ON THE USE OF AN AUXILIARY VARIABLE IN THE TRANSFORMATION OF DISCRETE DATA

by

ROBERT J. TAYLOR

Thesis submitted to the Graduate Faculty of the

Virginia Polytechnic Institute

in candidacy for the degree of

MASTER OF SCIENCE

in

STATISTICS

APPROVED:

Director of Graduate Studies

Dean of Applied Science and Business Administration

APPROVED:

Head of Department

Major Professor

May, 1955

Blacksburg, Virginia

TABLE OF CONTENTS

Chapter		Page
I.	INTRODUCTION.....	6
II.	PREVIOUS SOLUTIONS.....	12
III.	TRANSFORMATIONS USING AN AUXILIARY VARIABLE.....	14
	3.1 The z Transformation.....	14
	3.2 The w Transformation.....	15
IV.	DEVELOPMENT OF THE z TRANSFORMATION....	16
	4.1 Parameters of z.....	16
	4.2 Distribution of the Auxiliary Variable.....	20
	4.21 Constant.....	20
	4.22 Uniform Continuous Distribution.....	21
	4.23 Uniform Discrete Distribution.....	22
	4.3 Cumulative Probability Distribution of z.....	25
V.	DEVELOPMENT OF THE w TRANSFORMATION....	27
	5.1 Explanation of the w Transformation.....	27
	5.2 Mean and Variance of w_p.....	28

Chapter		Page
VI.	CONCLUSION.....	31
VII.	ACKNOWLEDGEMENTS.....	34
VIII.	BIBLIOGRAPHY.....	35
IX.	TABLES AND CURVES.....	38
X.	VITA.....	45

LIST OF TABLES

Page

Table I. First Four Terms of the Variance Expansion Disregarding the Factor One-Fourth, for Some Distributions of x 39

Table II. Cumulative Distribution of x Where x Has Uniform Discrete Distribution With n = 10 41

LIST OF FIGURES

Page

Figure 1. Cumulative Distribution of z x Has Uniform Discrete Distribution (n = 10)..... 43

Figure 2. Frequency Distribution of z When μ = 2.5 and x Has the Uniform Continuous Distribution..... 44

I. INTRODUCTION

In many of the applications of analysis of variance difficulties in the analysis arise when the conditions of the experiment are not ideal. Consider first the purposes of analysis of variance:

1. To perform tests of significance.
2. To estimate certain treatment differences that are of interest.
3. To obtain some idea of the accuracy of the estimates.

To see how these can be best accomplished will require understanding the assumptions underlying the analysis of variance. With an understanding of the assumptions, one knows:

1. Which are the most necessary, and what are likely to be the consequences if these are not fulfilled.
2. What can be done to utilize, for purposes of analysis, data which in their original form are not amenable to an analysis of variance.
3. Which can be dispensed with for certain purposes.

Churchill Eisenhart (1947) lists four assumptions for Model I where the parameters of interest are population means:

Assumption 1 (Random Variables). Observations are of random variables that are distributed about true mean values that are fixed constants.

Assumption 2 (Additivity). The parameters are related to over-all means in an additive manner. For example in a two way classification

$$m_{ij} = m_{..} + (m_{i.} - m_{..}) + (m_{.j} - m_{..})$$

Assumption 3 (Equal Variances and Zero Correlation). The random variables are homoscedastic and mutually uncorrelated; that is, they have a common variance and all covariances among them are zero.

Assumption 4 (Normality). The observations are jointly distributed in a multivariate normal distribution.

From a functional aspect Cochran lists for (a) treatment effects (b) environmental effects and (c) experimental errors, the following assumptions:

1. The treatment effects and the environmental effects must be additive.
2. The experimental errors must all be independent.

3. The experimental errors must have a common variance.
4. The experimental errors should be normally distributed.

According to Cochran (1947) the principal effect of the failure of the assumption of additivity is a loss of information. He goes on to state that this loss should be negligible when the treatment and replication effects do not exceed twenty per cent, as long as there is no very serious departure from additivity and experimental errors are not too low.

The effect of non-normality is not generally considered to be serious for moderate departures from normality, but any large departures in the region of the more outlying observations are likely to affect the validity of significance tests. In addition, there is likely to be a loss of efficiency in the estimation of treatment effects and a corresponding loss of power in the F-test and t-tests. Cochran states that in practice, the loss of efficiency due to non-normality is not often great. He also reports that a number of investigators (Bartlett, 1935; Hey, 1938; Pearson, 1931) working with non-normal populations that appear to cover the range of most cases encountered in practice, using a number of different methods give a consensus

that no serious error is introduced by non-normality in the significance levels of the F-test or the two tailed t-test.

When ordinary analysis of variance methods are used on data for which the true error variance changes from one observation to another there will be, as a rule, a loss of efficiency in the estimates of the treatment effects. Similarly there will be a loss of sensitivity in tests of significance. If the changes in the error variance are large these losses may be substantial. Usually a heterogeneous error variance is related to changes in the mean or to certain treatments or parts of the experiment. If this is the case, it affects the assumption of independence.

In situations occurring in practice then, one does not usually have to concern oneself with the consequences of non-additivity and non-normality. However, there are frequent cases where there is a heterogeneous error variance that causes loss of sensitivity in the tests of significance. To overcome this loss often one can stabilize the variance by a transformation. Keeping in mind the assumptions involved in analysis of variance, consider the

conditions for an ideal transformation. In this case:

1. The variance of the transformed variate should be unaffected by changes in the mean level.
2. The transformed variate should be normally distributed.
3. The transformed scale should be one for which an arithmetic average is an efficient estimate of the true mean level for any particular group of measurements.
4. The transformed scale should be one for which real effects are linear and additive.

This ideal, of course, is very rarely attained; however, there are a number of transformations (some for the Poisson distribution will be discussed in Part II) that are used for different kinds of data that are considered good enough for most practical purposes.

