Journal of Research of the National Bureau of Standards Vol. 48, No. 3, March 1952 Research Paper 2311

Properties of Certain Statistics Involving the Closest Pair in a Sample of Three Observations

Julius Lieblein

Triplicate readings are of wide occurrence in experimental work. Occasionally, however, only the closest pair of a triad is used, and the outlying high or low one discarded as evidencing some gross error. The present paper presents a mathematical investigation leading to precise determination of some of the biases that result from such selection. This project was suggested by certain experiments involving random sampling numbers and analysis of published chemical determinations. The theoretical findings agree closely with the empirical results and imply that selected pairs not only tend to overestimate considerably the precision of the experimental procedure, but also result in less accurate determinations.

1. Introduction

Triplicate determinations are fairly common in the chemical laboratory inasmuch as a third one is occasionally taken to indicate which of the other two is more likely to be off the mark. A corollary of this is that if only two of the three measurements are in close agreement the worker is under strong temptation to discard completely the remaining distant one on the ground that evidence of gross error is present. A similar practice also appears to be encouraged by instruction methods in quantitative chemical analysis which grade students not only upon the correctness of their results, made in duplicate, but also upon their precision as measured by the difference between the two results. Thus, a student might hope to improve his record by quietly making a third, uncalled-for analysis, give himself the advantage of the closest two of all three, and omit to mention the remaining one. This is a very striking case of the long-standing problem of the rejection of outlying observations and raises the statistical question of how estimates of the mean and variability of analyses are affected by such procedures. It is this question, rather than the rejection of outlying observations,[1] [2] that is emphasized in the present investigation, although the rejection problem is also touched on, in connection with the first of the three statistics, y_1, discussed below.

The author is indebted to W. J. Youden for drawing his attention to this question and suggesting its theoretical investigation when search of the statistical literature indicated that this apparently simple problem had not been considered heretofore.[3]

Accordingly, the present study was executed and resulted in the present paper, which is purely a mathematical treatment undertaken to verify and extend certain sampling results, obtained by Youden in an empirical investigation of the above question, which were reported in the National Bureau of Standards Technical News Bulletin for July, 1949 [3]. The method of treatment was to study some of the properties of two measurements that are selected out of a sample of three according to a stated criterion computed from the sample observations. The statistics that define such properties are of more general character than order statistics—that is, observations ordered according to size, such as the largest value in a sample, the sample median, etc. Whereas order statistics are widely treated in the literature,[4] the type of statistics being considered here, which depend on features other than size, have apparently received relatively little attention.[5]

This report is thus limited to the following three questions, answers to which will serve to throw light on the differences to be expected between taking two measurements at random ("true duplicates") and taking two measurements that are really part of a random sample of three.[6] (1) In a random sample of three observations from a single (continuous) population what values of the following ratio may be considered significant: ratio of the gap between the two closest values to the whole range of the sample? (2) How does the range in a sample of true duplicate measurements compare with the differ-

[1] Figures in brackets indicate the literature references at the end of this paper.

[2] For information on the many aspects of outlying observations that have been treated in the literature, the reader is advised to consult a recent article by F. E. Grubbs [1], in which, in addition to discussing several new criteria for testing discordant observations, he presents a detailed bibliography of the problem. A particularly comprehensive survey of developments prior to 1933 is provided in a study by P. R. Rider [2] published in that year. See also the two papers by W. J. Dixon [7, 8] and the one by G. R. Seth [10].

[3] After this paper was prepared, the author received a copy of a manuscript of an article by Franklin M. Henry of the University of California, Berkeley, entitled, "The loss of precision from discarding discrepant data". This article has since been published [11]. It presents no mathematical theory for triads, but gives, among other interesting points, a discussion of an experiment in judging 10-second time intervals by a series of triplicate measurements in which the two "closest" were averaged in each triad. The standard deviation of the mean of such averages for 50 triads was *0.131* sec, whereas theory (table 1b, Part B, col. 1) gives (since the standard deviation of the whole set of 150 readings is $\sigma=0.162$ instead of the $\sigma=1$ used in our table) the remarkably close value $0.7986\times0.162=$ *0.129* sec for samples from a normal population and $0.9083\times0.162=$*0.147* sec for samples from a rectangular population (col. 4). The author is obliged to Henry for his kindness in making his paper available in advance of publication.

Attention is also called to a note by G. R. Seth [10] on the distribution of the two closest among a set of three observations. Seth became interested in the problem in the course of a discussion with the author during his visit to the Statistical Engineering Laboratory in the spring of 1948. In this note he obtains in general terms some of the results also given in the present paper and applies them to the normal distribution. The author wishes to acknowledge that the present paper has benefited from correspondence with Seth on the problem. (In this connection see also footnotes 11 and 16.)

[4] For a comprehensive survey of the literature on order statistics, see Wilks [4].

[5] The most directly relevant article known to the author is by J. W. Tukey [5], in which he obtains tables relative to the distribution of the *largest* gap, rather than the smallest, in samples of from 2 to 10 by experimental sampling from a unit normal universe and also by analytical means.

[6] The answers to these questions are indicated in the tables as follows: (1), tables 2 and 1a; (2) and (3), table 1b. These tables, which are an attempt to condense the main results of this paper, are summarized in section 2.

ence between the two values in each of (a) the *closest* pair out of a sample of three measurements, (b) the lowest (or highest) pair out of such a sample, and (c) the pair of extremes (highest and lowest values) of the entire sample of three—as regards several types of universes? (3) How do the means compare in each case? It is not intended to consider other problems that can arise, such as drawing the sample from a mixed population, or adopting a rule to omit the extreme measurement only when the range of the three observations exceeds a specified value, and utilizing all three of them otherwise. Neither is it intended in this paper to go into any other statistical questions such as estimation and significance tests or more general decision problems.

2. Summary

The answers to the above questions involve primarily the investigation of the distributions of the three statistics, y_1, y_2, and y_3, whose main properties are summarized in tables 1a and 1b below and compared with the results of both actual sampling by making use of a table of random numbers, and data on chemical analyses that appeared in the chemical literature.[7]

The statistics y_i are defined as follows. Let x_1, x_2, x_3 be the sample of three observations arranged in order of increasing magnitude:

$$x_1 \leq x_2 \leq x_3.$$

Let now

$$x', x'', x'''$$

designate the same three observations rearranged so that x' and x'' are the two *closest* of the three and $x' \geq x''$. Then the selected statistics treated are

$$y_1 = \frac{x'-x''}{x_3-x_1}, \qquad y_2 = \frac{x'-x''}{2}, \qquad y_3 = \frac{x'+x''}{2}.$$

Results are presented, insofar as they have been obtained, for the three parent universes, rectangular, right triangular, and normal, though not necessarily in the same detail for each one.

The comparisons indicated in table 1 reveal the following facts for random samples of three measurements, where, unless otherwise stated, the statements apply to samples from a normal or a rectangular population:

1. The empirical sampling results, obtained prior to the theoretical calculations, show fairly substantial agreement with the theory. The chemical data from experimental determinations reported in a chemical journal and studied by W. J. Youden are likewise in agreement.[8]

2. The statistic y_1, which characterizes the partition of the range by the middle item in a random sample of three measurements, behaves remarkably alike for samples from three different basic populations, the normal, rectangular, and right triangular (table 1a). This suggests that this ratio statistic will not be very useful as a criterion for discriminating between a normal population and some other population.

3. A set of two observations selected by taking the closest two out of three from a normal or a rectangular population differs strikingly from other pairs taken from the three or from a pair of true duplicates, as shown by the following:

a. The average *difference* (as measured by y_2) between the selected pair is less than half that for the true duplicates, and the same is true of the variability of this distance as measured by the standard deviation (table 1b, Part A, Cols. 1, 3 and 4, 6). Furthermore, the difference between the selected pair behaves (again in an average sense) very much like *half* the difference between the two lowest (or highest) in the full sample of three, and (in the same sense) is similar to *one-quarter* the difference between the two most extreme measurements in the sample. The standard deviation of the difference between the closest pair is, however, comparable to the standard deviation of half the range (table 1b, Part A, Cols. 1, 2, and 4, 5).

b. The *mean* (y_3) of a selected pair varies somewhat more than the mean of a true duplicate pair, the average value of both these means being the

[7] For additional comparisons with experimental data that came to the author's attention too late for inclusion in the main body of the paper, see footnote 3.

[8] For other empirical evidence see footnote 3.

