KM 41 (1992)
pg.19-45

On the use of normal scores for ordinal and censored variables in PRELIS
An expository note

Anne Boomsma

1. Introduction

The purpose of this paper is to explain and exemplify the role of normal scores in the treatment of ordinal and censored variables in PRELIS. We largely follow the approach from the PRELIS manual, elaborate where necessary, and make corrections occasionally. We slightly depart from the PRELIS notation. Guidelines for further readings are provided. Right now, for general treatments of censored and truncated variables the reader is referred to Schneider (1986) and to Cohen (1991).

2. Ordinal variables

In the PRELIS manual Jöreskog & Sörbom (1988, p. 1-5ff.) handle ordinal variables as follows. First, for ordinal random variables X, observations $X = x$ are assumed to represent responses of subjects to a set of ordered categories. Next, for each ordinal variable X, it is assumed that there is an underlying continuous variable ξ, having a standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Thus throughout, for ordinal variables $\xi \sim N(0,1)$. Jöreskog (1991b, p. 1) notices that these underlying variables are not the same as latent variables in linear structural equation models.

In general, the ordinal variable X has k categories. The statement $X = i$ means that i is the realization of the random variable X, i.e. the observation of X belongs to category i, $i=1,2,\ldots,k$. Jöreskog & Sörbom (1988) stress that the actual values assigned to the categories of ordinal variables are often arbitrary, or even irrelevant as long as the ordinal information is retained. It thus does not matter, as they put it, which values are assigned to the categories as long as low scores correspond to low-order

---

Vakgroep Statistiek & Meettheorie, Rijkuniversiteit Groningen,
Grote Kruisstraat 2/1, 9712 TS Groningen, telefoon 050-636187.

categories of X, which are associated with relatively small values of ξ, and high scores correspond to high-order categories of X associated with larger values of ξ.

As an example, the relation between the observed ordinal variable X with k = 6 categories, and an underlying, latent normal variable ξ, is visualized in Figure 1, where parameters c_i, i=0,1,...,k, are so-called **threshold values** for the latent variable ξ.

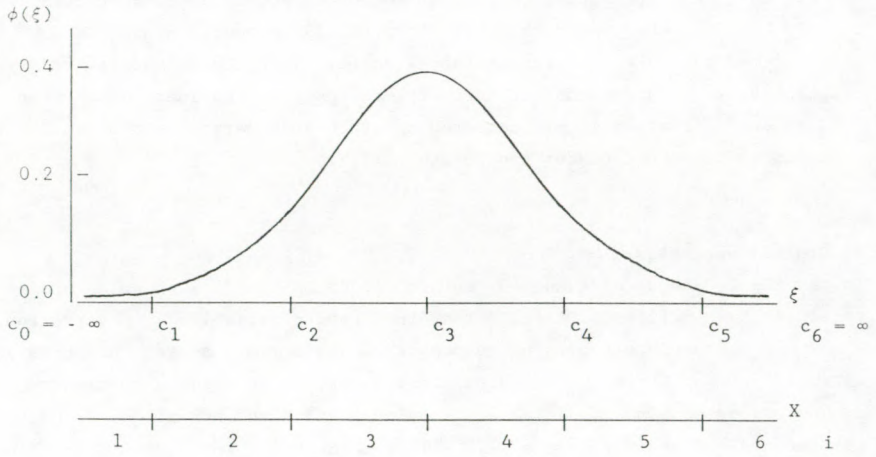

Figure 1. The relationship between a latent normal variable ξ and an associated, observed ordinal variable X.

Any observation of the ordinal random variable X may fall in one and just one of the six different categories. Once more, the statement X = i primarily means that the observation belongs to category i. The observed variable X could take arbitrary values i=1,2,...,6 on an ordinal scale, but those values could, for whatever reason, as well be 10, 12, 16, 20, 40, and 100.

