TRANSFORMATIONS FOR MIXTURES OF DISTRIBUTIONS

G. SADASIVAN

Institute of Agricultural Research Statistics, New Delhi

(Received in April, 1970)

1. INTRODUCTION

In setting up an analysis of variance we recognise three types of effects, (*a*) treatment effects, (*b*) environmental or block effects and (*c*) experimental error. The assumptions made in setting up the analysis of variance for the above three effects are (1) the treatment effects and environmental effects must be additive, (2) experimental errors must be independent of the other two, (3) the experimental errors must have a common variance (4) the experimental errors should be normally distributed. If any one of these assumptions is not satisfied suitable transformations are to be made to change the scale of measurements in order to make the analysis of variance valid. In most cases when the experimental errors do not have a homogeneous common variance, tranformation is applied to make it homoscedastic.

If the form of change in variance with mean level is known, the type of transformation to be used can be determined. Suppose we write $\sigma_x^2=f(m)$ where σ_x^2 is the variance on the original scale of measurements X with the mean of X equal to m ; then for any function $g(x)$ we have

$$\sigma_g^2=\left(\frac{dg}{dm}\right)^2 f(m)$$

so that σ_g^2 is constant say, c^2 and we have

$$g(m)=\int \frac{cdm}{\sqrt{f(m)}} \qquad \ldots(1)$$

This is an approximate formula which determines the type of transformation to be used for a particular type of data to make the analysis of variance valid. For the ideal transformation (*a*) the variance of the transformed variable should be unaffected by changes in the mean level, (*b*) the transformed variable should be normally distributed (*c*) the transformed scale should be one for which an arithmetic average is an efficient estimate of the mean level for any

particular group of measurements, (*d*) the transformed scale should be one for which real effects are linear and additive. Under these conditions Bartlett (1947) suggested :

(1) the square root transformation when data follows the Poison distribution ;

(2) logarithmic transformation of the simplest type when the mean level varies with the standard deviation. In such cases the variance is greater than the mean. This happens because the mean level itself fluctuates so that

$$\sigma_x^2 = m + \lambda^2 \sigma_m^2$$
$$= m + \lambda^2 m^2$$

for biological populations.

For large λ or m this variance law implies logarithmic transformation, a more exact transformation being

$$\lambda^{-1} \sinh^{-1} [\lambda \sqrt{x}] \quad \text{or} \quad \lambda^{-1} \log \{\sqrt{1+\lambda^2} x + \lambda \sqrt{x}.$$

This transformation holds for data following negative binomial distribution as well. For small x it becomes equivalent to $\sqrt{x}$ transformation, and for small number the transformation $\lambda^{-1} \sinh^{-1} [\sqrt{x+\frac{1}{2}}]$ is better. For large $\lambda\sqrt{x}$ it becomes equivalent to the log transformation. This transformation requires an approximate knowledge of λ. Whenever λ cannot be obtained $(1+x)$ can be used as an approximate transformation. It shows an approximate linear relationship with $\sinh^{-1}[\lambda\sqrt{x+\frac{1}{2}}]$ for the likely values of λ which appear in practice. "Beall", however, suggested that in entomological field experiments where an estimate of λ is required two plots of each treatments must be included in each randomised block. For such designs we can use the scale $\lambda^{-1} \sinh^{-1}[\lambda\sqrt{x}]$. The transformation is valid for the more general variance law $\sigma_x^2 = \mu^2(m + \lambda^2 m^2)$. When the distribution is log normal the change in the population variance is often proportional to the mean implying changes independent of the mean on logarithmic scale. In such cases transformations to the scale $\log\left(\frac{x}{1-x}\right)$ is useful. Correlations are transformed to $\frac{1}{2} \log \frac{1+r}{1-r}$. When the data follows the Binomial law, the angular transformation $g(x) = \sin^{-1}\sqrt{x}$ is used. The probit transformation is used for log-normal distributions. The transformation is given by $y = 5 + \left\{\frac{(x-\bar{x})}{\sigma}\right\}$. It converts the probability P in a normal distribution with mean 5 and variance 1 to the corresponding abscissa y. It is particularly useful when such a transformed quantity y is linearly dependent on another variable, x so that the transformation converts the functional relation between y and P to a straight line.

