AN ESTIMATE OF STANDARD DEVIATION OF NORMAL POPULATION BASED ON THE DIFFERENCE BETWEEN MEANS OF TWO GROUPS DIVIDED BY SAMPLE MEAN*)

By MINORU SIOTANI

(Received June 27, 1954)

1. Introduction

In estimating the standard deviation of a normal population, various ranges (simple range [1.2], linear combination of group ranges [3], and quasi ranges defined by F. Mosteller [4]) are often used instead of the sample standard deviation in view of their simplicity. However, when the sample size is greater than about 10, it is not desirable to use the simple range by reason of the practical loss in efficiency. The linear combination of the group ranges, which gives the best unbiased estimate of the population standard deviation, has been proposed by F. E. Grubbs and C. L. Weaver [3] in order to better the efficiency of the statistic range for sample sizes greater than 11. But the gain in efficiency is not still adequate and, moreover, the complexity of computation of this estimate is increased much as compared to the simple range. F. Mosteller's quasi ranges are very useful statistics for estimating the population standard deviation, but based on the large sample theory. Recently, J. H. Cadwell [5] has studied the approximate distribution of quasi ranges.

Instead of these statistics mentioned above, we consider the following statistic based on the difference between means of two groups devided by the sample mean. Let $x'_1, x'_2, \cdots, x_n$ be a sample of size n from a normal population with mean zero and variance σ^2. Let $\bar{\bar{x}}$ be the group mean which is calculated from observations smaller than the total sample mean, $\bar{x}$, and $\bar{\bar{\bar{x}}}$ the other group mean of observations larger than $\bar{x}$. Then we adopt the statistic U_n defined by

$$U_n=\bar{\bar{\bar{x}}}-\bar{\bar{x}} \tag{1}$$

for estimating σ. Furthermore, when the number of observations smaller thah $\bar{x}$ is ν, we define the conditional statistic $U_n(\nu)$ as

*) The details of this paper are given in [8] which was written in Japanese.

$$U_n(\nu)=\bar{\bar{x}}(n-\nu)-\bar{\bar{x}}(\nu) \qquad (2)$$

for a sample of size n, where $\bar{\bar{x}}(n-\nu)$ and $\bar{\bar{x}}(\nu)$ are respective means of two groups in the case of ν. Since, in practice, we have information on the number of observations smaller than $\bar{x}$ for a particular sample, we should positively make use of this information and thus conditional statistic $U_n(\nu)$ is a desirable one for estimating σ. Let N be the random variable denoting the number of observations smaller than $\bar{x}$ and x_1'', x_2'', $\cdots$, x_ν'' be the observations smaller than $\bar{x}$ when $N=\nu$. Then $U_n(\nu)$ may be rewritten as

$$U_n(\nu)=-\frac{n}{n-\nu}\frac{1}{\nu}\sum_{i=1}^{\nu}(x_i''-\bar{x}). \qquad (3)$$

that is, the sum of the negative deviations.

Our proposed statistic has a relation with the mean deviation,

$$W_n=\frac{1}{n}\sum_{i=1}^{n}|x_i'-\bar{x}|, \qquad (4)$$

in the following way; that is, since the conditional mean deviation $W_n(\nu)$ for $N=\nu$ is written as

$$W_n(\nu)=-\frac{2}{n}\sum_{i=1}^{\nu}(x_i''-\bar{x}), \qquad (5)$$

we have

$$U_n(\nu)=\frac{1}{2}\frac{n^2}{\nu(n-\nu)}W_n(\nu). \qquad (6)$$

The sampling distribution of W_n was obtained by H. J. Godwin [6], but the present author has studied the sampling distributions of $U_n(\nu)$ and U_n by paying his attention to the sum of the negative deviations from the total sample mean, *i.e.*, $\sum_{i=1}^{\nu}(x_i''-\bar{x})$, and obtained the distribution of W_n independently of Godwin. The efficiencies of the estimates based on U_n and $U_n(\nu)$ as compared to the sample standard deviation and the coefficients of unbiasedness of our estimate are given with the help of these distributions.

I express my gratitude to Professor J. Ogawa of Ōsaka University and Mr. Y. Utida for their invaluable encouragements, indications and criticisms.