In much statistical data that consists of integers such as the number of bacterial colonies, weed counts in a given area, etc., conditions will be such that the variate is approximately Poisson distributed. In this latter distribution the variance is equal to the mean so that if one has a large treatment mean then associated with it is a large variance; conversely if one has a small treatment mean, then there will be a

small variance associated with it. Hence, the variances are not equal. In addition to the fact that the Poisson distribution is a discrete distribution, it is also non-normal as to shape. This obviously does not fulfill the assumptions for analysis of variance; therefore, some sort of transformation should be made so that the necessary assumptions can be realized.

II. PREVIOUS SOLUTIONS

This problem has been discussed often and several solutions have been advanced, not only in the case of the Poisson distribution but in the similar cases of the binomial and negative-binomial distributions.

In the case of the Poisson, Bartlett (1936) showed that the square root of y has an approximately constant variance of one-quarter if μ is large (y is the Poisson variable and μ is the Poisson mean) and, in addition, is more nearly normal than the Poisson variable itself. He also showed that $\sqrt{y + 1/2}$ is a better transformation because it has a more nearly constant variance even for smaller values of μ. Anscombe (1948) reports that by considering the transformation $\sqrt{y + c}$ the most nearly constant variance occurs when $c = 3/8$, a result Anscombe credits to A. H. L. Johnson. Fieller (1947) reports an example of the use of the simple square root transformation in bio-assay which he found quite satisfactory.

Freeman and Tukey (1949) suggest the transformation $\sqrt{x} + \sqrt{x + 1}$ and maintain that if a $\pm$ 6% variation in the variance is permissible their transformation is usable for expectations of unity or more. In contrast, they say that under the same conditions Anscombe's

transformation is usable only for expectations of 2.2 or more. In addition, Freeman and Tukey note the ease that their transformation may be obtained from the square root table.

Although at present the only practicable method of examining the properties of a proposed transformation is by empirical sampling methods, such as Freeman and Tukey adopted, it is of interest to obtain theoretical formulas which will be valid at least for large values of μ.

III. TRANSFORMATIONS USING AN AUXILIARY VARIABLE

3.1 The z Transformation

In this thesis two transformations will be discussed. The first one is an extension of Anscombe's formula in that an auxiliary variable is used instead of a constant. Thus:

$$z = \sqrt{y + x}$$

where: y = Poisson variable with mean μ

x = auxiliary variable

The auxiliary variable will be limited in range so that knowing z there will be one and only one value of y related to it. Since the Poisson is a discrete distribution with values at the positive integers the range of x can be taken to be $0 \leq x < 1$.

Three specific distributions of x will be considered:

1. Constant (this special case is equivalent to Anscombe's work).
2. Uniform continuous distribution.
3. A uniform discrete distribution.

3.2 The w Transformation

The second transformation to be considered is an extension of the first:

$$w_p = \left[(1-x)y^{\frac{p}{2}} + x(y+1)^{\frac{p}{2}}\right]^{\frac{1}{p}}$$

where: x and y are as defined above (Section 3.1)

Note that $w_2 = z$

IV. DEVELOPMENT OF THE z TRANSFORMATION

4.1 Parameters of z

For ease in expanding, the formula

$$z = \sqrt{y + x} \tag{1}$$

was written

$$z = \sqrt{\mu + \mu v + x} \tag{2}$$

where: μ = Poisson mean

$$v = \frac{y - \mu}{\mu}$$

The cumulants of v are therefore:

$$\begin{aligned} \kappa_r(v) &= \frac{\kappa_r(y - \mu)}{\mu^r} \\ &= 0 \qquad r = 1 \\ &= \mu^{-r+1} \qquad r > 1 \end{aligned}$$

The moments about zero of v were obtained by substituting the cumulants in the expression (given in Kendall, p. 62) for the moments about zero in terms of the cumulants.

A factor μ was taken out of the expression (2) for z. The rest was then arranged into two terms one of which was unity and the other could be assumed to

be less than unity, though no proof is presented here that this assumption is true in probability. Thus:

$$z = \mu^{\frac{1}{2}}\left[1 + \left(v + \frac{x}{\mu}\right)\right]^{\frac{1}{2}} \tag{3}$$

The expression in the brackets can then be expanded in a binomial series. The expected value of z was found by taking the expected value of the resulting series term by term since the auxiliary variate is assumed to be independent of the Poisson variate. The expected value was then written in terms of negative powers of μ. All terms of order $\frac{1}{\mu^5}$ and higher were assumed small enough to be dropped. Anscombe (1948) demonstrated the convergence of a similar series. The rth moment about zero of the auxiliary variate was expressed as m_r'. Thus:

$$\begin{aligned} E(z) = \mu^{\frac{1}{2}}\Big\{1 &+ \frac{1}{\mu}\left(\frac{4m_1' - 1}{8}\right) + \frac{1}{\mu^2}\left(\frac{-16m_2' + 24m_1' - 7}{128}\right) \\ &+ \frac{1}{\mu^3}\left(\frac{64m_3' - 240m_2' + 260m_1' - 75}{1024}\right) \\ &+ \frac{1}{\mu^4}\left(\frac{-1280m_4' + 8960m_3' - 21280m_2' + 19600m_1' - 5509}{32768}\right) \\ &+ \cdots\Big\} \end{aligned} \tag{4}$$

Since

$$z^2 = \mu + \mu v + x \tag{5}$$

$$E(z^2) = \mu + m_1' = \mu\left(1 + \frac{m_1'}{\mu}\right) \tag{6}$$

The expression for the variance was obtained by subtracting the square of the series for $E(z)$ from $E(z^2)$

$$Var(z) = E(z^2) - [E(z)]^2 = \kappa_2 \tag{7}$$

$$= \frac{1}{4}\Big\{1 - \frac{1}{\mu}\Big(\frac{-8m_2' + 8(m_1')^2 + 8m_1' - 3}{8}\Big)$$

$$-\frac{1}{\mu^2}\Big(\frac{16m_3' - 56m_2' - 16m_2'm_1' + 24(m_1')^2 + 52m_1' - 17}{32}\Big)$$

$$-\frac{1}{\mu^3}\Big(\frac{-160m_4' + 1088m_3' + 128m_3'm_1' + 32(m_2')^2 - 2512m_2'}{512}$$

$$+\frac{-576m_2'm_1' + 592(m_1')^2 + 2128m_1' + 645}{512}\Big) + \cdots\Big\} \tag{8}$$