2. THE PROBLEM AND SOME RESULTS IN MIXTURES

One of the major problems attempted in the present paper is to find a suitable method of analysis when the data from a design follows different types of distributions in different ranges. The splitting in the practical situations would be automatic, as replicates, times etc. The distributions in such cases would be (*a*) Poisson (*b*) negative binomial, (*c*) log-normal, (*d*) contageous. Examples of such distributions are (1) numbers of field plots with 0, 1, 2, ...larva *e* (2) number of plates with 0, 1, 2 ... bacterial colonies etc. In such cases the distributions followed in each portion can be empirically checked by the chisquare test. For the same portion of the data different types of distributions can be checked up. The one giving the highest probability for x^2 can be selected as adequate. Further, the forms of the distribution in different portions may be the same ; but they may vary in the parameters concerned. In such cases also different transformations can be used in different portions of the design. Another way is to use different transformations which a priori seem to be appropriate for the data and later on check the transformed data for near normality or for rapid convergence to normality. If the data are near normal these specifications are good enough. If they are not normal we have to try other specifications. Some times it may happen that the distributions of the whole or part are the resultant of mixtures of distributions in the probability sense. In such cases the mixtures can be due to addition or multiplication of the component variables. In some other situations mixtures can be due to a weighted average of the cumulative distribution functions. They may even be the μ-mixtures of the cumulative distribution functions of Teicher. Some of these mixtures will asymptotically converge to standard distributions as (1) the negative binomial (2) the log-normal (3) the contageous types. The transformations to be used in such and other cases to render analysis of variance valid are investigated in this paper.

With this object in view we review some of the results from the theory of mixtures of distributions. The simplest mixture occurs when each observation is the sum or product of two component variables U_i and V_i. The distribution function $F(x)$ of the sum of two independent variables is given by

$$F(x) = F_1(x) * F_2(x).$$

This symbolic representation of the distribution functions corresponds to a genuine multiplication of characteristic functions. If both the components belong to the discrete type, the composite fuction is also discrete. When both functions are of the continuous type and at least one of the frequency functions F_1 is bounded for all x and may be expressed as a Riemann integral

$$f(x) = \int_{-\alpha}^{+\alpha} f_1(x-z)\, f_2(z)dz = \int_{-\alpha}^{+\alpha} f_2(x-z)\, f_1(z)dz$$

the compound distribution belongs to the continuous type and the compound frequency function is continuous everywhere. Depending on the basic distributions such compound distributions will take different forms. The mean for such distributions is m_1+m_2 and variance $\sigma^2=\sigma_1^2+\sigma_2^2$. For higher moments about the mean a general expression is :

$$\mu_\nu = E\{(\xi-m_1+\eta-m_2)^\nu\}.$$

It easily follows from the above that a compound distribution of two normal variables will itself be normal. Further, the bionomial distribution reproduces itself by the addition of independent variables.

The term mixture may also mean a genuine weighted average of cumulative distribution functions. Let $T=\{F\}$ be a family of one dimensional cumulative distribution functions and let $m=\mu$ be a class of measures defined on (a) a Borel field of subsets of F, with $\mu(F)=1$ for all $\mu\in m$. Then $\int g(F)d\mu(F)$ is defined in the usual manner for measurable mappings g of T into the realline.

If $g=g_x(F)=F(x)$ this becomes

$$H=H(x)=\int_T F(x)d\mu(F) \qquad \ldots (2)$$

The resultant distribution function H is called a mixture or more specifically μ=mixture of T, provided the mixing measure μ does not assign measure one to a particular member of T. For a stipulated T, the family $H=H(T)$ of mixtures H, swept out as μ varies over m, will be called the class of m mixtures of T or simply the class mixtures of T. In particular the family T may be indexed by a finite number of parameters $\alpha_1, \alpha_2, \ldots \alpha m$ each αi varying over the real line

$$T=\{F(x, \alpha_1, \alpha_2, \ldots \alpha m)\}.$$

If $G=\{G(\alpha_1, \alpha_2, \ldots \alpha m)\}$ denotes the class of m-dimensional c.d. f.s and $F(x; \alpha_1, \alpha_2 \ldots \alpha m)$ be measurable on $(m+1)$ dimensional Euclidean space R^{m+1}, then, m may be taken to be the class of measures $\{\mu_G\}$ on R^m induced by $G\in G$ and (2) becomes :

$$H(x)=\int_{R_m} F(x; \alpha_1, \alpha_2 \ldots \alpha m)\, d\, G(\alpha_1, \alpha_2 \ldots \alpha m).$$

If $G(\alpha_1, \alpha_2, \ldots \alpha m)=\prod_{i=1}^{m} Gi(\alpha_i)$ the mixture is termed a product measure mixture. We can speak of a discrete or absolutely continuous mixture according as $G(\alpha_1, \alpha_2 \ldots \alpha m)$ (or μ) is a discrete or absolutely continuous c.d.f. (measure).

We detail below the results pertaining to mixtures of specific distributions. Contageous distributions are mixtures of distributions.

