ON THE TESTING OF OUTLYING OBSERVATIONS

By A. KUDO

Kyusyu University, Fukuoka
and
Indian Statistical Institute, Calcutta

1. Introduction

The problem of testing outlying observations concerning the mean value of normal population has been treated by various authors such as W. R. Thompson (1942), E. S. Pearson and C. Chandrasekar (1936), K. R. Nair (1948), F. E. Grubbs (1950) and N. V. Smirnoff (1940). The statistics suggested by these authors are mostly based on the difference between the extreme value and the mean value, when the variances are assumed to be same and known to us. When the variance is unknown, the statistic is the difference divided by the estimate of the standard deviation. Concerning the latter case the statistics treated by Grubbs and Smirnoff seem to be essentially the same where the standard deviation is estimated from the sample in hand, whereas K. R. Nair suggested that the standard deviation should be estimated from another independent sample. The former is called the Pearson-Chandrasekar statistic. The characters of these statistics, like efficiency or optimum property of power function, have not been fully treated so far. In this paper, we propose statistics to meet the more general situations, where we have got other independent samples containing information about the population mean value and the variance of the population. The Pearson-Chandrasekar statistic is a special case of ours. We shall prove that our statistic is optimum in the sense that it is uniformly best for special class of alternative hypotheses among a certain reasonably restricted class of testing procedures.

In § 2, we shall formulate the problem in the case when the variance is unknown and we shall propose a testing procedure, which will be proved to be optimum in § 3. As a corollary, we shall observe that the Pearson-Chandrasekar statistic is optimum whereas Nair's statistic is not. In § 4, we shall discuss briefly the problem in the case when the variance is known to us.

Concerning the tables to be used for our test procedure, some results have been obtained by the author, which will be discussed in a forthcoming paper.

2. Formulation of the problem

Let $x_i^{(1)}$ $(i = 1, 2, \ldots, N_1)$ be distributed as $N(m_i, \sigma^2)$ $(i = 1, 2, \ldots, N_1)$ respectively, and $x_i^{(2)}$ $(i = 1, 2, \ldots, N_2)$ as $N(m^{(2)}, \sigma^2)$ and $x_i^{(3)}$ $(i = 1, 2, \ldots, N_3)$ as $N(m^{(3)}, \sigma^2)$. We assume that they are mutually independent and the values of these parameters $m_i (i = 1, 2, \ldots, N_1)$, $m^{(2)}$, $m^{(3)}$ and σ are unknown to us.

Our null hypothesis H_0 is $H_0 \equiv H(m_1 = m_2 = \cdots = m_{N_1} = m^{(2)})$ where $m^{(2)}$, $m^{(3)}$ and σ are free. We have N_1 alternative hypotheses $H_1, H_2, \ldots, H_{N_1}$ where $H_i (i = 1, 2, \ldots, N_1)$ is the hypothesis $H_i = H(m_1 = \cdots = m_{i-1} = m_i - \Delta = m_{i+1} = \cdots = m_{N_1} = m^{(2)}, \Delta > 0)$. In other words, under H_0 the observations in the first and the second groups are all from the same population, while under the hypothesis H_i the i-th observation in the first group only is from a different normal population with greater mean and the same variance.

Let D_i $(i = 0, 1, 2, \ldots, N_1)$ be the decision to accept H_i $(i = 0, 1, 2, \ldots, N_1)$. Our problem is to find out an optimum decision procedure as to these N_1+1 decisions.

At first, we have to make the meaning of optimum clear. We introduce the following requirements or criteria:

(1°). We want to accept the decision D_0 with the pre-assigned probability, say, $1-p$, when D_0 is true.

(2°) The decision procedure must be invariant (*a*) the addition of any constant to all the first N_1+N_2 observations and (*b*) under the addition of any constant to all the last N_3 observations.

(3°) The decision procedure must remain invariant when all the observations are multiplied by any positive constant.

