THE EFFICIENCY OF STATISTICAL TOOLS AND A CRITERION FOR THE REJECTION OF OUTLYING OBSERVATIONS.

BY E. S. PEARSON, ASSISTED BY C. CHANDRA SEKAR.

1. *Thompson's criterion.*

In an interesting paper recently published in the *Annals of Mathematical Statistics**, Dr William R. Thompson has suggested a new criterion for the rejection of outlying observations. If $x_1, x_2, \ldots x_N$ represent a series of observed values of a variable x, and

$$\bar{x} = \sum_{i=1}^{N} (x_i)/N, \quad s^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2/N \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1),$$

then Thompson writes

$$\tau_i = (x_i - \bar{x})/s \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2).$$

He then shows that if x_i is an observation arbitrarily selected from a random sample of N drawn from an infinite normal population, then the elementary probability distribution of τ is

$$p(\tau) = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{(N-1)\pi}\,\Gamma\left(\frac{N-2}{2}\right)} \left(1 - \frac{\tau^2}{N-1}\right)^{\frac{N-4}{2}} \ldots\ldots\ldots\ldots (3).$$

This is a symmetrical limited range distribution of Pearson's Type II; the probability integral of the τ-distribution may be found directly from appropriate tables†, or by noticing that if we write

$$\tau = t\sqrt{\frac{N-1}{N-2+t^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4),$$

then t follows Student's distribution having $N-2$ degrees of freedom, for which tables of probability levels are available‡. It will also be noted that τ is distributed as $r\sqrt{N-1}$, where r is the coefficient of correlation between N pairs of observations randomly drawn from two completely independent normal distributions.

It is clear—and consequences following from this will be discussed later—that the N values of τ given by the observations of a single sample are not independent. It is however true to say that if for any randomly chosen x_i we denote the probability that the absolute value of τ_i is greater than a specified value, say τ_0, by

$$P = P\{|\tau_i| > \tau_0\} = 2\int_{\tau_0}^{\sqrt{N-1}} p(\tau)\,d\tau \ldots\ldots\ldots\ldots\ldots\ldots (5),$$

* Vol. VI. (1935), pp. 214—219.

† *Tables for Statisticians and Biometricians*, Part II, Table XXV.

‡ See Student, *Metron*, Vol. v. (1925), pp. 105—120, R. A. Fisher, *Statistical Methods for Research Workers* (1935), Table IV.

and write $\phi = NP$, then ϕ is the expectation of the number of observations per sample of N drawn from a single normal population for which

$$|x_i - \bar{x}| > \tau_0 s \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6).$$

In other words, if we decide to reject all observations for which the inequality (6) is true, then in the long run we shall reject one observation, on the average, in every $1/\phi$ samples unnecessarily, i.e. when all the observations have in fact been drawn from a common normal population.

Thompson has given a Table of the limits, τ_0, for three values of ϕ, namely 0·2, 0·1 and 0·05, and for the following values of N,

3, 4, ... 21, 22, 32, 42, 102, 202, 502, 1002.

Thus using the second value of ϕ, (0·1), and the rule of rejection denoted by the inequality (6), we shall on the average reject unnecessarily one observation per 10 samples; this figure will be the same whatever the value of N.

An alternative form of Table would be one in which limits, τ_0, were given for fixed values of $P = \phi/N$; thus ϕ would increase and $1/\phi$ decrease with the sample size. If for example $P = {\cdot}01$, the observer would be likely, on the average, to discard unnecessarily one observation in every 10 samples when each of these contained 10 observations, one in every sample when it contained 100, and 10 in every sample when it contained 1000 observations. The fixing of ϕ or of P is, however, a matter of choice, the basis of Thompson's criterion being simply the equation (3) giving the probability distribution of τ.

It will be found that the great majority of criteria that have been invented for rejecting outlying observations contain as an initial condition the assumption that σ, the standard deviation in the hypothetical common population, is known*. Under certain circumstances this may be true, as for example when dealing with errors of observation where the value of σ may be estimated with great precision from past experience. Generally, however, in practice it is necessary to substitute into the formulae obtained the standard deviation calculated from the sample of observations under consideration, and this limitation, sometimes frankly recognised by the inventor of the test, sometimes apparently overlooked, renders inaccurate the probability basis upon which the criterion rests.

Dr Thompson's criterion is free from this defect; it provides, by a process which has sometimes been termed "studentizing," a true measure of the risk of rejecting an observation when in fact the whole sample has been drawn from a single normal distribution. In other words, if we describe the application of the criterion as that of testing the hypothesis, say H_0, that the sample has been drawn from a single normal population, then Thompson's method gives precise control of the risk of rejecting H_0 when it is true. This is what J. Neyman and E. S. Pearson† in their

* For an extensive survey of such tests see P. R. Rider, *Washington University Studies, Science and Technology*, No. 8 (1933).

† See, for example, *Biometrika*, Vol. xx^A. (1928), pp. 175—240, *Phil. Trans. Roy. Soc.* Vol. ccxxxi. A (1933), pp. 289—337 and *Statistical Research Memoirs* (issued by the Department of Statistics, University College, London), Vol. i. (1936), pp. 1—37.

general treatment of the theory of testing statistical hypotheses have termed the control of the first kind of error. It is, however, necessary to point out that by satisfying this condition alone it does not follow that an efficient tool has been placed in the hands of the experimenter.