The compound Poisson distributions are precisely mixtures of Poisson distributions which are necessarily discrete distributions with jumps at the non-negative integers. In the course of determining limit distributions of sums of interchangeable random variables mixtures of normal distributions are encountered. In certain situations one is interested in the distribution of a random variable X, but knows only the conditional distributions of X given the values of some auxiliary random variable Y. Then the desired distribution of X is a mixture of the known conditional distributions. A family $F=F(x, \alpha)=F(x, \alpha_1, \alpha_2,...\alpha_m)$ where α_j varies over an additive abelian group $Djg(j=1, 2, .. m)$ is called additively closed, if for every admissible α, β,

$$(F(x\,;\,\alpha)*F(x\,;\,\beta)=F(x, \alpha+\beta)$$

where * as usual denotes the convolution operation. The families of Normal, Poisson, Binomial and many other distributions are encompassed within this definition. Teicher (1960) obtained the following general results for mixtures of distributions (1) An infinitely divisible mixing (G) of an additively closed family (T) yields an infinitely divisible mixture (H). (2) The convolution of two Compound Poisson distributions is again a Compound Poisson distribution whose mixing c.d.f. is the convolution of two mixing c.d.f.'s (3) The convolution of two mixtures of symmetric stable distributions of fixed exponent β is again a mixture of the same type with mixing c.d.f. the convolution of the given mixing c.d.f.'s (4) No mixture of symmetric stable distributions with fixed exponent β, $(0<\beta\leqslant 2)$ is a symmetric stable distribution with exponent β.

Sometimes the underlying distribution of interest to the statistician will be generated by selecting one of a collection of alternative normal distributions according to some scheme. If the resulting mixture of normal distributions is itself normal regular analysis of variance can be performed. Considering mixtures of the two parameter family of normal distributions with parameters θ and σ^2 the following results emerge.

1. Suppose the conditional distribution $\mu\{\sigma^2 \mid \sigma^2>\sigma_0^2\}=0$. Then a sufficient but unnecessary condition that a μ-mixture of normal distributions be normal with mean θ_0 and variance σ_0^2 is that the conditional distributions of θ given σ^2 is normal with mean θ_0 and variance $\sigma_0^2-\sigma^2$ for all values of σ^2 for which it is defined.

2. Further in order that a μ-mixture of normal distributions be normal with mean $\theta_0=0$ and variance $\sigma_0^2=1$, it is necessary that :

(*i*) $\mu\{\sigma^2/\sigma^2>1\}=0=\mu\{\theta,\ \sigma^2/\sigma^2=1,\ \theta\neq 0\}$. Hence it may be supposed that $\mu\{\sigma^2/\sigma^2\geqslant 1\}=0$.

(*ii*) $\mu\left\{\theta, \sigma^2\Big/\frac{\theta}{1-\sigma^2}>c\right\}.\mu\left\{\theta, \sigma^2\Big/\frac{\theta}{1-\sigma^2}<-c\right\}>0$, all $c>0$.

(*iii*) $\int_{R^2} exp\left\{\frac{\theta^2}{2(1-\sigma^2)}\right\} d\mu=\infty$,

(*iv*) $\mu\{\theta/ \mid \theta \mid < e^{-\frac{1}{2}}\} : \mu\{\theta, \sigma^2/\theta^2 \leqslant \sigma^2 \log_e 1/\sigma^2\} > 0.$

(*v*) the θ-spectrum of μ be not confined to a subset of members in arithmetic progression ; further, for all integers m (all real b) and all integers $n \geqslant 1$,

$$\mu\left\{ \theta/\bar{r} \sum \bigcup_{j=0}^{n-1} \left[\frac{8j+1}{8n}, \frac{8j+3}{8n}\right]\right\} < 1,$$

where $\bar{r}$ significes the fractional part of r and either $r = \theta - m/n$ or $r = b\theta$.

3. A product-measure mixture of normal distributions is normal with mean θ_0 and variance σ_0^2 if and only if for some σ_1^2 in $(0, \sigma_0^2)$, $G_{\sigma_1^2}(\theta) = \phi(\theta ; \theta_0, \sigma_0^2 - \sigma_1^2)$ and $G_1(\sigma^2)$ is degenerate at σ_1^2. From this it follows that a mixture of normal distributions with identical means cannot be normal. It is intuitively plausible that no countable mixture of normal distributions is normal and if the variances σ_j^2 have a minimum this is indeed the case. Moreover, it can also be shown that a countable mixture of normal c.d.f.'s cannot be normal if the means are a set of numbers in arithmetic progression; but a countable mixture of normal distributions can be arbitrarily close to a normal distribution. The relationship:

$$\int_0^\infty \left[\frac{1}{2\sqrt{\pi\alpha}} e^{-x^2/4\alpha}\right] e^{-\alpha} d\alpha = \tfrac{1}{2} e^{-|x|}$$

reveals that an exponential mixture of normal distributions is a Lapalce distribution.

Some other results are as follows. The Gamma distribution is a negative binomial mixture of commonly scaled but differently exponented Gamma c.d.f.'s A mixture of Poisson distributions is linked with the moment problem. Under certain conditions the mixture follows the compound Poisson distribution.

