ON A CRITERION FOR THE REJECTION OF OBSERVATIONS AND THE DISTRIBUTION OF THE RATIO OF DEVIATION TO SAMPLE STANDARD DEVIATION

By William R. Thompson

Criteria for the rejection of outlying observations may be designed to reject a given fraction of all observations, or a proportion varying with the size of the sample. Irwin[1] has discussed several criteria based on sampling from a normal population which had been used previously, as well as one which he proposed. This is based on the principal of fixing the expectation of rejecting an observation from a sample independently of the aggregate number, N, of the sample. The criterion, λ, is $1/\sigma$ times the interval between successive observations in ascending order of magnitude, where σ is the standard deviation of the sampled population. In the same paper he gave, for different values of N, a table of $P_1(\lambda)$ and $P_2(\lambda)$, respectively probabilities of exceeding given values of λ for the first or second such interval from either end. In actual use, however, σ is estimated from the sample standard deviation, and we are left to decide whether observations in question are to be included or not in estimating the standard deviation as also whether or not to modify this by addition or subtraction of an estimate of its probable error. The object of the present communication is to develop a criterion free from defects of this nature, depending only on the assumption of random sampling from a normal universe. For this purpose we develop the distribution of τ defined by

$$\tau \equiv \frac{\delta}{s}, \tag{1}$$

where s is the sample standard deviation and δ is the deviation of an arbitrary observation of the sample from the sample mean. This leads to definite criteria, which are simple in application.

Accordingly, consider a sample $\{x_i\}$, $i = 1, \cdots, N$, to be drawn at random from a normal population of unknown mean and standard deviation, and that the order of enumeration is arbitrary. Then x_N is an arbitrary one of the elements or *observations*. Now, let

$$\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i, \qquad s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}, \tag{2}$$

and

$$\delta \equiv x_N - \bar{x}. \tag{3}$$

Then we will prove that the distribution of $\tau \equiv \delta/s$ in repeated sampling with a fixed aggregate number, N, is given by substitution of

$$\sqrt{n}\cdot z = t = \sqrt{n}\cdot\tau/\sqrt{n+1-\tau^2}$$

in the z or t distribution of "Student" and R. A. Fisher,[2] where $n = N - 2$. To this end let $N > 2$, and let $n = N - 2$, and

$$(4) \qquad (n+1)\bar{x}_1 = \sum_{i=1}^{n+1} x_i, \quad \text{and} \quad S_1(x-\bar{x}_1)^2 = \sum_{i=1}^{n+1}(x_i-\bar{x}_1)^2.$$

Obviously, the $(n+1)\bar{x}_1 + x_N = N\cdot\bar{x}$, whence

$$(5) \qquad \bar{x} - \bar{x}_1 = \frac{x_N - \bar{x}}{n+1} = \frac{\delta}{n+1}, \qquad \text{whence} \qquad x_N - \bar{x}_1 = \frac{n+2}{n+1}\cdot\delta.$$

Furthermore, $N\cdot s^2 = S_1(x-\bar{x}_1)^2 + (n+1)(\bar{x}_1-\bar{x})^2 + (x_N-\bar{x})^2$, whence

$$(6) \qquad N\cdot s^2 = S_1(x-\bar{x}_1)^2 + \frac{n+2}{n+1}\cdot\delta^2.$$

Now, considering the separate samples, $\{x_i\}$, $i = 1, \cdots, N-1$, and $\{x_N\}$, of aggregate number, $N-1$ and 1, respectively; Fisher has shown[2] that if we set

$$(7) \qquad t = \frac{(x_N - \bar{x}_1)\cdot\sqrt{n}}{\sqrt{S_1(x-\bar{x}_1)^2}}\cdot\sqrt{\frac{n+1}{n+2}},$$

then, for $t_0 > 0$, the probability, p, that $t < t_0$ is

$$(8) \qquad p = \frac{1}{2} + \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\cdot\sqrt{n\cdot\pi}}\int_0^{t_0}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}\cdot dt,$$

and $P = 2(1-p)$ is the probability that $|t| > t_0$.