TABLE 1a. *Characteristics of the ratio y_1 of the distance between the closest pair to the range in a sample of three measurements*

$$y_1 = \frac{x'-x''}{x_3-x_1},\ x'' \leq x',\ x_1 \leq x_2 \leq x_3$$

	Normal population		Rectangular population		Right triangular population		Published chemical data
	Theory	Sampling with random numbers	Theory	Sampling with random numbers	Theory	Sampling with random numbers	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
N, number of samples of 3	----------	400	----------	200	----------	200	75
Probability density function	$\frac{3\sqrt{3}}{\pi(y_1^2-y_1+1)}, 0 \leq y_1 \leq 1/2$	------	$2, 0 \leq y_1 \leq 1/2$	------	$2, 0 \leq y_1 \leq 1/2$	------	------
Expected or mean value	0.2621	0.2582	0.25	0.2506	0.25	0.2441	0.2573
Standard deviation	0.1428	0.1421	0.1443	0.1612	0.1443	0.1480	0.1565

same. Especially noteworthy is the fact that the true average of the population is more accurately estimated by using the two most discrepant observations of the three in forming an average, $\frac{1}{2}(x_1+x_3)$, than by taking the two that are most in agreement, although neither method is as accurate as taking the mean of all three (table 1b, Part B, cols. 1, 4). Thus, selection of measurements on the basis of close agreement *increases* rather than decreases the true error of measurement.

In addition to the above relationships the behavior of the outlying observation x''' is of interest

TABLE 1b. *Characteristics of other statistics related to the closest pair of measurements in a sample of 3*

$x_1 \leq x_2 \leq \ldots \leq x_n$ denote the measurements in a sample of n ordered according to size. If $n=3$, then x' and x'', $x' \geq x''$, denote the two *closest* measurements in the sample (x_1, x_2, x_3). The measurements are drawn independently at random from the populations designated. The rectangular population has been adjusted to unit variance and centered at the origin. Exact values and distribution functions are given where practical. Where the interval of nonzero probability density is omitted for a probability distribution, the variate is assumed to take all values from $-\infty$ to $+\infty$. For fuller explanation see text.

	Normal population, $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, -\infty<x<\infty$			Rectangular population with unit variance, $g(x)=\frac{1}{\sqrt{12}}, -\sqrt{3}\leq x\leq\sqrt{3}$		
	(1) Closest pair in a sample of 3	(2) Lowest pair [a] in a sample of 3	(3) Sample of 2 ("true duplicates")	(4) Closest pair in a sample of 3	(5) Lowest pair [a] in a sample of 3	(6) Sample of 2 ("true duplicates")
	A. Statistics relative to the DISTANCE between two values					
Statistic	$x'-x''=2y_2=y_2'$	$(x_2-x_1)/2=s$	$(x_2-x_1)/2=p$	$x'-x''=2y_2=y_2'$	$(x_2-x_1)/2=s$	$(x_2-x_1)/2=p$
Probability density function.	$\frac{3\sqrt{3}}{\pi}\int_{y_2'}^{\infty} e^{-\frac{1}{4}(3t^2+y^2)}dt$ $0\leq y_2'<\infty$	$\frac{6\sqrt{3}}{\pi}\int_{\frac{1}{2}s}^{\infty} e^{-(3t^2+s^2)}dt$ $0\leq s<\infty$	$\frac{2}{\sqrt{\pi}}e^{-p^2}$ $0\leq p<\infty$	$\frac{1}{\sqrt{3}}(\sqrt{3}-y_2')^2$ $0\leq y_2'\leq\sqrt{3}$	$\frac{1}{\sqrt{3}}(\sqrt{3}-s)^2$ $0\leq s\leq\sqrt{3}$	$\frac{2}{3}(\sqrt{3}-p)$ $0\leq p\leq\sqrt{3}$
Mean	$E(y_2')=0.4535$	$E(s)=\frac{3}{4\sqrt{\pi}}=0.4231$	$E(p)=\frac{1}{\sqrt{\pi}}=0.5642$	$E(y_2')=\frac{\sqrt{3}}{4}=0.4330$	$E(s)=\frac{\sqrt{3}}{4}=0.4330$	$E(p)=\sqrt{\frac{1}{3}}=0.5774$
Other means for comparison.	$E(y_2')=0.4451$ (experimental value [b]) $E(x_3-x_1)/4=\frac{3}{4\sqrt{\pi}}=0.4231$	$E(x_3-x_1)/4=\frac{3}{4\sqrt{\pi}}$		$E(x_3-x_1)/4=\frac{\sqrt{3}}{4}$	$E(x_3-x_1)/4=\frac{\sqrt{3}}{4}$	
Standard deviation	$\sigma(y_2')=0.3746$	$\sigma(s)=5.3379$	$\sigma(p)=\left(\frac{1}{2}-\frac{1}{\pi}\right)^{\frac{1}{2}}=0.4263$	$\sigma(y_2')=\frac{3\sqrt{5}}{20}=0.3354$	$\sigma(s)=\frac{3\sqrt{5}}{20}=0.3354$	$\sigma(p)=\sqrt{\frac{1}{6}}=0.4082$
Other standard deviations for comparison.	$\sigma(x_3-x_1)/2=0.4442$	$\sigma(x_3-x_1)/2=0.4442$		$\sigma(x_3-x_1)/2=\frac{\sqrt{15}}{10}=0.3873$	$\sigma(x_3-x_1)/2=\frac{\sqrt{15}}{10}=0.3873$	
	B. Statistics relative to the AVERAGE of two values					
Statistic	$(x'+x'')/2=y_3$	$(x_1+x_2)/2=q$	$(x_1+x_2)/2=m$	$(x'+x'')/2=y_3$	$(x_1+x_2)/2=q$	$(x_1+x_2)/2=m$
Probability density function.	$\frac{6}{\pi\sqrt{2\pi}}\int_0^{\infty}\int e^{-Q}dt\,dy_2$, where t ranges over $(-\infty, -3y_2+y_3)$ and $(3y_2+y_3, \infty)$; $Q=\frac{1}{2}t^2+y_2^2+y_3^2$	[c]	$\frac{1}{\sqrt{\pi}}e^{-m^2}$	[c]	[c]	$\frac{1}{3}(\sqrt{3}-\lvert m\rvert)$, $-\sqrt{3}\leq m\leq\sqrt{3}$
Mean	$E(y_3)=0$	$E(q)=-\frac{3}{4\sqrt{\pi}}=-0.4231$	$E(m)=0$	$E(y_3)=0$	$E(q)=-\frac{\sqrt{3}}{4}=-0.4330$	$E(m)=0$
Other means for comparison.	$E(y_3)=-0.0335$ (experimental value [b]) $E(x_1+x_3)/2=0$ $E(x_1+x_2+x_3)/3=E\bar{x}=0$			$E(x_1+x_3)/2=0$ $E(x_1+x_2+x_3)/3=0$		
Standard deviation	$\sigma(y_3)=\left(\frac{1}{2}+\frac{\sqrt{3}}{4\pi}\right)^{\frac{1}{2}}=0.7986$	$\sigma(q)=0.6244$	$\sigma(m)=\sqrt{\frac{1}{2}}=0.7071$	$\sigma(y_3)=\sqrt{\frac{33}{40}}=0.9083$	$\sigma(q)=\sqrt{\frac{33}{80}}=0.6423$	$\sigma(m)=0.5$
Other standard deviations for comparison.	$\sigma(y_3)=0.8098$ (experimental value [b]) $\sigma(x_1+x_3)/2=\left(\frac{1}{2}-\frac{\sqrt{3}}{4\pi}\right)^{\frac{1}{2}}=0.6018$ $\sigma(x_1+x_2+x_3)/3=\sigma_{\bar{x}}=\frac{1}{\sqrt{3}}=0.5774$	$\sigma(x_1+x_3)/2=0.6018$		$\sigma(x_1+x_3)/2=\sqrt{\frac{3}{10}}=0.5477$ $\sigma_{\bar{x}}=\frac{1}{\sqrt{3}}=0.5774$	$\sigma(x_1+x_2)/3=\sqrt{\frac{3}{10}}=0.5477$	

[a] The characteristics of the *highest* pair are obtainable from symmetry considerations.
[b] Values obtained by sampling experiments using a table of random normal deviates.
[c] These density functions have been omitted since they are rather complicated.

and will be briefly considered.