The last two conditions require that if we have two groups of observations $(x_1^{(1)}, \ldots, x_{N_1}^{(1)}, x_1^{(2)}, \ldots, x_{N_2}^{(2)}, x_1^{(3)}, \ldots, x_{N_3}^{(3)})$ and $(y_1^{(1)}, \ldots, y_{N_1}^{(1)}, y_1^{(2)}, \ldots, y_{N_2}^{(2)}, y_1^{(3)}, \ldots, y_{N_3}^{(3)})$ and there are the following relations

$$y_j^{(i)} = ax_j^{(i)} + b, \qquad (i = 1, 2;\ j = 1, 2, \ldots, N_i)$$

and

$$y_j^{(3)} = ax_j^{(3)} + c, \qquad (j = 1, 2, \ldots, N_3),$$

where a is positive and b and c are constant, our decision procedure must give the same decision.

(4°) The probability of a correct decision when the i-th population mean is shifted to the right by the same amount must be the same for all i.

(5°) We want to maximize the probability of making a correct decision when D_i is correct.

Now our problem is to find out a decision procedure satisfying the conditions (1°)–(5°).

Let

$$\bar{x}^{(i)} = \sum_{j=1}^{N_i} x_j^{(i)}/N_i, \quad \bar{x} = (N_1\bar{x}^{(1)} + N_2\bar{x}^{(2)})/(N_1+N_2)$$

$$\bar{m}^{(1)} = \sum_{i=1}^{N_1} \bar{m}_i/N_1, \quad \bar{m} = (N_1\bar{m}^{(1)} + N_2 m^{(2)})/(N_1+N_2),$$

$$
\left.
\begin{aligned}
S_i^2 &= \sum_{j=1}^{N_i} (x_j^{(i)} - \bar{x}^{(i)})^2 / N_i \\
S_{12}^2 &= \Big\{ \sum_{j=1}^{N_1} (x_j^{(1)} - \bar{x})^2 + \sum_{j=1}^{N_2} (x_j^{(2)} - \bar{x})^2 \Big\} / (N_1 + N_2) \\
S^2 &= \{(N_1 + N_2) S^2 + N_3 S_3^2\} / (N_1 + N_2 + N_3) \\
S_m^2 &= \Big\{ \sum_{j=1}^{N_1} (m_j^{(1)} - \bar{m})^2 + N_2 (m^{(2)} - \bar{m})^2 \Big\} / (N_1 + N_2) \\
S_m^2(a) &= \frac{(N_1 + N_2 - 1)}{(N_1 + N_2)^2} a^2 \\
& (i = 1, 2, 3).
\end{aligned}
\right\} \qquad \ldots (2.1)
$$

Let M be the suffix of the population which has the greatest sample value among the first N_1 observations, so that

$$x_M = \max_{j=1,2,\ldots,N_1} x_j^{(1)}.$$

Our optimum decision procedure will be shown to be as follows:

$$
\left.
\begin{aligned}
&\text{if } \frac{x_M - \bar{x}}{S} \leqslant \lambda_p, \quad \text{select } D_0, \\
&\text{if } \frac{x_M - \bar{x}}{S} > \lambda_p, \quad \text{select } D_M,
\end{aligned}
\right\} \qquad \ldots (2.2)
$$

where λ_p is a constant which depends on N_1, N_2, N_3 and p. Concerning the numerical calculation of λ_p, when $N_2 = N_3 = 0$, the values have been tabulated by ENIAC (see Grubbs, 1950).

3. Derivation of the optimum procedure

In the usual manner, our decision procedure can be expressed by the following N_1+1 functions, $d_0(\tilde{X}), d_1(\tilde{X}),\ldots, d_{N_1}(\tilde{X})$ where $\tilde{X}$ denotes the observation $(x_1^{(1)},\ldots, x_{N_1}^{(1)}, x_1^{(2)},\ldots, x_{N_2}^{(2)}, x_1^{(3)},\ldots, x_{N_3}^{(3)})$ and $d_i(\tilde{X})$, $(i = 0, 1,\ldots, N_1)$ denote the probabilities of selecting $D_i, (i = 0, 1,\ldots, N_1)$ respectively when the observation $\tilde{X}$ is given. We shall call this set of N_1+1 functions a decision function. Naturally these N_1+1 functions must satisfy the following relations $1 \geqslant d_i(\tilde{X}) \geqslant 0$, $(i = 0, 1,\ldots, N_1)$ and $\sum_{i=0}^{N_1} d_i(\tilde{X}) = 1$. To meet our mathematical necessity we shall assume that they are measurable functions.