3. Some new results

When data from a design is divided group-wise into different blocks or in a haphazard manner, where the group formations would be automatic or otherwise, the distribution followed by each may be specified and tested. If the specifications are of the standard types, appropriate transformations can be employed to render the distributions normal. The normal variates in each portion can be reduced to the standard normal form. Then the composite data can be analysed in the usual manner using analysis of variance and suitable inferences drawn. This follows from the fact that the joint distribution is standard normal in the whole range for if there be K groups each with density

$$p(x)dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

the density in the composite group would be :

$$p'(x)dx=\frac{1}{k}\cdot\frac{k}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

The result is true for composite mixtures of the type discussed here. If the distributions are normal in the different groups, the densities would take the form :

$$p(x)dx=\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(x-\mu_i)^2/2\sigma_i^2}dx \qquad (i=1, 2, \ldots k)$$

and the density for the composite group would be :

$$p'(x)dx=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i\sqrt{2\pi}}e^{(x-\mu_i)^2/2\sigma_i^2}dx$$

Thus the pooling at any particular variate value $x=x_1$ will not generate a normal distribution ; but if the pooling is about

$$t=\frac{x-\mu_1}{\sigma_1}=\frac{x-\mu_2}{\sigma_2}=\frac{x-\mu_k}{\sigma_k},$$

the distribution would be normal. If all the means are equal

$$p'(x)dx=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(x-\mu_i)^2/2\sigma_i^2}dx$$

and if all the variances are equal

$$p'(x)dx=\frac{1}{\sigma k\sqrt{2\pi}}\sum_{i=1}^{k}e^{-(x-\mu_i)^2/2\sigma^2}dx$$

These are all non-normal and hence direct transformations for mixtures of this type are difficult to obtain.

When the mixtures occur due to combination of different variables into a single observation we get the following :

1. If the individual distributions are Binomials with means np_1 and np_2 and variances $np_1(1-p_1)$ and $np_2(1-p_2)$, the mixture in the additive sense is a Binomial with mean $n(p_1+p_2)=m$ and variance

$=c^1+m-\frac{m^2}{n}$ where $c^1=2np_1p_2$. The transformation in this case is given by

$$g(m)=\int \frac{cdm}{\sqrt{c^1+m-\frac{m^2}{n}}}$$

$$=\int \frac{\sqrt{n}\, cdm}{\sqrt{\frac{n^2}{4}+nc^1-\left(m-\frac{n}{2}\right)^2}}$$

$$=\sqrt{n}\, c \sin^{-1}\left\{\frac{m-\frac{n}{2}}{\sqrt{\frac{n^2}{4}+nc^1}}\right\}.$$

Thus the angular transformation is useful for the compound Binomial distribution as well.

2. The Laplace distribution is given by :

$$p(x)=\tfrac{1}{2}e^{-|x|}\ (-\alpha<x<\alpha)$$

The mean of the distribution is 0 and the variance is 2. Since there is no functional relationship between the mean and variance, transformation by the approximate function above is not possible.

3. For the negative Binomial distribution

$$E(a_x)=N\frac{(k+x-1)!}{x!\ (k-1)!}\frac{p^x}{(1+p)^{k+x}}$$

is expressed in terms of parameters p and k. Here the mean is $\bar{x}=pk$ and variance $s^2=p(p+1)k$. Thus $s^2=c^1\bar{x}$. Hence square root transformation can be an approximation for this distribution. We can as well put $\mu_2=\mu_1'+\frac{\mu_1'^2}{k}$ which leads to Beall's transformation:

$$x^1=q^{-\frac{1}{2}}\ \sinh^{-1}(qx)^{\frac{1}{2}}\ .$$

4. A random variable ξ will be said to have a rectangular distribution, if the frequency function is a constant equal to $\frac{1}{2h}$ in a certain finite interval $(a-h,\ a+h)$ and 'o' outside this interval. If the distribution ranges from a to b, the mean is

$$m=\frac{a+b}{2}\ \text{and}\ \sigma^2=\frac{(b-a)^2}{12}=km.$$

If data follows this type of distribution, the transformation for analysis of variance works out to be :

$$g(m)=\frac{c}{\sqrt{b}}\sqrt{m}.$$

5. In Branching Processes there is the compound Poisson distribution. The generating function of the sum $S_N = X_1+X_2+\ldots+X_N$ is the compound function $g(f(s))$ where $f(s)=\Sigma f_i s^i$. Two special cases are of interest. (*a*) If the $X_i s$ are Bernoulli variables with $P\{Xi=1\}=P$ and $P\{xi=0\}=q$, then $f(s)=q+ps$. (*b*) If N has a Poisson distribution with mean θ, then : $h(s)=e^{-t+if(s)}$

The distribution with this generating function is called the compound Poisson distribution. If the X_i's are Bernoulli variables and N has a Poisson distribution, then $h(s)=e^{-tp+tps}$, the sum S_N has a Poisson distribution with mean tp. Thus $m=tp=s^2$. Hence

$$f(m)=\int \frac{cdm}{\sqrt{m}}=cm^{\frac{1}{2}}.$$

Hence in cases where the distribution of the observations conforms to the compound Poisson type, square root transformation will bring the data to the normal form approximately.