This consideration has so important a bearing on the choice of statistical tests in general that it has seemed to us worth while discussing in some detail the conditions under which it appears that Thompson's method provides an efficient criterion for the rejection of outlying observations.

2. *The efficiency of statistical tools.*

What requirements should the theoretical statistician bear in mind in constructing efficient working tools for the experimenter? We think he may usefully remember two considerations which are of general application in the construction of any scientific exploratory tool:

(*a*) A tool is devised for use under certain limited conditions and will only be fully efficient as long as these conditions are satisfied.

(*b*) To test whether these conditions hold good other tools are generally needed.

An illustration of the meaning of these points from a non-statistical field may be useful. The lead-line is used at sea to measure the depth of the ocean bottom. Soundings at fixed but discrete intervals of time may be taken from a moving ship either by the hydrographer who is engaged on a survey of the ocean bottom, or by the navigator who in a fog, using the hydrographer's chart, wishes to test the hypothesis that his ship is just entering a certain channel. In both cases an assumption is made that changes in the sea bottom are gradual; if there might exist unknown pinnacled rocks of great height, sudden changes in contour or sunken ships, the lead, cast at discrete intervals of time, would be an inefficient, and even dangerous, tool to rely upon. Further, some form of dragging operation, rather than sounding, would be required to test whether it was justifiable to assume the bottom to be free from sudden changes.

The same considerations will be found to apply in the case of nearly all the tools of physical and biological science, although in many cases the conditions are so universally satisfied that the worker hardly stops to remember that limitations to the efficiency of the tool exist.

Keeping these points in mind, we may turn to the problem of the theoretical statistician who is concerned with the design of statistical tools. In the first place it must be noted that, as in other branches of applied mathematics, it is necessary for him to construct a precise but probably a simplified model which he believes will represent the phenomena of observation with sufficient accuracy to provide useful results. In so far as there is a practical problem to solve, this model will contain certain unknowns, and what is required is to devise the most effective method of obtaining information from the data concerning these unknowns. The tool will only be efficient provided that the model is appropriate, or, in other words, provided certain conditions are satisfied. To determine whether this is the case, a different set of tools will generally be required.

Some of the chief procedures of statistical analysis may usefully be classified under two heads:

(1) *The estimation of characteristics of a population*, that is to say the estimation of the values of unknowns in our model. The old procedure was to record a single-valued estimate of the unknown parameter and attach to it a probable error. More recently the conception of the confidence or fiducial interval has been introduced*. In either case the procedures employed have only a precise meaning in so far as it is possible to specify in mathematical terms the alternative forms of population distribution that are considered possible. For example, the standard error of a sample correlation coefficient, r, as ordinarily calculated, or the confidence interval, (ρ_1, ρ_2) †, have no exact meaning unless we can assume that the population distribution is of the normal form. Further, to test the validity of this assumption a different form of test is required.

(2) *The testing of statistical hypotheses.* Here the problem is to determine whether it is likely that certain unknowns in the model have specified values. In Dr Thompson's case, the problem is to test whether the sample has been drawn from a single normal population, but it is not possible to devise an efficient test if we only bring into the picture this single normal probability distribution with its two unknown parameters. We must also ask how sensitive the test is in detecting failure of the data to comply with the hypothesis tested, and to deal with this question effectively we must be able to specify the directions in which the hypothesis may fail. In other words, an efficient test of a statistical hypothesis, H_0, must be associated with a set of admissible alternative hypotheses and not solely with H_0. This set provides the model on which the statistical tool-maker can set to work.

Even when the alternatives cannot be specified in such precise form as to allow mathematical methods to be applied to full advantage, we feel sure that much is gained by a review of the types of alternatives between which it is wished to discriminate.

3. *Limitations of the τ criterion.*

Approaching from this point of view the problem of testing the hypothesis, H_0, that a sample of N observations has been drawn from a single normal population, we must ask what are the possible alternatives. Unless it is possible to conceive some alternative to H_0 there would be no justification in rejecting outlying observations. In the first place we are presumably making the assumption that the majority of observations come from some single normal population‡; we believe

* R. A. Fisher, *Proc. Camb. Phil. Soc.* Vol. xxvi. (1930), p. 528; J. Neyman, *Journ. Roy. Stat. Soc.* Vol. xcvii. (1934), p. 589; E. S. Pearson and J. C. Clopper, *Biometrika*, Vol. xxvi. (1934), pp. 404—413.

† Tables which will facilitate the calculation of these intervals prepared by Miss F. N. David will be issued shortly as a *Biometrika* Publication.

‡ If an admissible alternative to H_0 were that the observations came from a single non-normal population, then what would be required would be a test of normality, not a test for the rejection of outlying observations.

however that it is possible that one or two of them do not belong to this set. We may therefore perhaps adequately represent the situation by using the following model: alternative to H_0 are hypotheses that k of the observations ($k \geqslant 1$) come from s normal populations ($s \leqslant k$) having different means or standard deviations (or both) from the single population from which the majority ($N-k$) of the observations have been drawn. The assumption of normality in the other populations is not necessary, but as far as the present discussion goes nothing is lost by making it.