Although the basic ideas present little difficulty, the explicit values and probability distributions needed in this paper often involve calculation of multiple integrals over quite complicated regions. The exact calculation of these integrals has usually required much tedious manipulation, too lengthy to warrant more than the briefest indication. A detailed manuscript of these procedures is in the possession of the author.

3. Derivation of Results; Descriptive Properties

3.1. The Statistic y_1

a. Distribution and moments in general

Let x_1, x_2, x_3 be the three observations, arranged in order of magnitude, in a random sample of three from a population with *pdf* (probability density function) $f(x)$, supposed continuous (and differentiable as often as necessary), and suppose $f(x)$ is nonzero in the interval (a, b) where either or both endpoints may be at infinity. Then the joint density function of x_1, x_2, x_3 is [6]

$$p(x_1,x_2x_3)dx_1dx_2dx_3=3!f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3,\quad a\le x_1\le x_2\le x_3\le b. \tag{1}$$

Letting $x'\ge x''$ be the two *closest* observations, the statistic y_1 may be written

$$y_1=\frac{x'-x''}{x_3-x_1}=\min\ (y_{11},y_{12}),$$

where

$$y_{11}(x_1,x_2,x_3)\equiv\frac{x_2-x_1}{x_3-x_1},\ y_{12}(x_1,x_2,x_3)\equiv\frac{x_3-x_2}{x_3-x_1}\equiv 1-y_{11}(x_1,x_2,x_3) \tag{1a}$$

are simply functions of the x's and will be used with the arguments often omitted for brevity. Thus it is required to find the distribution of the variate y_1, defined over $0\le y_1\le\frac{1}{2}$, which takes different functional forms, namely

$$y_1=\begin{cases} y_{11}(x_1,x_2,x_3) \text{ if } 0\le y_{11}(x_1,x_2,x_3)\le\frac{1}{2} \\ y_{12}(x_1,x_2,x_3) \text{ if } 0\le y_{12}(x_1,x_2,x_3)\le\frac{1}{2}, \end{cases}$$

where y_{11}, y_{12} are simply used as abbreviations for the fractions in (1a).

To find the distribution of y_1, we have (in the notation of the theory of probability), since the events indicated on the right are mutually exclusive,

$$P\{y_1\le Y\}=P\left\{0\le y_{11}\le Y,\ 0\le y_{11}\le\frac{1}{2}\right\}+P\left\{0\le y_{12}\le Y,\ 0\le y_{12}\le\frac{1}{2}\right\}, \tag{2}$$

which is equivalent to

$$P\{y_1\le Y\}=\begin{cases} 0, & \text{if } Y<0 \\ P\{0\le y_{11}\le Y\}+P\{0\le y_{12}\le Y\}, & \text{if } 0\le Y\le\frac{1}{2} \\ 1, & \text{if } Y>\frac{1}{2} \end{cases} \tag{2a}$$

The equation (2a) can be differentiated with respect to Y to give the probability density function in the form

$$p(y_1)=p_1(y_{11})+p_2(y_{12}),\ \text{if}\ 0\le y_{11},\ y_{12}\le\frac{1}{2}=0, \tag{2b}$$

$$=0,\ \text{otherwise},$$

with y_{11} and y_{12} replaced by y_1 in the result. Thus the required distribution is reduced to those of statistics of the usual type.

To find $p_1\ (y_{11})$, apply the transformation [9]

$$\begin{aligned} x_1&=r-q \\ x_2&=r-q(1-y_{11}) \\ x_3&=r \end{aligned} \tag{3}$$

to (1), obtaining

$$h(y_{11},q,r)dy_{11}dr\,dq=6qf(r-q)f[r-q(1-y_{11})]f(r)dy_{11}dr\,dq,\quad 0\le q\le r-a,\ a\le r\le b,\ 0\le y_{11}\le 1 \tag{4}$$

whence the *pdf* of the variate y_{11} is

$$p_1(y_{11})=\int_a^b\int_0^{r-a}6q\ \ f(r-q)f(r)f[r-q(1-y_{11})]\,dq\ \ dr$$

$$=p_1(1-y_{12}),\qquad 0\le y_{11}\le 1. \tag{5}$$

Since $y_{12}=1-y_{11}$, its density function is, similarily,

$$p_2(y_{12})=\int_a^b\int_0^{r-a}6q\ f(r-q)f(r)f(r-qy_{12})\,dq\ dr,\quad 0\le y_{12}\le 1. \tag{6}$$

Hence finally (2b) gives

$$p(y_1)=\int_a^b\int_0^{r-a}6q\ f(r-q)f(r)\phi(y_1)dq\,dr,\quad 0\le y_1\le\frac{1}{2}, \tag{7}$$

[9] This is obtained by putting $y_{11}=\frac{x_2-x_1}{x_3-x_1}$, $q=x_3-x_1$, and $r=x_3$.

where

$$\phi(y_1)=f(r-qy_1)+f[r-q(1-y_1)], \qquad (8)$$

as the general formula for the distribution of y_1 for a population with continuous *pdf* $f(x)$.

The above expressions appear to lend themselves to but few general statements. Thus, it may be seen from (7) and (8) that for a rectangular parent population $f(x)$ the distribution $p(y_1)$ of y_1 is rectangular. Furthermore, y_1 will evidently be rectangularly distributed for all parent distributions for which the function $\phi(y_1)$ does not depend upon y_1, that is, for which the function

$$\phi(y_1)=f(r-qy_1)+f[r-q(1-y_1)]$$

depends at most upon r and q. If f is a linear function (triangular or rectangular distribution), this is seen to be true, for the y_1's cancel out. Conversely, by differentiating with respect to y_1, it can be shown that if y_1 has a rectangular distribution, then f must be linear, if differentiable.

For future use, it is desirable to obtain general expressions for the moments, μ_k, of y_1. In view of (1a), (2b), and (6) these are given by

$$\mu_k=\int_0^{\frac{1}{2}} y_1{}^k p(y_1)\,dy_1=\int_0^{\frac{1}{2}} y_{11}^k p_1(y_{11})\,dy_{11}+$$

$$\int_0^{\frac{1}{2}} y_{12}^k p_1(1-y_{12})\,dy_{12}$$

which, under the transformations

$$y_{11}=\frac{1}{2}-t, \qquad y_{12}=\frac{1}{2}-t,$$

become

$$\mu_k=\int_0^{\frac{1}{2}}\left(\frac{1}{2}-t\right)^k\left[p_1\left(\frac{1}{2}-t\right)+p_1\left(\frac{1}{2}+t\right)\right]dt,$$

in which p_1 is the *pdf* of the ratio

$$y_{11}=\frac{x_2-x_1}{x_3-x_1}.$$

If the function $p_1(u)$ is one that is symmetrical about $u=\frac{1}{2}$, then μ_k may be written, putting $\frac{1}{2}-t=s$,

$$\mu_k=2\int_0^{\frac{1}{2}} s^k p_1(s)ds.$$

For certain symmetrical universes, the distribution of y_{11} has been investigated numerically by W. J. Dixon [7] for samples of three and various larger sizes as well.[10]

b. Rectangular universe

For the rectangular or uniform parent universe given by

$$f(x)=1, \qquad 0\le x\le 1$$

and zero elsewhere (this simple form is called the "square" universe), the general expression (7) becomes

$$p(y_1)=2\int_0^1\int_0^r 6q\,dq\,dr=2, \qquad 0\le y_1\le\frac{1}{2},$$

verifying that the ratio y_1 also is rectangular.

The first few moments are

$$E(y_1)=\frac{1}{4}, \qquad E(y_1^2)=\frac{1}{12}, \qquad \sigma(y_1)=\frac{\sqrt{3}}{12}=.1443.$$

It is interesting to see whether values of the ratio y_1 tend to depend on the spread, x_3-x_1, of the sample values.

It can be shown by the method used in obtaining (4) that, for the rectangular case, the joint probability density function of y_1, x_1, x_3 is

$$f(y_1,x_1,x_3)=12(x_3-x_1), \qquad 0\le x_1\le x_3\le 1, \qquad 0\le y_1\le\frac{1}{2}.$$

Since this is independent of y_1, it follows that the ratio y_1 is independent, not only of the range, but also of both sample extremes x_1 and x_3.

c. Triangular universe

In simplest form this is given by

$$f(x)=2x, \qquad 0\le x\le 1$$

and zero elsewhere. Formula (7) here gives

$$p(y_1)=\int_0^1\int_0^r 48qr(r-q)[r+(r-q)]dq\,dr=2, 0\le y_1\le\frac{1}{2},$$

so that the distribution of y_1 is identical to that of the previous case.