2. Distributions of $U_n(\nu)$ and U_n.

Let $x_1 \leqq x_2 \leqq \cdots \leqq x_n$ be an ordered sample from a normal population with mean zero and variance σ^2. The joint distribution of $x_1, x_2, \cdots, x_n$ under the condition that $N=\nu$ is

$$f(x_1, \cdots, x_n \mid \nu)=\frac{1}{P_n[N=\nu]}\, n!\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right] \qquad (7)$$

$$(x_1 \leqq x_2 \leqq \cdots \leqq x_\nu \leqq \bar{x} \leqq x_{\nu+1} \leqq \cdots \leqq x_n) .$$

The transformation

$$\left.\begin{aligned} &y_i=x_i-\bar{x} \quad (i=1, \cdots, n-1)\\ &y_n=\bar{x} . \end{aligned}\right. \qquad (8)$$

leads to the simultaneous frequency function of $y_1, \cdots, y_{n-1}$, after integreating out y_n, in the form

$$f(y_1, \cdots, y_{n-1} \mid \nu)=\frac{1}{P_n[N=\nu]}\sqrt{n}\, n!\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1}$$

$$\exp\left[-\frac{1}{2\sigma^2}\{y_1^2+\cdots+y_{n-1}^2+(y_1+\cdots+y_{n-1})^2\}\right], \qquad (9)$$

where y_i are restricted by the relations

$$\left.\begin{array}{lll} y_{i-1} \leqq \ y_i \ \leqq -\dfrac{1}{n-i+1}(y_1+\cdots+y_{i-1}) & & i=n-1, n-2, \cdots, \nu+2 \\ 0 \ \leqq y_{\nu+1} \leqq -\dfrac{1}{n-\nu}(y_1+\cdots+y_\nu) & & \\ y_{j-1} \leqq \ y_j \ \leqq 0 & & j=\nu, \nu-1, \cdots . 2 \\ -\infty < \ y_1 \ \leqq 0 & & \end{array}\right\} D_1 \qquad (10)$$

Furthermore, we make the following transformation

$$\left.\begin{aligned} \sqrt{2\cdot 1}\,\sigma\,\xi_1 &=-y_1+y_2\\ \sqrt{3\cdot 2}\,\sigma\,\xi_2 &=-y_1-y_2+2y_3\\ &\ \,\vdots\\ \sqrt{\nu(\nu-1)}\,\sigma\,\xi_{\nu-1} &=-y_1-y_2-\cdots-y_{\nu-1}+(\nu-1)y_\nu\\ \sqrt{\nu}\,\sigma\,\xi_\nu &=-y_1-y_2-\cdots-y_\nu\\ \sqrt{\frac{n-\nu-1}{n-\nu}}\,\sigma\,\xi_{\nu+1} &=-y_{\nu+1}-\frac{y_1+\cdots+y_\nu}{n-\nu}\\ \sqrt{\frac{n-\nu-1}{n-\nu-1}}\,\sigma\,\xi_{\nu+2} &=-y_{\nu+2}-\frac{y_1+\cdots+y_{\nu+1}}{n-\nu-1}\\ &\ \,\vdots\\ \sqrt{\frac{1}{2}}\,\sigma\,\xi_{n-1} &=-y_{n-1}-\frac{y_1+\cdots+y_{n-2}}{2} \end{aligned}\right\} \qquad (11)$$

Now

$$y_1^2+y_2^2+\cdots+y_{n-1}^2+(y_1+\cdots+y_{n-1})^2$$
$$=\sigma^2\left(\xi_1^2+\cdots+\xi_{\nu-1}^2+\frac{n}{n-\nu}\xi_\nu^2+\xi_{\nu+1}^2+\cdots+\xi_{n-1}^2\right).$$