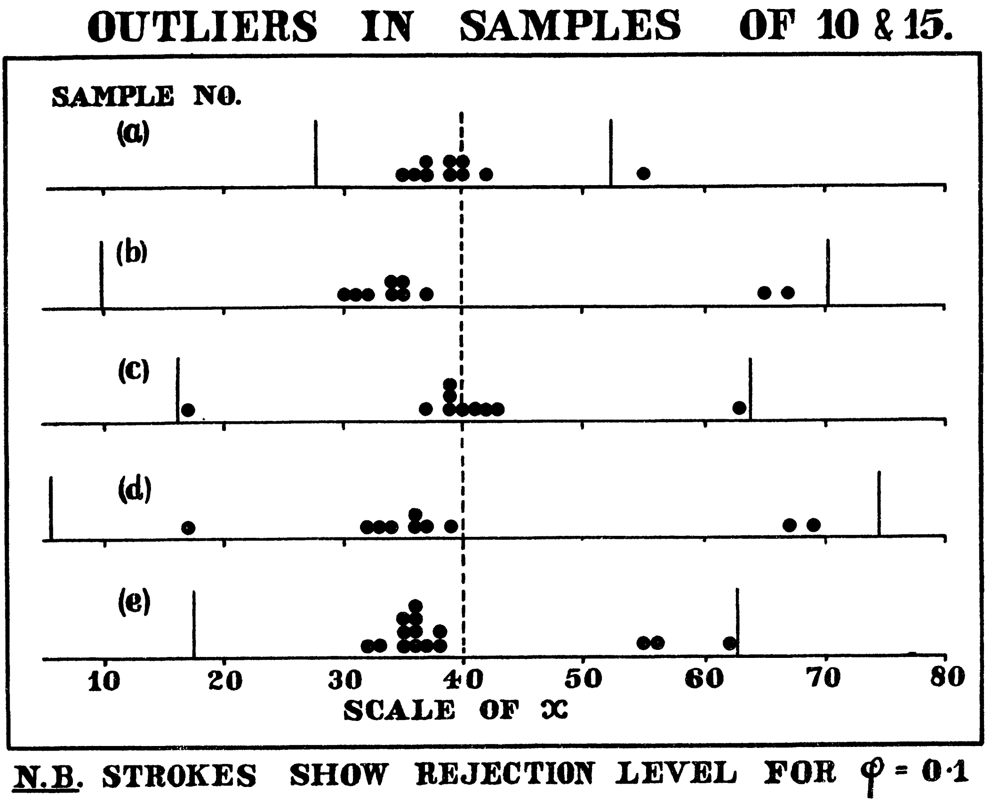

Fig. 1.

The point that we wish now to make is that if $k > 1$, Thompson's criterion is not very suitable for the purpose in view, at any rate for relatively small samples. In other words, it would appear that the criterion is only really efficient in the presence of a *single* outlying observation. To appreciate the reason for this criticism consider the data shown in Figure 1. They represent 4 different samples of 10 observations and 1 sample of 15, each having a mean of 40 on the scale shown. On intuitive grounds we should be inclined to say that in every case H_0 should be rejected, i.e. that there exist outlying observations not belonging to the main set. Yet if we apply Thompson's criterion, taking the level of significance $\phi = 0{\cdot}1$ (i.e.

running the risk of rejecting 1 observation in every 10 samples when H_0 is true), we find that it is only in the case of sample (*a*) that an outlier is picked out. In the remaining four cases none of the values of $|x_i - \bar{x}|$ exceed $\tau(\phi = 0{\cdot}1) \times s$. The limits $\bar{x} \pm \tau(\phi = 0{\cdot}1) \times s$ are marked by vertical strokes on the diagram.

We may next consider the data in Figure 2, which are based on 50 samples of 5 drawn randomly from a normal distribution with the help of Tippett's Random Sampling Numbers*.

We shall use the following notation:

(*a*) for the N values of τ in a sample arranged in descending order of absolute magnitude, write $\tau_1, \tau_2, \ldots \tau_N$; thus

$$|\tau_1| \geqslant |\tau_2| \geqslant \ldots \geqslant |\tau_N| \qquad \text{.........................(7)};$$

(*b*) for the N values arranged in magnitude taking account of sign, write $\tau^{(1)}, \tau^{(2)}, \ldots \tau^{(N)}$; thus

$$\tau^{(1)} \geqslant \tau^{(2)} \geqslant \ldots \geqslant \tau^{(N)} \qquad \text{.............................(8)}.$$

Since the probability distribution for any τ taken at random follows the law of equation (3), it follows that, although the 5 values of τ obtained from the same sample are not independent, the total distribution of the $50 \times 5 = 250$ values of τ obtained in the experiment may be graduated by the curve obtained by writing $N = 5$ in the equation. This is shown in the upper portion of Figure 2. Below are shown distributions of (*A*) τ_1, (*B*) τ_2 and (*C*) τ_3, τ_4 and τ_5, and again of (*A*) $\tau^{(1)}$, (*B*) $\tau^{(2)}$, (*C*) $\tau^{(3)}$, (*D*) $\tau^{(4)}$ and (*E*) $\tau^{(5)}$.

