TESTING FOR THE SINGLE OUTLIER IN A REGRESSION MODEL

By K. S. SRIKANTAN*

Indian Statistical Institute

SUMMARY. The general problem of testing a regression model against an alternative hypothesis of a single outlier is treated in this paper. Certain test criteria are developed and their 'nominal percentage points', which control the significance of the tests at or below prescribed levels, are obtained.

INTRODUCTION

An experimenter may be confronted with the hypothesis that the observations he has obtained are independent and normal with a linear regression on a known set of variables. He may contemplate an alternative in which one or a specified number of observations, *though he does not know precisely which*, can deviate from the assumed regression or can have increased variability. Here we have to test the hypothesis of the given regression against the alternative of a specified number of 'outliers'. The test should be formulated in such a manner that it is 'best' in detecting any of the alternatives when they happen to be true and, at the same time, controls the error of rejecting the hypothesis, when true, at a preassigned level of probability. This is the problem of testing for outliers.

This problem differs from the analysis of variance. In the latter the regression model is tested against all possible alternatives of shift in the means of the observations. But in the outlier problem the alternatives are restricted to the deviation in the mean of one or a specified number of any of the observations from the given type of regression.

It is seen, therefore, that the outlier test is appropriate to a situation in which we possess the additional information (over and above that required for the analysis of variance) that only a certain number of any of the observations can deviate, in their mean, from the assumed regression. In some problems such information is forthcoming. For example, in factories, it might be found, from past experience, that at a time only one or two machines went out of alignment.

When the number of outliers contemplated in the alternative hypothesis is just one, the test is called a single outlier test. This paper deals with such tests.

The problem of testing for a single outlier when the hypothesis is that all the observations come from identical and independent normal populations with the same but unknown mean and variance has been completely solved. The first significant contribution towards this problem was made by Pearson and Chandra Sekhar (1936) based on the work of Thompson (1935). They suggested certain test criteria and also obtained their percentage points in small samples. Later on percentage

* Now with Central Statistical Organisation, New Delhi.

points for these criteria were obtained for larger sample sizes also by Grubbs (1950). Percentage points for a slightly different criterion involving an external estimate of variance were obtained by a different method by Nair (1948).

The present paper deals with the more general problem of the single outlier in a regression model. This, no doubt, is a more difficult problem. In this case, the distributions of the test criteria, apart from being difficult to solve, should involve, when solved, as parameters, certain functions of the known variables on which the observations have a linear regression. In order to avoid this only the 'nominal percentage points' of these criteria have been obtained. The 'nominal percentage points' always control the error of the first kind at a level not exceeding the specified one; and under some conditions, which hold in restricted cases in small samples, these points control the error of the first kind actually at the specified level. 'Nominal percentage points' are used in this special sense throughout this paper.

The method followed in this paper, for obtaining the nominal percentage points, would be seen to resemble in some respects that adopted by Pearson and Chandra Sekhar (1936) and in certain others that of Cochran (1941). In the next section we formulate the problem of the single outlier and the criteria for its detection.

2. Single outlier in a regression model

2.1. *Preliminaries*. Let $\underline{y} = (y_1, y_2, \ldots, y_n)$ be n independently and normally distributed random variables with the same unknown variance σ^2 and means to be specified later according to the alternative hypotheses considered. Let $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ be m unknown parameters and $X = \{x_{ij}\}$ be a matrix of $m \times n$ known constants of rank m where $m < n$. Further let

$$\mu_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_m x_{mj}. \qquad \ldots \quad (2.1)$$

If $E(y_j) = \mu_j$, $j = 1(1)n$, then the least square estimate of $\underline{\beta}$ is given by

$$\underline{b} = \underline{y}.X'.(X.X')^{-1}. \qquad \ldots \quad (2.2)$$

We shall represent the deviations of the observed values from their regression estimates by the vector

$$\underline{e} = (e_1, e_2, \ldots, e_n);$$

thus

$$\underline{e} = \underline{y} - \underline{b}.X. \qquad \ldots \quad (2.3)$$

It could be shown that the dispersion matrix of $\underline{e}$ is $\Lambda\sigma^2$ where Λ is the $n \times n$ matrix

$$\{\lambda_{ij}\} = I - X'.(X.X')^{-1}.X. \qquad \ldots \quad (2.4)$$

The error sum of squares will reduce to

$$S^2 = \underline{e} \cdot \underline{e}' = \underline{y} \cdot \Lambda \cdot \underline{y}'. \qquad \ldots \quad (2.5)$$

