A computer program for determining optimal data transformations minimizing skew*

WILLIAM P. DUNLAP† and JOHN A. DUFFY
Tulane University, New Orleans, Louisiana 70118

A program is presented that solves for an optimal transformation of data that minimizes skew. Various possible transformations are represented by a function that depends upon a single parameter, λ. A process of iterative approximations is employed to find a value of λ that will transform the data so as to produce a third moment equal to zero. The transformation found by this process may be used as computed or used to indicate which of the commonly employed transformations will most adequately counteract skew. The use of this program in processing data before analysis of variance is discussed.

Transformation of experimental data may serve a number of purposes, the most common of which is to make data more amenable to statistical analysis. A general discussion of the use of transformation for data can be found in Mueller (1948); a discussion of transformation aimed specifically at counteracting problems of skew can be found in Burros (1951). Most of the so-called parametric statistical procedures assume certain basic properties of the data analyzed, such as normality and homogeneity of variance across groups. In addition to mathematical requirements for statistical tests, the interpretation of experimental outcomes may be greatly simplified if other properties of the data exist, such as linearity of relationships between variables or absence of interaction. Transformations have been proposed to attempt to remedy all of these possible deficiencies in data (Federer, 1963). By and large, an E has simply tried various recommended transformations until he found one that seemed to work. Some statisticians have approached the problem of finding workable transformations in a more orderly manner (e.g., Anscombe & Tukey, 1963; Box & Cox, 1964). These authors have first selected a family of possible transformations that can be represented by a one (or more) parameter function, then they have solved for the value of the parameter(s) that minimizes the undesirable aspect of the data.

The present paper and accompanying computer program employ this approach in finding an optimum transformation to minimize skew, a very common deviation from normality occurring in experimental data. Positive skew appears to be the rule rather than the exception when one measures latencies or frequencies of rarely occurring events; fortunately, a tendency for skew is easily detected by simple examination of the raw data. Although violation of the normality assumption does not usually produce serious departures in Type I error rates, skewed distributions often have the additional problem of nonindependence between the variance and the mean; thus, the assumption of homogeneity of variance is often violated in conjunction with violation of the assumption of normality. The simultaneous violation of several assumptions is more likely to have severe effects on Type I errors with parametric tests (Box, 1953; Boneau, 1960). Previously suggested methods of finding a satisfactory transformation for reducing skew have involved plotting of means against variances and empirically trying various transformations to reduce the relationship between them (Bartlett, 1947). The transformations commonly used to correct positive skew are the logarithm, the square root, and the reciprocal of the raw data.

The present procedure involves the use of a family of transformations, suggested by Box and Cox (1964), that depends upon the single parameter, λ.

$$Y = \frac{X^{\lambda} - 1}{\lambda}$$

where X represents the raw data and Y the transformed score. When $\lambda = -1$, this function produces the reciprocal transformation, when $\lambda = 0$ the logarithmic transformation, when $\lambda = \frac{1}{2}$ the square root, and when $\lambda = 1$ the data are unchanged, except in a linear manner that has no effect on skew.[1] The measure of skew employed is the dimensionless third moment (DM3 in the program), which is zero for symmetric distributions but takes on increasingly positive or negative values depending upon the severity and direction of skew:

$$DM3 = \frac{\Sigma (Y - \bar{Y})^3 N}{(N-1)(N-2) S^3}$$

where Y is the mean, S the standard deviation of Y, and N the sample size. The following program (Table 1), written in FORTRAN IV, converges iteratively on a value of λ, such that the third moment equals zero.

*The authors wish to acknowledge the support of the Tulane University Computer Center in developing and testing this program.

†Reprints requests should be addressed to William P. Dunlap, Department of Psychology, Tulane University, New Orleans, Louisiana 70118.