$E(z^3)$ and $E(z^4)$ were also obtained by the same manner. The third cumulant of z was obtained by the formula

$$\kappa_3 = E(z^3) - 3E(z^2)E(z) + 2[E(z)]^3 \tag{9}$$

and was found to be

$$\kappa_3 = \mu^{\frac{3}{2}}\Big\{\frac{-1}{16\mu^2} + \frac{1}{\mu^3}\Big(\frac{16m_3' - 48m_2' - 48m_2'm_1' + 32(m_1')^3 + 48(m_1')^2 + 36m_1' - 17}{128}\Big)$$

$$+\frac{1}{\mu^4}\Big(\frac{-192m_4' + 1248m_3' + 384m_3'm_1' + 192(m_2')^2 - 2736m_2' - 384m_2'(m_1')^2}{2048}$$

$$+\frac{-1824m_2'm_1' + 576(m_1')^3 + 1536(m_1')^2 + 2088m_1' + 1829}{2048}\Big) + \cdots\Big\} \tag{10}$$

Then

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \kappa_3\kappa_2^{-\frac{3}{2}} = \kappa_3\Big(\frac{1}{4}\Big)^{-\frac{3}{2}}\Big[1 - \frac{1}{\mu}\Big(\frac{-8m_2' + 8(m_1')^2 + 8m_1' - 3}{8}\Big) + \cdots\Big]^{-\frac{3}{2}}$$

The minus three-halves power of the variance was obtained by dividing the series into two terms (first dividing by an appropriate constant) one of which was unity and the other numerically less than one in probability. Therefore

$$\gamma_1 = -\frac{1}{2\mu^{1/2}} + \frac{1}{\mu^{3/2}}\left(\frac{32m_3' - 72m_2' - 96m_2'm_1' + 64(m_1')^3 + 72(m_1')^2 + 48m_1' - 25}{32}\right) + \cdots \quad (11)$$

This formula in terms of moments about the mean of the auxiliary variate is

$$\gamma_1 = -\frac{1}{2\mu^{1/2}}\left\{1 - \frac{32m_3 - 72m_2 + 48m_1' - 25}{16\mu} + \cdots\right\} \quad (12)$$

Similarly

$$\kappa_4 = E(z^4) - 4E(z^3)E(z) - 3[E(z^2)]^2 + 12E(z^2)[E(z)]^2 - 6[E(z)]^4 \quad (13)$$

$$= \mu^2\left\{\frac{1}{16\mu^3} + \frac{1}{\mu^4}\left(\frac{32m_4' - 192m_3' - 128m_3'm_1' - 96(m_2')^2 + 384m_2'(m_1')^2}{512}\right.\right.$$
$$\left.\left. + \frac{+576m_2'm_1' - 192(m_1')^4 - 384(m_1')^3 - 384(m_1')^2 - 256m_1' + 142}{512}\right) + \cdots\right\} \quad (14)$$

Then

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \kappa_4\kappa_2^{-2} = \kappa_4\left(\frac{1}{4}\right)^{-2}\left[1 - \frac{1}{\mu}\left(\frac{-8m_2' + 8(m_1')^2 + 8m_1' - 3}{8}\right) + \cdots\right]^{-2}$$
$$= \frac{1}{\mu}\left\{1 + \frac{16m_4' - 96m_3' - 64m_3'm_1' - 48(m_2')^2 + 160m_2' + 192m_2'(m_1')^2 + 288m_2'm_1'}{16\mu}\right.$$
$$\left. + \frac{-96(m_1')^4 - 192(m_1')^3 - 160(m_1')^2 - 96m_1' + 59}{16\mu} + \cdots\right\} \quad (15)$$

And in moments about the means of the auxiliary variate

$$\gamma_2 = \frac{1}{\mu}\left\{1 + \frac{16 m_4 - 48(m_2)^2 - 96 m_3 + 160 m_2 - 96 m_1' + 59}{16\mu} + \cdots\right\} \qquad (16)$$

4.2 Distribution of the Auxiliary Variate

4.21 Constant

When x is a constant, say c, then the moments about zero are:

$$m_r' = c^r \qquad (1)$$

On substituting these values for the moments into the formulae for the parameters of z the following were obtained

$$E(z) = \left\{1 + \frac{1}{\mu}\left(\frac{4c-1}{8}\right) + \frac{1}{\mu^2}\left(\frac{-16c^2+24c-7}{128}\right) + \frac{1}{\mu^3}\left(\frac{64c^3-240c^2+260c-75}{1024}\right)\right.$$
$$\left. + \frac{1}{\mu^4}\left(\frac{-1280c^4+8960c^3-21280c^2+19600c-5509}{32768}\right) + \cdots\right\} \qquad (2)$$

$$Var(z) = \frac{1}{4}\left\{1 - \frac{1}{\mu}\left(\frac{8c-3}{8}\right) - \frac{1}{\mu^2}\left(\frac{-32c^2+52c-17}{32}\right)\right.$$
$$\left. - \frac{1}{\mu^3}\left(\frac{512c^3-1920c^2+2128c+645}{512} + \cdots\right\}\right. \qquad (3)$$

$$\gamma_1 = -\frac{1}{2\mu^{1/2}}\left\{1 + \frac{25-48c}{16\mu} + \cdots\right\} \qquad (4)$$

$$\gamma_2 = \frac{1}{\mu}\left\{1 + \frac{59-96c}{16\mu} + \cdots\right\} \qquad (5)$$

Note that this last value disagrees with Anscombe (1948) who gives

$$\gamma_2 = \frac{1}{\mu}\left\{1 + \frac{945 - 1536c}{256\mu} + \cdots\right\}$$

In spite of a diligent double check the disagreement still stands. Percentagewise this disagreement is nearly insignificant.