The outer limits for τ_1 are $\pm\sqrt{N-1} = 2$; there are also outer limits for τ_2 and τ_3 which are shown by vertical strokes. Again limits exist for $\tau^{(i)}$. An explanation is at once suggested of the failure of the criterion to pick out the outlying observations in Figure 1; for certain sizes of sample the upper limit of $|\tau_i|$ $(i \geqslant 2)$ lies within the significance level given by the test. When this is so, only 1 observation can possibly be rejected however heterogeneous the data, and if the two extreme observations are close together none can be rejected at all.

To investigate this point more fully we must take the general case of a sample of N, and consider what limits there are to the value of the τ's. The problem may be put in this form:

If $\tau_1, \tau_2, \ldots \tau_N$ are any real numbers satisfying the equalities

$$\tau_1 + \tau_2 + \ldots + \tau_N = 0 \qquad \text{.............................(9)},$$

$$\tau_1^2 + \tau_2^2 + \ldots + \tau_N^2 = N \qquad \text{..........................(10)},$$

and the inequalities

$$\tau_1^2 \geqslant \tau_2^2 \geqslant \ldots \geqslant \tau_N^2 \qquad \text{............................(11)},$$

it is required to determine the maximum value of τ_i^2 $(i = 1, 2, \ldots N)$.

* *Tracts for Computers*, xv.

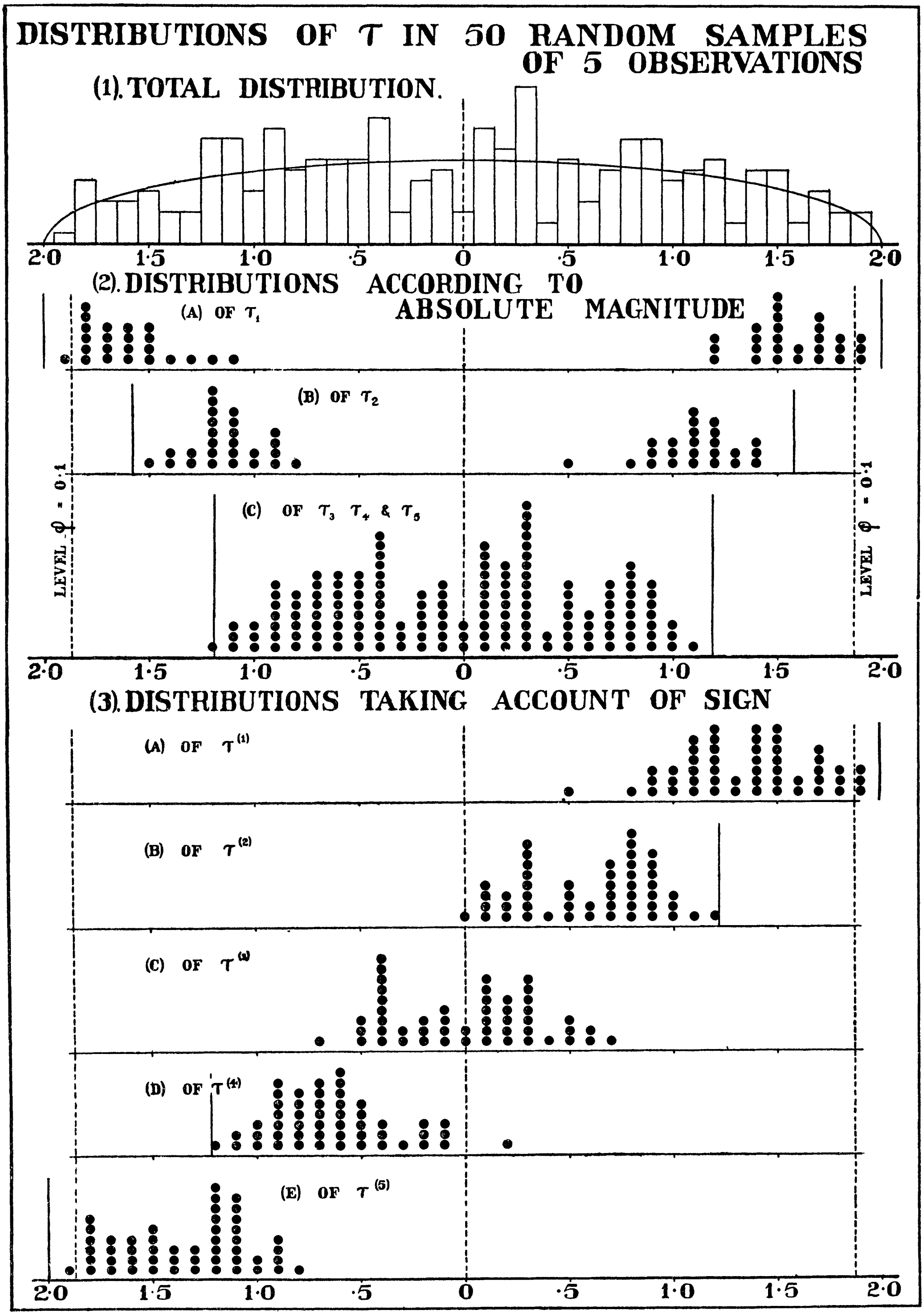
DISTRIBUTIONS OF τ IN 50 RANDOM SAMPLES OF 5 OBSERVATIONS
(1). TOTAL DISTRIBUTION.
2·0 1·5 1·0 ·5 0 ·5 1·0 1·5 2·0
(2). DISTRIBUTIONS ACCORDING TO ABSOLUTE MAGNITUDE
(A) OF τ1
(B) OF τ2
(C) OF τ3 τ4 & τ5
LEVEL φ = 0·1
LEVEL φ = 0·1
2·0 1·5 1·0 ·5 0 ·5 1·0 1·5 2·0
(3). DISTRIBUTIONS TAKING ACCOUNT OF SIGN
(A) OF τ(1)
(B) OF τ(2)
(C) OF τ(3)
(D) OF τ(4)
(E) OF τ(5)
2·0 1·5 1·0 ·5 0 ·5 1·0 1·5 2·0