We shall use the following statistics:

$$d_i = e_i.\{\lambda_{i,i} S^2\}^{-\frac{1}{2}}, \qquad \ldots \quad (2.6)$$

$$t_i = d_i^2, \qquad \ldots \quad (2.7)$$

$$u_i = d_i^2 \quad \text{if} \quad d_i \geqslant 0$$

and

$$= 0 \quad \text{if} \quad d_i < 0 \qquad \ldots \quad (2.8)$$

$$l_i = 0 \quad \text{if} \quad d_i \geqslant 0$$

and

$$= d_i^2 \quad \text{if} \quad d_i < 0. \qquad \ldots \quad (2.9)$$

It is well known from the general theory of least squares that d_i is symmetrically distributed and that t_i follows the Beta distribution with parameters

$$(\tfrac{1}{2}, (n-m-1)/2).$$

Consequently, for any $x > 0$ we have

$$\text{Prob } (u_i \geqslant x) = \text{Prob } (d_i \geqslant \sqrt{x}) = \tfrac{1}{2} - \tfrac{1}{2} I_x(\tfrac{1}{2}, (n-m-1)/2) \qquad \ldots \quad (2.10)$$

where

$$I_x(p, q) = \left\{ \int_0^x \xi^{p-1}(1-\xi)^{q-1} d\xi \right\} / B(p, q) \qquad \ldots \quad (2.11)$$

Similarly for $x > 0$,

$$\text{Prob } (l_i \geqslant x) = \text{Prob } (d_i \geqslant -\sqrt{x}) = \tfrac{1}{2} - \tfrac{1}{2} I_x(\tfrac{1}{2}, (n-m-1)/2). \qquad \ldots \quad (2.12)$$

2.2. *Test for a single specified outlier.* When it is specified that the i-th observation is an outlier, we can set up

$$E(y_j) = \mu_j, \; j \neq i,$$

and

$$E(y_i) = \mu_i + \delta_i \qquad (i \text{ specified and } \mu\text{'s given by } (2.1)),$$

and express the null hypothesis H_0 and the alternative hypothesis H in terms of δ_i as follows:

$$H_0 : \delta_i = 0$$

and

$$H \; : \delta_i = \delta \neq 0.$$

Let k_1 be the upper $200\alpha\%$ point and k_2 the upper $100\alpha\%$ point of the distribution of t_i. It could be easily shown that against one sided alternatives, uniformly most powerful similar regions exist and, corresponding to the level of significance α, these are given by

$$u_i \geqslant k_1 \quad \text{if} \quad \delta > 0 \qquad \ldots \quad (2.13)$$

and

$$l_i \geqslant k_1 \quad \text{if} \quad \delta < 0. \qquad \ldots \quad (2.14)$$

Against both sided alternatives, the best unbiassed test, which controls the error of the first kind at the level α, is

$$t_i \geqslant k_2. \qquad \ldots \quad (2.15)$$

2.3. *Test for a single unspecified outlier.* Let

$$t = \max(t_1, t_2, \ldots, t_n) = \max_i \; e_i^2/\lambda_{i,i} \; \underset{\sim}{y} \, \Lambda \, \underset{\sim}{y}' \qquad \ldots \quad (2.16)$$

$$u = \max(u_1, u_2, \ldots, u_n) = \max_i \; e_i|e_i|/\lambda_{i,i} \, \underset{\sim}{y} \, \Lambda \, \underset{\sim}{y}' \qquad \ldots \quad (2.17)$$

and

$$l = \max(l_1, l_2, \ldots, l_n) = \max_i -e_i/e_i|/\lambda_{i,i} \, \underset{\sim}{y} \, \Lambda \, \underset{\sim}{y}' \qquad \ldots \quad (2.18)$$

where λ's are defined by (2.4) and e's by (2.2) and (2.3).

When the outlier is not specified, the results of Section 2.2 immediately suggest the use of one of the criteria t, u or l, these having the common property of being the maximum of the studentised squared deviations in some appropriate sense. The criterion u should be used when the alternative hypothesis is that the expected value of the outlier exceeds that given by the regression model; l should be used against one sided alternatives in the other direction and t against both sided alternatives. Though from intuitive considerations these tests seem to be justifiable, their performance will not be examined in this paper.

