Nematol. medit. (1979), 7: 107-115.

Biometric Unit of the Agricultural Research Centre, Merelbeke, Belgium
State Nematology and Entomology Station, Merelbeke, Belgium

ESTIMATION OF THE POPULATION MEAN AFTER TRANSFORMATION OF THE DATA; AN APPLICATION FOR ZOOLOGICAL DATA

by

R. MOERMANS and W. A. COOLEN

Biological data are often transformed before statistical analysis is carried out on them (Cancela da Fonseca, 1965; Proctor and Marks, 1975). The purpose of such transformation is mainly to establish the homogeneity of the variances and the normality of the distribution.

A frequently used transformation is $X_i = \log_a Y_i$ ($i = 1, 2 \ldots n$) with Y_i the i^{th} original observation, X_i the corresponding transformed value and a the base of the logarithmic system. However, a difficulty arises when interpreting the calculated mean $\overline{X}$, because the scale in which it is expressed is not always very meaningful either for the research worker himself or for the practitioner. Therefore, besides the value of $\overline{X}$, the value obtained by inverse transformation is also given. The latter is calculated by taking antilog ($\overline{X}$) or $a^{\overline{X}}$, which is the geometric mean. This estimate of the true mean of the population is however biased. This bias can be very important, in particular when a number of estimates obtained from independent sets of observations are averaged in order to estimate the average effect (Neyman and Scott, 1960).

To obtain an unbiased estimate of the population mean, a correction should be carried out. The term « unbiased » can be defined as: « ... the expectation of the estimate is the true value » (Kempthorne, 1952). It should be noted that the mean of the original observations is indeed an unbiased estimate of the true mean, although this estimate is not efficient. Efficiency means: « ... the variance of the distribution of the estimate relative to some standard » (Kempthorne, 1952). Although from a statistical point of view the problem of inverse

transformation has already been studied, the practical application seems not to be used frequently by research workers in zoological disciplines.

The purpose of this paper is to demonstrate the procedure of an inverse transformation with correction. First a simple experimental design is theoretically described, after which the necessary mathematical relations are given.

It should be noted that Aitchison and Brown (1957) and Thöni (1969) have published tables for the transformation $X_i = \log_e Y_i$ and $X_i = \log_{10} Y_i$ respectively for the case of a single sample. These tables cannot be used when several samples are combined in one analysis. Furthermore, these tables are not always easily accessible for the research worker. Because small electronic calculators are now readily available, the calculations can easily be done by the research worker himself.

THEORETICAL ASPECTS

Experimental design

The described experimental design is a completely randomized one. Suppose that p treatments must be compared for their effects on populations of soil inhabitants (e.g. the effect of p nematicides on the number of nematodes infesting a plant). For the first treatment there are n_1 experimental units (plots, pots ...) in the experiment; for the second treatment the number of experimental units is n_2 ...

If the original counts are represented by Y_{ij} ($i = 1, 2, \ldots n_j$; $j = 1, 2, \ldots p$) and if a transformation $X_{ij} = \log_e Y_{ij}$ is carried out, the X_{ij} values can be shown as in table I. Furthermore suppose that the transformed data within each treatment follow a normal distribution with parameters m_j ($j = 1, 2, \ldots p$) and σ^2, the latter being constant for each treament.

The transformed data can be analyzed using an analysis of variance, for which the model is given in table II, with $N = n_1 + n_2 + \ldots n_j + \ldots n_p$.

Table I - *Experimental design.*

Treatment	1	2 . . .	j . . .	p
	X_{11}	X_{12}	X_{1j}	X_{1p}
	X_{21}	X_{22}	X_{2j}	X_{2p}
	.	.	.	.
	.	.	.	.
	X_{i1}	X_{i2}	X_{ij}	X_{ip}
	.	.	.	.
	.	.	.	.
	$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_j j}$	$X_{n_p p}$

Table II - *Model of analysis of variance.*

Source	Degrees of freedom	Sum of squares	Mean square
Between treatments	$p - 1$	$T_2 = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (\overline{X}_j - \overline{X})^2$	$s_2^2 = T_2/(p-1)$
Within treatments	$N - p$	$T_1 = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$	$s_1^2 = T_1/(N-p)$
Total	$N - 1$	$\sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{ij} - \overline{X})^2$	