In general the relationship between X and ξ is defined as

$$X = i \quad \text{if} \quad c_{i-1} < \xi \leq c_i \quad , \quad i=1,2,\ldots,k \quad , \tag{1}$$

with $c_0 = -\infty < c_1 < c_2 < \ldots < c_{k-1} < c_k = \infty$. Clearly, if the observed random variable X has k categories, there are k-1 unknown thresholds for the latent variable. It should be understood that equation (1) holds by assumption. It is precisely this type of relationship the researcher should have in mind, if he specifies the scale type of variables to be ordinal in PRELIS. At any time, he should also stand up for the plausibility of that model. In practice, this implies that it has to be reasonable to asssume that $\xi \sim N(0,1)$.

In this respect the question may arise whether it is possible to perform a statistical test of the hypothesis $\xi \sim N(0,1)$, given a sample of ordinal observations $X_1, X_2, \ldots, X_N$, where N is the sample size. The answer is that this hypothesis cannot be falsified on the basis of a sample of observed category frequencies of X, because the normality assumption of ξ in the ordinal variable case imposes no real restriction on the distribution of the ordinal variable X. As will be shown later on, by definition there is never any discrepancy between the sample of ordinal data $X_1, X_2, \ldots, X_N$ and this normality assumption, because it is always possible to estimate the k-1 unknown threshold values c_i, in such a way to assure a perfect correspondence between the sample data and the underlying standard normal variable ξ.

However, consider for example the case of two ordinal variables X_1 and X_2, each with its own underlying variable ξ_1 and ξ_2, assuming bivariate normality with correlation $\rho(X_1,X_2)$; i.e. the case of polychoric correlation. It is then possible to construct a likelihood ratio test statistic that can be used to test the assumption of underlying bivariate normality (for details, see Jöreskog, 1991b). PRELIS gives the associated χ^2-statistic to test such a model of bivariate normality (it also does so in the case of polyserial correlation, where one of the pair of variables is declared continuous).

In conclusion: a univariate normality model cannot, but a bivariate normality model can be falsified, when dealing with ordinal variables assuming underlying normal variables.

In the ordinal variable case a **new continuous** (normal score) variable can be created, conditional upon an assumed underlying standard normal ξ. In Section 8 an outline is given on how to compute these normal scores.

3. Censored continuous variables

Often researchers are dealing with continuous variables with a limited range of observed values. For example, the total score on a test consisting of 20 dichotomous items lies within the inclusive range from 0 to 20. Now consider the case where the discriminating power of the test is too small for a specific population of examinees. Within the range 1 to 19, the scale is taken to be continuous; the observed scores are rounded to integers, and the roundings are supposed to be negligible. If the test is too easy, there will be many examinees who respond correctly to all 20 items. If the test is too difficult, there will be many subjects having all items wrong. In the first case the total score variable is said to be **censored above**; although many examinees have an observed score of 20, it is not likely that the abilities measured are the same for all these individuals. In the latter case the test is **censored below**; subjects having a test score of 0 are not supposed to have all the same ability. Other examples of censored variables can be found in Maddala (1983, Chapters 1 and 6).

Variables may not only be censored above (sometimes the term ceiling effect is used), or below (floor effect), but also be **doubly censored**. The PRELIS program can deal adequately with these type of variables.

A sample of observations of a continuous censored variable Z_1, Z_2, ..., Z_N is called a **censored sample**, where N is the sample size.

Before giving a formal definition, both for singly and doubly censored variables, notice the following carefully. First of all, in PRELIS censored variables are supposed to be continuous, not ordinal, variables. Secondly, in all censored cases an observed random variable Z, which represents a latent random variable ξ, is considered. By assumption ξ has a normal distribution with unknown mean μ and unknown variance σ^2. Thus throughout $\xi \sim N(\mu,\sigma^2)$ for censored variables. Given the relationship between Z and ξ, as defined below, this assumption imposes a real constraint to the probability distribution of $Z = \xi$ on the range of observed values of Z. This implies that, within that range, there could very well be a discrepancy between the sample data Z_1, Z_2, ..., Z_N and the assumption of normality.

The normality assumption for continuous censored variables is a strong one: it imposes a direct constraint to variable Z. In contrast, the normality assumption on ξ in the ordinal variable case, as described in the

previous section, imposes no real restriction on the univariate distribution of the ordinal variable X.