The problem investigated here is the evaluation of the percentage points of these criteria. This will be discussed in the next section.

3. Evaluation of the percentage points of the distribution of a maximum

3.1. *Nominal percentage points.* Let $(v_1, v_2, \ldots, v_n)$ be n random variables and further let $v = \max(v_1, v_2, \ldots, v_n)$. Also let

$$P(V) = \text{Prob}\,(v \geqslant V) \qquad \ldots \quad (3.1)$$

$$P_1(V) = \sum_{i=1}^{n} \text{Prob}\,(v_i \geqslant V) \qquad \ldots \quad (3.2)$$

and

$$P_2(V) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{Prob}\,(v_i \geqslant V, v_j \geqslant V). \qquad \ldots \quad (3.3)$$

Then we could derive from the probability laws for the joint occurrence of n events, that

$$P_1(V)-P_2(V) < P(V) < P_1(V). \quad \ldots \quad (3.4)$$

If V_0 is the exact upper $100\alpha\%$ point of the distribution of v, then

$$P(V_0) = \alpha. \quad \ldots \quad (3.5)$$

The quantity V_1 defined by $\qquad P_1(V_1) = \alpha \quad \ldots \quad (3.6)$

will be called the nominal upper $100\alpha\%$ point of the distribution of v. From (3.4) it follows that

$$P(V_1) \leqslant \alpha. \quad \ldots \quad (3.7)$$

Thus, in testing any alternative hypothesis against the upper tail determined by the nominal percentage point V_1, the actual level of significance cannot exceed α. Further if $P_2(V_1)$ happens to be zero then the nominal percentage point coincides with the actual one; a sufficient condition for this is:

$$\text{Prob}\ (v_i \geqslant V_1,\ v_j \geqslant V_1) = 0 \quad \text{for all} \quad i \neq j. \quad \ldots \quad (3.8)$$

For the test criteria developed in Section 2.3, the actual percentage points are difficult to evaluate and would, in general, depend on certain parametric functions of the unknown variables (x's) on which the observations have a linear regression. In order to avoid this only nominal percentage points will be obtained in this paper. However, for special types of regression functions, it will be shown that for samples of sizes below a specified number, these nominal percentage points turn out to be the actual ones. For this purpose we require the two lemmas proved below.

3.2. *Two useful lemmas.* Let Z_1, Z_2 and W be three real valued random variables such that

$$\text{Prob}\ (W \leqslant 0/Z_1, Z_2) = 0 \quad \text{for all } Z_1 \text{ and } Z_2. \quad \ldots \quad (3.9)$$

Let ρ be a real number where $|\rho| < 1$ and let

$$Q^2 = \{(Z_1^2-2\rho Z_1Z_2+Z_2^2)/(1-\rho^2)\}+W \quad \ldots \quad (3.10)$$

and $R_i = Z_i/Q$ $(i = 1, 2,)$. Further let h be any given positive number. Then we have the following.

Lemma 1: Prob $(R_1 \geqslant h, R_2 \geqslant h) = 0$ if $2h^2 \geqslant 1+\rho$ *and*

Lemma 2: Prob $(|R_1| \geqslant h, |R_2| \geqslant h) = 0$ if $2h^2 \geqslant 1+|\rho|$.

Proof: Obviously,

$$U = R_1^2-2\rho R_1R_2+R_2^2 < 1-\rho^2 \text{ by (3.9) and (3.10).}$$

But $$U = (R_1-R_2)^2+2R_1R_2(1-\rho).$$

Hence when $$R_1 \geqslant h \text{ and } R_2 \geqslant h,$$

$$U \geqslant 2h^2(1-\rho) \geqslant 1-\rho^2 \quad \text{if} \quad 2h^2 \geqslant 1+\rho.$$

Therefore we have Lemma 1.

Again, $$U \geqslant |R_1|^2-2|\rho|\,|R_1|\,|R_2|+|R_2|^2 = (|R_1|-|R_2|)^2+2|R_1|\,|R_2|(1-|\rho|).$$

Thus when $$|R_1| \geqslant h \text{ and } |R_2| \geqslant h,$$

$$U \geqslant 2h^2(1-|\rho|) \geqslant 1-\rho^2 \quad \text{if} \quad 2h^2 \geqslant 1+|\rho|$$

and Lemma 2 follows.