Basic principle

Using the properties of the expectations and the properties of the normal distribution (e.g. Hoel, 1971) it can be shown that:

$$E\ [\bar{Y}_{ij}] = M_j = \exp\ (m_j + \sigma^2/2)$$

E stands for the expectation and M_j for the true mean of the j^{th} population. The abbreviation exp stands for the number e to the power. The problem is to find a relation, starting from $\bar{X}_j$ and s^2 as estimators of respectively m_i and σ^2, such that the expected value of this relation equals M_j. In mathematical expression:

$$E\ [\exp\ (\bar{X}_j) \cdot f(t)] = M_j \qquad (1)$$

For a single sample of n observations, Finney (1941) has shown that such a function can be defined as a series which can be expanded as:

$$1 + \frac{n-1}{n}(s^2/2) + \frac{(n-1)^3}{n^2 \cdot 2!} \cdot \frac{(s^2/2)^2}{n+1} + \frac{(n-1)^5}{n^3 \cdot 3!} \cdot \frac{(s^2/2)^3}{(n+1)\cdot(n+3)} \ldots \qquad (2)$$

Oldham (1965) has shown that if several independent samples are combined in the analysis then the expansion of the series is given by:

$$1 + s_1^2\,(1 - 1/n_j)/2 + \frac{1}{2!}\frac{V}{V+2}(s_1^2\,(1 - 1/n_j)/2)^2 + \frac{1}{3!}\frac{V}{(V+2)\,(V+4)}(s_1^2\,(1 - 1/n_j)/2)^3 \ldots \qquad (3)$$

In (3), V is the number of degrees of freedom on which s_1^2 is based. It can easily be shown that (3) is identical with (2) if only one single sample is considered, for in this case:

$$n_j = n\ (j = 1, 2 \ldots p) \text{ and } V = n - 1.$$

ESTIMATION OF THE NUMBER OF *RADOPHOLUS SIMILIS* NEMATODES IN *CALATHEA MAKOYANA* PLANTS

From four *Calathea makoyana* plants, 2 g roots were taken in 4, 8, 8 and 6 replicates respectively and the numbers of *Radopholus similis* (Cobb) Thorne determined. These numbers are represented in table III; the same table also shows the transformed data, using natural logarithms. The observations are taken from a larger set from Heungens *et al.* (1971). The problem is twofold:

(i) can a significant difference be shown between the plants?

(ii) what is the mean number of nematodes in each plant, as an unbiased estimate of the true number?

Table III - *Numbers of* Radopholus similis *in 2 g roots. Original and transformed data* ($\log_e Y_{ij} = X_{ij}$).

Plant	1	2	3	4
Number of observations	4	8	8	6
	6	13	4	49
	1.791759	2.564949	1.386294	3.891820
	80	42	24	50
	4.382027	3.737670	3.178054	3.912023
	125	53	43	124
	4.828314	3.970292	3.761200	4 820282
	778	85	52	160
	8.656727	4.442651	3.951244	5.075174
		116	62	542
		4.753590	4.127134	6.295266
		133	221	718
		4.890349	5.398163	6.576470
		175	234	
		5.164786	5.455321	
		200	1880	
		5.298317	7.539027	

The first problem can be solved by carrying out an analysis of variance according to the general outline of a completely randomized design with unequal numbers of replicates.

Before the analysis is carried out, the homogeneity of variances is tested by the procedure of Bartlett (1947). The calculated χ^2 yields a value of 21.85 with three degrees of freedom. The probability of exceeding such a value (under the hypothesis of equal variances) is 0.00007, which is too small to accept equality of variances. If the data are transformed, using natural logarithms, the calculated χ^2 yields a value of only 4.10. The corresponding probability of exceeding that value does not exceed 0.252. The reduction of χ^2 shows that the transformation of the data was adequate, as far as the homogeneity of variances is concerned. Table IV gives the variances of the samples for original and transformed variables, and the results of the Bartlett test. Table V shows the analysis of variance of the transformed data.

Table IV - *Variances and results of the Bartlett-test.*

	Original data	Transformed data
Plant 1	127606.13	4.026371
Plant 2	4328.32	0.818871
Plant 3	407592.86	3.314664
Plant 4	81054.09	1.311666
Calculated X^2	21.852	4.090
Degrees of freedom	3	3
Probability	0.000070	0.251890

Table V - *Analysis of variance.*

Source	D.f.	Mean square	$F_{calc.}$	Prob.
Plants	3	0.827038	0.38	0.769
Error	22	2.162350		

Although no significant differences between the plants can be shown, as the probability of the calculated F-value is too large, there still remains the problem of obtaining an unbiased and efficient estimate of the true mean value of the number of *R. similis* for each plant.

Using (3) with $s_1^2 = 2.1623$ and $V = 22$ and n being 4, 8, 8 and 6 respectively, the value of f(t) can be calculated. Substitution of f(t) in (1) finally gives the estimates of the population means. Table VI shows the successive terms of f(t).