Fig. 2.

A detailed solution given by Mr J. M. C. Scott is set out in an Appendix to this paper. Here we shall only quote the results.

(1) If i is odd, but $< N$,

$$\text{Maximum } |\tau_i| = \sqrt{\frac{N}{i + \frac{1}{N-i}}} \quad \text{.......................(12)},$$

e.g.
$$\begin{cases} i = 1, \text{ maximum } |\tau_1| = \sqrt{N-1}, \\ i = 3, \text{ maximum } |\tau_3| = \sqrt{N(N-3)/(3N-8)}, \\ i = 5, \text{ maximum } |\tau_5| = \sqrt{N(N-5)/(5N-24)}. \end{cases}$$

(2) If $i = N$ and is odd,

$$\text{Maximum } |\tau_N| = \sqrt{\frac{N-1}{N+1}} \quad \text{.......................(13)}.$$

(3) If i is even, $$\text{Maximum } |\tau_i| = \sqrt{\frac{N}{i}} \quad \text{.............................(14)},$$

e.g.
$$\begin{cases} i = 2, \text{ maximum } |\tau_2| = \sqrt{\tfrac{1}{2}N}, \\ i = 4, \text{ maximum } |\tau_4| = \sqrt{\tfrac{1}{4}N}. \end{cases}$$

(4) A further result will be useful. If we do not consider the absolute magnitude, but the maximum value of the second largest value of τ or of $\tau^{(2)}$, we find

$$\text{Maximum } \tau^{(2)} = \sqrt{\tfrac{1}{2}(N-2)} \quad \text{.....................}$$

Similarly $$\text{Minimum } \tau^{(N-1)} = -\sqrt{\tfrac{1}{2}(N-2)} \quad \text{.................} \quad \text{(15)}.$$

Putting $N = 5$, we obtain the limits shown in Figure 2.

As shown in the Appendix, the maximum value of $|\tau_1|$ occurs when $N-1$ observations have identical values, and the remaining observation any other different value; the situation in Figure 1 (a) is approaching this. The maximum $\tau^{(2)}$ occurs when $N-2$ observations have identical values and the other two have a different common value, i.e. $\tau^{(1)} = \tau^{(2)}$ (see Figure 1 (b)). The maximum $|\tau_2|$ occurs when $N-2$ observations have identical values, the other 2 differ and $\tau_1 = -\tau_2$ (see Figure 1 (c)). The maximum $|\tau_3|$ occurs when $N-3$ observations have identical values, the other 3 differ and $\tau_1 = \tau_2 = -\tau_3$ (see Figure 1 (d)).

In Figure 3 we have plotted for different values of N:

(a) the maximum values of $|\tau_i|$ ($i = 2$, 3 and 4) from equations (12) and (14);

(b) the levels of significance for τ from Thompson's tables with $\phi = 0{\cdot}20$, $0{\cdot}10$ and $0{\cdot}05$ with an additional limit $\phi = 0{\cdot}02$;

(c) the maximum value of $\tau^{(2)}$ from equation (15).

Thus if we use the significance level for ϕ of $0{\cdot}10$, involving the risk of rejection of 1 observation in every 10 samples when H_0 is true, we see that under no circumstances can we reject more than 1 observation until we reach a sample of 11; we cannot reject more than 2 observations until $N = 22$; and we cannot reject more than 3 observations until $N = 32$. This result corresponds to the extreme