---

[10] In addition, Dixon has published a paper [8] that gives a thorough treatment of a large number of measures that may be used in testing whether an outlying observation (or several such) should be rejected. Of these statistics, the only one that has any direct relationship to any studied in the present paper is, for $n=3$,

$$r_{10}=\frac{x_n-x_{n-1}}{x_n-x_1}=\frac{x_3-x_2}{x_3-x_1}.$$

This expression, which (for $n=3$) is the same as $y_{12}(=1-y_{11})$ in (1a) above, is mentioned by Dixon as a criterion for testing the upper outlier x_3. The author is obliged to Dixon for making his two papers available in advance of publication.

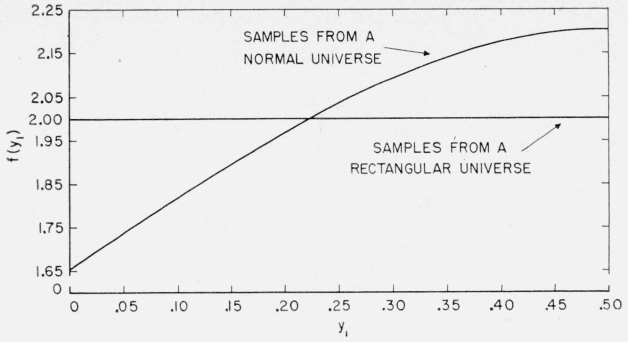


FIGURE 1. *Distribution of ratio y_1 in samples from a normal and from a rectangular universe.*

$$f(y_1)dy_1=\frac{3\sqrt{3}}{\pi}\cdot\frac{dy_1}{y_1^2-y_1+1}, \qquad 0\le y_1\le\frac{1}{2}$$

$$y_1=\frac{x'-x''}{x_3-x_1}$$

$$x'\ge x'', \qquad x_3\ge x_2\ge x_1$$

d. Normal Universe

For $f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, $-\infty\le x\le\infty$, (called the unit or standard normal distribution or universe), formula (7) becomes

$$p(y_1)=\int_{-\infty}^{\infty}\int_{0}^{+\infty}\frac{6}{(2\pi)^{3/2}}\,q\left\{\exp\left(-\frac{1}{2}[(r-q)^2+r^2+(r-qy_1)^2]\right)+\exp\left(-\frac{1}{2}[(r-q)^2+r^2+(r-q\sqrt{1-y_1})^2]\right)\right\}dq\,dr,\ 0\le y_1\le\frac{1}{2}, \tag{9}$$

or

$$p(y_1)=\frac{3\sqrt{3}}{\pi(1-y_1+y_1^2)}, \qquad 0\le y_1\le\frac{1}{2}.$$ [11]

This consists of the arc of a Cauchy distribution curve included between the left-hand inflection point and mode, and is shown in figure 1. Several percentage points obtained from the cumulative distribution are presented in table 2.

TABLE 2. *Percentage points of y_1 for the unit normal*

$$y_1=\frac{x'-x''}{x_3-x_1};\ Pr\{y_1\le y_1^0\}=\frac{6}{\pi}\arctan\left(\frac{2y_1^0-1}{\sqrt{3}}\right)+1$$

Probability, P, that y_1 does not exceed given value of y_1^0		Critical value, y_1^0, corresponding to given probability P	
y_1^0	P	P	y_1^0
0	0	0	0
1/11	0.1572	0.01	0.00603
1/6	0.2983	0.05	0.02979
1/3	0.6369	0.10	0.05874
		0.25	0.14128
		0.50	0.23205

[11] This distribution has also been obtained by G. R. Seth [10]. (See also footnotes 3 and 16).

The above table bears out the fact that the ratio y_1 is not a good criterion to use for the rejection of outlying observations. Thus, a ratio as marked as one-sixth or less, indicating that the outermost observation is at least five times as distant from the middle one as is the remaining one, may be expected (if the universe is normal) about 30 percent of the time; even when the distances are in the ratio 10 to 1 or more, one by no means has a rare event—it may be expected only a little less often than in one sample out of six.

The moments are given by

$$E(y_1)=\frac{1}{2}-\frac{3\sqrt{3}}{2\pi}\ln\frac{4}{3}=0.26209,$$

$$V(y_1)=\frac{3\sqrt{3}}{2\pi}-\frac{3}{4}-\frac{27}{4\pi^2}\left(\ln\frac{4}{3}\right)^2=0.020392,$$

$$\sigma(y_1)=\sqrt{V(y_1)}=0.14280.$$

The correlation between the range and y_1 is found as follows:

$$\rho(x_3-x_1,y_1)=\frac{E\left[(x_3-x_1)\cdot\dfrac{x'-x''}{x_3-x_1}\right]-E(x_3-x_1)\cdot E(y_1)}{\sigma(x_3-x_1)\cdot\sigma(y_1)}$$

$$=\frac{E(x'-x'')-E(x_3-x_1)\cdot E(y_1)}{\sigma(x_3-x_1)\cdot\sigma(y_1)},$$

$$\rho(x_3-x_1,y_1)=\frac{0.45352-(1.6926)(0.26209)}{(0.88837)(0.14280)}=0.0781,$$

on making use of the fact that

$$E(x'-x'')=E(2y_2)=\frac{6-3\sqrt{3}}{\sqrt{\pi}}=0.45352, \tag{10}$$

from a result obtained on page 263 and

$$E(x_3-x_1)=2E(x_3)=\frac{3}{\sqrt{\pi}}=1.69257$$

$$\sigma^2(x_3-x_1)=2\sigma^2(x_3)-2\sigma(x_1x_3)=2\left(1-\frac{9-3\sqrt{3}}{2\pi}\right)$$

$$=0.78920,$$

from the exact values given by Jones [9].

3.2. The Statistic $y_2=\frac{1}{2}(x'-x'')$

a. General Formula for Its Distribution

The development of section 3.1,a. can be used but will not be given here. It will be more fruitful, however, to pursue an alternative method adapted to the form of y_2 and y_3. This will readily yield the *joint* distribution of y_2, y_3 and thus simplify their study.

We first obtain the joint distribution of x', x'', x''', where it will be recalled that x' and x'' are the two *closest* observations, $x' \geq x''$, and x''' is the remaining one, the outlying value, either above or below the closest pair. Writing

$$x''=u, \quad x'=v, \quad x'''=w,$$

we have the transformation T given by

$$T: \begin{cases} x_1=u \\ x_2=v \\ x_3=w \end{cases} \text{when } x_2-x_1 \leq x_3-x_2, \text{ i.e. } u-2v+w \geq 0 \quad (R_1)$$

and

$$T: \begin{cases} x_1=w \\ x_2=u \\ x_3=v \end{cases} \text{when } x_2-x_1 \geq x_3-x_2, \text{ i.e. } v-2u+w \leq 0 \quad (R_2)$$

We know the joint distribution of x_1, x_2, x_3, namely

$$p(x_1,x_2,x_3)=3!f(x_1)f(x_2)f(x_3), a \leq x_1 \leq x_2 \leq x_3 \leq b, \quad (11)$$

and desire that of u, v, w resulting from the transformation T. Since the regions of definition become increasingly complex, we shall sacrifice some slight generality by taking $a=0$, $b=1$, and reworking the results whenever necessary. This will not be difficult once the general line of procedure has been indicated.

Since the function in (11) is symmetric, the density function for u, v, w remains of the same form. The only difficulty is determining the region over which it is different from zero. By somewhat tedious manipulations, this region may be shown to consist of the portions: [12]

$$R': \begin{cases} 2v-u \leq w \leq 1, & u \leq v \leq \frac{1}{2}(u+1), & 0 \leq u \leq 1 \quad (R_1') \\ 0 \leq w \leq 2u-v, & \frac{1}{2}v \leq u \leq v, & 0 \leq v \leq 1, \quad (R_2') \end{cases}$$

so that the *pdf* for (u, v, w) is

$$g(u,v,w)=6f(u)f(v)f(w) \text{ in } R' \quad (12)$$
$$=0 \text{ elsewhere.}$$

The joint distribution of $u(=x'')$ and $v(=x')$ may then be obtained by integration:

$$f_1(u,v)=\begin{cases} 6f(u)f(v)\int_{2v-u}^{1} f(w)dw, & u \leq v \leq \frac{1}{2}(u+1), 0 \leq u \leq 1 \\ 6f(u)f(v)\int_{0}^{2u-v} f(w)dw, & \frac{1}{2}v \leq u \leq v, 0 \leq v \leq 1 \end{cases} \quad (13)$$

It should be remembered that this formula holds only if the initial distribution $f(x)$ is non-zero in the range 0 to 1. For more general ranges a to b, the results would be rather complicated.

