TRANSFORMATIONS OF RANDOM VARIABLES

by

ERNEST WAYNE PHELPS

B. S., Kansas State University, 1959

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

Approved by:

A. M. Feyerherm

Major Professor

TABLE OF CONTENTS

INTRODUCTION 1

THEORETICAL CONSIDERATIONS 3

THE SQUARE ROOT TRANSFORMATION 7

THE LOGARITHMIC TRANSFORMATION 12

ANGULAR TRANSFORMATIONS 16

THE PROBIT TRANSFORMATION 20

TRANSFORMATIONS WITH FRACTIONAL POWERS OF THE VARIABLE 26

INVERSE HYPERBOLIC SINH TRANSFORMATIONS 28

MODIFIED AND SPECIAL TRANSFORMATIONS 33

SUMMARY AND CONCLUSIONS 35

INTRODUCTION

Conventional tests of hypotheses and methods for defining confidence intervals are not strictly valid unless the conditions and assumptions for specific statistical models are met. Such conditions and assumptions may be more nearly satisfied if the involved random variables are transformed to new forms. This procedure is especially useful when applied to analysis of variance for data derived from designs such as randomized blocks and Latin squares.

Transformations are especially appropriate when a random variable is such that its variance is a function of its mean μ so that $\sigma^2 = f(\mu)$. The proper change of variable would be one which produces a new random variable whose mean and variance are unrelated since homogeneity of variances is a common assumption in many statistical investigations. Other assumptions in linear models usually specify additivity of effects and normal and independently distributed random errors; that is $e \in N(0, \sigma^2)$. Since nonconformity to linear models is not uncommon, a need for developing appropriate transformations applicable to particular statistical data is present.

The usual purpose for transforming a variable in analysis of variance data is to change the scale of measurement in order to make the analysis more valid. One condition required for assessing accuracy in the ordinary unweighted analysis of variance includes the important one of a constant residual or error variance. If the variance tends to change with the mean level of the measurements, the variance can only be stabilized by a suitable change of scale. The functional relationship of variance to mean level determines the appropriate transformation to be used. For some data a transformation which achieves homogeneity of error variance has an added beneficial effect. Nonnormality often occurs with heterogeneity of error variances, and a transformation may partially correct both difficulties simultaneously. However, the normality assumption is somewhat less important than homogeneous variances because it can be shown that the sampling distribution of the F ratio is relatively insensitive to moderate departures from normality.

A third reason for transforming may be to achieve additivity of effects, which implies a linear model which does not contain interaction terms. An additive linear model, free of interactions, has particular advantages in the case of fixed and random effects in the same experiment. Tukey's test for non-additivity can be used in part to determine the appropriate transformation to obtain additivity of effects. However, in some cases there is an intrinsic interaction between the factors which cannot be considered a function of the choice of the scale of measurement and therefore it is not always possible to find a transformation which will eliminate non-additivity in a given situation.

Bartlett (1947) summarizes by stating that the ideal transformation will be one in which:

(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.

THEORETICAL CONSIDERATIONS

Consider a variate x whose mean $\mu = E(x)$ is a real variable with a range R of possible values, and a standard deviation $\sigma_x = f(\mu)$ is a non-constant function of μ. In order to stabilize variances, it is necessary to find a function $Y = f(x)$ such that both $f(x)$ and

$$\sigma_Y^2 = E\left\{Y - E(Y)\right\}^2$$

are functionally independent of μ for μ on R. This would require that

$$\frac{\partial f}{\partial \mu} = 0 \qquad \text{and} \qquad \frac{\partial \sigma_Y^2}{\partial \mu} = 0 \quad ,$$

for μ on R.

One approach to the problem is from the relationship

$$dY = f'(x)dx \quad .$$

An approximate solution is available by a summation process such that

$$\sigma_Y = f'(\mu)\sigma(\mu) \quad .$$

From this a rough approximation can be derived by setting this expression equal to a constant, say c, thus obtaining

$$f'(\mu) = \frac{c}{\sigma(\mu)} \quad .$$

Then $f(x)$ is an indefinite integral of $\frac{c}{\sigma(x)}$. The results obtained can only be useful if the application is found to be satisfactory.

A transformation $Y = f(X)$, which produces a variate Y such that its distribution is exactly normal, depends upon the original variate X. One can never exactly transform a discrete variate X into a variate Y with a continuous distribution. However any variate X with a continuous distribution function $F(X)$ can be transformed into a normally distributed variate Y by the transformation $Y = f(X)$ defined by the equation

$$F(X) = \int_{-\infty}^{Y} \frac{1}{\sqrt{2\pi}} e^{-1/2\, y^2} dy \quad .$$

However, the resulting function $Y = f(X)$ will not generally be functionally independent of the mean of X and there is the practical difficulty of solving this equation for Y. Therefore it seems somewhat more reasonable to seek asymptotic solutions to the problems of normalization and stabilization of variance.

Asymptotic solutions depend on the assumption that the distribution of X, with mean $\mu = \mu_n$ of X and range R_n, is determined by a parameter n which tends to infinity. Other independent parameters are present and represented collectively by θ with a range Ω. The variate of interest is $Y = f(X)$, in which $f(X)$ is functionally independent of μ and the parameter θ ranges on R_n, or θ in Ω, and such that the distribution of

$$f(X) - f(\mu_n)$$

tends as $n \to \infty$, to a normal distribution, while

$$\lim_{n \to \infty} \sigma_T^2 = C^2 \quad ,$$

where C^2 is an absolute constant. This does not deny the presence of the additional parameters θ. The function $f(X)$ may depend non-trivially on n, but since n is the only parameter on which the distribution of X depends, then $f(X)$ must be functionally independent of n.

The following theorems, due to Curtiss (1943), are helpful in determining asymptotic solutions.

Theorem 1 Let $\psi_n(x)$ be a non-negative function of x and n, defined almost everywhere and Lebesque integrable with respect to x over any finite interval of the x-axis for each $n > 0$. Let

$$Y = f(x) = \int_a^Y \psi_n(x)dx$$

where a is an arbitrary constant. Let $F_n(z)$ be the distribution function of the variate

$$Z = (X - \mu_n)\psi_n(\mu_n) \quad .$$

Suppose further that a continuous distribution function $F(z)$ exists such that

$$\lim_{n \to \infty} F_n(z) = F(z)$$

for all values of z. Then either one of the following two conditions is a sufficient condition for the distribution function $H_n(w)$ of the variate

$$W = f(x) - f(\mu_n)$$

to tend to $F(w)$, $-\infty < w < \infty$:

(a) To each w for which $0 < F(w) < 1$, there corresponds for all n sufficiently large at least one root $X = X_n$ to the equation

$$\int_{\mu_n}^X \psi_n(u)du = w$$

and this root X_n has the property that

$$\lim_{n \to \infty} (X_n - \mu_n)\psi_n(\mu_n) = w \quad .$$

(b) For all n sufficiently large, $\psi_n(\mu_n) > 0$, and

$$\lim_{n \to \infty} q_n(w) = 1$$

uniformly in any closed finite subinterval of the open interval defined by $0 < F(w) < 1$, where

$$q_n(v) = \frac{\psi_n v[t_n(\mu_n)]^{-1} + \mu_n}{\psi_n(\mu_n)} \quad .$$

The following lemma is useful in showing condition (a) in the theorem is a consequence of condition (b).