6. Some data from Entonological, Micological and Microbiological experiments will follow different types of contageous distributions. This may be so in different sub-divisions of the experiment, where the subdivision would be according to time, block etc. Transformations are required for such data in order to make analysis of variance valid. Contageous distributions are also mixtures of distribtions. The generating function of the contageous distribution is:

$$\psi(z)=e^{-m_1}e\left\{m_1 n\,!\sum_{s=0}^{\infty}\frac{m_2^s(z-1)^s}{(s+n)\,!}\right\}$$

Putting $n=0, 1, 2$, successively we get the generating functions for types A, B and C respectively. They are :

$$\psi_1(z)=e^{-m_1}e^{m_1[e^{m_2(z-1)}-1]} \text{ for type } A.$$

$$\psi_2(z)=e^{-m_1}e^{m_1[e^{m_2(z-1)}-1-m_2(z-1)]}/m_2(z-1)$$

for type B and

$$\psi_3(z)=e^{-m_1}e^{2m1[e^{m_2(z-1)}-1-m_2(z-1)]}/m_2^2(z-1)^2$$

for type C.

For the general characteristic function the moments are obtained as :

$$\mu_1{}^1 = m_1 m_2/(n+1)$$

$$\mu_2 = \frac{m_1 m_2}{n+1}\left(1+\frac{2m_2}{n+2}\right).$$

From these

$$\mu_2 = \mu_1' + \frac{2(n+1)}{(n+2)m_1}\,\mu_1'^2 \qquad \ldots(3)$$

Moreover,

$$\mu_2 - \mu_1' = 2m_2\mu_1'/(n+2),$$

so that the second moment μ_2, will always the greater than $\mu_1{}^1$ under our condition that n and m_2 are positive. We have the further result that :

$$\mu_3 = \frac{m_1 m_2}{n+1}\left\{1+\frac{6m_2}{n+2}+\frac{6m_2{}^2}{(n+2)(n+3)}\right\}$$

so that, since $m_1 \geqslant 0$, $\mu_3 > \mu_2$. Further

$$m_2 = (n+2)(\mu_2 - \mu_1{}^1)/2\mu_1{}^1$$

$$m_1 = (n+1)\mu_1'/m_2$$

and again n which gives the type of the distribution is seen to be

$$n = \frac{6(\mu_2{}^2 + \mu_1{}^1\mu_2 - \eta_1'\mu_3 - \mu_1'^2)}{\mu_1'^2 + 2\mu_1'\mu_3 - 3\mu_2{}^2}.$$

Relation (3) shows that Beall's transformation

$$x^1 = q^{-\frac{1}{2}} \sinh^{-1}(qx)^{\frac{1}{2}}$$

where $$q = \frac{2(n+1)}{(n+2)m_1}$$

can be used under the assumption that the variance law

$$\sigma x^2 = m + \lambda m^2$$

holds approximately. This assumes that the distributions are uni-parametric. But we also find the following results :

(*a*) For type A distribution

$$\mu_2 - \mu_1' = 2m_2\mu_1'/2 \qquad m_2 > 0 \qquad \mu_1' > 0.$$

Thus $\mu_2 \propto \mu_1'$ and hence square root transformation can be used as an approximation.

(*b*) Type B

$$\mu_2 - \mu_1' = 2m_2\mu_1'/3$$

$$\mu_2 = \mu_1'(1+\tfrac{2}{3}m_2).$$

Here also square root transformation can be used as an approximation. The exact transformation is ;

$$g(x)=\frac{-c}{2\sqrt{1+\frac{2}{3}m_2}}\,x^{1/2}.$$

(*c*) Type C

$$\mu_2-\mu_1'=2m_2\mu_1'/4 \text{ whence}$$

$$\mu_2=\mu_1'\left(1+\frac{m_2}{2}\right).$$

Hence here also square root transformation is feasible under the assumption that the distribution is uniparametric.

Again for type A distribution :

$$\mu_1'=m_1m_2\ ;\ \mu_2=m_1m_2(1+m_2)$$

$$=\mu_1'+\lambda^2\mu_1'^2 \text{ where } \lambda=\frac{1}{\sqrt{m_1}}.$$

Thus type A obeys the second variance law if we assume that $\frac{1}{\sqrt{m_1}}$ =a constant. This gives the transformation

$$\lambda^{-1}\sinh^{-1}[\lambda\sqrt{x}] \text{ or } \lambda^{-1}\log\{\sqrt{1+\lambda^2}x+\lambda\sqrt{x}\}.$$

We further find that $\mu_3=\mu_1'\{1-3m_2+m_2^2\}$ which gives the skewness of Type A distribution in terms of the mean and the parameter m_2, Further,

$$\mu_3>\mu_2 \quad \text{and it is positive and}$$

$$\mu_3-\mu_2=\mu_1'm_2(2+m_2).$$

Transformations can also be based on this property which is left out as futher problem for investigation. It may be noted that a transformation designed to make the third moment zero will normalise the distribution.

7. When the different types of mixtures of normal distributions approach the normal form exactly or approximately, no transformation of data follwing such distributions are necessary to render analysis of variance valid. But in cases of non-normality appropriate transformations are to be used. As a pre-requisite the exact forms of the mixtures are to be first evaluated.

