D. E. Barton and F. N. David, "Some notes on ordered random intervals," *J. R. Statist. Soc.* Vol. 18, B (1956), pp. 79-94.

[2] D. A. Darling, "On a class of problems related to the random division of an interval," *Ann. Math. Statist.*, Vol. 24 (1953), pp. 239-253.

[3] R. A. Fisher, "On the similarity of the distributions found for the test of significance in harmonic analysis and in Stevens's problem in geometric probability," *Ann. Eugen. Lond.*, Vol. 10 (1929).

[4] J. O. Irwin, "A unified derivation of some well-known frequency distributions of interest in biometry and statistics" *J. R. Statist. Soc.*, Vol. 118, A. (1955), pp. 389-398.

[5] W. L. Stevens, "Solution to a geometrical problem in probability," *Ann. Eugen. Lond.*, vol. 9 (1939).

[6] W. A. Whitworth, *Choice and Chance*, 1887 Cambridge Univ. Press.