The program works as follows. First, the sample size and data are read, and the data are searched for zero or negative values; if any are found, the absolute value of the minimum data value plus one is added to each score before proceeding, since at certain values of λ the transformation is defined only for positive data. Next, the third moment of the raw data (λ = 1) is calculated and printed along with the associated normal deviate.[2] All scores are then converted to logarithms (λ = 0) and the third moment recomputed. Having computed the third moment associated with two values of λ, the next step is to predict, by way of a straight line fit to these points, the value of λ that will produce a third moment of zero. The data are transformed by this predicted λ value and the corresponding third moment is calculated. The λ values associated with the last two third moments calculated are then used for a second approximation of λ. This process of approximation is repeated until a λ is found whose associated third moment has an absolute value less than 0.0001. When this converging process was tested with 276 samples of 100 scores, each having various extremes of skew, the number of iterative approximations required per sample did not exceed 5, and the average number was 3.35 (SD = 0.79). Thus, convergence was rapid, requiring only seconds of computer time.

In practice, this program may be used in several ways. The actual λ value produced may be used to transform the data for purposes of analysis with the assurance that skew has been minimized. Another use of the λ value, however, is to indicate which of the commonly recommended transformations will best counteract skew. If λ is near unity, the data do not need correction for skew; if λ is in the vicinity of −1, then the reciprocal transformation is recommended; a λ near 0.5 would argue for the square-root transformation; and a λ near zero would suggest logarithms as the most effective transformation.

When applied to data from experiments where analysis of variance (ANOVA) is the appropriate statistical technique, an additional problem is encountered: the fact that several groups of data are available to be corrected for skew, yet only one transformation can be used for the entire data set. One approach to this problem is to calculate the residuals from each group or cell of the design, then pool these residuals for the purpose of finding a value of λ. Examination of the residuals from ANOVA designs has been discussed by Anscombe and Tukey (1963). Another way of approaching the ANOVA problem is to find λs group by group. If all group λ values fall in the vicinity of a commonly used transformation, then the problem is solved. If they are more widely separated, then some average λ value should be used to determine the appropriate transformation. If the values of λ change drastically from group to group (or cell to cell), it may indicate that the effect of the treatment(s) is to influence the shape of the distribution rather than simply the mean; thus, a test that is sensitive to distribution shape rather than to just central tendency would be recommended. Certain nonparametric tests might suit the data more adequately.

The important reason that recommends the present approach for discovering an appropriate transformation to counteract skew is that it avoids the commonly employed practice of trying different transformations until one works. The criterion in the latter approach is often a transformation that produces a significant test statistic rather than the desired criterion, that of minimizing skew.