8. For the Gamma distribution given by :

$$\phi(x)=\frac{e^{-x}\,x^{L-1}}{\sqrt{L}}$$

$$E(x)=L\ ;\ \sigma_x^2=L.$$

Hence square root transformation is valid. For the β distribution of the first kind given by the density

$$\phi(x)=\frac{x^{L-1}(1-x)^{m-1}}{\beta(L,\ m)}$$

$$E(x)=\frac{L}{L+m} \text{ and } \sigma_x^2=\frac{Lm}{(L+m)^2(L+m+1)}$$

Thus $$\mu'_2=\frac{\mu_1'^2}{L\left(1+\frac{L}{\mu_1'}\right)}$$

If we assume that L is a constant $\mu_2 \propto \mu_1'^2$ and hence as a first approximation we can use logarithmic transformation for such cases.

An Example

The method suggested above has been applied to the data of an experiment conducted at the Indian Agricultural Research Institute, New Delhi on two different varieties of Jowar *viz* : Co-1 and I.S.-84, in order to Study the effect of 15 insecticides on the control of Sorghum Shoot fly. The treatments are given in Appendix I. The design used was R.B.D. with four replications for each variety. Appendix II gives the total plant counts per plot and Appendix III, the counts of dead hearts, Appendix IV gives the percentage of dead hearts. From this, frequency distributions for percentage dead hearts are formed for each variety and graphs drawn. From the graphs (Appendix VI) and by appropriate testing for specifications it is seen that the distributions under each variety belong to two negative binomial populations with means 9·82 and 17·23 and variances 117·00 and 230·39 respectively. Although the distributions in the two parts are of the same form. it is a case of two different negative binomials and is hence a mixture. The mixture of the two negative binomials is non-normal and so we proceed on the following lines for the analysis.

By using the transformation $\lambda^{-1}\log\sqrt{1+\lambda^2}x+\lambda x$ the two distributions have been transformed to bring them to normality by calculating the value of λ from the relation.

$\sigma_x^2=m+\lambda^2m^2$. Appendix VII shows the transformed data. The transformed data have been standardised and analysed. The results are shown below :

Analysis of Variance Table

Source	*d.f.*	*S.S.*	*M.S.*	*F*
Between replications within groups	6	1·4802	0·2467	0·647
Between varieties	1	0·0000	0·0000	—
Between treatments	14	79·3615	5·6686	14·87
Varieties x treatments	14	5·2565	0·3754	0·984
Error	84	32·0260	0·3812	—
Total	119	118·1242	—	—

Conclusion : The treatments differ significantly.

Table of treatment means in descending order of magnitude is shown below along with the critical difference.

T_{15}	T_1	T_{14}	T_9	T_8	T_7
—1·207	—1·194	—1·156	—0·931	—0·837	—0·145

T_{12}	T_3	T_4	T_2	T_5	T_{10}	T_{13}	T_6	T_{11}
0·941	0·915	0·896	0·788	0·755	0·376	0·358	0·221	0·208

$C.D. = 0·617$

Conclusions

From the above practical example it can be concluded that in situations where the data belong to different distributions the methods described in this paper can be applied for analysing the data. Thus if the first group had followed contageous distribution and the second binomial distribution, the corresponding transformations could have been used and the data standardised as a prelude to analysis of variance. When the data follow composite mixtures of distributions the appropriate transformations suggested for such cases can be utilised. The Laplace distribution is not amenable to any transformation. For contageous distributions square-root transformations are adequate. For the α-function square-root transformation and for the β-function logarithmic transformation can be used. The angular transformation is useful for the compound binomial distribution. Square-root transformation is valid also for compound Poisson distributions.

Abstract

Frequently, data collected from Entomological and Mycological experiments will often conform to different types of distributions in

different sub-divisions of the design. In certain other situations the distributions for whole or part of the design will be mixtures of component distributions. The paper treats some of the types of transformations which can be used in such situations. A numerical application is also given.

REFERENCES

1. Eisenhart (1947) : Assumptions underlying analysis of variance. Biometrics Vol : 3.
2. W.G. Cochran (1947) : Some consequences when the assumptions for analysis of variance are not satisfied. Biometrics Vol : 3.
3. M.S. Bartlett (1947) : The use of transformations Biometrics Vol : 3.
4. H. Cramer (1946) : Mathematical Methods of Statistics.
5. Henry Teicher : On the mixtures of distributions (1960). Annals of Mathematical Stattstics. Vol. 31, No. 1.
6. Neyman (1939) : On a new class of contageous distributions applicable in Entomology and Bacteriology. Annals of Mathematical Statistics. Vol. 10.