By condition (2°) we see that our decision procedure must depend only on the $N_1+N_2+N_3-2$ values of

$$
\left.
\begin{aligned}
y_1^{(1)} &= x_1^{(1)} - x_{N_1}^{(1)},\ldots, y_{N_1-1}^{(1)} = x_{N_1-1}^{(1)} - x_{N_1}^{(1)} \\
y_1^{(2)} &= x_1^{(2)} - x_{N_2}^{(1)},\ldots, y_{N_2}^{(1)} \quad = x_{N_2}^{(2)} - x_{N_1}^{(2)} \\
y_1^{(3)} &= x_1^{(3)} - x_{N_3}^{(2)},\ldots, y_{N_3-1}^{(3)} = x_{N_3-1}^{(3)} - x_{N_3}^{(3)}.
\end{aligned}
\right\} \qquad \ldots (3.1)
$$

Further by condition (3°) we see that our decision procedure must depend only on the $N_1+N_2+N_3-2$ values of

$$\left.\begin{aligned} &Z_1^{(1)} = \frac{y_1^{(1)}}{|y_{N_1-1}^{(1)}|}, \ \cdots, \ Z_{N_1-2}^{(1)} = \frac{y_{N_1-2}^{(1)}}{|y_{N_1-1}^{(1)}|}, \ Z_{N_1-1}^{(1)} = \frac{y_{N_1-1}^{(1)}}{|y_{N_1-1}^{(1)}|}, \\ &Z_1^{(2)} = \frac{y_1^{(2)}}{|y_{N_2-1}^{(2)}|}, \ \cdots, \ Z_{N_2}^{(1)} = \frac{y_{N_2}^{(2)}}{|y_{N_1-1}^{(1)}|}, \\ &Z_1^{(3)} = \frac{y_1^{(3)}}{|y_{N_1-1}^{(1)}|}, \ \cdots, \ Z_{N_3-1}^{(2)} = \frac{y_{N_3-1}^{(3)}}{|y_{N_1-1}^{(1)}|} \end{aligned}\right\} \quad \ldots \ (3.2)$$

Clearly the joint distribution of $Z_i^{(j)}$'s does not depend on any parameter of the population when D_0 is true, while it depends on the value of Δ/σ and the suffix i when D_i is correct.

Let $\tilde{Z}$ denote $N_1+N_2+N_3-2$ ($=M$ say) sample values; $f_0(\tilde{Z})$ be the frequency function of $\tilde{Z}$ when D_0 is correct; $f_i(\tilde{Z}|a)$ be the frequency function of $\tilde{Z}$ when D_i is correct; and $a = \Delta/\sigma$. As we have seen the decision function of our procedure must be a function of $\tilde{Z}$ only, and hence we can write our decision function as $d_0(\tilde{Z})$, $d_1(\tilde{Z}),\ldots, d_{N_1}(\tilde{Z})$.

Then our problem can be formulated as follows. We want to find the decision function $(d_0(\tilde{Z}),\ d_1(\tilde{Z}),\ldots,\ d_{N_1}(\tilde{Z}))$ such that

$$\int d_0(\tilde{Z})\, f_0(\tilde{Z})\, d\tilde{Z} = 1-p \qquad \ldots \ (3.3)$$

and

$$\int d_i(\tilde{Z})\, f_i(\tilde{Z})\,|a)\, d\tilde{Z} \qquad \ldots \ (3.4)$$

is independent of i and is maximum.

Here we need the frequency function of $\tilde{Z}$. At first we shall find the probability density function of $y_j^{(i)}$'s. As $f_0(\tilde{Z})$ does not depend on any parameter and $f_i(\tilde{Z}|a)$ depends only on a, we can assume that $\sigma = 1$, $m^{(2)} = 0$. $m^{(3)} = 0$.