4.22 <u>Uniform Continuous Distribution</u>

If the distribution of x is the uniform continuous distribution over the unit interval $0 < x < 1$ then

$$m_r' = \int_0^1 x^r dx = \frac{1}{r+1} \qquad (1)$$

Substituting these values for the moments about zero into the general parameter values for z one obtains:

$$E(z) = \mu^{\frac{1}{2}}\left\{1 + \frac{0.125}{\mu} - \frac{0.00260}{\mu^2} - \frac{0.00878}{\mu^3} + \frac{0.025}{\mu^4} + \cdots\right\} \qquad (2)$$

$$Var(z) = \frac{1}{4}\left\{1 - \frac{0.375}{\mu} + \frac{0.0729}{\mu^2} + \frac{1.92}{\mu^3} + \cdots\right\} \qquad (3)$$

$$\gamma_1 = -\frac{1}{2\mu^{1/2}}\left\{1 + \frac{7}{16\mu} + \cdots\right\} \qquad (4)$$

$$\gamma_2 = \frac{1}{\mu}\left\{1 + \frac{93}{240\mu} + \cdots\right\} \qquad (5)$$

This case is considered by David (1955) whose theoretical method is well adapted to this special form of the auxiliary variate but is not easily extended to more general cases.

4.23 Uniform Discrete Distribution

When x is distributed according to the uniform discrete distribution in the range $0 \leq x < 1$ the moments about zero are

$$m_r' = \left(\frac{1}{n}\right)^{r+1} \sum_{k=0}^{n-1} k^r \qquad (1)$$

Durell (p. 39,40) and many other algebra textbooks give expressions for the sum of the first n consecutive integers, the sum of their squares, the sum of their cubes, and methods of finding the sum of their higher powers. Adjusting these values for the range under consideration one obtains for the moments:

$$m_1' = \frac{n-1}{2n} \qquad (2)$$

$$m_2' = \frac{(n-1)(2n-1)}{6n^2} \qquad (3)$$

$$m_3' = \frac{(n-1)^2}{4n^2} \qquad (4)$$

$$m_4' = \frac{(n-1)}{30n^4}\left(6n^3 - 9n^2 + n + 1\right) \qquad (5)$$

Using the expression for the variance 4.1(7), the expression for the third cumulant 4.1(9) which is the same as the third moment about the mean, and

$$m_4 = m_4' - 4\, m_3' m_1' + 12\, m_2' (m_1')^2 - 6 (m_1')^4 \tag{6}$$

the following values for the moments about the mean were obtained:

$$Var(x) = m_2 = \frac{n^2-1}{12n^2} \tag{7}$$

$$m_3 = 0 \tag{8}$$

$$m_4 = \frac{39n^4 - 120n^3 + 40n^2 - 60n + 11}{60n^4} \tag{9}$$

Substituting the above values into the expression for parameters of z one obtains:

$$E(z) = \mu^{\frac{1}{2}} \left\{ 1 + \frac{1}{\mu}\left(\frac{n-2}{8n}\right) + \frac{1}{\mu^2}\left(\frac{-n^2-12n-8}{384n^2}\right) + \frac{1}{\mu^3}\left(\frac{-9n^2-42n-24}{1024n^2}\right) \right.$$

$$\left. + \frac{1}{\mu^4}\left(\frac{2865n^4 - 19640n^3 - 10520n^2 + 128}{98304n^4}\right) + \cdots \right\} \tag{10}$$

$$Var(z) = \frac{1}{4} \left\{ 1 - \frac{1}{\mu}\left(\frac{n^2-12n+2}{24n^2}\right) - \frac{1}{\mu^2}\left(\frac{-7n^3-34n^2-14n+4}{96n^3}\right) \right.$$

$$\left. - \frac{1}{\mu^3}\left(\frac{10649n^4 - 3480n^3 - 1660n^2 + 336n + 56}{4608n^4}\right) + \cdots \right\} \tag{11}$$

$$\gamma_1 = -\frac{1}{2\mu^{1/2}}\left\{1 - \frac{-7n^2-24n+6}{16n^2\mu} + \cdots\right\} \qquad (12)$$

$$\gamma_2 = \frac{1}{\mu}\left\{1 + \frac{-310n^4+240n^3-30n^2-24n+44}{240n^4\mu} + \cdots\right\} \qquad (13)$$

The above formulae are for a general n. It should be worth-while checking whether some n is better than others. Consider the quadratic expression from the second term of the variance (11) and set it equal to zero.

$$n^2 - 12n + 2 = 0$$

One root of this equation is 11.83 which suggests that the nearest possible value for n which is 12 might come close to eliminating this term of the variance. The other root, which is approximately 0.18 is too small to be of use. Accordingly when one substitutes n = 10, 12, 100, 00 the following results are obtained for the variance:

$n = 10$

$$Var(z) = 0.25\left[1 + \frac{0.0075}{\mu} + \frac{0.1099}{\mu^2} - \frac{2.24}{\mu^3} + \cdots\right]$$

$n = 12$

$$Var(z) = 0.25\left[1 - \frac{0.000579}{\mu} + \frac{0.1035}{\mu^2} - \frac{2.24}{\mu^3} + \cdots\right]$$

$n = 100$

$$Var(z) = 0.25\left[1 - \frac{0.0367}{\mu} + \frac{0.0765}{\mu^2} - \frac{2.23}{\mu^3} + \cdots\right]$$

$n = \infty$

$$Var(z) = 0.25\left[1 - \frac{0.0416}{\mu} + \frac{0.0729}{\mu^2} - \frac{2.31}{\mu^3} + \cdots\right]$$

Notice that the value n = 12 does in fact come closest of these to eliminating the second term in the variance, suggesting that ideally the auxiliary variate should consist of a duo-decimal random digit. However, the second term does become quite small when n = 10 thus allowing the use of a single random decimal digit as the variate x in the original transformation 4.1(1). Similarly n = 100 could be realized by taking two random decimal digits. This is seen to be a close approximation to the use of the continuous auxiliary variate (n = ∞).

