ON THE STUDENTIZED EXTREME DEVIATE FROM THE SAMPLE MEAN*

by

Aurora B. Abesamis

Introduction

In a random sample of size n drawn from a normal population with mean μ and variance σ^2, let

$$x_1 \leq x_2 \leq \cdots \leq x_n \,. \tag{1}$$

Define the standardized extreme deviate

$$u = \frac{x_n - \bar{x}}{\sigma} \quad \text{or} \quad \frac{\bar{x} - x_1}{\sigma} \tag{2}$$

where $\bar{x}$ is the mean of the sample.

The distribution of u was obtained by Mckay [4] and tables of the probability integral by Nair [5] and Grubbs [2] have been published.

To evaluate the probability of an extreme observation using the tables of u, one must know σ. This paper considers the analogous statistic

$$\frac{x_n - \bar{x}}{s} = t_n, \quad \text{or} \quad \frac{\bar{x} - x_1}{s} = t_1 \tag{3}$$

* M.S. Thesis submitted to the Graduate School, U.P., April 1960, in partial fulfillment of the requirements for the degree of Master of Science (Statistics).

when σ is replaced by an estimate s derived from v degrees of freedom. By symmetry, the distribution of t_n or t_1 must be the same.

Tables for the upper percentage points of the distribution of t_n or t_1 have been published for various sample sizes and degrees of freedom, v =10(1)20 and selected values to 120 and ∞. (See Nair [5] and David [1]).

Upper percentage points for degrees of freedom less than 10 were not attempted until Tienzo [6] tackled the distribution problem of t_n and developed in series form this distribution for sample sizes 3, 4, and 5. By evaluating the series, the upper 5% and 1% significance levels for t_n were then obtained for n = 3(1)5 and $1 \leq v \leq 10$. As suggested by E. S. Pearson the accuracy of these approximations is investigated here.

As an extension of Tienzo's work, revised 5% and 1% points and those for 10% 2.5%, 0.5%. and 0.1% are obtained here by the alternative method of numerical integration used by David [1] making use of the tables of the distribution of u_n published by Grubbs [2].

Distribution Function.

If x_i's are normally distributed with mean u and variance σ^2 then these parameters of location and scale are eliminated from the distribution of the standardized variate $z_i = (x_i - u)/\sigma$.

Generally, if the x_i's are independently selected from several normal populations with $E(x_i) = u_i$ and common standard deviation σ, then the distribution of any function of several $z_i = (x_i - u_i)/\sigma$ does not involve the u_i and σ.

Assume that the x_i are drawn from normal distributions with mean u and standard deviation σ. Also, let

$$V = \frac{1}{v} \sum_{i}^{n} (x_j - \mu_j)^2 \qquad (4)$$

where x_j are independently chosen from normal distributions with $E(x_j) = u_j$ and common standard deviation σ_o such that $\sigma_o^2 = k\sigma_o^2$ with k a known constant; $\hat{u}_j$ are estimators of u_j and are linear functions of a set of p statistics which are themselves linear functions of the n observations; and $v = m - p$.

In the simplest case, the n x_i's constitute a random sample from a single normal population $N(u, \sigma^2)$; the m x_j's are independently drawn elements of a sample from either the same population or another with the same standard deviation, say $N(u + \lambda, \sigma^2)$. In the case $k = 1$ and V is an unbiased estimate of σ^2 with $v = m - 1$.

A generalization may be illustrated in a randomized block experiment with k blocks and n treatments. The x_i's may be the treatment means and V is the mean square for error in the relevant analysis of variance so the V/K is an unbiased estimate of the variance of x_i with degrees of freedom $v = (k-1)(n-1)$. Assuming normally distributed observations with common variance, the x_i and V are statistically independent. In what follows we will always assume the null

hypothesis to be true, that is, $E(x_i) = u$ for all i and write

$s^{*2} = V/k$.