The joint distribution of y_2 and y_3 may be obtained from that of u, v in (13) by the transformation U

$$U: \quad y_2=\frac{1}{2}(x'-x'')=\frac{1}{2}(v-u)$$
$$y_3=\frac{1}{2}(x'+x'')=\frac{1}{2}(v+u)$$

with Jacobian $-\frac{1}{2}$, and inverse,

$$U^{-1}: u=-y_2+y_3, \qquad v=y_2+y_3. \quad (14)$$

Substitution into (13) presents no problem. The two partial regions in (13) are transformed as follows:

$$\text{first sub-region into} \begin{cases} 0 \leq y_2 \leq \frac{1-y_3}{3}, & \frac{1}{4} \leq y_3 \leq 1 \\ 0 \leq y_2 \leq y_3, & 0 \leq y_3 \leq \frac{1}{4}, \end{cases} \quad (15)$$

$$\text{second sub-region into} \begin{cases} 0 \leq y_2 \leq \frac{1}{3}y_3, & 0 \leq y_3 \leq \frac{3}{4} \\ 0 \leq y_2 \leq 1-y_3, & \frac{3}{4} \leq y_3 \leq 1. \end{cases}$$

Discussion of moments and other properties is most easily carried out in connection with the specific populations discussed below.

To find the distribution of x''' $(=w)$ the region R' must first be expressed by changing the order of the variables u, v, w so that the condition involving w is written last, permitting u and v to be integrated out. The procedure is the same as determining new limits when transforming variables or changing the order of integration.

The result of transforming the region and integrating out u and v is

$$p(w)=\left[\int_0^{\frac{1}{2}w}\int_0^{v} + \int_{\frac{1}{2}w}^{w}\int_{2v-w}^{v}\right] g \, du \, dw + \left[\int_{\frac{1}{2}(w+1)}^{1}\int_{u}^{1} + \int_{w}^{\frac{1}{2}(w+1)}\int_{u}^{2u-w}\right] g \, dv \, du, \quad (16)$$
$$0 \leq w \leq 1,$$

where $g=g(u,v,w)$ is given by (12).

[12] Note that the variables u, v appear in reverse order in (R_1') compared with (R_2'). If the order is kept the same, it will be found that (R_2') will need to be further broken into 2 parts, (R_{21}') and (R_{22}'). The present order will therefore be retained in the interest of simplicity. This need occasion no difficulty if care is used when integrating.

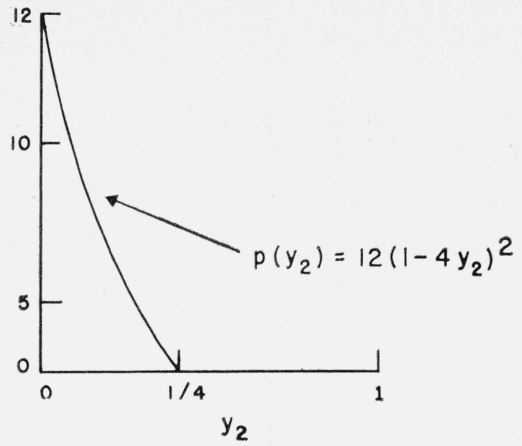


FIGURE 2. *Frequency function for* y_2.

b. Rectangular Universe

For a rectangular (square) universe

$$f(x)=1, \qquad 0\leq x\leq 1,$$

(13) becomes, with the aid of (14), (15),

$$p(y_2,y_3)=\begin{cases} 12(1-3y_2-y_3),\ y_2\leq y_3\leq 1-3y_2,\ 0\leq y_2\leq \frac{1}{4} \\ 12(-3y_2+y_3),\ 3y_2\leq y_3\leq 1-y_2,\ 0\leq y_2\leq \frac{1}{4} \end{cases} \tag{17}$$

so that the *pdf* of y_2 is

$$\begin{aligned} p(y_2)&=12\int_{y_2}^{1-3y_2}(1-3y_2-y_3)dy_3+ \\ &\quad 12\int_{3y_2}^{1-y_2}(y_3-2y_2)dy_3, \text{ if } 0\leq y_2\leq\frac{1}{4}, \\ &=12(1-4y_2)^2, \text{ if } 0\leq y\leq\frac{1}{4}. \end{aligned}$$

Its graph is sketched in fig. 2. It is seen that small values of the difference $(x'-x'')$ appear to be overwhelmingly frequent in samples of three from a rectangular population, thus giving a possible intuitive explanation of the fact that the dispersion is much less than in the case of true duplicates.

Moments of this distribution are

$$\begin{aligned} &E(y_2)=\frac{1}{16},\ E(y_2^2)=\frac{1}{160},\ \ldots, \\ &E(y_2^k)=6\cdot 4^{-k}(k+1)(k+2)(k+3), \\ &\sigma(y_2)=\frac{1}{16}\sqrt{\frac{3}{5}}=0.04841. \end{aligned}$$

It should be recalled that these moments apply only to sampling from the rectangular (square) population, in the form $f(x)=1$, $0\leqq x\leqq 1$; and 0 elsewhere. For the case of a symmetrical rectangular population with unit standard deviation, see sec. 3.2, d, (4).

c. Normal Universe

Since the limits are no longer 0 to 1, the distribution of y_2 has to be worked out anew.

For the sample of size three, the two functional forms of $y_2=\frac{1}{2}(x'-x'')$ are

$$y_2=\begin{cases} \frac{x_2-x_1}{2}, \text{ when } x_2-x_1\leq x_3-x_2 \\ \frac{x_3-x_2}{2}, \text{ when } x_2-x_1\geq x_3-x_2. \end{cases} \tag{18}$$

This becomes, putting $x_2-x_1=s_1$, $x_3-x_2=s_2$,

$$y_2=\begin{cases} \frac{1}{2}s_1, \text{ when } s_1\leq s_2 \\ \frac{1}{2}s_2, \text{ when } s_1\geq s_2. \end{cases}$$

The desired distribution will then be obtained from the joint *df* of s_1 and s_2 by integrating out the above conditions (18) separately and replacing the "free" s_i by $2y_2$.

The joint *df* of s_1 and s_2 is found by the usual method of transformation from the joint *df* of the basic ordered variables x_1, x_2, and x_3 as follows: The transformation

$$\begin{aligned} x_2-x_1&=s_1 \\ x_3-x_2&=s_2 \\ x_1+x_2+x_3&=s_3 \end{aligned}$$

carries the joint *df*

$$f(x_1,x_2,x_3,)\,dx_1dx_2dx_3=\frac{3!}{(2\pi)^{3/2}}e^{-\frac{1}{2}(x_1^2+x_2^2+x_3^2)}\,dx_1dx_2dx_3.$$

$$-\infty<x_1\leq x_2\leq x_3<\infty$$

into

$$g(s_1,s_2,s_3)\,ds_1ds_2ds_3=\frac{1}{\pi\sqrt{2\pi}}e^{-\frac{1}{6}s_3^2}\cdot e^{-\frac{1}{3}(s_1^2+s_1s_2+s_2^2)}ds_1ds_2ds_3, 0\leq s_1<\infty,$$

$$0\leq s_2<\infty, -\infty<s_3<\infty$$

Integrating out s_3 from $-\infty$ to ∞ gives

$$h(s_1,s_2)\,ds_1ds_2=\frac{\sqrt{3}}{\pi}e^{-\frac{1}{3}(s_1^2+s_1s_2+s_2^2)}ds_1ds_2,\ 0\leq s_1<\infty,$$

$$0\leq s_2<\infty.$$

We can then obtain the distribution of y_2 as

$$f(y_2)\,dy_2=\left[\int_{s_2>s_1} h(s_1, s_2)\,ds_2\cdot ds_1\right]_{s_1=2y_2}+$$
$$\left[\int_{s_1\geq s_2} h(s_1, s_2)\,ds_1\cdot ds_2\right]_{s_2=2y_2}$$
$$=\frac{\sqrt{3}}{\pi}\left\{\int_{2y_2}^{\infty} e^{-\frac{1}{3}\left(s_2^2+2s_2y_2+4y_2^2\right)}ds_2\cdot 2y_2dy_2+\right.$$
$$\left.\int_{2y_2}^{\infty} e^{-\frac{1}{3}\left(s_1^2+2s_1y_2+4y_2^2\right)}\,ds_1\cdot 2y_2\,dy_2\right\}$$
$$=\frac{4\sqrt{3}}{\pi}\,e^{-y_2^2}\int_{3y_2}^{\infty} e^{-\frac{1}{3}t^2}\,dt.\,dy_2,\ 0\leq y_2<\infty,$$

on using the transformations $t=\frac{1}{3}(s_2+y_2)$, $t=\frac{1}{3}(s_1+y_2)$ in the first and second terms respectively, and combining. This change of variable in the infinite integral is legitimate, for both old and new integrals converge, and the transformation $s=3t-y_2$ from s to t has a continuous derivative (unity) which does not vanish in the range of integration.