4.3 Cumulative Probability Distribution of z

Because of the one-to-one correspondence of a pair of values x and y with the resulting value of z the cumulative probability can be written as follows:

$$P(Z < z) = P(Y \leq y) + P(X < x)\left[P(Y \leq y+1) - P(Y \leq y)\right]$$

$$= P(Y \leq y) + P(X < x)\, P(Y = y+1)$$

Then the frequency distribution of z for the uniform continuous distribution of x is

$$f(z) = \frac{dP(Z<z)}{dz} = \frac{\partial P(Z<z)}{\partial x}\,\frac{\partial x}{\partial z}$$

$$= \frac{\dfrac{\partial P(Z<z)}{\partial x}}{\dfrac{\partial z}{\partial x}}$$

$$= 2z\,P(Y=y+1)$$

V. DEVELOPMENT OF THE W TRANSFORMATION

5.1 Explanation of the W Transformation

Consider the second transformation 3(2)

$$\omega_p = \left[(1-x)y^{\frac{p}{2}} + x(y+1)^{\frac{p}{2}}\right]^{\frac{1}{p}}$$

when y is the Poisson variate

x is an auxiliary variate

Note that for any value of p if x = 0 then the simple square root transformation suggested by Bartlett (1936) results

$$w_p(x=0) = \sqrt{y}$$

When p = 2

$$\omega_2 = \left[(1-x)y + x(y+1)\right]^{\frac{1}{2}}$$

$$= (y+x)^{\frac{1}{2}}$$

$$= z$$

And when p = 1

$$\omega_1 = \left[(1-x)y^{\frac{1}{2}} + x(y+1)^{\frac{1}{2}}\right]$$

which can be considered an extension of the Freeman-Tukey transformation (Freeman-Tukey, 1949). If x = 1/2 in this expression then exactly one-half of the Freeman-Tukey expression results.

Taking the limit as p approaches infinity

$$\lim_{p \to \infty} (w_p) = (y+1)^{\frac{1}{2}} \quad \text{if} \quad x \neq 0$$

If p approaches minus infinity

$$\lim_{p \to -\infty} (w_p) = y^{\frac{1}{2}} \quad \text{for } 0 \leq x < 1 \text{ and for all } y \geq 0$$

$$w_p = (y+1)^{\frac{1}{2}} \quad \text{when } x = 1$$

If p approaches zero then

$$\lim_{p \to 0} (w_p) = y^{\frac{1}{2}} \left(1 + \frac{1}{y}\right)^{\frac{x}{2}}$$

5.2 Mean and Variance of W_p

As was done with the z transformation, the y was written as $\mu + \mu v$, then

$$w_p = \left[(1-x)(\mu + \mu v)^{\frac{p}{2}} + x(1 + \mu + \mu v)^{\frac{p}{2}}\right]^{\frac{1}{p}} \qquad (1)$$

$$= \mu^{\frac{1}{2}} \left[(1-x)(1+v)^{\frac{p}{2}} + x\left(1 + \frac{1}{\mu} + v\right)^{\frac{p}{2}}\right]^{\frac{1}{p}} \qquad (2)$$

This latter arrangement was made for ease in expansion in a binomial series. The terms $(1+v)^{\frac{p}{2}}$ and $\left(1 + \frac{1}{\mu} + v\right)^{\frac{p}{2}}$ were expanded separately. The resulting series of the first was multiplied by 1-x and the series resulting from the second by x. These values were then added together. This series was then divided into two parts, unity and a rather

complex series called u. This was then expanded as a binomial. Terms that would be of order $\frac{1}{\mu^3}$ or higher when expected values were taken were assumed to be small enough to be omitted.

The first few terms of this expansion were as follows:

$$w_p = \mu^{\frac{1}{2}}(1 + p_1' u + p_2' u^2 + p_3' u^3 + \cdots) \qquad (3)$$

where $p_r' = \binom{1/p}{r}$

Therefore

$$E(w_p) = \mu^{\frac{1}{2}}[1 + p_1' E(u) + p_2' E(u^2) + p_3' E(u^3) + \cdots] \qquad (4)$$

Squaring the series (3) and taking expected values, term by term gives $E(w_p^2)$. Squaring series (4) one obtains $[E(w_p)]^2$. Substituting this series and the series for $E(w_p^2)$ into the variance formula 4.1(7) produces

$$\begin{aligned} Var(w_p) = \mu\{&(p_1')^2 E(u^2) + 2 p_1' p_2' E(u^3) - (p_1')^2 [E(u)]^2 \\ &- 2 p_1' p_2' E(u) E(u^2) - 2 p_1' p_3' E(u) E(u^3) \\ &- (p_1')^2 [E(u^2)]^2 - 2 p_2' p_3' E(u^2) E(u^3) - \cdots \} \end{aligned}$$

Substituting the values for $E(u^r)$, p_r' in terms of p and collecting terms results in

$$Var(w_p) = \frac{1}{4}\left(1 + \frac{8m_2' - 8(m_1')^2 - 8m_1' + 3}{8\mu} + \cdots\right)$$

Note that as far as the expansion goes the variance of w_p is independent of the value of p. Since w_p represents, in some way, all of the previous transformations discussed, the difference between the variances of these transformations must appear in the third and/or subsequent terms. Unfortunately, this expansion turned out to be very involved, and the third term has not been evaluated. Some work has been done on it but it is not completely reliable owing to the complicated algebraic operations entailed.

VI. CONCLUSION

Note in Table 1 that the last term in the variance in the different distributions was for the most part approximately constant for a given μ. If it can be assumed that this property holds true for higher terms, then the variances obtained by this expansion can be compared between the different distributions in spite of their not having converged to the true value. Such a comparison shows the variance of z, with x distributed by the uniform discrete distribution with n = 10, approaches a constant, one-fourth, before any other transformation specifically considered. This would allow the use of a single digit taken from a random number table as the auxiliary variate x.