Lemma - If $\gamma_n(w)$ is a non-negative function integrable over any finite interval of values of w; and if

$$\lim_{n \to \infty} \gamma_n(v) = 1$$

uniformly in any finite closed subinterval of an interval $w_1 < w < w_2$, then for every value of w in this interval there exists for all n sufficiently large a solution $z = z_n$ of the equation

$$\int_0^z \gamma_n(v)dv = w \quad ,$$

and the solution z_n has the property that

$$\lim_{n \to a} z_n = w \quad .$$

<u>Theorem 2</u> Let Y (or $f(x)$), Z, $F_n(z)$ and $F(z)$ be defined as in Theorem 1. Let the mean and variance of the distribution defined by $F(z)$ exist and have respective values 0 and c^2. Then the following three conditions, taken together, are sufficient that

$$\lim_{n \to \infty} [E(Y) - f(\mu_n)] = 0$$

$$\lim_{n \to \infty} \sigma_Y^2 = c^2$$

(i) $E(Y^2)$ exists for $n > 0$, and $\lim\limits_{n \to \infty} E(Z^2) = c^2$

(ii) Condition (b) of Theorem 1 holds.

(iii) $f(Z[\psi_n(\mu_n)]^{-1} + \mu_n) - f(\mu_n) = O|Z|$ uniformly in n as $|Y| \to \infty$.

Theorem 3 Let the distribution of a variate Z depend upon a parameter n, let $F_n(z)$ be the distribution function of Z, and let $F(z)$ be a continuous distribution function with the property that

$$\lim_{n \to \infty} F_n(z) = F(z) \quad .$$

Let a_n be a function of n such that

$$\lim_{n \to \infty} a_n = a \neq 0.$$

Then the distribution function of the variate $v = a_n Z$ tends as $n \to \infty$ to the distribution function $F(v/a)$ if $a > 0$, and to the distribution function $1 - F(v/a)$ if $a < 0$. If the variance of Z exists and tends to c^2 as $n \to \infty$, then the variance of $a_n Z$ tends to $a^2 c^2$ as $n \to \infty$.

THE SQUARE ROOT TRANSFORMATION

The variance of a Poisson distributed variate is equal to its mean. Theory suggests that a square root transformation may prove useful. For enumeration data, where values of the random variable arise by counts, the variate may tend to be distributed in a Poisson fashion. If the mean m is large, and actually equal to the variance, we have:

$$\sigma^2(\sqrt{x}) \doteq 1/4 \; . \; . \; .$$

or more generally $(1/4)\,\lambda$ if

$$\sigma^2(x) = \lambda\, m \quad .$$

The above formula is an approximation formula only, and it is of some revelance to see how far the variance of $\sqrt{X}$ for a Poisson variate X is constant, when the mean m becomes small. The variance of $\sqrt{X}$ can also be calculated for a continuous distribution for which the variance equals the mean, where p is proportional to

$$X^{m-1} e^{-X} dx \quad .$$

Bartlett (1936) has investigated numerically the degree of approximation for values of m from .5 to 15.0 in the cases $\alpha = 0$ and $\alpha = 1/2$. He found that the variance of $\sqrt{X + (1/2)}$ is considerably closer to the limit $(1/4)$ for $1 \leq m \leq 10$ than is the variance of $\sqrt{X}$. At $m = 15$ the variance of $\sqrt{X}$ is .256 and that of $\sqrt{X + (1/2)}$ is .248. For the continuous curve, the variance of the transform variate approaches its limit suprisingly quickly, showing a peak about $m = 1$, which disappears when $\sqrt{X + (1/2)}$ is used. Thus $\sqrt{X}$ may be considered above $m = 0$, and from 10 to about 2 or 3, $\sqrt{X + (1/2)}$ is preferable. Below a mean m of about 2 or 3, an analysis is not useful unless a large number of replications is available from experimentation.

For example, let a variate X have a Poisson exponential distribution with parameter n. If X is an arbitrary constant, and if

$$Y = f(X) = \begin{cases} \sqrt{X + \alpha} & X \geq -\alpha \\ 0 & X < -\alpha \end{cases}$$

then the distribution of $Y - \sqrt{n + X}$ tends as $n \rightarrow \infty$ to a normal distribution which has mean zero and variance $1/4$, and

$$\lim_{n \rightarrow \infty} \sigma_Y^2 = 1/4 \quad .$$

For $\mu_n = n$, $\sigma_X = \sqrt{n}$, it is known, Curtiss (1943), that the distribution of the reduced variate $(X - n)/\sqrt{n}$ tends to the reduced normal distribution as $n \rightarrow \infty$. By Theorem 3, the distribution of the variate

$$Z = \frac{X - n}{2\sqrt{n + \alpha}} = \frac{1}{2} \cdot \sqrt{\frac{n}{n + \alpha}} \cdot \frac{X - n}{\sqrt{n}}$$

will tend to normality as $n \rightarrow \infty$, and the variance of Y will tend to the value $1/4$, which is also the variance of the limiting distribution. Let

$$\psi_n(x) = \begin{cases} \frac{1}{2\sqrt{X + \alpha}} & X > -\alpha \\ 0 & X \leq -\alpha \end{cases}$$

then

$$Y = f(X) = \int_{-\alpha}^{X} \psi_n(X)dx \quad .$$

The proof of the above statement can be verified by conditions (ii) and (iii) of Theorem 2. Therefore it is justifiable to use an appropriate square root transformation on a Poisson variate X.

In the analysis of variance for a Latin square it is standard practice to assume that treatment, row, and column effects are additive. The expected yield X_i of the ith plot, which receives the jth treatment and occurs in the kth row and the lth column is written

$$X_{gijk} = \mu + T_g + R_i + C_j + \epsilon_{gijk}$$

where μ is the average yield in the experiment and T_g, R_i and C_j represent the effects of treatment, row and column for each plot. Because the T, R and C constants are required only to measure differences between different treatments, rows and columns, we may put

$$\sum_g T_g = \sum_i R_i = \sum_j C_j = 0 \quad .$$

With normally and independently distributed experimental errors, simple equations are available for estimating unknown parameters. Therefore, the analysis should be a good description of the combined action of treatment and block effects.

Any set of yields X_i with expectation m_i may be written in a probability statement as:

$$\frac{{}_i\Pi\, e^{-m_i}\, m_i^{X_i}}{X_i!}$$

The logarithm of the likehood is

$$L = \sum_i (X_i \log m_i - m_i) - \sum_i \log X_i!$$

This gives the maximum likehood equation of estimating any parameter θ in the form

$$\frac{\Sigma(X_i - m_i)}{m_i} \frac{\partial m_i}{\partial \theta} = 0$$

where the summation extends over all plots whose expectations involve θ.