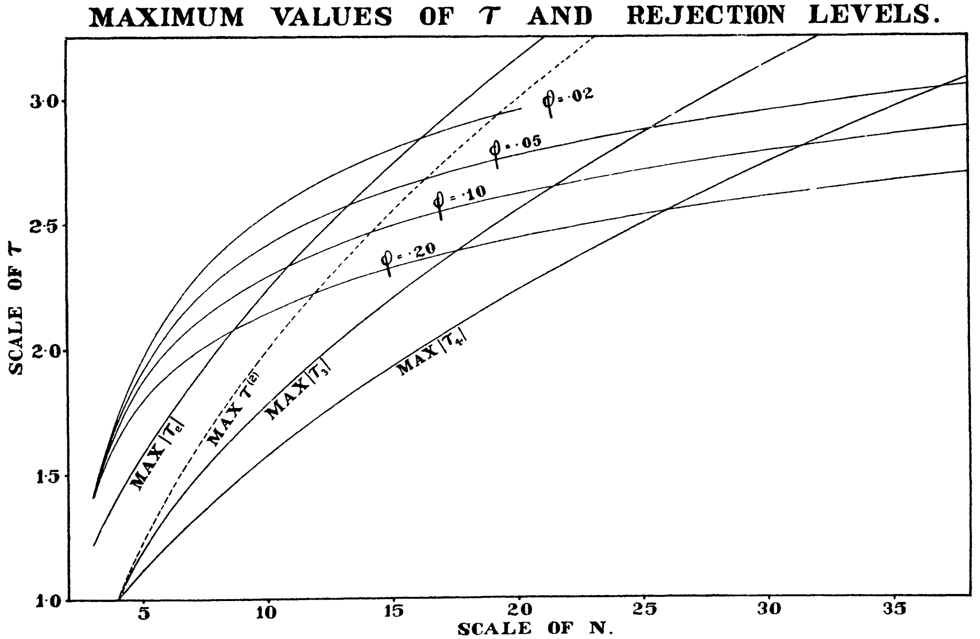

Fig. 3.

cases in which the $N-i$ observations are identical; for cases more likely to occur, such as those shown in Figure 1, the sample sizes must of course be larger.

It seems to follow, therefore, that if the situation we are faced with can be represented by a model in which the alternative to H_0 is that *one* observation has been drawn from a population with divergent mean (and possibly different standard deviation) to the remaining $N-1$ observations, Thompson's criterion will be very useful. If however the alternatives are that $k > 1$ observations come from another or other systems, then the test may be quite ineffective, particularly if the sample contains not more than 30 or 40 observations.

No doubt suitable criteria could be devised for each type of alternative; thus if it were possible that 2 observations had come from a population with different mean to the remainder (Figure 1 (*b*)), the criterion might involve the calculation of the ratio of, (i) the deviation of the mean of two extreme observations from the mean of the whole to, (ii) the standard deviation. But this test would be quite unsuitable if the divergent observations had come from two populations with means diverging in opposite directions from the mean of the remainder, as in Figure 1 (*c*).

The statistician who does not know in advance with which type of alternative to H_0 he may be faced, is in the position of a carpenter who is summoned to a house to undertake a job of an unknown kind and is only able to take one tool with him! Which shall it be? Even if there is an "omnibus" tool, it is likely to be far less sensitive at any particular job than a specialised one; but the specialised tool will be quite useless under the wrong conditions.

Following the lines that Student has suggested when using "range" as a criterion*, it might be thought that a suitable "omnibus" tool for rejecting outlying observations could be obtained as follows: (1) apply Thompson's criterion to the N observations; (2) if it rejects $k \geqslant 1$ outliers, apply the criterion again to the remaining $N-k$ observations, calculating $\bar{x}$ and s afresh for the reduced sample; (3) repeat this process until a stage is reached when no further observations are rejected. Two points must however be remembered:

(*a*) Even when there are obvious outliers, the process may never get started at all for reasons already discussed. This is the case with the samples shown in Figure 1 (*b*)—(*e*).

(*b*) In order to control the first kind of error (in Neyman and Pearson's sense), i.e. the risk of rejecting H_0 when it is true, considerable development of theory would be required involving the determination of the simultaneous distribution of τ_1, τ_2, τ_3, etc.

In conclusion, since it is sometimes held that the appropriate test can be chosen *after* examining the data in the sample, a final word of caution is necessary. To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis is true. In the present instance it might appear, for example, that the 1st and 2nd observations (in order of magnitude) were unusually far apart, or a gap might occur between the 2nd and 3rd or between the 5th and 6th: again, the standard deviation might be large compared to the range, or there might appear to be too few observations near the mean.

By choosing the feature most unfavourable to H_0 out of a very large number of features examined, it will usually be possible to find some reason for rejecting the hypothesis. It must be remembered, however, that the point now at issue will not be whether it is exceptional to find a given criterion with so unfavourable a value. We shall need to find an answer to the more difficult question. Is it exceptional that the most unfavourable criterion of the n, say, examined should have as unfavourable a value as this?

4. *The percentage limits of the extremes, $\tau^{(1)}$ and $\tau^{(N)}$.*

An examination of Figure 2 for the case $N=5$ shows how the form of the total τ-distribution at its extreme depends only on the distributions of $\tau^{(1)}$ and $\tau^{(5)}$. This is because the upper limit of $\tau^{(2)}$ lies at $+1{\cdot}225$ and the lower limit of $\tau^{(4)}$ at $-1{\cdot}225$. In general for $\tau^{(1)} > \sqrt{\tfrac{1}{2}(N-2)}$ we have for the probability law of $\tau^{(1)}$,

$$p(\tau^{(1)}) = Np(\tau) \qquad \ldots\ldots (16),$$

where $p(\tau)$ is given in equation (3). A similar form holds for the lowest τ if $\tau^{(N)} < -\sqrt{\tfrac{1}{2}(N-2)}$. Since $\phi = NP$, where P is defined in equation (5), it follows that the limits of τ which Thompson has tabled for $\phi = 0{\cdot}20$, $0{\cdot}10$ and $0{\cdot}05$ will