Let

$$M_1 = m_1 - m_{N_1},\ldots, M_{N_1-1} = m_{N_1-1} - m_{N_1}. \qquad \ldots \ (3.5)$$

Then $y_1^{(1)},\ldots, y_{N_1-1}^{(1)}$, $y_1^{(2)},\ldots, y_{N_2}^{(2)}$ are normally distributed with mean values $M_1,\ldots, M_{N_1-1}$, $-m_{N_1}, \ldots, -m_{N_1}$ and common variances equal to 2 and common covariances equal to 1; while $y_1^{(3)},\ldots, y_{N_3-1}^{(3)}$ are distributed with mean values equal to zero and with

same variances and covariances as those of the former group, and these two groups are independent. Therefore the probability density function is given by

$$C \exp\Big[-\frac{1}{2}\frac{1}{N_1+N_2}\Big\{(N_1+N_2)\Big(\sum_{\alpha=1}^{N_1-1}(y_\alpha^{(1)}-M_\alpha)^2+\sum_{\alpha=1}^{N_2}(y_\alpha^{(2)}+m_{N_1})^2\Big)-$$

$$-\Big(\sum_{\alpha=1}^{N_1-1}(y_\alpha^{(1)}-M_\alpha)+\sum_{\alpha=1}^{N_2}(y_\alpha^{(2)}+m_{N_1})^2\Big)\times$$

$$\times \exp\Big[-\frac{1}{2}\frac{1}{N_2}\Big\{N_2\sum_{\alpha=1}^{N_2-1}y_\alpha^{(2)2}-\Big(\sum_{\alpha=1}^{N_2-1}y_\alpha^{(2)}\Big)^2\Big\}\Big] \quad \ldots \quad (3.6)$$

where C is some constant the exact value of which is not needed for our purpose.

By a simple transformation we get the frequency function of

$$f(\bar{Z}) = CP_r(y_{N_1-1}^{(1)}>0)\int_0^\infty t^M \exp\Big[-\frac{1}{2}\frac{1}{N_1+N_2}\Big\{(N_1+N_2)\Big\{\sum_{\alpha=1}^{N_1-2}(Z_\alpha^{(1)}t-M_\alpha)^2+(t-M_{N_1-1})^2+$$

$$+\sum_{\alpha=1}^{N_2}(Z_\alpha^{(2)}t+m_{N_1})-\Big(\sum_{\alpha=1}^{N_1-2}(Z_\alpha^{(1)}t-M_\alpha)+(t-M_{N_1-1})+\sum_{\alpha=1}^{N_2}(Z_\alpha^{(2)}t+m_{N_1})\Big)^2\Big\}\Big]\times$$

$$\times \exp\Big[-\frac{1}{2}\frac{t^2}{N_2}\Big\{N_2\sum_{\alpha=1}^{N_2-1}Z_\alpha^{(2)2}-\Big(\sum_{\alpha=1}^{N_2-1}Z_\alpha^{(2)}\Big)^2\Big\}\Big]dt+$$

$$+CP_r(y_{N_1-1}^{(1)}<0)\int_0^\infty t^M \exp\Big[-\frac{1}{2}\frac{1}{N_1+N_2}\Big\{(N_1+N_2)\Big\{\sum_{\alpha=1}^{N_1-2}(-Z_\alpha^{(1)}t-M_\alpha)^2+(-t-M_{N_1-1})^2+$$

$$+\sum_{\alpha=1}^{N_2}(-Z_\alpha^{(2)}t+m_{N_1})^2\Big\}-\Big(\sum_{\alpha=1}^{N_1-2}(-Z_\alpha^{(1)}t-M_\alpha)-(-t-M_{N_1-1})+\sum_{\alpha=1}^{N_2}(-Z_\alpha^{(2)}t+m_{N_1})^2\Big)\Big\}\Big]\times$$

$$\times \exp\Big[-\frac{1}{2}\frac{t^2}{N_2}\Big\{N_2\sum_{\alpha=1}^{N_2-1}Z_\alpha^{(2)2}-\Big(\sum_{\alpha=1}^{N_2-1}Z_\alpha^{(2)}\Big)^2\Big\}\Big]dt. \quad \ldots \quad (3.7)$$