Because Poisson variation contains a weighted sum of the deviations from the observed expected values, the factor $\frac{1}{m_i}$ is introduced by the weights $\frac{1}{m_i} \frac{\partial m_i}{\partial \theta}$. The presence of this weight factor makes a linear model unsuitable. However, this weight can be stabilized and made constant by assuming a linear prediction formula in the square roots and transforming the data to square roots. For the Latin square, this prediction formula would be

$$\sqrt{m_i} = y_i = \mu + T_g + R_i + C_j$$

where

$$\sum_g T_g = \sum_i R_i = \sum_j C_j = 0 \quad .$$

Through the use of the method of Lagrangian multipliers, the maximum value of the logarithm of the likelihood is found by maximizing

$$L + \lambda(\sum_g T_g) + \psi(\sum_i R_i) + \gamma(\sum_j C_j) \quad .$$

Estimation of a typical constant yields a transformation of the form

$$y_i = \sqrt{x_i} \quad .$$

When m_i is small, an approximation procedure similar to the rules for finding square roots is available and sometimes more satisfactory.

The suitability of the linear prediction formula in a square root transformation must be considered when this analysis is being employed. The prediction formula can be evaluated by

$$\frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}$$

which has a sufficiently close χ^2 distribution. A high value of this χ^2 means that the prediction formula is not satisfactory, or that the experimental errors are higher than the Poisson distribution indicates or that both causes are operating. These effects can sometimes be separated by examining whether the observed yields deviate from the expected yields in a systematic or a random manner. If the deviation is systematic, the prediction formula is probably not satisfactory.

Although the square root transformation is especially applicable for a variate with a Poisson distribution, it is often times a good transformation in other situations. For instance at times a useful square root transformation is sometimes available for a variate with a Γ distribution.

THE LOGARITHMIC TRANSFORMATION

A logarithmic transformation of a random variable X is appropriate if effects of an experiment are proportional to the mean instead of additive. The standard deviation of the variable also tends to vary directly as the mean. A situation where

$$\sigma_x^2 = \mu + \sigma_\mu^2$$

or sometimes

$$\sigma_x^2 = k^2\mu$$

may call for a logarithmic transformation. Also frequently there is a proportional relationship between the means, variances, and ranges of the original random variable. A logarithmic transformation frequently makes the ranges almost equal and uncorrelated with the means. A transformation that tends to make the ranges more uniform will also tend to make the variances more uniform. In an experimental situation one can apply Tukey's test for non-additivity to each of the interaction terms to check the adequacy of the logarithmic transformation. Graphically one might observe a curvilinear relationship between the sample mean and variance which suggests that the relationship might be quadratic. If this is the situation, a logarithmic transformation will frequently make the mean and variance unrelated, thus satisfying the assumptions for the analysis of variance and normal theory.

The asymptotic theory for a logarithmic transformation differs in some respects from other transformations. Consider a variate X with mean μ and standard deviation σ and with functional relationship

$$\sigma = k_n(\mu_n + \alpha),$$

where α is an arbitrary constant, $k_n > 0$, and the $\lim_{n \to \infty} k_n$ exists and is finite. Stabilization of the variance of X at k^2 leads us to the function

$$Y = \log (X + \alpha), \qquad X > -\alpha \quad .$$

If X is a variate such that

$$P(X \leq -\alpha) = 0 \quad ,$$

then the corresponding reduced variate

$$Z = \frac{X - \mu_n}{k_n(\mu_n + \alpha)}$$

has a distribution function $F_n(Z)$ such that

$$F_n(-1/k_n) = 0 \quad .$$

Therefore, if

$$\lim_{n \to \infty} k_n = k > 0 \quad ,$$

the limiting distribution of Z, if it exists, must have a distribution function $F(Z)$ such that

$$F(-1/k - 0) = 0 \quad .$$

Hence, the limiting distribution of Z can never be normal if $k > 0$.

Curtiss (1943) showed that if the reduced variate Z does have a limiting distribution, the variate

$$W = \frac{1}{k_n \log (X + \alpha)} - \frac{1}{k_n \log (\mu_n + \alpha)} = \int_{\mu_n}^{X} \frac{1}{k_n(u + \alpha)} du, \quad X > -\alpha$$

may have a limiting distribution which is not the same as that of W. More specifically he states:

<u>Theorem 4</u> Let

$$P(X \leq -\alpha) = 0 \quad ,$$

and

$$\lim_{n \to \infty} k_n = k \geq 0 \quad ,$$

and $F_n(z)$ be the density function of the reduced variate

$$Z = \frac{X - \mu_n}{k_n(\mu_n + \alpha)} \quad .$$

Furthermore let $H_n(w)$ be the distribution function of the variate W given above. If a continuous distribution function $F(z)$ exists such that

$$\lim_{n \to \infty} F_n(y) = F(y)$$

for all y, then

$$\lim_{n \to \infty} H_n(w) = \begin{cases} F[\frac{e^{kw} - 1}{k}] , & k > 0 \\ F(w) , & k = 0 \end{cases}$$

And in considering the distribution of the transform variate Y the following theorem is useful.

Theorem 5 Under the hypotheses of theorem 4, and under the conditions that the improper integral

$$\int_{-\infty}^{0} w^2 dH_n(w) \qquad \left(\text{or} \int_{-1/k_n}^{0} k_n^{-2} [\log (1 + k_n w)]^2 dF_n(w) \right)$$

converges uniformly in n and that

$$\int_{-\infty}^{\infty} w^2 dF(w) = 1, \quad = E(w)^2 \quad ,$$

the following relations hold:

$$\lim_{n \to \infty} E(w) = \begin{cases} \int_{-1/k}^{\infty} \frac{1}{k} \log (1 + k_z) dF(z) & k > 0 \\ 0 & k = 0 \end{cases}$$

$$\lim_{n \to \infty} E(w^2) = \begin{cases} \int_{-1/k}^{\infty} \frac{1}{k^2} [\log (1 + k_z)]^2 dF(z) & k > 0 \\ 1 & k = 0 \end{cases}$$

From this it can be shown that if $F(z)$ is independent of any unknown parameters θ, and if k is positive and has the same value for all variates in a given problem, the transformation

$$Y_1 = \log (X + \alpha)$$

will yield an asymptotic stabilization of the variance under the conditions of Theorem 5. However, if

$$k = 0, \qquad Y = \log (X + \alpha)$$

converges stochastically to $\log (\mu_n + \alpha)$. The variance of the variate

$$Y = \log (X + \alpha)$$

can be stated by the equation

$$\sigma_Y^2 = k_n^2 \left\{ E(w^2) - [E(w)]^2 \right\} \quad .$$

The specific application of a logarithmic transformation depends somewhat upon the experimeter's insight into the particular type of variation present in the original variate X. If from a large number of observations the variate is essentially bounded from below and has a standard deviation which is proportional to the mean μ, the use of the transformation

$$Y = \log (X + \alpha) \quad ,$$

with $-\alpha$ less than the lower bound of X, will usually yield a normally distributed variate with the variance independent of the value of μ. In general there are a large number of distributions where the mean and standard deviation are proportional. The use of a logarithmic transformation in the analysis of variance should follow-up with tests for departure from normality for the observed distribution of transformed values. In some situations the logarithmic transformation may be somewhat more successful in stabilizing the variance rather than normalizing the data. However, it will be found that the logarithmic transformation is particularly useful in normalizing distributions which contain a positive skewness in the original data.