Table 1

```
C
C     DETERMINATION OF LAMDA VALUE MINIMIZING SKEW.
C
C     INPUT.
C          CARD 1 - SAMPLE SIZE PUNCHED IN COLUMNS 1 - 4.
C          CARD 2 - VARIABLE FORMAT CARD.
C          DATA CARDS.
C           (THIS SEQUENCE OF INPUT MAY BE REPEATED AS OFTEN AS DESIRED.)
C     A FINAL BLANK CARD WILL TERMINATE THE PROGRAM.
C
C     OUTPUT.
C          THIRD MOMENT OF THE RAW DATA.
C          STANDARD NORMAL DEVIATE FOR THE SKEW IN THE RAW DATA.
C          LAMDA.
C
      DIMENSION X(1000),Y(1000),FMT(16)
    1 READ(5,2)N
    2 FORMAT(I4)
      IF (N.EQ.0) CALL EXIT
      XN = N
      READ(5,3)FMT
    3 FORMAT(16A5)
      READ(5,FMT)(X(I),I=1,N)
C          CHECK FOR ZERO OR NEGATIVE VALUES IN THE RAW DATA.
      XMIN = X(1)
      DO 4 I = 2,N
      IF (XMIN.GT.X(I)) XMIN = X(I)
    4 CONTINUE
      IF (XMIN.GT.0.0) GO TO 7
      XMIN = 1.0-XMIN
      WRITE(6,5)XMIN
    5 FORMAT(1X,10H THE VALUE,F12.3,44H HAS BEEN ADDED TO PRODUCE ALL PO
     1SITIVE DATA)
      DO 6 I = 1,N
    6 X(I) = X(I)+XMIN
C          FIND THE THIRD MOMENT OF THE RAW DATA (LAMDA = 1).
    7 TM1 = DM3(X,N)
      STM = SQRT(6.0*XN/((XN-1.0)*(XN-2.0)))
      Z = TM1/STM
      WRITE(6,8) TM1,Z
    8 FORMAT(29HODIMENSIONLESS THIRD MOMENT =,F12.3/26H STANDARD NORMAL
     1DEVIATE =,F12.3)
C          TAKE LOGS OF RAW DATA (LAMDA = 0).
      DO 9 I = 1,N
    9 Y(I) = ALOG(X(I))
      TM2 = DM3(Y,N)
      XL1 = 1.0
      XL2 = 0.0
C          SOLVE FOR THE PREDICTED LAMDA.
   10 APL = (TM1*XL2 - TM2*XL1)/(TM1 - TM2)
C          TRANSFORM THE DATA VIA THE PREDICTED LAMDA.
      DO 11 I = 1,N
   11 Y(I) = X(I)**APL
C          FIND THE THIRD MOMENT OF THE TRANSFORMED DATA.
      TMA = DM3(Y,N)
C          CHECK THE THIRD MOMENT AGAINST THE CRITERION.
      IF (ABS(TMA).LT.0.0001) GO TO 12
      XL1 = XL2
      XL2 = APL
      TM1 = TM2
      TM2 = TMA
      GO TO 10
   12 WRITE(6,13) APL
   13 FORMAT(8HOLAMDA =,F12.3///)
C          IF THE USER DESIRES THE TRANSFORMED DATA, WRITE OUT ARRAY Y.
      GO TO 1
      END
      FUNCTION DM3(X,N)
C          CALCULATES THE DIMENSIONLESS THIRD MOMENT OF N VALUES IN ARRAY X.
      DIMENSION X(1)
      XB = 0.0
      SD = 0.0
      DM3 = 0.0
      XN = N
      DO 1 I = 1,N
    1 XB = XB+X(I)
      XB = XB/XN
      DO 2 I = 1,N
      D = X(I)-XB
      SD = SD+D**2
    2 DM3 = DM3+D**3
      SD = SQRT(SD/(XN-1.0))
      DM3 = DM3*XN/((XN-1.0)*(XN-2.0)*SD**3)
      RETURN
      END
```

REFERENCES

Anscombe, F. J., & Tukey, J. W. The examination and analysis of residuals. Technometrics, 1963, 5, 141-160.

Bartlett, M. S. The use of transformations. Biometrics, 1947, 3, 39-52.

Boneau, C. A. The effects of violation of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Box, G. E. P. Non-normality and tests on variances. Biometrika, 1953, 40, 318-335.

Box, G. E. P., & Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society, Series B, 1964, 26, 211-243.

Burros, R. H. Three rational methods for the reduction of skewness. Psychological Bulletin, 1951, 48, 505-511.

Federer, W. T. *Experimental design, theory and application.* New York: MacMillan, 1963.

Fisher, R. A. *Contribution to mathematical statistics.* New York: Wiley, 1950.

Mueller, C. G. Numerical transformations in the analysis of experimental data. Psychological Bulletin, 1948, 46, 198-223.

NOTES

1. That the transformation $(x^{\lambda} - 1)/\lambda$ produces $\log_e x$ when $\lambda = 0$ can be seen by replacing the term x^{λ} with its exponential series equivalent.

$$\frac{x^{\lambda}-1}{\lambda} = \frac{1}{\lambda}\left[1 + \lambda \log_e x + \frac{(\lambda \log_e x)^2}{2!} + \frac{(\lambda \log_e x)^3}{3!} + \ldots . - 1\right]$$

$$= \log_e x + \frac{\lambda (\log_e x)^2}{2!} + \frac{\lambda^2 (\log_e x)^3}{3!} + \ldots$$

When $\lambda = 0$, this expression equals $\log_e x$. In the program, division by λ was omitted, since this extra step results only in a linear transformation of data that has no influence on skew.

2. Fisher (1950) has reported that the sampling distribution of the third moment about zero is approximately normal, with a standard error of $[6N/(N-1)(N-2)]^{1/2}$ for fairly large samples; therefore, the normal deviate allows an approximate test for the significance of skew.

(Received for publication October 27, 1973;
revision received December 1, 1973.)