Appendix I

The following are the particulars of the experiment considered –

Design	–R.B.D.
Number of Treatment	–15
Varieties	–(2)C.O. I and IS – 84

Treatments

(1)	Themet Grannels	0·5 gms/furrow.
(2)	Control	fungicide alone
(3)	Menazon	0·25 gms/100 gms of seeds
(4)	Menazon	1·50 gms/100 gms of seeds
(5)	Menazon	1·00 gms/100 gms of seeds
(6)	Themet L.C. 80	3·00 gms/100 gms of seeds
(7)	Themet L.C. 80	6·00 gms/100 gms of seeds
(8)	Solberox Gr. 5%	0·75 gms/furrow
(9)	Solberox Gr. 5%	1·50 gms/furrow
(10)	Telodrin 20%	10·00 gms/100 gms of seeds
(11)	Telodrin 20%	15·00 gms/100 gms of seeds
(12)	Dieldrin 20%	8·00 gms/100 gms of seeds
(13)	Aldrin 40%	8·00 gms/100 gms of seeds
(14)	47470 25%	2 kgs/hectare
(15)	47470 25%	4 kgs/hectare

Appendix II

Total Plant Count of 2 different varieties of shoragum as affected by treatments

	Co−1				IS−84			
Treatment Number	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
1.	17	10	32	3	43	44	24	34
2.	52	58	28	36	59	89	76	39
3.	46	48	48	53	68	52	74	60
4.	44	43	32	60	77	77	32	57
5.	59	40	50	44	55	57	56	31
6.	9	14	6	7	31	25	40	37
7.	9	4	7	11	30	40	14	6
8.	24	15	26	33	55	81	54	55
9.	12	19	20	28	37	70	48	29
10.	31	41	65	30	25	72	58	25
11.	44	52	65	53	70	74	58	63
12.	13	6	11	18	30	15	31	31
13.	54	49	57	48	55	64	62	56
14.	51	54	57	50	78	86	71	76
15.	37	55	48	53	76	59	66	64

Appendix III

Total dead hearts in different replications of 2 varieties of Jowar under different treatments

	Co–1				IS–84			
Treatment Number	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
1.	0	0	0	0	0	0	0	2
2.	11	19	6	17	24	20	25	4
3.	12	14	6	11	20	25	23	20
4.	12	15	9	16	32	22	5	18
5.	16	5	8	23	18	15	5	6
6.	2	5	1	0	8	5	9	6
7.	0	1	0	0	2	11	1	1
8.	0	1	0	2	0	2	4	0
9.	0	0	0	1	0	1	2	0
10.	1	13	8	3	7	25	10	0
11.	13	12	7	7	28	35	32	20
12.	0	0	1	3	14	6	4	3
13.	5	0	10	2	11	14	16	6
14.	0	0	2	0	0	1	0	1
15.	1	1	0	0	0	0	0	0

Appendix IV

Percentage of total dead hearts caused by A Indica observations in each replication of the two soragham varieties

Treatment Number	*CO-1*				*IS-84*			
	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
1.	0·00	0.00	0·00	0 00	0·00	0 00	0·00	5·88
2.	21·15	17·24	12·00	19·44	40·67	22·47	32·89	9·52
3.	28·57	20·83	8·33	20·75	29·41	39·68	31·08	33·33
4.	27 27	23·91	22·50	26·66	81·56	28·57	13·88	31·58
5.	28·07	12·19	16·00	28·88	32·72	26·31	8·93	18·18
6.	20 00	33·33	0·00	0·00	40·00	19·23	4·25	16·21
7.	0·00	16·66	0·00	0·00	8·00	27·50	5·55	16·66
8.	0·00	0·00	0·00	5·00	0·00	2·74	7·41	0·00
9.	0·00	0 00	0·00	3·33	0·00	1·41	4·50	0·00
10.	2·94	31·70	7·35	8·57	23·07	34·72	20·63	0·00
11.	29·54	22·64	10·77	11·29	40·00	47·30	47·06	31·74
12.	0·00	0 00	9·09	13·33	46·66	37 50	8·33	12·90
13.	9·25	0·00	16·66	4·17	16·92	23·33	25·80	10·71
14.	0·00	0·00	0·00	0·00	0·00	1·30	0·00	1·39
15.	0·00	0·00	0·00	0·00	0 00	0·00	4·76	0 00