As we have relations (3.1), (3.2) and (3.5) we can easily see that this is equal to

$$C\,P_r\Big(Z_{N_1-1}=\frac{y_{N_1-1}^{(1)}}{y_{N_1-1}^{(1)}}\Big)\int_0^\infty t^M \exp\Big[-\frac{1}{2}\frac{1}{N_1+N_2}\Big[(N_1+N_2)\Big\{\sum_{\alpha=1}^{N_1-2}\Big(\frac{y_\alpha^{(1)}}{y_{N_1-1}^{(1)}}t-M_\alpha\Big)^2+$$

$$+(t-M_{N_1-1})^2+\sum_{\alpha=1}^{N_2}\Big(\frac{y_\alpha^{(2)}}{y_{N_1-1}^{(1)}}t+m_{N_1}\Big)^2\Big\}-$$

$$-\Big(\sum_{\alpha=1}^{N_1-2}\Big(\frac{y_\alpha^{(1)}}{y_{N_1-1}^{(1)}}\Big)t-M_\alpha\Big)+(t-M_{N_1-1})+\sum_{\alpha=1}^{N_2}\Big(\frac{y_\alpha^{(2)}}{y_{N_1-1}^{(1)}}t+m_{N_1}\Big)\Big\}\Big]\times$$

$$\times \exp\left[-\frac{1}{2}\frac{t^2}{N_2}\left\{N_2\sum_{\alpha=1}^{N_1-1}\left(\frac{y_\alpha^{(1)}}{y_{N_1-1}^{(1)}}\right)^2-\left(\sum_{\alpha=1}^{N_1-1}\frac{y_\alpha^{(1)}}{y_{N_1-1}^{(1)}}\right)^2\right\}\right]dt$$

$$= C P_1\left(Z_{N_1-1}=\frac{y_{N_1-1}^{(1)}}{|y_{N_1-1}^{(1)}|}\right)\int_0^\infty t^M \exp\left[-\frac{1}{2}\frac{1}{N_1+N_2}\frac{1}{y_{N_1-1}^{(1)2}}\left\{(N_1+N_2)\times\right.\right.$$

$$\times\left\{\sum_{\alpha=1}^{N_1-2}(y_\alpha^{(1)}t-y_{N_1-1}^{(1)}M_\alpha)^2+y_{N_1-1}^{(1)2}(t-M_{N_1-1})^2+\sum_{\alpha=1}^{N_2}\left(y_\alpha^{(2)}t+y_{N_1-1}^{(1)}m_{N_1}\right)^2\right\}-$$

$$\left.\left.-\left(\sum_{\alpha=1}^{N_1-1}y_\alpha^{(1)}t-y_{N_1-1}^{(1)}\left(\sum_{\alpha=1}^{N_1-1}M_\alpha\right)+\sum_{\alpha=1}^{N_2}y_\alpha^{(2)}t+y_{N_1-1}^{(1)}m_{N_1}\right)^2\right\}\right]\times$$

$$\times\exp\left[-\frac{1}{2}t^2\frac{1}{N_2\,y_{N_1-1}^{(1)2}}\left\{N_2\left(\sum_{\alpha=1}^{N_2-1}y_\alpha^{(2)2}-\left(\sum_{\alpha=1}^{N_2-1}y_\alpha^{(2)}\right)^2\right)\right\}\right]dt$$

$$= C|y_{N_1-1}^{(1)}|^{M+1}\int_0^\infty t^M\exp\left[-\frac{1}{2}\frac{1}{N_1+N_2}\left\{t^2\left\{(N_1+N_2)\left(\sum_{\alpha=1}^{N_1-1}y_\alpha^{(1)2}+\sum_{\alpha=1}^{N_2}y_\alpha^{(2)2}\right)-\right.\right.\right.$$

$$\left.\left.\left.-\left(\sum_{\alpha=1}^{N_1-1}y_\alpha^{(1)}+\sum_{\alpha=1}^{N_2}y_\alpha^{(2)}\right)^2\right\}-\right.$$