From the graph of cumulative probability for z (Figure 1) with x distributed by the uniform discrete distribution n= 10 it can be seen that the function is nearly a straight line except for values of z less than approximately unity. In the usual interpretation this would indicate approximate normality with a tendency for negative skewness, although the approximation to normality seems very good in the upper tails of the distributions. Note also that the straight line parts of these curves are parallel. Since the slope of the

cumulative probability plotted on normal probability paper represents the variance, one can see that the variances are nearly constant even for μ quite small. The points on this graph (Figure 1) were joined by a smooth curve for convenience. The true cumulative distribution follows a finely detailed step function. The tendency to normality can be seen from the frequency distribution of z (Figure 2) where x is distributed by the uniform continuous distribution (a limiting case of the uniform discrete distribution). This concurs with work done by Eudey (1949) and David (1955) who discuss the use of the z transformation for testing hypotheses. Eudey mentions the use of an auxiliary variable to give the smallest confidence intervals for a given level of significance. His tests for the one sided alternative are uniformly most powerful, the proof of which he gives in his paper.

Further work on a general form of a transformation of Poisson data seems to be suggested. Since the first two terms of the expansion of the variance of w_p are independent of p (and therefore are the same for a large family of transformations) the "best" type of transformation might be found by investigating the effects of changing p in the third or subsequent terms. Since the complexity (of which there is a

considerable amount involved in the first two terms) multiplies considerably with each term, the number of terms practically attainable by this particular analytic expansion seems severely limited. Possibly an alternative and more practical method of expansion might be found by further study.

VII. ACKNOWLEDGEMENTS

The author wishes to express his sincere appreciation to Professor M. C. K. Tweedie for his guidance during the course of this work. His many ideas and suggestions have made this paper possible.

The author is also indebted to Mrs. L. W. Claffey for preparing the final copies for presentation.

VIII. BIBLIOGRAPHY

1. Anscombe, F. J.: "The Transformation of Poisson, Binomial and Negative-Binomial Data." Biometrika 35, 246-254, 1948.

2. Bartlett, M. S.: "The Effect of Non-Normality on the t Distribution." Proceedings of the Cambridge Philosophical Society 31, 223-231, 1935.

3. Bartlett, M. S.: "The Square Root Transformation in the Analysis of Variance." Supplement to the Journal of the Royal Statistical Society 3, 68, 1936.

4. Bartlett, M. S.: "The Use of Transformations." Biometrics 3, 39-52, 1947.

5. Cochran, W. G.: "Some Consequences When the Assumptions for the Analysis of Variance Are Not Satisfied." Biometrics 3, 22-38, 1947.

6. David, F. N.: "The Transformation of Discrete Variables." Annals of Human Genetics Vol. 19, Part 3; 174-182, 1955.

7. Durell, C. V.: Advanced Algebra Vol. I, p. 38ff, G. Bell and Sons, Ltd., London, England, 1936.

8. Eisenhart, Churchill: "The Assumptions Underlying the Analysis of Variance." Biometrics 3, 1-21, 1947.

9. Eudey, M. W.: "On the Treatment of Discontinuous Random Variables." Technical Report Number 13, Statistical Laboratory, University of California, Berkeley, California, 1949.

10. Fieller, E. C.: "Some Remarks on the Statistical Background in Bio-Assay." The Analyst 72, No. 821; 37-43, 1947.

11. Freeman, M. F. and Tukey, J. W.: "Transformations Relative to the Angular and the Square Root." Memorandum Report 24, Statistical Research Group, Princeton University.

12. Hey, G. B.: "A New Method of Experimental Sampling Illustrated on Certain Non-Normal Populations." Biometrika 30, 68-80, 1938.

13. Kendall, M. G.: The Advanced Theory of Statistics Vol. I, Fourth Edition, p. 63, Charles Griffin and Company Limited, London, England, 1948.

14. Pearson, E. S.: "The Analysis of Variance in Cases of Non-Normal Variation." Biometrika 23, 114-133, 1931.

15. Rao, C. R.: Advanced Statistical Methods in Biometric Research 209-210, John Wiley and Sons, Inc., New York, 1952.

IX. TABLES AND CURVES

Table I. First Four Terms of the Variance Expansion Disregarding the Factor One-Fourth, for Some Distributions of x

μ	First Term	Second Term	Third Term	Fourth Term	Ratio of Variance to its Limiting Value
x is uniform discrete distribution (n = 10)					
1.0	1	0.0075	0.1099	-2.24	-
1.5	1	0.0050	0.0488	-0.633	0.391
2.0	1	0.0038	0.0274	-0.279	0.752
2.5	1	0.0030	0.0176	-0.143	0.878
3.0	1	0.0025	0.0122	-0.0817	0.933
x is uniform discrete distribution (n = 12)					
1.0	1	-0.000579	0.1035	-2.24	-
1.5	1	-0.000386	0.0460	-0.663	0.383
2.0	1	-0.000290	0.0259	-0.279	0.747
2.5	1	-0.00232	0.0166	-0.143	0.874
3.0	1	-0.000193	0.0115	-0.0817	0.930
x is uniform discrete distribution (n = 100)					
1.0	1	-0.0367	0.0765	-2.32	-
1.5	1	-0.0244	0.0339	-0.687	0.323
2.0	1	-0.0184	0.0191	-0.290	0.711
2.5	1	-0.0147	0.0122	-0.148	0.849
3.0	1	-0.0122	0.0085	-0.0860	0.910
x is uniform discrete distribution (n = 00)					
1.0	1	-0.0416	0.0729	-2.31	-
1.5	1	-0.0305	0.0324	-0.685	0.307
2.0	1	-0.0208	0.0182	-0.289	0.708
2.5	1	-0.0166	0.0117	-0.149	0.846
3.0	1	-0.0139	0.0081	-0.0856	0.9086

Table I. First Four Terms of the Variance Expansion Disregarding the Factor One-Fourth, for Some Distributions of x

(Continued)