Now, (5) and (6) in (7) give

$$(9) \qquad t = \frac{\frac{n+2}{n+1}\cdot\delta}{\sqrt{(n+2)\left(s^2 - \frac{\delta^2}{n+1}\right)}}\cdot\sqrt{\frac{n(n+1)}{n+2}} = \frac{\tau\cdot\sqrt{n}}{\sqrt{n+1-\tau^2}},$$

whence

$$(10) \qquad \tau = t\sqrt{\frac{n+1}{n+t^2}}, \quad \text{or} \quad \frac{\tau}{\sqrt{n+1}} = \sin\theta, \qquad \tan\theta = \frac{t}{\sqrt{n}} = z.$$

Accordingly, P is the probability that $|\tau| > \tau_0 \equiv t_0\sqrt{\frac{n+1}{n+t_0^2}}$.

Thus, if we want to determine τ_0 so that by rejecting all observations deviating from the sample mean by more than $s\cdot\tau_0$ we shall have an average relative

frequency of rejections per sample which is fixed, say ϕ; then we need only to set $P = \phi/N$. This follows at once from the hypothesis as x is a random element of the random sample of N elements drawn from the same normal universe (of unknown mean and standard deviation). The criterion of rejection, $s \cdot \tau_0$, is uniquely determined from the sample standard deviation and

TABLE I

N	τ for given ϕ			t for given ϕ			n
	$\phi = 0.2$	0.1	0.05	0.2	0.1	0.05	
3	1.40646	1.41228	1.41373	9.51	19.08	38.19	1
4	1.6454	1.6887	1.7103	4.30	6.20	8.84	2
5	1.791	1.869	1.917	3.48	4.54	5.84	3
6	1.895	1.997	2.067	3.19	3.97	4.84	4
7	1.973	2.093	2.182	3.04	3.68	4.38	5
8	2.041	2.170	2.274	2.97	3.51	4.12	6
9	2.099	2.237	2.348	2.93	3.42	3.94	7
10	2.144	2.295	2.413	2.89	3.36	3.83	8
11	2.190	2.343	2.472	2.88	3.31	3.76	9
12	2.229	2.388	2.521	2.87	3.28	3.70	10
13	2.262	2.425	2.567	2.86	3.25	3.66	11
14	2.296	2.463	2.598	2.86	3.24	3.60	12
15	2.325	2.497	2.636	2.86	3.23	3.58	13
16	2.357	2.522	2.670	2.87	3.21	3.56	14
17	2.382	2.553	2.699	2.87	3.21	3.54	15
18	2.404	2.576	2.733	2.87	3.20	3.54	16
19	2.429	2.601	2.759	2.88	3.20	3.53	17
20	2.448	2.625	2.783	2.88	3.20	3.52	18
21	2.471	2.647	2.800	2.89	3.20	3.50	19
22	2.487	2.661	2.819	2.89	3.19	3.49	20
32	2.636	2.819	2.985	2.944	3.216	3.479	30
42	2.737	2.925	3.093	2.991	3.248	3.489	40
102	3.047	3.233	3.407	3.182	3.397	3.603	100
202	3.266	3.448	3.621	3.347	3.546	3.736	200
502	3.528	3.704	3.872	3.569	3.752	3.927	500
1002	3.714	3.881	4.047	3.737	3.908	4.078	1000

$P = \phi/N$.

Note: τ is computed to 0.5 unit in the last place given from the given t which is believed correct to 1 unit in the last place.

number of elements, N, for any prescribed ϕ. Dropping the subscript, critical values of τ are given in *Table I* (together with corresponding values of t) for $\phi = 0.2$, 0.1, and 0.05 and values of $n \equiv N - 2$ which should be sufficient for most practical purposes. The normal deviate (for unit standard deviation and the same P) lies between these values and is approached by τ and t (in the

tabulated range of ϕ) from opposite sides as n increases, the approximation to τ being the closer of the two. Accordingly Sheppard's tables may be used with good approximation for $n > 1000$, with $\phi/N = P$, the probability of exceeding numerically the given deviate. They may be used to advantage also in interpolation between $n = 100, 1000$ by means of differences at the tabulated points.

A crude rejection system where we reject an observation if it deviate from the mean of all others by more than a *fixed constant* times the standard deviation of such a difference in terms of σ as estimated from the variance of these others by $\tilde{\sigma} = \sqrt{\frac{S_1 (x - \bar{x}_1)^2}{N - 2}}$ amounts to taking a fixed value of t as criterion. The intention is usually to fix the probability (P) of rejection of observations rather than the expectation of rejections per sample (ϕ); and this, of course, is the expected approximate result for *large samples*. For small samples, however, say $4 < N < 32$, by rejection of observations deviating thus by more than $3 \cdot \tilde{\sigma} \cdot \sqrt{\frac{N}{N-1}}$, it appears from (7) and *Table I* that approximately ϕ would be fixed rather than P.