$$\left.-2t\left\{(N_1+N_2)\left(\sum_{\alpha=1}^{N_1-1}y_\alpha^{(1)}M_\alpha-\sum_{\alpha=1}^{N_2}y_\alpha^{(2)}m_{N_1}\right)-\left(\sum_{\alpha=1}^{N_1-1}y_\alpha^{(1)}+\sum_{\alpha=1}^{N_2}y_\alpha^{(2)}\right)\left(\sum_{\alpha=1}^{N_1-1}M_\alpha-N_2m_{N_1}\right)\right\}+\right.$$

$$\left.+\left\{(N_1+N_2)\left(\sum_{\alpha=1}^{N_1-1}M_\alpha^2+N_2m_{N_1}^2\right)-\left(\sum_{\alpha=1}^{N_1-1}M_\alpha-N_2m_{N_1}\right)^2\right\}\right\}\right]\times$$

$$\times\exp\left[-\frac{1}{2}\frac{t^2}{N_2}\left\{N_2\left(\sum_{\alpha=1}^{N_2}y_\alpha^{(2)2}\right)-\left(\sum_{\alpha=1}^{N_2}y_\alpha^{(2)}\right)^2\right\}\right]dt$$

$$= C|x_{N_1-1}^{(1)}-x_{y_1}^{(1)}|^{M+1}\exp\left[-\frac{1}{2}(N_1+N_2)S^2\right]\times$$

$$\times\int_0^\infty\exp\left[-\frac{t^2}{2}(N_1+N_2+N_3)S^2\right]\exp\left[t(N_1+N_2)\left\{\sum_{\alpha=1}^{N_1}x_\alpha^{(1)}m_\alpha-\bar{x}\left(\sum_{\alpha=1}^{N_1}m_\alpha\right)\right\}\right]dt. \quad \ldots \quad (3.8)$$

It should be noted here that this is not the joint density function of $\tilde{X}$, but is only an expression for the frequency function of $\tilde{Z}$ in terms of $\tilde{X}$. This enables us to make our discussion simple.

Now let us consider

$$\int A d_0(\tilde{Z}) f_0(\tilde{Z}) + \sum_{i=1}^{N_1} d_i(\tilde{Z}) f_i(\tilde{Z}|a) d\tilde{Z}, \qquad \ldots (3.9)$$

This value depends on the decision function and the values of A and a. The decision function for which (3.9) attains the maximum value when A and a fixed, is easily proved to be the following

$$d_0(\tilde{Z}) = 1 \qquad \text{if } A f_0(\tilde{Z}) > f_j(\tilde{Z}) \text{ for all } j\,; \qquad \ldots (3.10)$$

$$d_i(\tilde{Z}) = 1 \qquad \text{if } f_i(\tilde{Z}) > A f_0(\tilde{Z}),\ f_i(\tilde{Z}) > f_j()\tilde{Z} \qquad (3.11)$$

for all $j(i \neq j)$.

For other points we define the value of decision function and arbitrarily, because it does not affect the value of the integral (3.9).

We shall prove that for any positive value a and probability p, $0 < p < 1$, there exists a positive number A such that (3.9) attains its maximum for the decision function (2.2) with probability $1-p$ of selecting D_0 when D_0 is the correct decision.

It will then immediately follow that the procedure (2.2) is optimum. Because if it is not optimum, there must be another decision procedure with the same value for the integral (3.3) but with a greater value for (3.4). Therefore, the value of (3.9) must increase. This leads us to a contradiction.

It is sufficient to prove that the following relations hold for some suitably chosen A.

$$A f_0(\tilde{Z}) \geqslant f_i(\tilde{Z}|a) \text{ if and only if } (x_i - \bar{x})/s \leqslant \lambda_p. \qquad \ldots (3.12)$$

$$f_i(\tilde{Z}|a) \geqslant f_j(\tilde{Z}|a) \text{ if and only if } x_i \geqslant x_j. \qquad \ldots (3.13)$$

This is so since these relations show that the decision function (3.2) makes the value of (3.9) maximum because of the relations (3.10) and (3.11). As

$$\{\tilde{X};\ f_i(\tilde{Z}|a) \geqslant f_j(\tilde{Z}|a)\} = \{\tilde{X}\,;\ g_i(\tilde{X}|a) \geqslant g_j(\tilde{X}|a)\}$$

where $\quad g_i(\tilde{X}|a) = g(\tilde{X}|m_1 = \cdots = m_{i-1} = m_{i+1} = \cdots = m_{N_1} = 0,\ \ m_i = a).$