ANGULAR TRANSFORMATIONS

In a situation where a random variable tends to have a binomial distribution, an angular or inverse sine transformation is often effective in stabilizing the variances. For such quantal data one would have a binomial population in which a proportion P possesses a given attribute. Sample estimates of P, p_i, can be taken and will give sample evidence of the parameter values. If these proportions cover a wide range of value, the inverse sine transformation of the original variate would be particularly relevant prior to the analysis of the data. This wide range of proportions cannot always be accounted for since the expectation of p equals P and the variance of p equals PQ/n, where

$$Q = 1 - P \quad .$$

The inverse sine transformation is probably the most common angular transformation, but there are other angular transformations that would be particularly applicable if additional information is available about the original variate.

The inverse sine square root transformation is of the general form:

$$q(x) = \sin^{-1}\sqrt{x} \quad .$$

If the transformation is effective, the approximately constant variance on the transform is $821/n$, provided that the inverse sine, which denotes an angle is measured in degrees.

Suppose the basic observations of the original variate are proportions and the mean and variance are related in the following form

$$\sigma^2 = \mu(1 - \mu) \quad .$$

The following transformation is effective in stabilizing the variances in such instances,

$$Y_i = 2 \arcsin \sqrt{X_i}$$

where X_i is a proportion. For values of X close to zero or close to unity, the following transformation is recommended,

$$Y_i = 2 \arcsin \sqrt{X_i + (\tfrac{1}{2} n)}$$

where n is the number of observations upon which X is based. Both of the above transformations are variations of the general form and the analysis is facilitated by an insight into the original variate X_i.

For a negative binomial variable X, with mean m and exponent k, the latter being constant and known, the appropriate transformation would be

$$y = \sinh^{-1} \sqrt{\frac{X + c}{k - 2c}} \quad .$$

The optimum value of c is roughly $3/8$ if m is large and $k > 2$, and the variance is approximately $\frac{1}{4} \psi'(k)$, where $\psi'(t)$ denotes the second derivative of $\ln \Gamma(t)$, a gamma function determined by the cumulant-generating function, with respect to t.

Consider a variate X with a binomial relative frequency distribution and parameter p which has n possible values $0, 1/n, 2/n, \ldots n/n$. For an arbitrary constant α, if

$$Y = f(x) = \begin{cases} \sqrt{n} \sin^{-1} \sqrt{X + \alpha/n} \ , & -\alpha/n \leq X \leq 1 - \alpha/n \\ 0 \ , & X < -\alpha/n \ , \\ & X > 1 - \alpha/n \end{cases}$$

where Y is measured in radians, then the distribution of

$$Y - \sqrt{n} \sin^{-1} \sqrt{p + (\alpha/n)}$$

tends as $n \to \infty$ to a normal distribution which has mean zero and variance $1/4$ and

$$\lim_{n \to \infty} \sigma_Y^2 = 1/4 \quad .$$

Since

$$\mu_n = p \quad \text{and} \quad \sigma_X^2 = pq/n \ , \quad \text{where} \quad q = 1 - p \ ,$$

it can be shown that the distribution of the reduced variate

$$\sqrt{n}(X - p)/\sqrt{pq}$$

will tend to normality as $n \to \infty$. By Theorem 3, the distribution of

$$Z = \frac{\sqrt{n}\,(X - p)}{2\sqrt{(p + \frac{\alpha}{n})(q - \frac{\alpha}{n})}}$$

will tend to normality with a limiting variance of $1/4$, which is also the limiting distribution. Let

$$\psi_n(x) = \begin{cases} \dfrac{\sqrt{n}}{2\sqrt{(x + \frac{\alpha}{n})(1 - x - \frac{\alpha}{n})}} & -\frac{\alpha}{n} < x \quad 1 - \frac{\alpha}{n} \\ 0 & x \leq -\frac{\alpha}{n}, \ x \geq 1 - \frac{\alpha}{n} \end{cases}$$

then one obtains the integral

$$Y = \int_{-\alpha/n}^{X} \psi_n(x)dx \quad .$$

From here the conditions (ii) and (iii) of Theorem 2 can be satisfied justifying the asymptotic considerations of this angular transformation.

Sometimes it is convenient to express X as a percentage, which has the effect of multiplying Z by 100. In this case the quantity

$$\sqrt{n} \sin^{-1} \sqrt{X + 100\frac{\alpha}{n}} - \sqrt{n} \sin^{-1} \sqrt{100p + 100\frac{\alpha}{n}}$$

has a distribution approaching normality, and $\sigma_Y \to 50$ instead of $1/2$.

The choice of $\alpha = 1/2$ is somewhat more useful for p near 0 or 1, but the choice $\alpha = 0$ is more suitable if the estimated p lies between .3 and .7 for $n = 10$ which is somewhat common. However, for other values of n, the question as to the choice of α for the above transformation for convergence to normality is somewhat arbitrary.

The analysis following the angular transformation

$$Y = \sin^{-1}\sqrt{P} \;, \quad \text{where} \quad 0 \le P \le 1, \quad \text{and} \quad 0 \le Y \le 90^\circ$$

may be carried out in one of two ways, (1) as a multiple regression or (2) as an analysis of variance, both methods being iterative. If n is not constant, weight analysis may be carried out, but due to the loss of orthogonality the advantages of the analysis of variance are to a large extent lost.

The angular transformation

$$Y = f(x) = 2 \arcsin \sqrt{x} \;, \quad \text{where} \quad 0 \le X \le 1 \;, \quad \text{and} \quad 0 \le Y \le \pi \quad \text{(radian measure)}$$

is very effective for the simplification of variance of random variables where the variance is dependent on the mean. However, other transformations such as

$$X = \sin^2 Y \quad \text{where} \quad 0 \le X \le 1 \quad \text{and} \quad 0 \le Y \le 90^\circ \quad \text{(degree measure)}$$

are equally as effective for stabilization of variance and will have

$$V(Y) = \frac{820.7}{n} + \frac{1}{n} \;.$$

However, both are limited with the problem of the variance largely dependent upon X even after transformation in the extreme range of proportions. Judicious design of an experiment and use of a design such as a parallelogram design will somewhat eliminate this problem. It is frequently useful to look at past experimental material or preliminary pilot experiments to give prior indication of this problem. Bartlett (1936) has proposed an adjustment to give corrections for such extreme responses. This adjustment is

$$Y(x) = \arcsin \sqrt{x} \qquad \text{where} \quad 0 < x < 1, \quad 0 < Y < 90^{\circ}$$

$$Y(0) = \arcsin \sqrt{\tfrac{1}{4} n}$$

$$Y(1) = 90 - Y(0)$$

For practical purpose this adjusted transformation fulfills the requirement that the variance is independent of the mean provided

$$0.05 < P < 0.95 \qquad \text{and} \qquad n \geq 10$$

and is constant for all samples. In conclusion, tables are readily available for application of the angular transformation to experimental data.

THE PROBIT TRANSFORMATION

The probit transformation is particularly useful when applied to dosage-mortality and time-survival data. One is concerned with the variables of the percentage mortality to the concentration of treatment. The graphical representation of these two variables with the percentage of dead organisms on the ordinate to the concentration of dosage or treatment on the abscissa is usually a sigmoid curve. This curve resembles the letter S, with the change in percentage kill per unit of the abscissa being smallest near mortalities of 0 and 100 per cent, and largest near 50 per cent.