3.3. *Evaluation of nominal percentage points of u and l.* Since $u = \max(u_1, u_2, \ldots, u_n)$ it follows from (3.6) and (2.10) that the nominal upper $100\alpha\%$ point of u is given by the following equation in x:

$$(n/2)\,.\,[1-I_x(\tfrac{1}{2},\,(n-m-1)/2)] = \alpha. \qquad \ldots \quad (3.11)$$

It also follows from (3.8) that a sufficient condition for this nominal percentage point to coincide with the actual one is that

$$\text{Prob }(u_i \geqslant x,\, u_j \geqslant x) = 0 \quad \text{for all } i \neq j,$$

that is, $$\text{Prob }\; (d_i \geqslant \sqrt{x},\;\; d_j \geqslant \sqrt{x}) = 0.$$

By an appropriate real linear transformation from

$$(y_1, y_2, \ldots, y_n) \quad \text{to} \quad (z_1, z_2, \ldots, z_n),$$

d_i and d_j could be represented as

$$d_i = z_1/S \quad \text{and} \quad d_j = z_2/S$$

where $$S^2 = \{(z_1^2-2\rho_{i,j}z_1z_2+z_2^2)/(1-\rho_{i,j}^2)\}+z_3^2+z_4^2+\ldots+z_{n-m}^2 \qquad \ldots \quad (3.12)$$

with $\rho_{i,j} = \lambda_{i,j}\{\lambda_{i,i}\lambda_{j,j}\}^{-\frac{1}{2}}$ (λ's being defined by (2.4)), and Prob $(S \leqslant 0/z_1, z_2) = 0$ for all z_1 and z_2. Applying Lemma 1 we get the following necessary condition for the nominal percentage point x to coincide with the actual one :

$$2x \geqslant 1+\rho_{i,j} = 1+\lambda_{i,j}\{\lambda_{i,i}\lambda_{j,j}\}^{-\frac{1}{2}} \qquad \ldots \quad (3.13)$$

(λ's being defined by (2.4)).

The nominal percentage points of l turn out to be the same as those of u in virtue of (2.12); (3.13) becomes a sufficient condition for the coincidence of the nominal and actual percentage points of l.

3.4. *Evaluation of the nominal percentage points of t.* Since $t = \max(t_1, t_2, \ldots, t_n)$ where t_i follows the B distribution with parameters $(\frac{1}{2}, (n-m-1)/2)$ (Section 2.1), it is seen from (3.6) that the nominal upper $100\alpha\%$ point of the distribution of t is given by the equation in x:

$$n[1-I_x(\tfrac{1}{2}, (n-m-1)/2)] = \alpha. \qquad \ldots \quad (3.14)$$

A sufficient condition for x to be the actual percentage point is that

$$\text{Prob} \quad (t_i \geqslant x,\ t_j \geqslant x) = 0 \text{ for all } i \neq j;$$

that is,
$$\text{Prob} \quad (|d_i| \geqslant \sqrt{x},\ \ |d_j| \geqslant \sqrt{x}) = 0.$$

Proceeding as in Section 3.3 and using Lemma 2, we find that a sufficient condition for this is

$$2x \geqslant 1+|\rho_{i,j}| = 1+|\lambda_{i,j}|\{\lambda_{i,i}\lambda_{j,j}\}^{-\frac{1}{2}} \qquad \ldots \quad (3.15)$$

where λ's are defined by (2.4).

4. Scope of the tables of nominal percentage points

Tables of the nominal upper 5% and 1% points of the distribution of the criteria u(or l) and t are presented in Tables 1 to 3 for sample sizes up to $n = 20$ and regression variables $m = 1, 2$ and 3. These were computed by solving equations (3.11) and (3.14) for $\alpha = 0.05$ and 0.01. For this purpose Karl Pearson's Incomplete Beta Function Tables were used and Newton's divided difference formula was applied to carry out inverse interpolation of the second order. The results are expected to be correct to within one digit in the fourth decimal place.