We have

$$g_i(\tilde{X}|a) - g_j(\tilde{X}|a) = C\left|x_{N_1-1}^{(1)} - x_{N_1}^{(1)}\right|^{M+1} \exp\left(-\frac{1}{2}(N_1+N_2)S_m^2(a)\right) \times$$

$$\times \int_0^\infty t^M \exp\left(-\frac{t^2}{2}(N_1+N_2+N_3)S^2\right) \exp\left[\left\{at(x_i-\bar{x})\right\} - \exp\left\{-at(x_j-\bar{x})\right\}\right] dt. \ldots (3.14)$$

This is non-negative if and only if $x_i \geq x_j$. Therefore (3.13) is proved.

Similarly, $Af_0(\tilde{Z}) \geq f_1(\tilde{Z}|a)$ is equivalent to

$$Ag_0(\tilde{X}) - g_1(\tilde{X}|a)$$

$$= C|x_{N_1-1}^{(1)} - x_{N_1}^{(1)}|^{M+1} \int_0^\infty t^M \exp\left(-\frac{t^2}{2}(N_1+N_2+N_3)S^2\right) \times$$

$$\times \left[A - \exp\left[-\frac{1}{2}(N_1+N_2)S_m^2(a)\right] \exp\left[a(x_i-\bar{x})\right]\right] dt. \quad \ldots \quad (3.15)$$

By a simple transformation this is equivalent to

$$\int_0^\infty \xi^M \exp\left[-\frac{1}{2}(N_1+N_2+N_3)\xi^2\right] \times$$

$$\times \left[A - \exp\left(-\frac{(N_1+N_2+N_3)}{2} S_m^2(a)\right) \exp\left(a\xi \frac{x_i-\bar{x}}{S}\right)\right] d\xi > 0. \quad \ldots \quad (3.16)$$

As the integral in (3.16) is a continuous increasing function of A and a continuous decreasing function of $(x_i-\bar{x})/S$, for any a and p there exists a positive number A such that (3.16) holds if and only if $(x_i-\bar{x})/S < \lambda_p$. Therefore (3.12) is proved. This concludes the proof.

4. AN OPTIMUM DECISION PROCEDURE WHERE σ IS KNOWN

In case when σ is known to us, everything becomes quite simple. Under the same notation as in § 2, we can easily see that the procedure will not depend on the values of $x_1^{(2)} - x_{N_2}^{(2)}$ but it will depend only on $x_1^{(1)}, \ldots, x_{N_1}^{(1)}, x_1^{(2)}, \ldots, x_{N_2}^{(2)}$. Naturally we must cancel the condition (2°). Our optimum solution is found to be the following:

$$\text{If } \frac{x_M - \bar{x}}{\sigma} < \lambda_p, \text{ select } D_0,$$

$$\text{If } \frac{x_M - \bar{x}}{\sigma} > \lambda_p \text{ select } D_M, \quad \ldots \quad (4.1)$$

where λ_p is a constant which depends on N_1, N_2 and p.

We are not going into the details of the derivation of the optimum solution, as it is exactly similar to that of § 3.

In this case, our decision procedure will depend only on N_1+N_2-1 values of $y_1^{(1)}, \ldots, y_{N_1-1}^{(1)}$ and $y_1^{(2)}, \ldots, y_{N_2}^{(2)}$ defined by (3.1). Instead of the joint density function $(\tilde{Z}|M_1, \ldots, M_{N_1-1}, m_{N_1})$ which was required in § 3 and given in (3.7), we used the density function of $y_1^{(1)}, \ldots, y_{N_1}^{(1)}, y_1^{(2)}, \ldots, y_{N_2}^{(2)}$.