Table VI - *Successive terms of* f(t).

Plant 1	Plant 2 and 3	Plant 4
1.000000	1.000000	1.000000
0.810881	0.946028	0.900978
0.301367	0.410194	0.372058
0.068925	0.109451	0.094548
0.010978	0.020339	0.016733
0.001305	0.002822	0.002211
0.000121	0.000305	0.000228
0.000009	0.000026	0.000019
.	.	.
.	.	.
.	.	.
2.193586	2.489165	2.386775

These terms rapidly converge to very small values, e.g. values which have no longer a significant impact on the value of f(t). Moreover, if an experimental design is used with equal numbers of replicates, the calculations are considerably reduced as only one series of f(t) must be calculated.

CONCLUSIONS

Table VII gives the different calculated means, respectively the arithmetic mean of the original data, the arithmetic mean of the transformed data, the geometric mean, and the inverse transformed

mean with correction. The geometric mean always yields a value less than the arithmetic mean of the original data and the corrected mean. As already mentioned in the introduction, the geometric mean is a biased estimation and in such way that it is an underestimation.

Table VII - *Mean values.*

Plant	Original data $\overline{Y}$	Transformed data $\overline{X}$	Geometric mean $\exp(\overline{X})$	Corrected mean $\exp(\overline{X}).f(t)$
1	247	4.414707	82	181
2	102	4.352825	77	193
3	315	4.349556	77	193
4	274	5.095173	163	390

Furthermore it should be noted that due to an extreme value in the sample (e.g. plant 3) the mean of the original value is large. The effect of this can be reduced by using a logarithmic transformation.

Returning to the original scale, the correction deals with the bias. Extreme values also can produce mean values which are quite different when working with the original data although the geometric means have the same value, e.g. plants 2 and 3. The fact that the corrected means also yield the same value is explained by the equal numbers of replicates with plants 2 and 3.

SUMMARY

The basic principles are given for estimating the mean of a population after transforming the data to a logarithmic scale. These principles are applied to an experiment dealing with *Radopholus similis* and *Calathea makoyana* plants. The different mean values are compared and it is shown that the geometric mean, which is frequently used, is an underestimation of the true mean.

RIASSUNTO

Stima di popolazioni per mezzo di trasformazione dei dati: un esempio pratico per dati zoologici.

Vengono illustrati i princìpi di base per stimare popolazioni per mezzo di trasformazione dei dati in scala logaritmica. Viene dato un esempio pratico

sulla interrelazione esistente tra il nematode *Radopholus similis* e la specie vegetale *Calathea makoyana*. I diversi valori medi sono raffrontati tra loro e viene dimostrato che la media geometrica, frequentemente usata, dà una stima per difetto rispetto alla media reale.

LITERATURE CITED

AITCHISON J. and BROWN I. A., 1957 - The lognormal distribution. Cambridge University Press. 176 pp.

BARTLETT M. S., 1947 - The use of transformation. *Biometrics, 3*: 39-52.

CANCELA DA FONSECA J. P., 1965 - L'Outil Statistique en Biologie du Sol. I. Distribution de Fréquences et Tests de Signification. *Rev. Ecol. Biol. Sol, 2*: 299-332.

FINNEY D. J., 1941 - On the distribution of a variate whose logarithm is normally distributed. *R. Statist. Soc., Suppl.* 7: 155-161.

HEUNGENS A., MOERMANS R. and DE GRISSE A., 1971 - Spreiding van *Radopholus*-aaltjes binnen de wortels van *Calathea*-planten. *Mededel. Fakul. Landbouw, 36*: 1319-1321.

HOEL P. G., 1971 - Introduction to Mathematical Statistics. 4th ed. Wiley. 409 pp.

KEMPTHORNE O., 1952 - The design and analysis of experiments. Wiley. 631 pp.

NEYMAN J. and SCOTT E. L., 1960 - Correction for bias by a transformation of variables. *Ann. mathemat. statis., 31*: 643-655.

OLDHAM P. D., 1965 - On estimating the arithmetic means of lognormally distributed populations. *Biometrics, 21*: 235-239.

PROCTOR J. R. and MARKS F. C., 1975 - The determination of normalizing transformations for nematode data from soil samples and of efficient sampling schemes. *Nematologica, 20*: 395-406.

THÖNI H. - A table for estimating the mean of a lognormal distribution. *J. Am. Statist. Ass. (Amer), 64*: 632-636.

Accepted for publication on 6 January, 1979.