Using the subscript i to denote the order of the observations as in (1),

$$t_i = \frac{x_i - \bar{x}}{s^*} \equiv \left[\frac{x_i - \mu}{\sigma} - \frac{\bar{x} - \mu}{\sigma} \right] \cdot \frac{\sigma}{s^*}$$

$$\equiv \frac{z_i - \bar{z}}{s} = \frac{u_i}{s} \qquad (5)$$

where $s^2 = \left[\frac{s^*}{\sigma} \right]^2$

Let the distribution of $u = ts$ be

$$F(u) = \int_0^u f(\xi)\, d\xi \,. \qquad (6)$$

For the purpose of this paper it need not be explicitly evaluated since only its numerical evaluation as given by Grubbs [2] is used here.

Since u and s are independently distributed,

$$f(u, s)\,du\,ds = f(u)f(s)\,du\,ds, \qquad 0 \leq u < \infty, \quad 0 \leq s < \infty. \qquad (7)$$

Let $F(t_a)$ be the probability that $t \leq t_a$. Since the cumulative distribution of $t = u/s$ is the probability that $t = u/s \leq t_a$, and $s \geq u/t$, then $F(t_a)$ can be obtained from (7) by integrating over the region for which $s \geq u/t_a$.

Thus

$$F(t_\alpha) = \int_0^\infty f(u) \int_{u/t_\alpha}^\infty f(s)\,ds\,du . \qquad (8)$$

The Method of Computation

The problem of obtaining the upper percentage points of $t = u/s$ is approached by means of the relationship (8). If we denote by t_α the α percentage points for t, then these values are the solutions t_α of the equation

$$1 - F(t_\alpha) = P(t \geq t_\alpha) = \int_0^\infty f(u) \int_0^{u/t_\alpha} f(s)\,ds\,du = \alpha . \qquad (9)$$

f(u) cannot be given explicit algebraic formulation; see, for example, Nair [5]. For the purpose of numerical integration, we follow David [1] in reading integrals of f(u) at intervals of .05 from the table of Grubbs [2] and these values of the integral are given against the center points determined by

$$.05(i - \tfrac{1}{2}), \qquad i = 1, 2, \ldots$$

The inner integral of (9) may be expressed conveniently as an incomplete gamma function and tables of the incomplete gamma function were used by David [1] to obtain his table of the upper α percentage points of t_n for $\nu \geq 10$. However, the precision to be obtained from interpolating in the tables of the incomplete gamma function is not considered adequate enough for the corresponding α percentage points of t_n for $\nu < 10$. In preference to interpolation from the incomplete gamma tables, (9) was evaluated for even values of $\nu < 10$ from the well known Poisson series expansion

(given, for example in the Introduction to Fisher and Yates' Tables).

For the case $v = 1$, the inner integral of (9) was evaluated from the cumulative distribution function of the normal variate. The values t_a for $v = 3, 5, 7$, and 9 were computed by Lagrangian interpolation [3] from the values obtained for even values of v.

The Appendix gives the values for t_a thus evaluated for the upper 10%, 5%, 2.5%, 1%, 0.5%, and 0.1% points of distribution of the studentized extreme deviate from the sample mean in a normal population for sample size $n = 3(1)10,12$ and degrees of freedom $v = 1(1)9$. The 5% and 1% points for $n = 3$ and 4 are reproduced from Tienzo [6]; his values for $n = 5$ (derived by extrapolation) have been revised.

Application of the Test.

It is recognized by those who collect and analyze sample data that some values in a sample of n observations are so far removed from the remaining values that the analyst is not willing to believe that these values have come from the same population. Many times the analyst is dubious about these outlying values and feels that he should make a decision as to whether to accept or reject these values as part of his sample. In another case, the analyst may not be looking for an error, but may wish to recognize a situation when an occasional observation occurs which is from a different population. He may wish to discover whether a significant F in an analysis of variance is due to an extreme value significantly different from the other values. Or, he may wish to discover whether an extreme value differs significantly from the other values, although the F in the analysis of variance is not significant. This apppears to have been the application envisioned by McKay [4] and Nair [5]

Example: To illustrate how the statistic may be applied, a hypothetical example is given here:

EXTREME DEVIATE FROM THE SAMPLE MEAN

An F test in an experiment has shown that the effects of related drugs A, B, and C on the rate of finger tapping of a group of trained subjects are significantly different. However, the experimenter suspects from a closer scrutiny of the observations that the significant F may have been due to the marked effect of drug A.