The first two moments of y_2 involve integrals of the following types:

$$\int_0^{\infty}\int_{ky}^{\infty} ye^{-Q}\,dx\,dy=\frac{\sqrt{\pi}}{\Delta}\left(\sqrt{a}-\frac{ak'}{\sqrt{q}}\right)$$

$$\int_0^{\infty}\int_{ky}^{\infty} y^2e^{-Q}\,dx\,dy=\frac{2a}{\Delta^{3/2}}\arctan\left(\frac{\sqrt{\Delta}}{2ak'}\right)-\frac{ak'}{\sqrt{q}},$$

in which

$$Q=ax^2+bxy+cy^2,\ \Delta=4ac-b^2>0,\qquad a,\ c>0,$$
$$k'=k+\frac{b}{2a},\qquad q=ak^2+bk+c.$$

These values give

$$E(y_2)=\frac{6-3\sqrt{3}}{2\sqrt{\pi}}=0.22676$$

$$V(y_2)=\frac{1}{2}-\frac{63-33\sqrt{3}}{4\pi}=0.03508$$

$$\sigma(y_2)=0.18730.$$

d. Comparison of y_2 with other measures of two observations

(1) "True duplicates" (sample size $n=2$).

For samples of 2, the closest pair (x'', x') is simply the entire sample:

$$x'=x_2,\ x''=x_1,$$

and

$$y_2=\frac{x'-x''}{2}=\frac{x_2-x_1}{2}=\frac{1}{2}R_2,$$

where R_n will be used to denote the range of a sample of n. In table 1b, y_2 (for samples of 2) is denoted by p (Part A, cols. 3, 6).

Since a main objective is to make comparisons for samples from rectangular and from normal populations, it is first necessary to put them on a comparable basis.[13] The normal population studied is symmetrical and has standard deviation unity. The rectangular population with these same characteristics of location and scale is

$$g(x)=\frac{1}{\sqrt{12}},\qquad -\sqrt{3}\leq x\leq\sqrt{3},$$
$$=0 \text{ otherwise},$$

since the standard deviation of the rectangular (square) population previously considered is $\sqrt{12}$. The quantities needed in the comparisons below involving the rectangular distribution will be most conveniently obtained by computing them for the simple case of a square distribution and then multiplying by the scale magnifying factor $\sqrt{12}$. Evidently the statistic y_2 will not be affected by the shift in location of the population.

The results are (here $2y_2$ is simply the range, x_2-x_1):

Rectangular universe $g(x)=1/\sqrt{12}, -\sqrt{3}\leq x\leq\sqrt{3}$; and 0 elsewhere.

$$E(y_2)=E\left(\frac{x_2-x_1}{2}\right)=E\left(\frac{1}{2}R_2\right)=\frac{1}{2}\cdot\frac{1}{3}\cdot\sqrt{12}=0.5774$$

$$\sigma(y_2)=\sigma\left(\frac{x_2-x_1}{2}\right)=\sqrt{\frac{1}{72}}\cdot 2\sqrt{3}=0.4082.$$

(From the distribution of the range $p(R_n)=n(n-1)R_n^{n-2}(1-R_n)$ for $n=2$, combined with a transformation which multiplies the scale of the variable by $\sqrt{12}$.)

Normal universe $f(x)=(1/\sqrt{2\pi})e^{-x^2/2},\ -\infty<x<\infty$

$$E(y_2)=E\left(\frac{x_2-x_1}{2}\right)=1/\sqrt{\pi}=0.5642$$

$$\sigma(y_2)=\sigma\left(\frac{x_2-x_1}{2}\right)=\left(\frac{1}{2}-\frac{1}{\pi}\right)^{\frac{1}{2}}=0.4263$$

(From Jones [9].)

(2) Lowest (or highest)[14] pair out of three ($n=3$).

For samples of three, $x_1\leq x_2\leq x_3$, we have the following results for $s=(x_2-x_1)/2$:

Rectangular universe $g(x)=1/\sqrt{12},\ -\sqrt{3}\leq x\leq\sqrt{3}$; and 0 elsewhere

[13] This consideration did not arise when studying the statistic y_1, because, being a ratio of lengths, it is unaffected by changes in scale of the parent population.

[14] Since the parent distribution is in each case symmetrical, the results for the lowest and highest pair are identical.

$$E\left(\frac{x_2-x_1}{2}\right)=\frac{1}{8}\cdot\sqrt{12}=0.4330$$

$$\sigma\left(\frac{x_2-x_1}{2}\right)=\sqrt{\frac{3\times 12}{320}}=0.3354$$

(From Wilks [6].)

Normal universe $f(x_1)=(1/\sqrt{2\pi})e^{-x^2/2}, -\infty<x<\infty$

$$E\left(\frac{x_2-x_1}{2}\right)=\frac{3}{4\sqrt{\pi}}=0.4231$$

$$\sigma\left(\frac{x_2-x_1}{2}\right)=\left(\frac{1}{2}-\frac{9+6\sqrt{3}}{16\pi}\right)^{\frac{1}{2}}=0.3379$$

(From Jones [9].)

(3) Half-range, $\frac{1}{2}(x_3-x_1)(n=3)$.

The analogous quantities which describe the spread in the set of 3 are:

Rectangular universe $g(x)=1/\sqrt{12}, -\sqrt{3}\leq x\leq\sqrt{3}$; and 0 elsewhere.

$$E\left(\frac{x_3-x_1}{4}\right)=E\left(\frac{1}{4}R_3\right)=\frac{1}{4}\cdot\frac{1}{2}\cdot\sqrt{12}=0.4330$$

$$\sigma\left(\frac{x_3-x_1}{2}\right)=\sqrt{(1/80)\cdot 12}=0.3873$$

(From the distribution of the range $p(R_n)$ for $n=3$.)

The reason for using one-fourth rather than one-half the range runs somewhat as follows. The distance (x_2-x_1) between two adjacent values in a sample of 3 can take values from zero all the way up to the range of all three. Thus, in a rough average sense, this distance represents some fraction of the range, and it happens that in the cases we have considered, this fraction is remarkably closely given by one-half, so that half this distance, namely $s=\frac{1}{2}(x_2-x_1)$, is, in the same sense, given by *one-fourth* the range.

Normal universe $f(x)=(1/\sqrt{2\pi})e^{-x^2/2}, -\infty<x<\infty$.

$$E\left(\frac{x_3-x_1}{4}\right)=\frac{3}{4\sqrt{\pi}}=0.4231$$

$$\sigma\left(\frac{x_3-x_1}{2}\right)=0.4442$$

(From Jones [9].)

(4) Closest pair in samples of three.

For comparison, moments of $2y_2=x'-x''=y_2'$ for samples of three are presented here based on the moments of y_2 found above (secs. 3.2, b, and 3.2, c.) and also adjusted, in the case of the rectangular universe, for moving the mean of the distribution to the origin and increasing the scale by the factor $\sqrt{12}$.

Rectangular universe $g(x)=1/\sqrt{12}, -\sqrt{3}\leq x\leq\sqrt{3}$; and 0 elsewhere.

$$E(y_2')=2E(y_2)=2\cdot\frac{1}{16}\sqrt{12}=0.4330$$

$$\sigma(y_2')=2\sigma(y_2)=2\cdot\frac{1}{16}\sqrt{\frac{3}{5}\cdot 12}=0.3354$$

Normal universe $f(x)=(1/\sqrt{2\pi})e^{-x^2/2}$, $-\infty<x<\infty$.

$$E(y_2')=2E(y_2)=2(0.226761)=0.4535$$

$$\sigma(y_2')=2\sigma(y_2)=2(0.18730)=0.3746$$

The reason for using twice y_2, rather than y_2, for comparison with the previous values is analogous to that given in section 3.2, d, (3) for using one-fourth rather than one-half the range. The restriction to the *closest* pair means that the distance $x'-x''$ cannot vary to the same extent as x_2-x_1, for its size is limited at most to *half* the range, while x_2-x_1 can take values up to the range of the sample. Thus it is to be expected that $x'-x''$, that is, $2y_2$, is the quantity comparable to $\frac{x_2-x_1}{2}$, which in turn, by the argument in section 3, 3.2, d., (3), is comparable to $\frac{x_3-x_1}{4}$. It turns out that these relationships are *exactly* true in the case of the parent (adjusted) rectangular distribution, and remarkably close in the case of the parent unit normal.