Since the Jacobian of the above transformation is $(-1)^{n-1}\sigma^{-(n-1)}\sqrt{n-\nu}$, the joint frequency function of $\xi_1, \cdots, \xi_{n-1}$ when $N=\nu$ is given by

$$f(\xi_1, \cdots, \xi_{n-1}|\nu)=\frac{1}{P_n[N=\nu]}\left(\frac{n}{n-\nu}\right)^{\frac{1}{2}} n!\left(\frac{1}{\sqrt{2\pi}}\right)^{n-1}$$
$$\exp\left[-\frac{1}{2}\left\{\xi_1^2+\cdots+\xi_{\nu-1}^2+\frac{n}{n-\nu}\xi_\nu^2+\xi_{\nu+1}^2+\cdots+\xi_{n-1}^2\right\}\right], \quad (12)$$

where the domain of the variables is, from (10) and (11), as follows:

$$D_2\begin{cases} 0\leqq \xi_{n-i}\leqq\sqrt{\dfrac{i+2}{i}}\,\xi_{n-i-1} & (i=1, \cdots, n-\nu-2)\\ 0\leqq\xi_{\nu+1}\leqq\sqrt{\dfrac{\nu}{(n-\nu)(n-\nu-1)}}\,\xi_\nu & \\ 0\leqq\xi_\nu\quad<\infty & \\ 0\leqq\xi_{\nu-1}\leqq\dfrac{1}{\sqrt{\nu-1}}\,\xi_\nu & \\ 0\leqq\xi_j\quad\leqq\sqrt{\dfrac{j+2}{j}}\,\xi_{j+1} & (j=\nu-2, \nu-3, \cdots, 1)\end{cases} \quad (13)$$

Noticing the relation $\xi_\nu=-\dfrac{1}{\sigma\sqrt{\nu}}\sum_{i=1}^{\nu}(x_i-\bar{x})$, then in order to obtain the frequency function of the sum of the negative deviations under the condition that $N=\nu$, it would be seen that we need to integrate out $\xi_1, \cdots, \xi_{\nu-1}, \xi_{\nu+1}, \cdots, \xi_{n-1}$ from (12). Using the F. E. Grubbs' $F_r(x)$ [7] defined in connection with the distribution of the difference between the extreme and sample mean, we easily obtain

$$f(\xi_\nu)=\frac{1}{P_n[N=\nu]}\binom{n}{\nu}\frac{1}{\sqrt{2\pi}}\left(\frac{n}{n-\nu}\right)^{\frac{1}{2}}$$
$$F_\nu\left(\frac{1}{\sqrt{\nu}}\xi_\nu\right)F_{n-\nu}\left(\frac{\sqrt{\nu}}{n-\nu}\xi_\nu\right)\exp\left[-\frac{1}{2}\frac{n}{n-\nu}\xi_\nu^2\right]. \quad (14)$$

From this, the distributions of our statistics $U_n(\nu)$, and $W_n(\nu)$ are given, respectively, in the following way:

$$f(U_n(\nu))=\frac{1}{P_n[N=\nu]}\binom{n}{\nu}\frac{1}{\sqrt{2\pi}\,\sigma}\left(\frac{\nu(n-\nu)}{n}\right)^{\frac{1}{2}}$$
$$F_\nu\left(\frac{n-\nu}{n}\frac{U_n(\nu)}{\sigma}\right)F_{n-\nu}\left(\frac{\nu}{n}\frac{U_n(\nu)}{\sigma}\right)\exp\left[-\frac{1}{2\sigma^2}\frac{\nu(n-\nu)}{n}U_n^2(\nu)\right] \quad (15)$$

and

$$f(W_n(\nu))=\frac{1}{P[N=\nu]}\binom{n}{\nu}\frac{1}{\sqrt{2\pi}\,\sigma}\left(\frac{n^2}{4\nu(n-\nu)}\right)^{\frac{1}{2}}$$
$$F_\nu\left(\frac{n}{2\nu}\frac{W_n(\nu)}{\sigma}\right)F_{n-\nu}\left(\frac{n}{2(n-\nu)}\frac{W_n(\nu)}{\sigma}\right)\exp\left[-\frac{1}{2\sigma^2}\frac{n^3}{4\nu(n-\nu)}W_n^2(\nu)\right] \quad (1\cdot6)$$

Then the distribution of the statistics U_n and W_n, which are eliminated the condition that $N=\nu$, are obtained by

$$f(U_n)=\sum_{\nu=1}^{n-1}f(U_n(\nu))P_n[N=\nu], \quad f(W_n)=\sum_{\nu=1}^{n-1}f(W_n(\nu))P_n[N=\nu] \quad (17)$$

Especially, it is easily seen that $f(W_n)$ obtained above agrees with Godwin's result.