General theory implies that the dosage-mortality curve is primarily descriptive of the variation in susceptibility between individuals of a population. One would suspect that the individual lethal dose would vary from individual to individual and would suspect that this distribution curve would be normal under uniform conditions. However, it is not easy to apply experimental techniques to determine this lethal dosage because all animals more susceptible than those which could be killed with a smaller dosage, would be killed along with those requiring a given dosage. This requires a technique where the dosage is varied and determines a series of percentage kills for proportionate areas, $(\frac{p}{p + q})$. If these percentage kills are then plotted on the ordinate and dosage on the abscissa, one would obtain a cumulative normal frequency distribution. The proportionate areas are those of the categories killed or not killed. The assumption of individual susceptibility can be easily checked by experiment and comparison with the theoretical normal curve of error. The expected dosage is then given in terms of a standard deviation about a mean or median at the origin. However this system would introduce negative expected dosages which would be inconvenient.

A system of statistical units called probits or probability units is available for transforming the sigmoid dosage-mortality curve to a straight line. This transformation will not modify the proof or disproof of basic assumptions and is one obtained by equating 0 of the usual statistical table of deviates to the digit 5 and then the deviates of the normal curve, given in terms of σ or standard deviations are added to this unit to obtain the probit corresponding to each percentage kill. Sometimes one must take the probit transformation to one additional step if the variation shows a geometrical rather than arithmetical distribution. This possibility can be tested by converting the observed dosages to logarithms and again plotting the dosages inferred from mortality or probits against those secured experimentally. This double-transformation will frequently give a straight line relationship and often times gives a more adequate measure of susceptibility.

If the transformation of dosages to logarithms completes the transformation of the dosages-mortality curve to a straight line, it is an index to the inherent susceptibility of the individual animal to the poison. There must then be a direct proportionality between the concentration of the poison in the dose administered and the amount of poison fixed by the essential tissues of the animal. Frequently the poisoning of an individual multicellular animal can be attributed to the death of a proportion of the essential cells, and hence the susceptibility of the animal as a whole can be determined by the average susceptibility of its essential cells. Therefore the logarithmic transformation may be sought in the relation between the dosage administered and the amount of poison fixed by the essential cells or tissues.

One of the principal empirical formulas for describing the absorption or fixation process is

$$KC^{1/n} = \frac{x}{m}$$

where, C is the concentration of the drug (or dosage), X = the amount fixed in the organism, m = mass of the absorbing constituents within the organism, and K and n are constants. Since m is essentially constant from animal to animal one can combine constants and reduce the above formulas to

$$\log \ C = n \log X + K'$$

From this it is apparent that there is a linear relation between the logarithm of the concentration and the logarithm of the amount fixed by the cells of the animal. Hence the observed logarithmic conversion of the dosage-mortality curve is not due, to use of the amount fixed in the tissue as the true individual lethal dose. Another absorption equation is somewhat more consistent as far as the middle and higher kills and dosages are concerned in the logarithm-probit transformation.

The first estimate of the transformed dosage-mortality curve, the provisional regression line, is ordinarily not calculated but represents the best judgement of the experimenter. Sometimes the experimenter will want to calculate this first estimate if the observations appear to be widely scattered. Occasionally the initial transformed dosage-mortality curve will serve the needs of the experimenter if the data is uniform.

Fitting of a dosage-mortality curve is an attempt to infer from a given experiment the conditions obtaining in a class or species of organisms, the calculated regression line of the dosage probit diagram is the most accurate estimate which can be drawn from the data, provided that basic assumptions are correct. In some cases a first approximation is sufficient but frequently it will only represent a rather important correction of initial estimates. Each separate observation can be weighted accurately and limits can be determined about the calculated regression line. The regression line has the form

$$Y = a + b(X - \bar{x})$$

where, Y = the mortality in probits on the transformed dosage mortality curve corresponding to any given dosage X, $a = \bar{y}$ = numerical average probit for all determinations in part of experiment being fitted by a straight line, $\bar{x}$ = the average dosages in logarithms, and b is the regression coefficient or slope of the line. The necessary calculations are:

$$\bar{x} = \frac{\Sigma(wx)}{\Sigma(w)}, \quad \bar{y} = \frac{\Sigma(wy)}{\Sigma(w)}, \quad b = \Sigma(wxy) - \bar{x}\,\Sigma(wy) \quad \text{and}$$

$$A = \Sigma(wx^2) - \bar{x}\,\Sigma(wx)$$

where:

w = weight of a given observation, the product of the weighting coefficient multiplied by the no. of killed plus survived.

x = a function of the experimental dosage, usually its logarithm.

y = the probit corresponding to the observed percentage mortality.

The regression line will always pass through the point $\bar{x}$, $\bar{y}$ and hence fix the degree of the susceptibility to a toxic agent shown by the population as a whole. From a statistical viewpoint, b is the slope or the tangent of the angle with which the regression line will pass through the point established by $\bar{x}$ and $\bar{y}$. From a biological viewpoint, b measures how closely the individual organisms in the experiment agree with one another in their sensitivity to the toxic agent. It is convenient to express this toxicological characteristic as the percentage increase in dosage that is required to increase kill by one probit. This is the ratio of $100 \log_e 10$ to b or $\frac{230.26}{b}$.

One is of course concerned about how close the sample evidence agrees with the true situation. There is a χ^2 test available for comparing how well this true relationship is approximated. It involves few computations beyond those required for determining the regression equation. The first consideration is to determine whether the observed mortalities agree with the assumption of a rectilinear relationship on the logarithmic probability scale within limits of sampling error. Each dosage observed is compared with the expected dosage from the regression equation, but instead of calculating separately each expected probit (mortality) and then subtracting it from the observed probit (mortality), a short cut method adopted by Fisher may be used. The χ^2 may be calculated as

$$\chi^2 = [\,\Sigma(wy^2) - \bar{y}\,\Sigma(wy)] - b[\,\Sigma(wxy) - \bar{x}\,\Sigma(wy)]$$

and it depends upon its degrees of freedom n'. Sometimes determination of the proper degree of freedom is not always straight-forward. Procedures are also available for determining the variances of position and slope. Additionally one can calculate a zone of error of the regression line to supplement the calculated regression line which is the best available estimate of the true dosage-mortality curve. One then can make probability statements about whether the zone encloses the true dosage-mortality curve when transformed to the logarithmic-probit diagram. The case of zero survivors is an important problem in the case of survival data. This problem has been considered by R. A. Fisher. For experiments with no survivors, $x = \infty$ with weight

$$\frac{Z^2}{pq} \to Zx \to 0$$

for large samples. The equations appropriate for plotting the points on the probit diagram, namely

$$a = \frac{s}{n} \quad \text{and} \quad \frac{1}{2\pi} \int_x^{\infty} e^{-\frac{1}{2}t^2} = q$$

cannot be used in this form for fitting the regression line when the number of survivors is small. One must apply the Method of Maximum Likelihood to such cases. By introduction of a weighting deviate a procedure is then available so experiments with few or no survivors may exert their proper influence in adjusting the regression line. This correction is necessary because the omission of experiments because they show no survivors would constantly bias the estimates in the sense of exaggerating the number of survivors to be expected.