3.3. The Statistic y_3

As for y_2, the distribution of

$$y_3=\frac{x'+x''}{2}$$

in the general situation involves a complicated argument, not only because of the complexity of the distribution function, but because of the involved character of the region over which the integration must be performed. Therefore it is not considered profitable to discuss the properties of y_3 from a general viewpoint, but its properties will be illustrated for individual universes, to show how they may be derived in any given case.

a. Rectangular Universe

We cannot use the joint *df* $p(y_2, y_3)$ in the form (17), because y_2 cannot be integrated out of the region as written. It is therefore necessary to reverse the order of the variables in the expression for the region. The result is

$$p(y_3)=\begin{cases}4y_3(3-7y_3), & 0\leq y_3\leq \frac{1}{4}\\ 2(1-2y_3+2y_3^2), & \frac{1}{4}\leq y\leq \frac{3}{4}\\ 4(1-y_3)(7y_3-4), & \frac{3}{4}\leq y_3\leq 1,\end{cases}$$

whose graph is sketched in figure 3.

For the moments we have

$$E(y_3)=\frac{1}{2},\qquad \sigma(y_3)=\frac{1}{4}\sqrt{\frac{11}{10}}=0.2622.$$

As in section 3.2, b, the rectangular distribution to which these apply is the square form $f(x)=1$, $0\leq x\leq 1$, and 0 elsewhere. For the form with standard deviation unity, see section 3.2, c, (4).

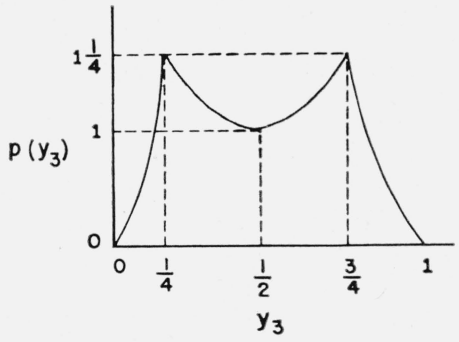


FIGURE 3. *Frequency function of y_3 for a rectangular universe.*

b. Normal Universe

The statistic y_3 takes the functional forms

$$y_3=\frac{x'+x''}{2}=\begin{cases}\frac{x_1+x_2}{2}, \text{ when } x_2-x_1\leq x_3-x_2 (i.e.\ x_3\geq 2x_2-x_1)\\ \frac{x_2+x_3}{2}, \text{ when } x_2-x_1\geq x_3-x_2 (i.e.\ x_1\leq 2x_2-x_3).\end{cases} \tag{19}$$

As a first step in obtaining the distribution of y_3, the joint *df* of x' and x'' is determined from that of x_1, x_2, x_3.

Writing

$$f(x_1,x_2,x_3)dx_1dx_2dx_3=6f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3,$$
$$-\infty<x_1\leq x_2\leq x_3<\infty$$

where, on the right-hand side, $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, we have, on integrating over the above conditions,

$$\begin{aligned}g(x',x'')dx'dx''&=6\left[\int_{x_3>2x_2-x_1}f(x_1)f(x_2)f(x_3)dx_3\cdot dx_1dx_2\right]_{\substack{x_1=x''\\x_2=x'}}\\&\quad+6\left[\int_{x_1\leq 2x_2-x_3}f(x_1)f(x_2)f(x_3)dx_1\cdot dx_2dx_3\right]_{\substack{x_2=x''\\x_3=x'}}\\&=6f(x')f(x'')[1-F(2x'-x'')+F(2x''-x')]\,dx'dx'',-\infty<x''\leq x'<\infty,\end{aligned} \tag{20}$$

where

$$F(x)=\int_{-\infty}^{x}f(t)dt.$$

The desired distribution is then derived by means of the transformation

$$y_3=\frac{1}{2}(x'+x'')$$

$$y_2=\frac{1}{2}(x'-x''),$$

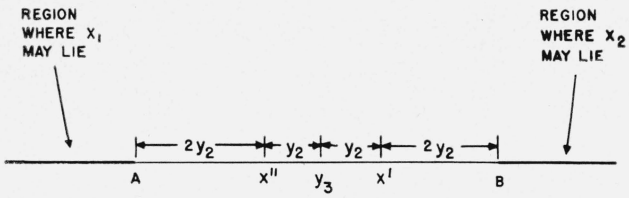
giving

$$f(y_3)dy_3=\frac{6}{\pi\sqrt{2\pi}}\left[\int_0^\infty\int_{y_3+3y_2}^\infty+\int_0^\infty\int_{-\infty}^{y_3-3y_2}\right]e^{-(\frac{1}{2}t^2+y_2^2+y_3^2)}dt\,dy_2\cdot dy_3,\qquad -\infty<y_3<\infty. \tag{21}$$

Alternative derivation. The author is indebted to Professor J. Wolfowitz of Cornell University for the following interesting method of deriving the above result.

The method is to take two of the three observations (which may be done in C_2^3 ways) and express the fact that they are the closest two by writing the condition that the third is at a greater distance from either one than the interval between the two selected.

This may be schematically shown as follows:



If x' and x'' are the closest pair, then

(i) Half the distance between them is $y_2=\dfrac{x'-x''}{2}$;

(ii) The abscissa of the mid-point between them is $y_3=\dfrac{x'+x''}{2}$;

(iii) The condition that x'' and x' are the closest two is equivalent to the condition: x_1 lies to the left of A *or* x_3 lies to the right of B.

Combining condition (iii) with the fact that, either from the diagram or by inverting the transformation in (i) and (ii), $x'=y_3+y_2$ and $x''=y_3-y_2$, gives for the joint df of y_3, y_2,

$$f(y_3,y_2)\,dy_3\,dy_2=$$

$$\frac{3}{\pi}e^{-\frac{1}{2}[(y_3+y_2)^2+(y_3-y_2)^2]}\left(\int_{-\infty}^{y_3-3y_2}\frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}}\,dt+\int_{y_3+3y_2}^{\infty}\frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}}\,dt\right)2dy_3\,dy_2,\ 0\le y_2<\infty,-\infty<y_3<\infty$$

The odd moments of y_3 vanish by symmetry.

The even moments of y_3 require the evaluation of integrals of the type

$$\phi_{2k}(a,p)=\int_{-\infty}^{\infty}\int_0^\infty\int_{-\infty}^{py+z}z^{2k}e^{-(ax^2+by^2+cz^2)}\,dx\,dy\,dz.$$

This may be accomplished by first putting $k=0$, differentiating [15] with respect to p, and obtaining an integral of the form

$$\frac{\partial}{\partial p}\phi_0(a,p)=\int_{-\infty}^{\infty}\int_0^\infty y\,e^{-Q}\,dy\,dz=\frac{2\sqrt{\pi m}}{\Delta},$$

where $Q=kx^2+lxy+my^2$, $\Delta=4km-l^2>0$, k, $m>0$. Integrating back yields the value of $\phi_0(a,p)$. Next, differentiating this value with respect to c gives the even moments of y_3.

We thus obtain the results

$$E(y_3^k)=0,\ k \text{ odd}$$

$$V(y_3)=E(y_3^2)=\frac{1}{2}+\frac{\sqrt{3}}{4\pi}=0.6378\ [16]$$

$$\sigma(y_3)=0.7986$$

c. Comparisons with Other Measures of Two Observations

Since for samples of two, y_3 is merely the midrange, $m=\frac{1}{2}(x_1+x_2)$, we have the following results:[17]

(1) "True duplicates" (sample size $n=2$)

Rectangular universe: $g(x)=1/\sqrt{12}$, $-\sqrt{3}\le x\le\sqrt{3}$; and 0 elsewhere.

$$E(y_3)=0$$

$$\sigma(y_3)=\frac{1}{2}$$

Normal universe: $f(x)=(1/\sqrt{2\pi})e^{-x^2/2}$, $-\infty<x<\infty$.

$$E(y_3)=0$$

$$\sigma(y_3)=\frac{1}{2}\sqrt{2}=0.7071$$

[15] This and the other steps of the analysis in the case of multiple integrals may be shown to be valid by methods analogous to the usual ones for simple integrals.