A further experiment is undertaken with the following hypothesis in mind:

$H_0 : u_A = u_B = u_C$ That is, the effect of drug A on the rate of finger tapping is the same as those of the other drugs.

The alternative hypothesis would be:

$H_1 : u_A < u_B = u_C$. That is, drug A affects the rate of finger tapping to a higher degree than drugs B and C.

The appropriate criterion in this situation is $t_1 = \frac{\bar{x} - x_1}{s}$

Using a randomized block experiment on the effect of the same dose of drugs A, B, and C upon the rate of finger tapping of four trained subjects, suppose the following data are obtained:

Drug	Subject No.				Totals	Means
	1	2	3	4		
A	11	56	15	6	88	22
B	26	83	34	13	156	39
C	20	71	41	32	164	41
Totals	57	210	90	51	$G = 408$	$\bar{x} = 34$

Analysis of Variance:

Sources of Variation	d.f.	Sum of Squares	Mean Square	F
Subjects	3	5,478	1,826	
Treatments	2	872	436	7.88*
Error	6	332	55.3	
Total	11	6,682		

An interesting feature would then be the closeness of the two means for B and C in contrast to the mean of A. Applying the extreme deviate test.

$$t_3 = \frac{(34 - 22)\sqrt{4}}{7.44} = 3.23 ,$$

with $n = 3$, $k = 4$, and $v = 6$. Referring to the Appendix, we find that $t_3 = 3.23$ is beyond the tabular value $t_3(.025) = 3.03$ and $t_3(.05) = 2.55$ showing that it is an exceptional value.

BIBLIOGRAPHICAL REFERENCES

[1] David, H. A., "Revised Upper Percentage Points of the Studentized Deviate from the Sample Mean, "**Biometrika,** Vol. 43, 1956, pp. 449-451.

[2] Grubbs, F. E., "Sample Criteria for Testing Outlying Observations," **Annals of Mathematical Statistics, Vol.** 21, 1950, pp. 27—57.

[3] **Interpolation and Allied Tables,** London: Her Majesty's Stationery Office, 1936.

[4] McKay, A. T,, "The Distribution of the Difference Between the Extreme Observation and the Sample Mean in a Sample of n from a normal Universe, "**Biometrik,** Vol, 27, 1935, pp, 466—471,

[5] Nair, K, R,, "The Distribution of the Extreme Deviate from the Sample Mean and its Studentized Form, "**Biometrik,** Vol. 35, 1948, pp. 118—144.

[6] Tienzo, B, P,, "On the Distribution of the Extreme Studentized Deviate from the Sample Mean," Unpublished Thesis, 1958 University of the Philippines.

APPENDIX

UPPER PERCENTAGE POINTS OF THE STUDENTIZED EXTREME DEVIATE FROM THE SAMPLE MEAN

$$(x_n - \bar{x})/s \text{ or } (\bar{x} - x_1)/s$$

10% Points

ν \ n	3	4	5	6	7	8	9	10	12
1	7	8	9	10	11	11	12	12	13
2	2.9	3.4	3.8	4.1	4.4	4.6	4.7	4.9	5.2
3	2.24	2.62	2.90	3.10	3.27	3.42	3.55	3.66	3.85
4	2.03	2.35	2.59	2.77	2.92	3.05	3.15	3.25	3.41
5	1.90	2.20	2.41	2.57	2.70	2.82	2.91	3.00	3.14
6	1.82	2.10	2.30	2.45	2.57	2.68	2.77	2.85	2.98
7	1.77	2.03	2.22	2.36	2.48	2.58	2.66	2.74	2.87
8	1.73	1.98	2.17	2.31	2.42	2.51	2.59	2.67	2.79
9	1.70	1.95	2.13	2.27	2.37	2.46	2.54	2.61	2.73

APPENDIX (Continued)