Let us write it as $f(\bar{Y}\,|\,M_1,\ldots,M_{N_1-1},\ m_{N_1})$. This is obviously given by

$$f(\bar{Y}\,|\,M_1,\ldots,\ M_{N_1-1},\ m_{N_1})$$

$$= C\exp\left[-\frac{1}{2}\frac{1}{(N_1+N_2)}\left\{(N_1+N_2)\left(\sum_{\alpha=1}^{N_1-1}(y_\alpha^{(1)}-M_\alpha)^2+\sum_{\alpha=1}^{N_2}(y_\alpha^{(2)}+m_{N_1})^2\right)-\right.\right.$$

$$\left.\left.-\left(\sum_{\alpha=1}^{N_1-1}(y_\alpha^{(1)}-M_\alpha)+\sum_{\alpha=1}^{N_2}(y_\alpha^{(2)}+m_{N_1})\right)^2\right\}\right] \qquad \ldots\ (4.2)$$

where C is some constant. This is equal to the following function of $x_1^{(1)},\ldots,\ x_{N_1}^{(1)}$, $x_1^{(2)},\ \ldots,\ x_{N_2}^{(2)}$),

$$g(x_1^{(1)},\ldots,\ x_{N_1}^{(1)},\ x_1^{(2)},\ldots,\ x_{N_2}^{(2)}\,|\,m_1,\ldots,\ m_{N_1})$$

$$= C\exp\left[-\frac{1}{2}(N_1+N_2)\left\{S_{12}^2-2\left(\sum_{\alpha=1}^{N_1}m_\alpha x_\alpha^{(1)}-\left(\sum_{\alpha=1}^{N_1}m_\alpha\right)\bar{x}\right)+S_m^2\right\}\right]. \qquad \ldots\ (4.3)$$

Proceeding as in § 3 we can easily prove that the decision procedure given in (4.1) is optimum. The details are left to the reader.

In the case $N_2 = 0$ tables of the values of λ_p have been given by Grubbs (1950) and Nair (1948).

5. Further discussions

The arguments given in this paper can be generalized in various ways. For instance, if we know that the population mean values of first N_1 observations are all equal to that of the second group of observations or only one is different i.e. shifted to the right or the left, then the optimum solution (in the same sence as before) will be found to be based on the statistic $|x_M-\bar{x}|/s$ which is the maximum of $|x_i-\bar{x}|/s$ $(i = 1, 2,\ldots, N_1)$. On the other hand, if we know that 2 mean values are shifted to the right by the same amount, the optimum procedure will be based on the statistic $(x_{M_1}+x_{M_2}-2\bar{x})/s$ which is the maximum of $(x_i + x_j-2\bar{x})/s$ $(i\neq j.\ \ i, j = 1, 2,\ldots, N_1)$. This does not seem to coincide with the statistic proposed by Grubbs for testing two outlying observations.

As we should not select our probability set-up and the decision procedure after getting our samples, we must be very careful in using these procedures. Further investigations on the following problem are obviously necessary. Suppose we have a group of observations and we know *a priori* that they are from the same population or from

two different populations and we have only a vague knowledge about the actual values of the parameters of these two possible cases. For instance, in the notations of § 2, we may have the alternation hypotheses that the $i_1, i_2, \ldots, i_{n-1}$ and i_n-th ($1 \leqslant i_1 < i_2 < \ldots < i_n \leqslant N_1$, $n = 1, 2, \ldots, N_1$) observations are from another normal population with a different mean and the same variance. In this case we have $2^{N_1} - 1$ alternative hypotheses. Our problem is how to make decision concerning these hypotheses.

6. ACKNOWLEDGEMENT

The author is deeply grateful to the Indian Statistical Institute, where this work was completed and to Mr. D. Basu for his kind suggestions.

REFERENCES

GRUBBS, F. E. (1950) : Sample criteria for testing outlying observations. *Ann. Math. Stat.*, 21, 27-58.

NAIR, K. R. (1948) : The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, 35, 118-144.

PEARSON, E. S. and CHANDRASEKAR, C. (1936) : The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28, 308-320.

SHIRNOFF, N. V. (1940) : On the estimation of the maximum term in a series of observations. *C. R. (Doklady). Acad. Sci.*, New Series, 28, 346-350.

THOMPSON, W. R. (1942) : On a criterion of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, 32, 301-310.

Paper received : September, 1954.