The τ-criterion not only affords a precise extension of such a rejection system, but also a reduction of the actual process of application to a minimum, with one noteworthy exception for the case, $N = 3$. Here we may use as criterion with identical effect the ratio, $\frac{d_2}{d_1}$; where $x_1 \leqslant x_2 \leqslant x_3$, $d_2 = x_3 - x_2$, $d_1 = x_2 - x_1$, and $d_2 \geqq d_1$. This order can always be adopted for the test, and it is readily verified that

$$\frac{d_2}{d_1} = \frac{\sqrt{3} \cdot t - 1}{2}, \tag{11}$$

whence for $\phi = 0.2, 0.1$, and 0.05, respectively we have $\frac{d_2}{d_1} \cong 7.74, 16.0$, and 32.6. Thus, for $N = 3$, we may take merely the ratio of the greater to the other numerical deviation from the median observation as criterion.

Section 2

Although not required in connection with the rejection criterion developed above, there is a simple generalization of τ with a closely related distribution which may be valuable in somewhat different circumstances. Consider the same situation as given above, except that $\{x_i\}$ is divided into two subsets, where $i = 1, \cdots, N - k$, and $i = N - k + 1, \cdots, N$, respectively; giving two random samples of aggregate number, $N - k$ and k. Let the means of these be $\bar{x}_1$ and $\bar{x}_2$, respectively; and s and $\bar{x}$ be as before. Then in general let

$$\delta \equiv \bar{x}_2 - \bar{x} \quad \text{and} \quad \tau \equiv \frac{\delta}{s}. \tag{12}$$

TABLE II

$\tau_{(P,N,1)}$

N	$P = 0.9$	0.8	0.7	0.6	0.5	0.4	$P = 0.3$	0.2	0.1	0.05	0.02	0.01	N
3	.221	.437	.643	.832	1.000	1.144	1.260	1.3450	1.3968	1.4099	1.41352	1.414039	3
4	.173	.347	.520	.693	.866	1.039	1.212	1.386	1.559	1.6080	1.6974	1.7147	4
5	.158	.316	.476	.639	.808	.983	1.170	1.374	1.611	1.757	1.869	1.9175	5
6	.149	.300	.453	.612	.777	.952	1.143	1.360	1.631	1.814	1.973	2.0509	6
7	.144	.290	.440	.594	.757	.932	1.125	1.349	1.640	1.848	2.040	2.142	7
8	.141	.284	.431	.583	.744	.918	1.111	1.340	1.644	1.870	2.087	2.207	8
9	.139	.280	.425	.575	.734	.907	1.102	1.334	1.647	1.885	2.121	2.256	9
10	.137	.276	.420	.569	.727	.899	1.094	1.328	1.648	1.895	2.146	2.294	10
11	.136	.274	.416	.564	.721	.893	1.088	1.324	1.648	1.904	2.166	2.324	11
12	.135	.272	.413	.560	.717	.888	1.083	1.320	1.649	1.910	2.183	2.348	12
13	.134	.270	.411	.557	.713	.884	1.080	1.317	1.649	1.915	2.196	2.368	13
14	.134	.269	.408	.554	.710	.881	1.076	1.314	1.649	1.919	2.207	2.385	14
15	.133	.268	.407	.552	.707	.878	1.073	1.312	1.649	1.923	2.216	2.399	15
16	.133	.267	.405	.550	.705	.875	1.071	1.310	1.649	1.926	2.224	2.411	16
17	.132	.266	.404	.548	.703	.873	1.069	1.309	1.649	1.928	2.231	2.422	17
18	.132	.265	.403	.547	.701	.871	1.067	1.307	1.649	1.931	2.237	2.432	18
19	.131	.264	.402	.546	.699	.869	1.065	1.305	1.649	1.932	2.242	2.440	19
20	.131	.264	.401	.544	.698	.868	1.063	1.304	1.649	1.934	2.247	2.447	20
21	.130	.263	.400	.543	.697	.867	1.062	1.303	1.649	1.936	2.251	2.454	21
22	.130	.263	.399	.542	.696	.865	1.061	1.302	1.649	1.937	2.255	2.460	22
23	.130	.262	.398	.541	.695	.864	1.059	1.301	1.649	1.938	2.259	2.465	23
24	.130	.262	.398	.541	.694	.863	1.058	1.300	1.649	1.940	2.262	2.470	24
25	.130	.261	.397	.540	.693	.862	1.057	1.299	1.649	1.941	2.264	2.475	25
26	.130	.261	.397	.539	.692	.861	1.056	1.299	1.648	1.942	2.267	2.479	26
27	.129	.261	.397	.538	.691	.860	1.056	1.298	1.648	1.942	2.269	2.483	27
28	.129	.261	.396	.538	.691	.860	1.055	1.297	1.648	1.943	2.272	2.487	28
29	.129	.260	.396	.537	.690	.859	1.054	1.297	1.648	1.944	2.274	2.490	29
30	.129	.260	.395	.537	.690	.859	1.054	1.296	1.648	1.944	2.275	2.493	30
31	.129	.260	.395	.536	.689	.858	1.054	1.296	1.648	1.945	2.277	2.495	31
32	.129	.260	.394	.536	.689	.858	1.053	1.295	1.648	1.945	2.279	2.498	32
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582	∞