μ	First Term	Second Term	Third Term	Fourth Term	Ratio of Variance to its Limiting Value	True Ratio (Freeman-Tukey 1949)
$x = c$	$(c = 0)$		$z = \sqrt{y}$			
1.0	1	0.275	0.531	-1.26	0.65	1.608
1.5	1	0.250	0.236	-0.373	1.113	1.641
2.0	1	0.187	0.133	-0.158	1.152	1.560
2.5	1	0.150	0.085	-0.084	1.151	-
3.0	1	0.125	0.059	-0.0467	1.137	1.360
$x = c$	$(c = 3/8)$		$z = \sqrt{y + 3/8}$			
1.0	1	0	0.0625	-2.34	-	0.717+
1.5	1	0	0.0278	-0.693	0.335	-
2.0	1	0	0.0156	-0.293	0.723	0.294+
2.5	1	0	0.0100	-0.150	0.860	-
3.0	1	0	0.00695	-0.0867	0.920	0.983+

Table II. Cumulative Distribution of z Where x Has Uniform Discrete Distribution With n = 10

z	$\mu = 1$	$\mu = 2$	$\mu = 2.5$	$\mu = 5$
0	0.405	0.162	0.102	0.010
0.316	0.442	0.189	0.123	-
0.447	0.479	0.216	0.144	-
0.548	0.515	0.243	0.164	0.021
0.632	0.552	0.271	0.184	-
0.707	0.589	0.298	0.205	-
0.774	0.626	0.325	0.226	0.031
0.836	0.662	0.352	0.246	-
0.895	0.699	0.379	0.266	-
0.948	0.736	0.406	0.287	-
1.000	0.754	0.433	0.313	0.048
1.09	0.791	0.487	0.364	-
1.14	-	-	-	0.074
1.18	0.828	0.542	0.415	-
1.26	0.865	0.596	0.466	0.099
1.34	0.902	0.650	0.517	-
1.41	0.926	0.695	0.564	0.139
1.51	0.944	0.749	0.629	0.181
1.61	0.963	0.803	0.693	0.222
1.73	0.963	0.866	0.770	0.283
1.79	-	-	-	0.317
1.82	-	0.893	0.881	-
1.84	-	-	-	0.353
1.87	0.990	-	-	-
1.90	-	0.920	0.851	0.387
1.95	-	-	-	0.423
2.00	0.996	0.951	0.898	0.458
2.05	-	-	-	0.493
2.07	-	-	0.918	-
2.10	-	-	-	0.528
2.12	0.996	0.969	-	-
2.14	-	-	0.938	0.563
2.19	-	-	-	0.598
2.23	0.999	0.984	0.961	0.631
2.28	-	-	-	0.660
2.32	-	-	-	0.689

Table II. Cumulative Distribution of z Where x Has Uniform Discrete Distribution With n = 10

(Continued)

z	$\mu = 1$	$\mu = 2$	$\mu = 2.5$	$\mu = 5$
2.34	-	0.990	0.975	-
2.36	-	-	-	0.718
2.41	-	-	-	0.747
2.45	1.000	0.995	0.987	0.772
2.49		-	-	0.793
2.53		-	-	0.814
2.55		0.997	0.992	-
2.57		-	-	0.835
2.60		-	-	0.856
2.64		0.999	0.996	0.873
2.68		-	-	0.887
2.72		-	-	0.899
2.74		1.000	0.998	-
2.76			-	0.913
2.79			-	0.925
2.83			0.999	0.936
2.88			-	0.946
2.93			1.000	0.957
3.00				0.970
3.05				0.975
3.10				0.981
3.13				0.984
3.16				0.987
3.24				0.991
3.32				0.995
3.39				0.997
3.46				0.998
3.53				0.999