In general the probit technique involves the assumption that susceptibility is distributed normally, with inferred dosages, in terms of units called probits, plotted against the logarithms of their corresponding observed dosages giving a straight line relationship. The procedure or transformation is particularly adaptable and statistical methods are available for considering 0 or 100 per cent kills, for determining a weight proportional to its reliability, for computing the position and slope of the transformed dosage-mortality curve, for measuring the goodness of fit of the regression line to the observations by the χ^2

test, and for calculating the error in position and in slope and their combined effect at any log-dosage. Tables are available for determining the probit values in the probit or log-probit transformation.

TRANSFORMATIONS WITH FRACTIONAL POWERS OF THE VARIABLE

Frequently it is advantageous to transform a non-normal variable to approximately a normal variable for certain forms appearing in the analysis of variance. This is convenient because the normal distribution has been so exhaustively investigated and tabulated and also because certain models specify a normal distribution of variates. A transformation of importance is

$$y = X^r \qquad \text{where} \qquad 0 < r < 1$$

and X is a non-negative variate. One might also consider the transformation to a Type III variate and the use of $(X + \alpha)^r$ where α is a random variable in the interval $[0, 1]$. One can approach the problem by observing the effect of taking the $1/r$th power of a normal variate rather than considering the effect of the rth power of a general variate X. Therefore we would have

$$X = m(1 + k^{1/2} Z) \quad ,$$

where Z is a normal variable with mean zero and standard deviation unity. This would make X a normal variable with mean m and coefficient of variation (the standard deviation divided by the mean) equal to $k^{1/2}$. One assumes that m is large and positive and k is small so that

$$P\left\{X \leq 0\right\} < \epsilon$$

and ϵ is extremely small. We can find the moments by evaluation of

$$E(X^t) = m^t E(1 + k^{1/2} Z)^t$$

where t is greater than unity and has both integral and fractional values.

Thus for transformations of

$$r = 1/2,\ 1/3,\ 1/4,\ \ldots$$

the values of t needed will always be integral and thus exact values of the moments X^2, X^3, X^4, . . . may be obtained. From here the moments about the mean and the values of the moment ratios can be obtained. If r is not of this simple form, the numerator of the fraction is not unity, and hence t will be fractional and the moments of the transform variate can only be found as a power series in k. This limitation is not of serious consequence if k is small and one can omit the higher powers of the expansion. The results of the previous discussion have been carried out by (Moore, 1957) and are given in the following table form as functions of k. The first cases are approximations while the last three are exact. These can be graphed and from diagramatic form one would see that a change of r would possibly radically alter the value of k even though the (β_1, β_2) points of the transformed variate may be very close. For a practical application of the transformation one would make the first two moments of the transformed variate equal to those of a normal curve to insure a correct value of k was being utilized. It is frequently useful to examine the coefficient of variation of the transformed variate, $X^{1/r}$. One can then determine if the variable has both the correct momental ratios and the correct coefficient of variation to be transformed into an approximate normal distribution. Essentially the same procedure can be used for transformation to a X^2 or Type III distribution. One would have to utilize the moments for determining the proper constants. One should also consider the coefficient of variation c as well as the appropriate values of β_1 and β_2 for determination of the proper particular transformation. Many other distributions can be handled in a similar fashion.

Table 1

Momental Ratios of $X^{1/r}$ if X is Normal

r	β_1	β_2
3/4	$k(1 + 0.7963k + 0.336763k^2)$	$3(1 + 0.148148k + 0.021602k^2 + 0.003086k^3)$
2/3	$2.25k(1 + 0.291667k + 0.06423k^2)$	$3(1 + 0.6k + 0.16k^2)$
1/2	$8k(3 + k)^2(2 + k)^{-3}$	$3(4 + 20k + 5k^2)(2 + k)^{-2}$
1/3	$108k(3 + 16k + 15k^2)^2(3 + 12k + 5k^2)^{-3}$	$3 + 72k(7 + 48k + 75k^2 + 15k^3)(3 + 12k + 5k^2)^{-2}$
1/4	$162k(2 + 32k + 149k^2 + 198k^3 + 33k^4)^2 \times (2 + 21k + 48k^2 + 12k^3)^{-2}$	$3 + (4124k + 94201k^2 + 70735k^3 + 2042501k^4 + 1880525k^5 + 235008k^6) \times 8(2 + 21k + 48k^2 + 12k^3)^{-2}$

INVERSE HYPERBOLIC SINH TRANSFORMATION

In certain entomological experiments the variate may be one such that it cannot strictly be subjected to the analysis of variance. It is of concern then to determine how the variable may be transformed so that the analysis of variance becomes applicable.

One problem in entomological experiments is with a variate X which would be the number of insects in one of a group of small contiguous areas, say plots, within a larger area, say a block. If we let the expectation of X over all these plots be M and the standard deviation be σ; then over a number of these larger areas, the counts of insects are distributed in a completely random fashion, from the Poisson distribution, $\sigma^2 = M$. However, situations occur due to nature and changes are present in expectation from plot to plot within a block. Therefore, σ^2 will tend to be greater than M and we can only describe this variance by the statement

$$\sigma^2 = f(M) \quad .$$

This function, $f(M)$, must be considered carefully since it determines the transformation which may be developed to make the standard deviation independent of the mean.

One proposal to the problem is to state,

$$\sigma^2 = KM \quad ,$$

where K is a constant. However, in field data there does not tend to be a linear relationship between σ^2 and M or their respective estimates s^2 and $\bar{x}$, but there tends to be a disproportionately greater departure as $\bar{x}$ increases.

The preceding considerations would suggest that

$$\sigma^2 - M \propto M$$

is generally untrue. The curvilinearity might be represented by

$$\sigma^2 - M \propto M^2$$

which leads to the relationship that

$$\sigma^2 = M + kM^2$$

where k is a suitable constant. It will then be noted that

$$K = (\sigma^2 - M)M^{-2}$$

is the Charlier coefficient of disturbance from a Poisson distribution. If comparisons are made to consider the suitability of the relationship

$$\sigma^2 = M + kM^2 \quad ,$$

they will involve finding how they fit observations on s^2 and $\bar{x}$. The exact fit is difficult to determine and it is necessary to fall back upon pairs of estimates $\bar{x}$ and s^2 from the data for the empirical determination of K and k. Beull (1942) estimates

$$K = \Sigma s^2 / \Sigma \bar{x} \quad , \qquad k = (\Sigma s^2 - \Sigma \bar{x}) / \Sigma \bar{x}^2$$

where Σ represents the summation over all pairs.

Upon application to field data Beull has found that

$$\sigma^2 = M + kM^2$$

is generally a better approximation to the form of $f(M)$ and makes it preferable to proceed with the analysis of data from this assumption. The transformation appropriate to this type of data

$$X' = k^{-1/2} \sinh^{-1} (kx)^{1/2}$$

was developed by the method of Tippett (1934).