* *Biometrika*, Vol. XIX. (1927), pp. 161- 162.

correspond to the upper 10 %, 5 % and 2·5 % probability levels of $\tau^{(1)}$ (and the corresponding lower levels for $\tau^{(N)}$), as long as these levels fall beyond

$$\text{Max. } \tau^{(2)} = \sqrt{\tfrac{1}{2}(N-2)}.$$

In Figure 3 a dotted line has been drawn showing the changing value of this latter function and, by observing where it crosses the successful levels of ϕ, we can tell up to what size of sample the general distribution of τ of equation (3) may be used to give the percentage limits of the τ's calculated from the two extreme observations in a sample. Besides Thompson's levels at $\phi = 0{\cdot}20$, 0·10 and 0·05 an additional level at $\phi = 0{\cdot}02$ has been calculated*. From these we obtain the following table, in which columns 2, 3 and 4 repeat Thompson's results. While it would be useful to find the percentage limits of the extreme τ for values of N beyond those shown in the table, these cannot be obtained from the probability integral of the general τ-distribution, owing to the overlap of the distributions of $\tau^{(2)}$ and $\tau^{(1)}$ for larger values of N.

Upper probability limits for $\tau = \dfrac{x - \bar{x}}{s}$, *for the highest observation in a sample of* N, *i.e. for* $\tau^{(1)}$.

N	10 %	5 %	2·5 %	1 %
3	1·4065	1·4123	1·4137	1·4142
4	1·6454	1·6887	1·7103	1·7234
5	1·791	1·869	1·917	1·955
6	1·895	1·997	2·067	2·130
7	1·973	2·093	2·182	2·265
8	2·041	2·170	2·274	2·374
9	2·099	2·237	2·348	2·464
10	2·144	2·295	2·413	2·540
11	2·190	2·343	2·472	2·606
12	—	2·388	2·521	2·663
13	—	2·425	2·567	2·713
14	—	2·463	2·598	2·759
15	—	—	2·636	2·800
16	—	—	2·670	2·837
17	—	—	—	2·871
18	—	—	—	2·903
19	—	—	—	2·932

N.B. The same limits, with negative sign, will apply to the τ calculated from the lowest observation, i.e. to $\tau^{(N)}$.

In concluding this paper we should like it to be clear that we consider Dr W. R. Thompson has suggested a useful, practical criterion which the experi-

* This was done by backward interpolation in Table XXV of *Tables for Statisticians and Biometricians*, Part II, for $N=5$ to 19. For the case $N=4$ the law of equation (3) gives a rectangular distribution and for $N=3$ a simple result follows on noting that if we write $\tau=\sqrt{2}\cos\theta$, θ is distributed uniformly between 0 and π.

menter may employ provided that he recognises that it may be inefficient in the presence of more than one outlying observation. The criterion possesses a great advantage which so many criteria that have been invented before lack, in that it provides complete control over the risk of rejecting the hypothesis tested when it is true. Our purpose has been to show that even when this control is assured, other difficulties exist which appear to be inevitably inherent in the problem of the rejection of outlying observations.

APPENDIX.

BY J. M. C. SCOTT.

THE object of this note is to find the values of $\tau_1, \ldots \tau_N$ which satisfy the following relations:

$$\begin{cases} |\tau_1| \geqslant |\tau_2| \geqslant \ldots \geqslant |\tau_i| \ldots \geqslant |\tau_N| & \text{(17)}, \\ \tau_1+\tau_2+\ldots+\tau_N=0 & \text{(18)}, \\ \tau_1^2+\tau_2^2+\ldots+\tau_N^2=N & \text{(19)}, \\ |\tau_i| \text{ to be as great as possible} & \text{(20)}. \end{cases}$$

Line (20) may be replaced by the requirement

$$\frac{\tau_1^2+\tau_2^2+\ldots+\tau_N^2}{\tau_i^2} \text{ to be as small as possible} \qquad \text{(21)}.$$

Now (17), (18) and (21) are homogeneous, and their solutions are the solutions of (17), (18), (19) and (21)—i.e. of (17), (18), (19) and (20)—multiplied by an arbitrary factor. If this arbitrary factor is determined by

$$\tau_i=1 \qquad \text{(22)},$$

instead of by (19), we are led to consider the following problem which is equivalent to the original one:

$$\begin{cases} |\alpha_1| \geqslant |\alpha_2| \geqslant \ldots \geqslant |\alpha_i| \ldots \geqslant |\alpha_N| & \text{(23)}, \\ \alpha_1+\alpha_2+\ldots+\alpha_N=0 & \text{(24)}, \\ \alpha_i=1 & \text{(25)}, \\ \alpha_1^2+\alpha_2^2+\ldots+\alpha_N^2 \text{ to be as small as possible} & \text{(26)}. \end{cases}$$

We have to find i numbers (α_1 to α_i) numerically $\geqslant 1$, and $N-i$ numbers (α_{i+1} to α_N) numerically $\leqslant 1$, satisfying (24), (25) and (26). The order within these sets can be settled afterwards by (23), and is irrelevant at present.