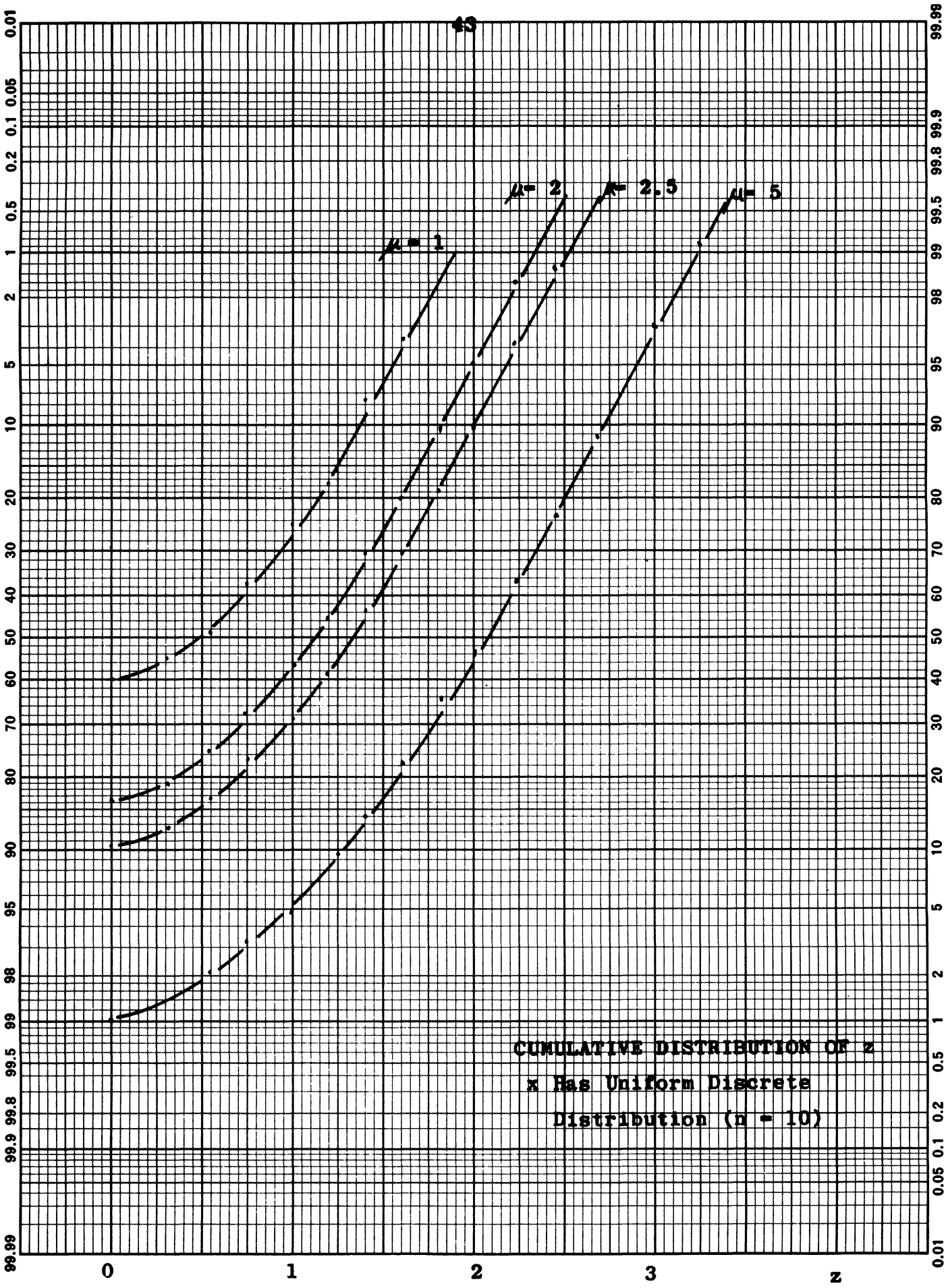
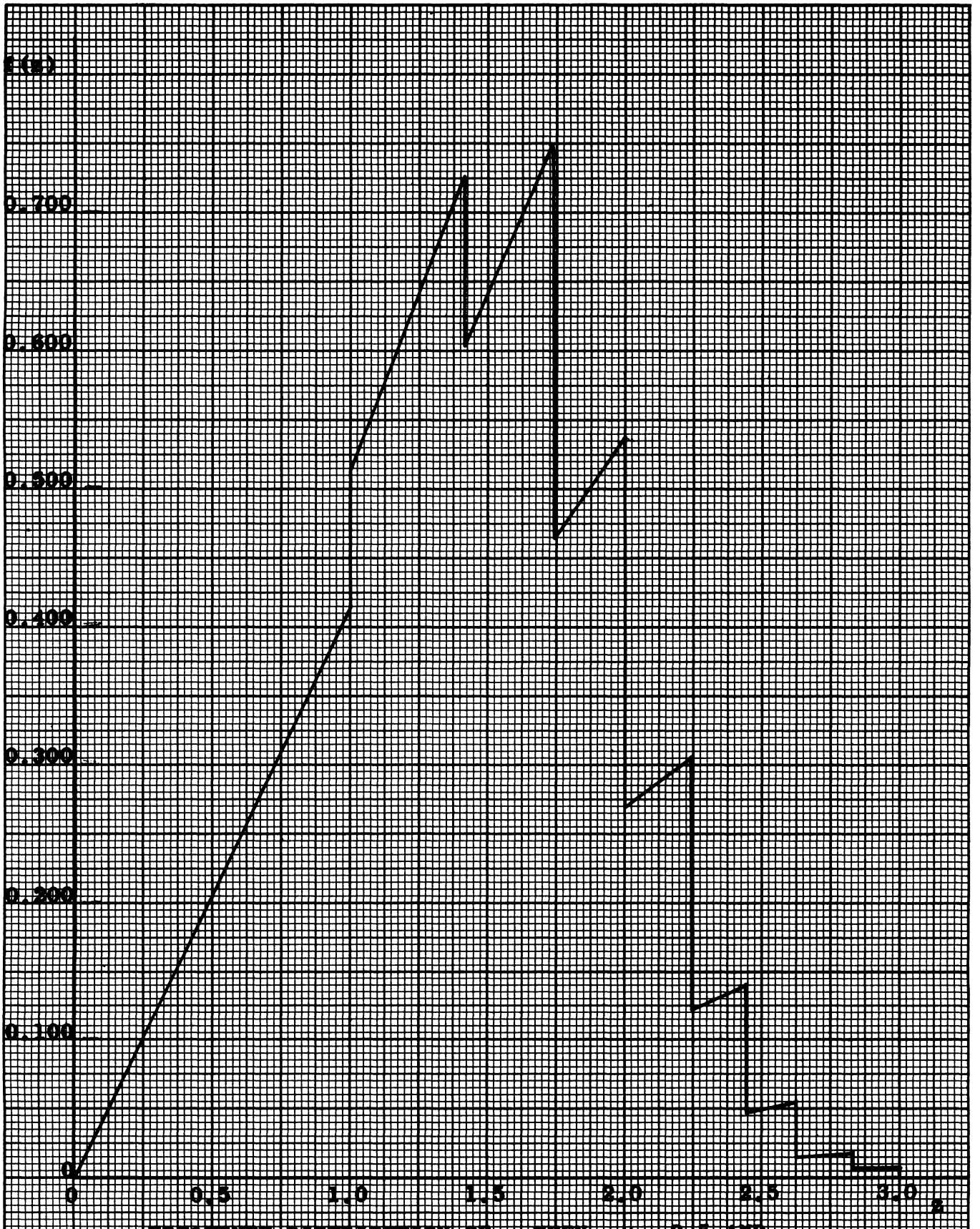

μ = 1
μ = 2
μ = 2.5
μ = 5
CUMULATIVE DISTRIBUTION OF z
x Has Uniform Discrete
Distribution (n = 10)
0
1
2
3
z
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.5
99.8
99.9
99.99
99.98
99.9
99.8
99.5
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
0.01

FIGURE 1



FREQUENCY DISTRIBUTION OF x WHEN $\mu = 2.5$ AND

x HAS THE UNIFORM CONTINUOUS DISTRIBUTION

FIGURE 2

The vita has been removed
from the scanned document