The adequacy of this transformation is judged by the extent to which it stabilizes variability. If in the transformation we express

$$\sinh^{-1} (kx)^{1/2} \qquad \text{when} \qquad kx < 1 \quad ,$$

as a well known series, we have

$$X' = x^{1/2} - \frac{1}{6} kx^{3/2} + \frac{3}{40} k^2 x^{5/2} - \frac{5}{112} k^3 x^{7/2} + \ldots ,$$

where for

$$k = 0 \quad , \qquad X' = x^{1/2} \quad .$$

For large values if kx, X' varies almost as $\log X$ or as $\log (X + 1)$ so the above expansion for practical purposes embraces the root and logarithmic transformations. Beall (1942) gives a table for the transformation for a probable range of observations of X and k at intervals close enough for practical purposes. Values of X' outside those can be calculated from

$$X' = k^{-1/2} \log_e \left\{ (kx)^{1/2} + (1 + kx)^{1/2} \right\} \quad .$$

When using field data in making the transformation, it is necessary to estimate the value of k empirically by

$$k = (\Sigma s^2 - \Sigma \bar{x}) / \Sigma \bar{x}^2$$

for which estimates $\bar{x}$ of the mean and s of the standard deviation must be found. The proposed transformation besides making the variability within a block for a repeated treatment the same for all treatments and block, should also provide quantities satisfying the assumptions underlying the analysis of variance. That is in the analysis of variance the chance variability for each plot shall be, when the effects of block and of treatment are removed, normally distributed with a standard deviation common to all plots, in which of course the standard deviation of the chance variability for a given plot is independent of the expectation for that plot. Diagramatically this can be compared and the proposed transformation should tend to make the standard deviation independent of the mean in accordance with the assumptions underlying the analysis of variance.

Several results may be noted or compared to suggest that the transformation proposed did tend to make the variability within a given treatment independent of the mean for that treatment: Some are: 1) Indication after transformation that the residual sum of squares about the regression was greater than the reduction in squares due to regression, whereas it would be consistently less before transformation. 2) As can be seen above, the regression generally did not effect a significant reduction in variability after transformation but did before (the small number of degrees of freedom made high significance difficult of attainment). 3) After transformation the sign of the regression will be a chance matter, whereas before transformation it was consistently positive. This result suggests that the proposed transformation did tend to make the variability within a given treatment independent of the mean for that treatment.

The mathematical basis for the preceeding transformation was suggested by the method used by Tippett (1934). The procedure is as follows. It is required to find $X' = f(x)$ such that the standard deviation σ_x' of X', shall be approximately constant. We could have

$$X' = f(M) + f'(M)(X - M) + \ . \ . \ .$$

where M is the expectation of X and hence approximately

$$(X' - M') = f'(M)(X - M)$$

where M' is the expectation of X'. Hence

$$\sigma^2_{X'} = \left\{f'(M)\right\}^2\sigma^2$$

where σ is the standard deviation of the observations, X. Replacing $\sigma_{X'}$ in the above equation by a constant, C, and substituting for σ we have

$$f'(M) = C(M + kM^2)^{-1/2}$$

where k is as has been previously discussed a constant peculiar to our data. Integrating the above equation one obtains

$$f(M) = 2Ck^{-1/2} \sinh^{-1} (kM)^{1/2} \quad .$$

From this integration, the proper transformation appears to be of the form

$$\sinh^{-1} (kx)^{1/2}$$

but it is wise instead to use

$$k^{-1/2} \sinh^{-1} (kx)^{1/2}$$

because this transformation becomes identical with the established transformation $X^{1/2}$, when $k = 0$. The derivation is such that the transformation can only be justified if the application is found satisfactory. The transformation should then put the data in a form where the standard deviation approaches a constant independent of the mean. The estimation of the constant k can be facilitated by the design of an experiment with a repetition of treatments within blocks.

MODIFIED AND SPECIAL TRANSFORMATIONS

Frequently biological populations change proportional to the mean, which infer that the changes would be independent of the mean on the logarithmic scale. This situation might be encountered in field plot experiments where the variate of interest would be the percentage of area covered by a specific crop. Often times an efficient comparison will involve a regression of this percentage area covered with time. If the percentage areas are less variable when near 0 (or 100) than when in the middle of the range, a suitable transformation might lead to improvement of the data. Bartlett (1947) has found that the transformation

$$y = \log (x/(100 - x))$$

where x is a percentage variate, summarized the time change and has the effect of changing the regression lines at 0 on the original scale. Analysis of variance would then be made upon linear regressions of the percentage variate with time. The transformation is an empirical one and rationalization of its use is implied by noticing that the change in percentage area depends upon the amount of plant species as well as the available area for expansion, or the competition of species causing it to decrease. For a constant cause of change, the percentage area X would depend upon an equation

$$\frac{dx}{dt} = Ax(100 - x) \quad .$$

By integration

$$\log \frac{x}{100 - x} = Bt + C \quad \text{or} \quad X = \frac{100}{1 + \exp -(Bt + C)}$$

which is the formula for the law of population growth.

Transformation of the sample correlation r will often make its disturbance less skew and more stable in variance. Because the variance of a correlation coefficient is approximately

$$(1 - \rho^2)^2 / n - 1 \quad ,$$

where ρ is the true value of the coefficient and n is the number of observations in the sample, one obtains the transformation

$$Z = 1/2 \log \left\{(1 + r)/(1 - r)\right\}$$

to make the variance independent of the mean. Although rare, it is possible to analyze correlation coefficients by the analysis of variance, and the above transformation would be the appropriate one.

The above transformation is of use because often reliable results are useful to:

(1) Test if an observed correlation differs significantly from a theoretical value.

(2) Test if two observed correlations are significantly different.

(3) Combine a number of independent estimates (if available) into an improved estimate.

(4) Perform tests (1) and (2) with such average values.

This transformation has the following characteristics. It leads approximately to a normal distribution where tests may be carried out without difficulty. As r changes from 0 to 1, Z will pass from 0 to ∞. For small values of r, Z is nearly equal to r, but as r approaches unity, Z increases without limit. For negative values of r, Z is negative. One advantage of this transformation of r into Z lies in the distribution of these two quantities in random samples. The standard deviation of r depends on the true value of the correlation ρ, as is seen from the formula

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n' - 1}} \quad .$$

Since ρ is unknown, we have to substitute for it the observed value r, and this value is not a very accurate estimate of ρ in small samples. The standard error of Z is simpler in form, approximately

$$\sigma_Z = \frac{1}{n' - 3}$$

and is practically independent of the value of the correlation from which the sample was drawn. Secondly, the distribution of r is not normal in small samples, and even remains far from normal for large samples and high correlations. The distribution of Z tends to normality for large samples and any value of the correlation. Finally, Z changes rapidly in distribution as ρ changes, whereas, Z is nearly constant in distribution and accuracy can be improved by small corrections for departure from normality although they are not necessary. Therefore one can assume that Z is normally distributed with sufficient accuracy in many cases.

SUMMARY AND CONCLUSIONS

The selection of a transform scale of measurement will depend upon (1) The nature of the data, and (2) the statistical procedures to be used. Choice of an appropriate transformation depends on the nature of the original variate. Often times one has to apply a transformation and then make appropriate tests to determine the effectiveness of the transformation. Sometimes it is not possible to normalize the data for certain types of variates.

It may be impossible to find a transformation that results in homogeneous error variances. However, if enough sample evidence is available, it is usually possible to determine a relationship between means and variances and to make them independent by an appropriate transformation. The standard or basic transformation are the square root, logarithmic, and angular or arcsine transformation. Many variations of these basic transformations are available. Transformations of this type have achieved new importance when considered with experimental designs and the analysis of variance. Bartlett (1947) has the following suggested summary of appropriate transformations.