3. $P_n[N=\nu]$

$P_n[N=\nu]$ are the probabilities that the number of observations smaller than $\bar{x}$ is ν in a sample of size n. If we know the values of $P_n[N=\nu]$, we can simply test whether or not our population is normal and also the randomness of samples drawn from a normal population. In the quality control, when we record the number of observations smaller than the total sample mean, $\bar{x}$, in each sample drawn from the population of manufacturing process and find that we have frequently samples of too high or too low values of the number of the smaller observations, it is necessary to doubt and examine the assumption of normality or the existence of the external disturbances in the manufacturing process.

The numerical values of $P_n[N=\nu]$ for $n=2\sim20$ have been tabulated in my paper [8] (in Japanese).

4. **Relative efficiencies of our estimation as compared to the sample standard deviation**

In this section we consider the relative efficiencies, as compared to the sample standard deviation, of our method to estimate the population standard deviation σ based on the statistics $U_n(\nu)$ and U_n. Let $E(x\mid\nu)$ be the mathematical expectation of the random variable x under the condition that $N=\nu$ and let $E(x)$ be the one without condition. Also, we denote the operator of averaging with respect to N by E_N and the variance of a statistic T by $D^2(T)$.

We can consider the following two situations:

(a) To obtain the unbiased estimate of σ as a whole.

(b) To obtain the unbiased estimate of σ in each sample by making use of the knowledge of the number.

Obviously, the method (b) of estimation is more efficient than the method (a), since, in the former case, we use the more information in estimating.

In the method (a), setting

$$E(U_n)=k_n\sigma\,, \tag{18}$$

the statistics U_n/k_n becomes a unbiased estimate of σ as a whole. Then the relative efficiency, as compared to the sample standard deviation, of using the statistic U_n/k_n for estimating σ may be evaluated as the ratio of the variance

$$D^2(U_n/k_n)=\frac{1}{k_n^2}D^2(U_n) \tag{19}$$

to the variance

$$D^2(S_n/c_n)=\frac{1}{c_n^2}\,D^2(S_n), \tag{20}$$

where $S_n=\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}}$ and $c_n=\sqrt{\frac{2}{n}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}=E(S_n)/\sigma\,.$

In the method (b), we use the statistic $U_n(\nu)/k_{n,\nu}$ in order to obtain the unbiased estimate of σ in each sample, where the coefficient $1/k_{n,\nu}$ is calculated from

$$E(U_n\mid\nu)=k_{n,\nu}\,\sigma\,. \tag{21}$$

In this case the efficiency on the whole of our estimation must be evaluated by comparing the variance

$$E_N E\left(\frac{U_n^2}{k_{n,\nu}^2}\mid\nu\right)-\left[E_N E\left(\frac{U_n}{k_{n,\nu}}\mid\nu\right)\right]^2=\sum_{\nu=1}^{n-1}P_n[N=\nu]\,E\left(\frac{U_n^2}{k_{n,\nu}^2}\mid\nu\right)-1 \tag{22}$$

with the variance of S_n/c_n. The tables of the values of the coefficients of unbiasedness $1/k_n$ and $1/k_{n,\nu}$ and the efficiencies of both methods of estimation are inserted in [8] for sample sizes $n=2\sim20$. The general behavior of efficiencies of our estimates is illustrated by Fig. 1 with other estimates.

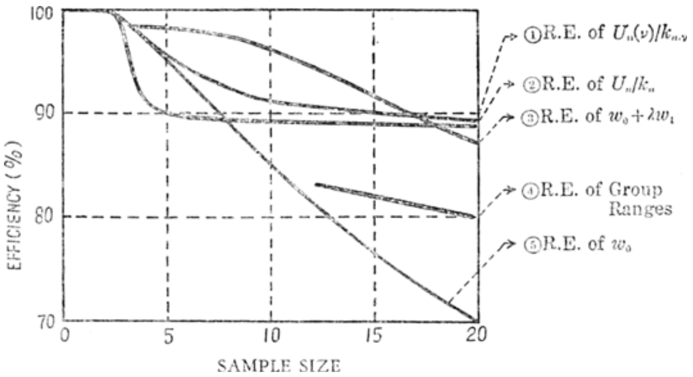

Fig. 1. Efficiencies of the various estimates

In Fig. 1, w_0 and w_1 represent the simple range and the first quasi-range, $x_{n-1}-x_2$, respectively. Curves ③ and ④ are drawn by the results of [3] and [5], respectively.