Note: $\tau_{(P,N,k)} \equiv \sqrt{\frac{N-k}{k\,(N-1)}} \cdot \tau_{(P,N,1)}$

Further, let $n_1 + 1 = N - k$, $n_2 + 1 = k$, $S_1(x - \bar{x}_1)^2$ be the sum of squared deviations in the first sub-sample and similarly $S_2(x - \bar{x}_2)^2$ be that for the second. Then Fisher has shown[2] that the generalized

$$t = \frac{(\bar{x}_2 - \bar{x}_1)\sqrt{n_1 + n_2}}{\sqrt{S_1(x - \bar{x}_1)^2 + S_2(x - \bar{x}_2)^2}} \sqrt{\frac{(n_1 + 1)\,(n_2 + 1)}{n_1 + n_2 + 2}} \tag{13}$$

is distributed as before for $n = n_1 + n_2$. Obviously,

$$N \cdot \bar{x} = (n_1 + 1)\bar{x}_1 + (n_2 + 1)\bar{x}_2\,,$$

whence

$$\delta \equiv \frac{(n_1+1)(\bar{x}_2-\bar{x}_1)}{N} \equiv \frac{(n_1+1)(\bar{x}-\bar{x}_1)}{n_2+1}, \tag{14}$$

and

$$S_1(x-\bar{x}_1)^2 + S_2(x-\bar{x}_2)^2 = N\cdot s^2 - (n_1+1)(\bar{x}_1-\bar{x})^2 - (n_2+1)(\bar{x}_2-\bar{x})^2 = N\left(s^2 - \frac{n_2+1}{n_1+1}\cdot\delta^2\right), \tag{15}$$

whence

$$t = \tau\sqrt{\frac{n\cdot k}{n+2-k-k\cdot\tau^2}}, \qquad \text{where } n = N-2; \tag{16}$$

i.e., $t = \sqrt{n}\cdot\tan\theta$, $\sqrt{n+2-k}\cdot\sin\theta = \sqrt{k}\cdot\tau$.

In connection with analysis of variance where the total sample may be divided into several subsets of observations, the generalized τ may be used, accordingly, to indicate in a simple manner which (if any) of the means of subsets differ significantly from the general mean where the equivalent t-test is applicable.

In general let $\tau_{(P,N,k)} \geqq 0$ be a number such that P is the probability that $|\tau| > \tau_{(P,N,k)}$; where, as above, N is the total number of observations in the whole sample, k is the number of these in the subsample and τ is defined by (12). Then by (16), obviously,

$$\tau_{(P,N,k)} \equiv \sqrt{\frac{N-k}{k(N-1)}}\cdot\tau_{(P,N,1)}. \tag{17}$$

In *Table II* are given values of $\tau_{(P,N,1)}$ for a range of values of the arguments, N and P. The critical values of τ in *Table I* are simply values of this function for $P = \phi/N$ where ϕ is taken as parameter, i.e., $\tau_{(\phi/N,N,1)}$.

Rider[3] has given an interesting review of rejection criteria previously proposed.

BIBLIOGRAPHY

1. Irwin, J. O., *Biometrika, 17* (1925), pp. 100–128; *17* (1925), pp. 238–250.
2. Fisher, R. A., *Metron, 5,* (1925), No. 3, pp. 90–104, 109–112.
"Student," *Metron, 5,* (1925), No. 3, pp. 105–108, 113–120.
3. Rider, Paul R., St. Louis, Washington University Studies, new series, Science and Technology, No. 8 (1933), 23 pp.

Yale University.