Variance in terms of mean m	Transformation	Approximate variance on new scale	Revelant distribution
m	$\sqrt{X}$ or $\sqrt{X + \frac{1}{2}}$	0.25	Poisson
$\lambda^2 m$	for small integers	$0.25\ \lambda^2$	Empirical
$\frac{2m^2}{n - 1}$	$\log_e X$	$\frac{2}{n - 1}$	Sample variances
$\lambda^2 m^2$	$\log_e X$, $\log_e(X + 1)$	λ^2	Empirical
	$\log_{10} X$, $\log_{10}(X + 1)$	$0.189\ \lambda^2$	
$\frac{m(1 - m)}{n}$	$\sin^{-1}\sqrt{x}$ (radians)	$\frac{0.25}{n}$	Binomial
	$\sin^{-1}\sqrt{x}$ (degrees)	$\frac{821}{n}$	
$\frac{m(1 - m)}{n}$	Probit	not constant	
$\frac{m(1 - m)}{n}$	$\log_e\left[\frac{x}{1 - x}\right]$	$\frac{1}{nm(1 - m)}$	
$\lambda^2 m^2(1 - m)^2$	$\log_e\left[\frac{x}{1 - x}\right]$	λ^2	Empirical
$\frac{(1 - m^2)^2}{n - 1}$	$\frac{1}{2}\log_e\left[\frac{1 + x}{1 - x}\right]$	$\frac{1}{n - 3}$	Sample correlations
$m + \lambda^2 m^2$	$\lambda^{-1}\sinh^{-1}\left[\lambda\sqrt{x}\right]$ or	0.25	Negative binomial
$\mu^2(m + \lambda^2 m^2)$	$\lambda^{-1}\sinh^{-1}\left[\lambda\sqrt{x + \frac{1}{2}}\right]$ for small integers	$0.25\ \mu^2$	Empirical
. . . .	To expected normal scores	1 for large n	Ranked data

BIBLIOGRAPHY

Anscombe, F. J. "The Transformation of Poisson, binomial and negative-binomial data." Biometrics. 35:246. 1948.

Baker, G. A. "Transformations of Non-normal Frequency Distributions into Normal Distributions." Annals of Math. Stat. 5:113-123. 1934.

Bartlett, M. S. "A Modified Probit Technique for Small Probabilities." Journal of Roy Stat. Soc. Suppl. 8:113-117. 1946.

________. "The Square Root Transformation in the Analysis of Variance." Journal of Roy Stat. Soc. Suppl. 3:68-78. 1936.

________. "The Use of Transformations." Biometrics. 3:39-52. 1947.

Bartlett, M. S. and Kendall, D. G. "The Statistical Analysis of Variance-heterogeneity and the Logarithmic Transformation." Journal of Roy Stat. Soc. Suppl. 8:128-138. 1946.

Beall, Geoffrey. "The Transformation of Data from Entomological Field Experments so that the Analysis of Variance Becomes Applicable." Biometrika. 32:243-262. 1942.

Bliss, C. I. "The Calculation of the Dosage-mortality Curve." Annals of Applied Biology. 22:134-167. 1935.

________. "The Comparison of Dosage-mortality Data." Annals of Applied Biology. 22:307-333. 1935.

Box, G. E. P. "Non-normality and Tests on Variances." Biometrika. 40:318-335. 1953.

Claringbold, P. J., Biggers, J. D. and Emmens, C. W. "The Angular Transformation in Quantal Analysis." Biometrics. 9(4):467-483. 1953.

Cochran, W. G. "The Analysis of Variance When Experimental Errors Follow the Poisson or Binomial Laws." Annals of Math. Stat. 11:335- . 1940.

Curtiss, J. H. "On Transformations Used in the Analysis of Variance." Annals of Math. Stat. 14:107-122. 1943.

Federer, Walter T. Experimental Design. New York: Macmillan Company, 1955.

Fisher, R. A. Statistical Methods for Research Workers. London: Oliver and Boyd, 1950.

Fraser, D. A. S. Statistics, An Introduction. New York: John Wiley and Sons, 1958.

Graybill, Franklin A. An Introduction to Linear Statistical Models. Volume 1. New York: McGraw-Hill Book Company, 1961.

Healy, M. J. R. "The Planning of Probit Assays." Biometrics. 6(4):424-431. 1950.

Hotelling, Harold and Frankel, Lester R. "The Transformation of Statistics to Simplify Their Distribution." Annals of Math. Stat. 9(2):87-96. 1938.

Moore, P. G. "Transformations to Normality Using Fractional Powers of the Variables." Journal of the American Statistical Association. 52(278): 237-246. 1957.

Olshen, A. C. "Transformations of the Pearson Type III Distribution." Annals of Math. Stat. 9(3): 1938.

Rao, C. Rodhakrishna. Advanced Statistical Methods in Biometric Research. New York: John Wiley and Sons, 1952.

Snedecor, George W. Statistical Methods. Ames: Iowa State College Press, 1956.

Tukey, John W. "One-degree of Freedom for Non-additivity." Biometrics. 5(3): 232-242. 1949.

Wadsworth, George P. and Bryan, Joseph G. Introduction to Probability and Random Variables. New York: McGraw-Hill Book Company, 1960.

Wilks, Samuel S. Mathematical Statistics. New York: John Wiley and Sons, 1962.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

ACKNOWLEDGEMENT

The author wishes to express his appreciation to Dr. A. M. Feyerherm of the Department of Statistics, Kansas State University, for his helpful suggestions and assistance during the preparation of this report.

TRANSFORMATIONS OF RANDOM VARIABLES

by

ERNEST WAYNE PHELPS

B. S., Kansas State University, 1959

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

Approved by:

____
Major Professor

ABSTRACT

The method of transforming random variates arises from a need to satisfy certain assumptions used in the analysis of variance for numerical data. The procedures are particularly applicable if the variance and the mean of the original variate are functionally related. Application of an appropriate transformation is facilitated if this functional relationship is known.

This report considers some of the mathematical theory and expected results when a transformation is applied. A transformation considered and of interest is the square root transformation for a variate with a Poisson distribution and with the variance proportional to the mean. This type of variate is encountered in enumeration data, with counts for the variable. The square root transformation should, in this situation, produce variances independent of means.

A logarithmic transformation of the original variate is useful if the standard deviation of the variable tends to vary directly as the mean or if the effects in an experiment are proportional to the mean instead of additive. Graphical comparisons and applications of Tukey's test for non-additivity are helpful to determine if the transformation has provided the desired result.

An angular transformation is of consequence if the original variate tends to be distributed in a binomial fashion and if the proportions tend to vary over a rather wide range. An angular or inverse sine transformation should produce a variable with rather constant variance if it is effective. The Probit transformation is applicable to dosage-mortality and time-survival data in biological experiments. The probit technique is used in relation to regression techniques to determine lethal dosage in such experiments.

A fractional power transformation is useful in transforming a variate to a normal or other revelant distribution about which results are obtainable. The inverse hyperbolic sine transformation is applicable in entomological experiments where there is a curvilinear relationship between the mean and variance. This transformation should produce a standard deviation that approaches a constant independent of the mean if an analysis of variance is to be meaningful.

Modified and special transformation are applied in special situations. A number of these transformations are considered in this report. Finally there is a summary of transformations used in particular situations.