Lemma 1. *The numbers* α_{i+1} to α_N *are equal; the positive numbers among* α_1 *to* α_i *are equal; and the negative numbers among* α_1 *to* α_i *are equal.*

Proof. If in one of these sets there are two unequal numbers, replace them by their mean. This gives a permissible set of α's but diminishes $\Sigma(\alpha^2)$, contradicting (26).

Further, the positive numbers among α_1 to α_i must be $=1$, since $\alpha_i=1$. Thus we have

$$\begin{cases} r & \text{numbers equal to } A \leqslant -1, \text{ say,} \\ i-r & \quad ,, \qquad ,, \quad ,, \; +1, \\ N-i & \quad ,, \qquad ,, \quad ,, \; C, \text{ where } |C| \leqslant 1. \end{cases}$$

Henceforward we will suppose $i<N$.

Lemma 2. $A=-1$, *if* $r>0$.

Proof. Suppose the lemma is not true; then there is an $\alpha_k<-1$. Change α_k from A to $A+\epsilon$ where $\epsilon>0$, and change α_N from C to $C-\epsilon$. This is possible if $C>-1$. It decreases $\Sigma(\alpha^2)$ by

$$A^2+C^2-(A+\epsilon)^2-(C-\epsilon)^2=2\epsilon(C-A-\epsilon)>0 \text{ for small enough } \epsilon,$$

which contradicts (26). This argument fails if $C=-1$. In this case, change the signs of α_1 to

α_{i-1}, and change α_{i+1} to α_N from -1 to $1-\frac{2}{N-i}$. This gives a permissible set of α's. If now $i<N-1$, $\Sigma(\alpha^2)$ is reduced, contrary to (26); if $i=N-1$, $\Sigma(\alpha^2)$ is left unchanged and some of the α's are equal to $-A>1$, which is contrary to Lemma 1. Thus Lemma 2 is established.

It follows that either $A=-1$ or $r=0$. In each case

$$C=\frac{2r-i}{N-i} \qquad\qquad (27),$$

leading to

$$\Sigma(\alpha^2)=i+\frac{(2r-i)^2}{N-i} \qquad\qquad (28).$$

We have now to determine r so that this expression is a minimum, and to remember that

$$\tau_s^2=\alpha_s^2\,\frac{\Sigma(\tau^2)}{\Sigma(\alpha^2)}=\frac{\alpha_s^2 N}{\Sigma(\alpha^2)} \qquad\qquad (29).$$

Case 1. If i is even, $r=\frac{1}{2}i$, $C=0$ and therefore

$$\Sigma(\alpha^2)=i.$$

It follows that

$$\text{Maximum } |\tau_i|=\sqrt{\frac{N}{i}} \qquad\qquad (30),$$

and that of $\tau_1, \ldots \tau_i$ half are equal to $\sqrt{N/i}$ and half equal to $-\sqrt{N/i}$, while the remaining $N-i$ values are zero. This result is otherwise obvious, for if $\tau_i^2>N/i$, then

$$\sum_{s=1}^{N}(\tau_s^2)\geqslant\sum_{s=1}^{i}(\tau_s^2)>N.$$

Case 2. If i is odd, then $r=\frac{1}{2}(i\pm 1)$, $C=\pm 1/(N-i)$ and

$$\Sigma(\alpha^2)=i+\frac{1}{N-i}.$$

It follows that

$$\text{Maximum } |\tau_i|=\sqrt{\frac{N}{i+\frac{1}{N-i}}} \qquad\qquad (31).$$

Of $\tau_1, \ldots \tau_i$,

$$\begin{cases} \frac{1}{2}(i\pm 1)\text{ values}=-\sqrt{\dfrac{N}{i+\frac{1}{N-i}}}, \\ \frac{1}{2}(i\mp 1)\text{ values}=+\sqrt{\dfrac{N}{i+\frac{1}{N-i}}}, \end{cases}$$

while

$$\tau_{i+1}=\ldots=\tau_N=\pm\sqrt{\frac{N}{(N-i)(Ni-i^2+1)}}.$$

It finally remains to consider the position where $i=N$. If N is even, the result is clearly the same as for case 1 above. If N is odd, $\frac{1}{2}(N-1)$ of the α's will equal $-(N+1)/(N-1)$ and $\frac{1}{2}(N+1)$ will equal $+1$. Therefore

$$\Sigma(\alpha^2)=\frac{N(N+1)}{N-1}.$$

Hence

$\frac{1}{2}(N-1)$ of the τ's will equal $\pm\sqrt{(N+1)/(N-1)}$ and $\frac{1}{2}(N+1)$ will equal $\mp\sqrt{(N-1)/(N+1)}$;

thus

$$\text{Maximum } |\tau_i|=\sqrt{\frac{N-1}{N+1}} \qquad\qquad (32).$$

A further result used in the main paper is that of the maximum value of $\tau^{(2)}$, the second τ in order of magnitude. Following a similar line of proof, it can be readily shown that this occurs when

$$\left.\begin{array}{l} \tau^{(1)}=\tau^{(2)}=\sqrt{\frac{1}{2}(N-2)} \\ \tau^{(3)}=\ldots=\tau^{(N)}=-\sqrt{2/(N-2)} \end{array}\right\} \qquad\qquad (33).$$