[16] Acknowledgment is due G. R. Seth, who first discovered and communicated this value to the author after deriving it by a different method, which the author has found useful at other points of this paper. (See also footnote 3.)

[17] As in section 3.2, d, the reference for the case of the rectangular universe is Wilks [6]; for the normal universe, Jones [9]. The values for the rectangular universe are computed by finding the moments of y_3 for the square universe and then adjusting by the location and scale factors described in section 3.2, d, (1).

(2) Lowest[18] pair out of three ($n=3$)

Rectangular universe:[19]

$$E\left(\frac{x_1+x_2}{2}\right)=-\sqrt{3}/4=-0.4330$$

$$\sigma\left(\frac{x_1+x_2}{2}\right)=\sqrt{\frac{11\cdot 12}{320}}=0.6423$$

Normal universe:

$$E\left(\frac{x_1+x_2}{2}\right)=\frac{1}{2}E(x_1)=\frac{1}{2}\left(-\frac{3}{2\sqrt{\pi}}\right)=-0.4231$$

$$\sigma\left(\frac{x_1+x_2}{2}\right)=\left(\frac{1}{2}-\frac{9-2\sqrt{3}}{16\pi}\right)^{\frac{1}{2}}=0.6244$$

(3) Midrange of all three measurements, $\frac{1}{2}(x_1+x_3)$, ($n=3$)

Rectangular universe:[19]

$$E\left(\frac{x_1+x_3}{2}\right)=0$$

$$\sigma\left(\frac{x_1+x_3}{2}\right)=\sqrt{\frac{3}{10}}=0.5477$$

Normal universe:

$$E\left(\frac{x_1+x_3}{2}\right)=0$$

$$\sigma\left(\frac{x_1+x_3}{2}\right)=\left(\frac{1}{2}-\frac{\sqrt{3}}{4\pi}\right)^{\frac{1}{2}}=0.6018$$

(4) Closest pair in samples of three

For comparison, moments of $y_3=(x'+x'')/2$, the average of the closest pair out of three, are presented here, based on the moments of y_3 found above (secs. 3.3, a and 3.3, b.) and also adjusted, in the case of the rectangular distribution, to the location and scale factors used several times previously.

Rectangular universe:[19]

$$E(y_3)=0$$

$$\sigma(y_3)=\sqrt{\frac{33}{40}}=0.9083$$

Normal universe:

$$E(y_3)=0$$

$$\sigma(y_3)=\left(\frac{1}{2}+\frac{\sqrt{3}}{4\pi}\right)^{\frac{1}{2}}=0.7986$$

[18] Analogous results for the highest pair are obtainable from symmetry considerations.

[19] Adjusted as already mentioned in previous sections: $g(x)=1/\sqrt{12}, -\sqrt{3}\leq x\leq\sqrt{3}$; and 0 elsewhere.

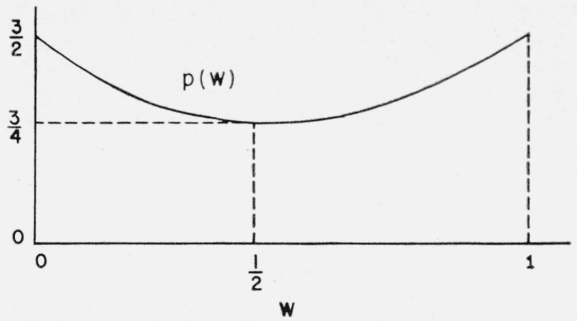


FIGURE 4. *Distribution of the outlying value in a sample of three from the rectangular distribution*

3.4. The Extreme Value x'''

For the rectangular distribution we obtained the joint density function (12) in section 3.2, a, above, of x', x'', x'''. This consisted of just the product of the individual density functions (unity for the rectangular universe), but defined over a complicated appearing region. Following the principles elucidated in that section, we obtain the *df* of x''' by first expressing the region suitably, then integrating out x' and x''. The result is given by equation (16) which, for the rectangular distribution, becomes, with $x'''=w$,

$$p(w)=3\left(w^2-w+\frac{1}{2}\right), \qquad 0\leq w\leq 1, \qquad (22)$$

which is the parabola sketched in figure 4.

Although considerable attention has been devoted to the anomalous values or outliers x_1 or x_3 (x_n for samples of n) separately, these have not, so far as is known to the author, been united into a single statistic of the type x'''. *Thus, (22) actually exhibits a distribution of an* OUTLIER *as distinct from a ("one-end") extreme value* x_1 *(or* x_3*).*
The moments of w are

$$E(w)=\frac{1}{2}, \qquad \sigma(w)=\frac{1}{10}\sqrt{10}=0.3162$$

The joint distribution of w and y_3, given without proof, is as follows:

$$f(y_3,w)=\begin{cases} 4(w-y_3), & \frac{1}{4}w\leq y_3\leq w, \qquad 0\leq w\leq 1 \\ 12y_3, & 0\leq y_3\leq\frac{1}{4}w, \qquad 0\leq w\leq 1 \\ 4(y_3-w), & w\leq y_3\leq\frac{1}{4}(w+3), \qquad 0\leq w\leq 1 \\ 12(1-y_3), & \frac{1}{4}(w+3)\leq y_3\leq 1, \qquad 0\leq w\leq 1 \end{cases}$$

This distribution may be used to obtain the correlation between x''' and $y_3=\frac{1}{2}(x'+x'')$, which turns out to be $-.37689$, or slightly under $-3/8$. This seems

to imply a slight tendency for a pair of close, small values in a sample of 3 from a rectangular population to be associated with a relatively large outlying value, and conversely, for close high values.

The distribution of the extreme value x''' would also be of interest for the normal and other distributions. Although the analytical methods necessary for handling the integrals encountered could be developed and extended on the basis of the procedures thus far given, this would not appear to be warranted for the purposes of this paper.

4. Conclusion

This paper has developed methods for deriving the exact distributions and related properties of certain statistics not heretofore considered which throw light on some aspects of the behavior of very small samples encountered in experimental laboratory work. These statistics, designated y_1, y_2, y_3 depend not solely on the order of the observations but also take their relative *closeness* into account. The aim was to provide only the mathematical theory, for samples of three, and present only the more interesting results and comparisons (summarized in table 1) and not attempt to use the results as a basis for setting up criteria for the rejection of observations.

The results have some bearing on the old question of the rejection of outlying observations. They show that at least for the normal, rectangular, and right triangular universes, for a sample as small as three a rejection criterion based on the relative sizes of the two gaps formed by the three measurements is hardly a satisfactory one, for high ratios between these gaps occur with surprising frequency, as indicated in table 2, even when all three observations come from the same universe.

5. References

[1] F. E. Grubbs, Sample criteria for testing outlying observations, Ann. of Math. Stat. **21,** 27 to 58 (March 1950).

[2] P. R. Rider, Criteria for rejection of observations, Washington University Studies—New Series, Science and Technology, No. 8, St. Louis, (1933).

[3] The fallacy of the best two out of three, NBS Technical News Bulletin **33,** 77 (July 1949).

[4] S. S. Wilks, Order statistics, Bull. Am. Math. Soc. **54,** 6 to 50 (1948).

[5] J. W. Tukey. Comparing individual means in the analysis of variance, Biometrics **5,** 99 to 114 (June 1949).

[6] S. S. Wilks, Mathematical Statistics (Princeton University Press, 90 to 93, 1943).

[7] W. J. Dixon, Ratios involving extreme values, Ann. Math. Stat. **21,** 488 to 506 (Dec. 1950).

[8] W. J. Dixon, Analysis of extreme values, Ann. Math. Stat. **22,** 68 to 78 (Mar. 1951).

[9] H. L. Jones, Exact lower moments of order statistics in small samples from a normal distribution, Ann. Math. Stat. **19,** 270 to 273 (1948).

[10] G. R. Seth, On the distribution of the two closest among a set of three observations, Ann. Math. Stat. **21,** 298 to 301 (1950).

[11] F. M. Henry, The loss of precision from discarding discrepant data, Research Quarterly of the Am. Assoc. for Health, Phys. Ed., and Recreation **21,** No. 2, 145 to 152 (May 1950).

WASHINGTON, September 29, 1950.