5% Points

ν \ n	3	4	5	6	7	8	9	10	12
1	13.5	16.4	19	20	22	23	24	25	26
2	4.23	4.98	5.5	6.0	6.3	6.6	6.9	7.1	7.5
3	3.03	3.50	3.88	4.15	4.36	4.55	4.72	4.86	5.11
4	2.58	2.98	3.26	3.48	3.65	3.80	3.93	4.05	4.24
5	2.37	2.71	2.95	3.15	3.30	3.43	3.54	3.64	3.80
6	2.24	2.55	2.78	2.95	3.09	3.21	3.31	3.39	3.54
7	2.15	2.45	2.66	2.82	2.95	3.06	3.15	3.23	3.37
8	2.09	2.37	2.57	2.72	2.85	2.95	3.04	3.12	3.25
9	2.04	2.32	2.51	2.65	2.78	2.87	2.96	3.03	3.15

APPENDIX (Continued)

2.5%

v \ n	3	4	5	6	7	8	9	10	12
1	27	33	37	40	43	45	47	49	52
2	6.1	7.1	7.9	8.5	9.0	9.4	9.8	10.1	10.7
3	4.2	4.9	5.4	5.8	6.1	6.4	6.7	6.9	7.2
4	3.22	3.68	4.02	4.28	4.49	4.67	4.82	4.96	5.19
5	2.88	3.28	3.57	3.79	3.96	4.11	4.23	4.34	4.52
6	2.68	3.03	3.29	3.48	3.63	3.77	3.88	3.98	4.14
7	2.55	2.88	3.11	3.28	3.43	3.55	3.65	3.74	3.89
8	2.46	2.77	2.98	3.15	3.29	3.40	3.49	3.58	3.72
9	2.39	2.69	2.90	3.05	3.18	3.29	3.38	3.46	3.59

APPENDIX (Continued)

1% Points

ν \ n	3	4	5	6	7	8	9	10	12
1	68	82	93	101	108	114	119	123	130
2	9.9	11.33	12.6	13.6	14.4	15.0	15.6	16.1	16.9
3	5.5	6.29	6.9	7.3	7.7	8.1	8.4	8.6	9.0
4	4.23	4.81	5.23	5.54	5.80	6.03	6.22	6.39	6.68
5	3.65	4.11	4.45	4.70	4.93	5.11	5.26	5.39	5.62
6	3.32	3.72	4.02	4.24	4.43	4.58	4.71	4.82	5.01
7	3.11	3.48	3.74	3.94	4.11	4.25	4.37	4.46	5.63
8	2.96	3.31	3.56	3.74	3.89	4.02	4.13	4.22	4.38
9	2.86	3.19	3.41	3.59	3.73	3.86	3.95	4.04	4.19

APPENDIX (Continued)
0.5% Points

ν \ n	3	4	5	6	7	8	9	10	12
1	135	164	186	202	216	227	237	245	260
2	13.7	16.1	17.9	19.2	20.3	21.3	22.1	22.8	24.0
3	7.5	8.7	9.4	10.0	10.4	10.7	11.0	11.3	11.8
4	5.09	5.79	6.29	6.68	6.99	7.26	7.49	7.69	8.04
5	4.23	4.79	5.19	5.50	5.75	5.97	6.15	6.32	6.60
6	3.83	4.30	4.63	4.88	5.08	5.26	5.40	5.53	5.75
7	3.56	3.98	4.27	4.50	4.68	4.84	4.96	5.07	5.25
8	3.36	3.75	4.02	4.22	4.38	4.52	4.63	4.73	4.90
9	3.22	3.58	3.84	4.02	4.17	4.30	4.40	4.49	4.64

APPENDIX (Continued)
0.1% Points

ν \ n	3	4	5	6	7	8	9	10	12
1	675	821	928	1011	1079	1136	1185	1227	1300
2	30.7	36.1	40.0	43.1	45.6	47.7	49.5	51.1	53.8
3	14.5	16.7	17.9	18.7	19.4	20.1	20.7	21.2	22.1
4	7.8	8.8	9.6	10.1	10.6	11.0	11.4	11.7	12.2
5	6.1	6.9	7.5	7.9	8.2	8.5	8.8	9.1	9.5
5	5.2	5.9	6.3	6.6	6.9	7.1	7.3	7.5	7.8
7	4.7	5.3	5.6	5.9	6.1	6.3	6.4	6.6	6.9
8	4.4	4.9	5.2	5.4	5.6	5.8	5.9	6.1	6.3
9	4.2	4.6	4.9	5.1	5.3	5.5	5.6	5.7	5.9