5. Remarks on computation and further works

In practical calculation of the statistic $U_n(\nu)$ for each sample drawn from the population may be carried out from two means $\bar{x}$ and $\bar{\bar{x}}(\nu)$ by the relation

$$U_n(\nu)=\bar{\bar{x}}(n-\nu)-\bar{\bar{x}}(\nu)=\frac{n}{n-\nu}(\bar{x}-\bar{\bar{x}}(\nu)),$$

hence if we tabulate the coefficients $\dfrac{1}{k_{n,\nu}}\dfrac{n}{n-\nu}$ for each n and ν we can easily obtain the individual value of the unbiased statistic $U_n(\nu)/k_{n,\nu}$ in each sample. If we stand on the viewpoint of making use of the information of the number of observations smaller than $\bar{x}$, we could also efficiently and easily adopt the mean deviation $W_n(\nu)$ for estimating the population standard deviation since

$$W_n(\nu)=\frac{2\nu}{n}(\bar{x}-\bar{\bar{x}}(\nu)).$$

Like the modified t-test using the range instead of the sample standard deviation, we can make another modified t-test using our statistic U_n or $U_n(\nu)$. Since $\bar{x}$ and $U_n(\nu)$ are stochastically independent, it will be easily found that the frequency function of the statistic

$$T_\nu = \frac{\bar{x}-m}{U_n(\nu)/d_{n,\nu}} \tag{23}$$

is given by

$$f(T_\nu) = K(n,\nu)\int_0^\infty x\,F_\nu(x)\,F_{n-\nu}\left(\frac{\nu}{n-\nu}x\right) \exp\left[-\frac{1}{2}\,\frac{n\nu}{n-\nu}x^2\left(1+\frac{n}{d_{n,\nu}^2\,\nu(n-\nu)}T_\nu^2\right)\right]dx, \tag{24}$$

where $d_{n,\nu}=k_n$ or $k_{n,\nu}$ and $K(n,\nu)=\frac{1}{P[N=\nu]}\,\frac{1}{2\pi}\,\frac{1}{d_{n,\nu}}\binom{n}{\nu}\frac{\sqrt{\nu}\,n^2}{(n-\nu)^{3/2}}$. As the frequency function of unconditional statistic T is obtained as

$$f(T)=\sum_{\nu=1}^{n-1}P[N=\nu]f(T_\nu),$$

it is necessary to tabulate the α percent values of T_ν for each n, ν and appropriate significance levels α $(0<\alpha<1)$.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] L. H. C. Tippett, "On the extreme individuals and the range of samples taken from a normal population" Biometrika, Vol. 17, P. 364.

[2] E. S. Pearson and H. O. Hartley, "The probability intergral of the range in samples of n observations from a normal population." Biometrika, Vol. 32, Part III and IV, (1942).

[3] F. E. Grubbs and C. L. Weaver, "The best unbiased estimate of population standard deviation based on group ranges" Journal of the American Statistical Association, Vol. 42, (1947), pp. 224-241.

[4] F. Mosteller, "On some useful 'inffieient' statistics" Ann. Math. Stat., Vol. 17 (1946), pp. 377-408.

[5] J. H. Cadwell, "The distribution of quasi-ranges in samples from a normal population", Ann. Math. Stat., Vol. 24 (1953), pp. 603-613.

[6] H. J. Godwin, "On the distribution of the estimate of mean deviation obtained from samples from a normal population, "Biometrika, Vol. 33 (1945), pp. 254-265.

[7] F. E. Grubbs, "Sample criteria for testing outlying observations", Ann. Math. Stat., Vol. 21 (1950), pp. 27-58.

[8] M. Shiotani, "On the distribution of the sum of the positive or negative deviations from the mean in the sample drawn from the normal population", The Proceedings of the Institute of Statistical Mathematics, (in Japanese), Vol. 2, No. 1 (1954), pp. 63-74.