5. Special regressions: coincidence of nominal and actual percentage points

5.1. *Regression on one variable* ($m = 1$). For the simple case $E(y_j) = \beta$, $j = 1(1)n$, the test criteria u and l reduce to those formulated by Pearson and Chandra Sekhar (1936) and Grubbs (1950). We have, in fact,

$$u = 1-S_n^2/S^2 \quad \text{(Grubbs)} \qquad \ldots \quad (5.1)$$

$$= \{T^{(n)}\}^2/(n-1) \quad \text{(Pearson and Chandra Sekhar)} \qquad \ldots \quad (5.2)$$

and

$$l = 1-S_1^2/S^2 \quad \text{(Grubbs)} \qquad \ldots \quad (5.3)$$

$$= \{T^{(1)}\}^2/(n-1) \quad \text{(Pearson and Chandra Sekhar).} \qquad \ldots \quad (5.4)$$

It was found that the nominal percentage points of u and l agreed to all four places of decimals with the actual ones obtained by Grubbs.

Since, in this case,

$$\rho_{i,j} = (-1)/(n-1) \quad \text{for all} \quad i \neq j,$$

the nominal 5% points of u (or l) are actual up to sample size 14 and the 1% points up to 19. The nominal 5% points of t are up to sample size 13 and 1% points up to 18.

5.2. *Regression on two variables* ($m = 2$). An important special regression is

$$E(y_j) = \beta_1+\beta_2 . j, \quad j = 1(1)n. \qquad \ldots \quad (5.5)$$

In this case we get after simplification,

$$\Lambda = I-\{\alpha_{i,j}\} \qquad \ldots \quad (5.6)$$

where $\alpha_{i,j} = \{(n^2-1)+12(i-[n+1]/2)(j-[n+1]/2)\}/n(n^2-1)$. Therefore, max $\rho_{i,j} = 2/(n-1)$. Thus for this regression the nominal 5% points of u (or l) are actual up to sample size 10, and the 1% points are actual up to 15. Again,

$$\max\ |\rho_{i,j}| = 4.\{n^2-2n+13\}^{-1}.$$

Therefore, the nominal 5% points of t are actual up to sample size 9 and the 1% points up to 14.

5.3. *Regression on three variables* ($m = 3$). The special regression considered is

$$E(y_j) = \beta_1+\beta_2 \sin\ (2\pi j/\lambda)+\beta_3 \cos\ (2\pi j/\lambda) \qquad \ldots \quad (5.7)$$

where λ is a known positive integer and $j = 1(1)n$.
When n is a multiple of λ, we have

$$\Lambda = I-\{\alpha_{i,j}\} \qquad \ldots \quad (5.8)$$

where $$\alpha_{i,j} = \{1+2 \cos [2\pi(i-j)/\lambda]\}/n.$$

Thus $$\rho_{i,j} = -\{1+2 \cos [2\pi(i-j)/\lambda]\}/(n-3) \leqslant 1/(n-3). \qquad \ldots \quad (5.9)$$

So the nominal 5% points of u(or l) are actual up to sample size 13 and the 1% points up to 18 for this regression.

ACKNOWLEDGEMENT

The author acknowledges, with gratitude, his debt to Dr. P. B. Patnaik of the Indian Statistical Institute for his kind encouragement and useful suggestions in the preparation of this paper; to Dr. Des Raj of the same Institute who proposed the problem of the single outlier; to Shri Sivaramakrishnan who helped in the computations; and to the referee of this paper who suggested many improvements and made the presentation more lucid.

NOMINAL UPPER PERCENTAGE POINTS IN UNITS OF 10^{-4}

TABLE 1*

Regression : $E(y_j) = \beta_1 x_{1,j},\ j = 1(1)n$

sample size	u or l 5%	u or l 1%	t 5%	t 1%
(1)	(2)	(3)	(4)	(5)
3	9973	9998	9993	10000
4	9506	9900	9752	9950
5	8730	9558	9102	9721
6	7968	9072	8547	9340
7	7304	8553	7934	8697
8	6739	8052	7384	8446
9	6258	7589	6899	8011
10	5846	7169	6474	7606
11	5489	6789	6099	7233
12	5178	6446	5768	6890
13	4903	6136	5472**	6576
14	4660	5855	5208	6289
15	4441	5599	4970	6026
16	4245	5366	4754	5784
17	4067	5152	4558	5561
18	3905	4956	4379	5356**
19	3757	4775	4215	5165
20	3621	4607	4063	4989