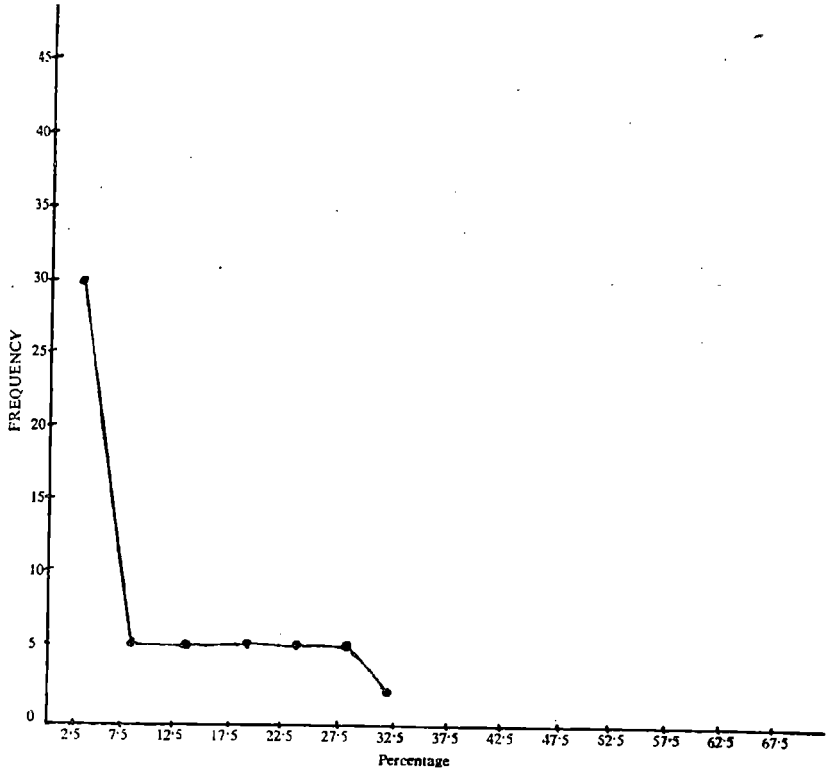
Appendix V

Frequency distribution of the percentage data

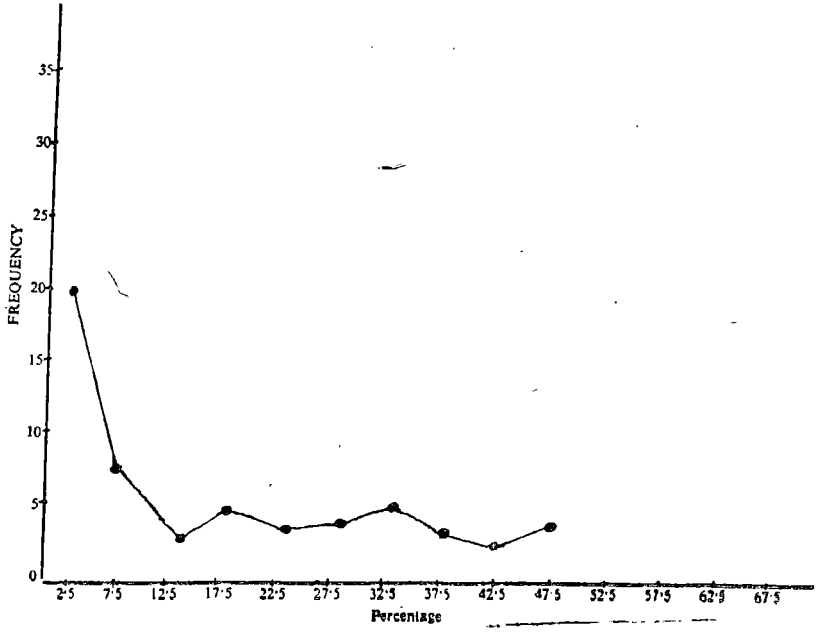
Variety I (Co I)		*Variety II (IS-84)*	
Class limits %	*Frequency*	*Class limits %*	*Frequency*
0-5	30	0-5	20
5-15	5	5-10	7
10-16	5	10-15	3
15-20	6	15-20	5
20-25	6	20-25	4
25-30	6	25-30	5
30-35	2	30-35	7
	—	35-40	4
	60	40-45	2
	—	45-50	3
			—
			60

Appendix VI

Frequency Polygons for the Variety CO-I



Frequency Polygons for Variety TS-84



Appendix VII

Table showing the transformed variates

Mean for first variety=1·055

Mean for the second variety=0·847

Treatment Number	*CO-1*				*IS-84*			
	R_1	R_2	R_3r	R_4	R_1	R_2	R_3	R_4
1.	0·000	0·000	0 000	0 000	0.000	0·000	0 000	1·732
2.	2.162	2·068	1·902	2·123	2·819	2·478	2 697	1 994
3.	2·303	2·155	1·736	2·155	2·632	2·805	2·664	2·703
4.	2·282	2·220	2·192	2·273	2·830	2·616	2·204	2·673
5.	2·296	1·909	2·033	2·307	2·694	3·568	1·957	2·356
6.	2·137	2·374	0·000	0·000	2·809	2·388	1·563	2·291
7.	0 000	2·052	0·000	0·000	1·897	2·593	1·702	2·307
8.	0·000	0·000	0·000	1·508	0·000	1·345	1·856	0·000
9.	0·000	0 000	0·000	1·335	0·000	1·045	1·531	0 000
10.	1·283	2·351	1·679	1·748	2·492	2·809	2 464	0·000
11.	2·319	2·194	1·851	1·874	2·809	2·908	2·904	2·676
12.	0·000	0 000	1·775	1·950	2·899	2·773	1·921	2·162
13,	1·782	0·000	5·052	1·430	2·317	2·499	2·556	2·059
14.	0·000	0·000	6·000	0·000	0·000	1·011	0.000	1·038
15.	0·000	0 000	0 000	0 000	0·000	0 000	1·621	0·000