TABLE 2

$E(y_j) = \beta_1 x_{1,j} + \beta_2 x_{2,j},\ j = 1(1)n$

sample size	u or l 5%	u or l 1%	t 5%	t 1%
(1)	(2)	(3)	(4)	(5)
4	.9985	.9999	.9996	10000
5	9604	9920	9801	9960
6	8872	9608	9283	9753
7	8114	9140	8652	9389
8	7438	8626	8038	8954
9	6858	8125	7481**	8504
10	6363**	7660	6987	8080
11	5938	7233	6553	7661
12	5570	6848	6170	7284
13	5249	6500	5831	6938
14	4967	6185	5530	6621**
15	4716	5900**	5259	6330
16	4493	5640	5011	6064
17	4291	5404	4797	5820
18	4109	5187	4597	5504
19	3943	4988	4416	5386
20	3792	4805	4248	5094

TABLE 3

$E(y_j) = \beta_1 x_{1,j} + \beta_2 x_{2,j} + \beta_3 x_{3,j},\ j = 1(1)n$

sample size	u or l 5%	u or l 1%
(1)	(2)	(3)
5	.9993	10000
6	9660	9933
7	8979	9645
8	8231	9105
9	7551	8688
10	6961	8188
11	6455	7720
12	6020	7291
13	5644**	6901
14	5315	6549
15	5026	6231
16	4769	5942
17	4540	5679
18	4334	5439**
19	4140	5220
20	3979	5018

*Table 1 provides actual percentage points of u and l for the regression $E(y_j) = \beta$ (Section 5.1).

**Percentage points are actual up to and including this sample size for the regression Table 1 : $E(y_j) = \beta$; Table 2 : $E(y_j) = \beta_1 + \beta_2 . j$ (Section 5.2); Table 3 : $E(y_j) = \beta_1 + \beta_2 \sin(2\pi j/\lambda) + \beta_3 \cos(2\pi j/\lambda), j = 1(1)n$, where λ is a positive integer and n is a multiple of λ (Section 5.3).

Note :

1. u, l and t are given by equations (2.16), (2.17) and (2.18).
2. Nominal percentage points of u and l satisfy (3.11) and those of t satisfy (3.14).
3. Condition for coincidence of the nominal and actual percentage points of u and l is given by (3.13) and for t by (3.15).
4. In virtue of (3.13) and (3.15) the nominal upper 100α percent points of u (or l) are the same as the 200α percent points of t. Thus the nominal 1% and 5% points of u could be used as the 2% and 10% points of t and nominal 1% and 5% points of t could be used as 0.5% and 2.5% points of u(or l).

References

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

Cochran, W. G. (1941): The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen.*, **II**, 47-52.

Godwin, H. J. (1945): On the distribution of the estimate of mean deviation obtained from samples from a normal population. *Biometrika*, **33**, 254-256.

Grubbs, F. E. (1950): Testing outlying observations. *Ann. Math. Stat.*, **21**, 27-58.

Hartley, H. O. and Pearson, E. S. (1942): The probability integral of the range in samples of n observations from a normal population. *Biometrika*, **32**, 301-310.

——— (1943): Tables of the probability integral of the studentized range. *Biometrika*, **33**, 89-99.

Irwin, J. O. (1925): On a criterion for the rejection of outlying observations. *Biometrika*, **17**, 238-250.

MCKAY, A. T. (1935): The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, **27**, 466-471.

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

NEYMAN, J. and PEARSON, E. S. (1933): On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc.*, (Lond.) **231**, 289-337.

PEARSON, E. S. (1926): A further note on the distribution of range in samples taken from a normal population. *Biometrika*, **17**, 151-164.

PEARSON, K. (1934): *Tables of the Incomplete Beta-Function*. Biometrika Office, University College, London.

———— (1931) : *Tables for Statisticians and Biometricians*, Part II, edited by Karl Pearson, CX-CXIX.

SMIRNOFF, N. V. (1941): On the estimation of the maximum term in a series of observations. C. R. (Doklady). *Acad. Sci.* (N.S.), **33** 346-50.

"Student", (1927): *Biometrika*, **19**, 151-164.

THOMPSON, W. R. (1935) : On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. *Ann. Math. Stat.*, **6**, 214-219.

TIPPET, L. H. C. (1925): The extreme individuals and the range of samples taken from a normal population. *Biometrika*. **17**, 151-184.

W. P. A. (1945): Tables of probability integral of the mean deviation in normal samples. *Biometrika*, **33**, 250-265.

Paper received : March, 1956.

Revised : October, 1960.