Three cases of censoring are described now. In each case the observed random variable Z, representing a latent, underlying random variable $\xi \sim N(\mu,\sigma^2)$, is considered.

a. Censored below. Above a **lower threshold value B** for ξ, variable Z is observed on an interval scale; above B it is assumed that Z reflects the latent variable ξ accurately. Below B, the value Z = B is observed. In principle, the value of the threshold B is supposed to be known: it equals the smallest value of the observed variables $Z_1, Z_2, \ldots, Z_N$ in a sample of size N (with high probability, unless N is small). Thus a variable Z that is censored from below (as determined by the threshold value B) is defined as

$$\begin{aligned} Z &= B \quad \text{if} \quad \xi \leq B \ . \\ Z &= \xi \quad \text{if} \quad \xi > B \ . \end{aligned} \tag{2}$$

It follows that for the observations Z = B, all that is known is that the latent variable $\xi \leq B$, i.e. $Pr\{Z = B\} = Pr\{\xi \leq B\}$.

b. Censored above. Below an **upper threshold value A** for ξ, variable Z is observed on an interval scale; below A it is assumed that Z reflects the latent variable ξ accurately. Above A, the value Z = A is observed. Again, the value of the threshold A is supposed to be known: it equals the largest value of the observed variables $Z_1, Z_2, \ldots, Z_N$. Thus a variable Z that is censored from above (as determined by the threshold value A) is defined as

$$\begin{aligned} Z &= \xi \quad \text{if} \quad \xi < A \ , \\ Z &= A \quad \text{if} \quad \xi \geq A \ . \end{aligned} \tag{3}$$

It follows that for the observations $Z = A$, all that is known is that the latent variable $\xi \geq A$, i.e. $\Pr\{Z = A\} = \Pr\{\xi \geq A\}$.

c. **Doubly censored**. Between a lower threshold value B and an upper threshold value A, variable Z is observed on an interval scale; between B and A it is assumed that Z reflects the latent variable ξ accurately. Below B the value $Z = B$, above A the value $Z = A$ is observed. The values of the thresholds B and A are supposed to be known: they equal the smallest and the largest values, respectively, of the observed variables $Z_1, Z_2, \ldots, Z_N$. Thus a variable Z that is doubly censored (as determined by the threshold values B and A) is defined as

$$\begin{aligned} Z &= B \quad \text{if} \quad \xi \leq B \ , \\ Z &= \xi \quad \text{if} \quad B < \xi < A \ , \\ Z &= A \quad \text{if} \quad \xi \geq A \ . \end{aligned} \tag{4}$$

4. Truncated variables

In the literature (e.g. Johnson & Kotz, 1970a; Maddala, 1983) a distinction is made between censored and truncated variables. In PRELIS only censored variables are dealt with. The LISCOMP program (Muthén, 1988) explicitly distinguishes between censored and truncated variables, and handles both. The difference between both types of variables can be described as follows (cf. Maddala, 1983, Chapters 1 and 6).

For censored variables no restriction is being laid on the range of the population distribution of the latent variable $\xi \sim N(\mu,\sigma^2)$, from which N observations are drawn. In practice, a random sample $\xi_1, \xi_2, \ldots, \xi_N$ is drawn from $-\infty < \xi < \infty$, thus from the entire range of ξ. In the censored variable case, for some observations it is only known whether or not they are above and/or below certain thresholds.

For truncated variables, also consider the population distribution of the latent variable ξ. However, suppose that before a sample $\xi_1, \xi_2, \ldots, \xi_N$ of size N is drawn, the distribution of ξ is truncated from above at the point

$\xi = A$, say. This means that no observations are drawn from $\xi > A$. If the latent variable ξ has a singly truncated normal distribution with upper truncation point A, all sample observations come from $\xi \leq A$. Thus, if there is a restricted (or truncated) sampling range on the variable of interest we are dealing with a truncated variable. An example would be a sample drawn from children with an intelligence quotient smaller than 120, or a sample drawn from the male population within an age range between 21 and 40 years. In these examples observations outside the indicated range are not included in the sample, though the researcher wants to draw conclusions for a broader range of ξ.

A sample of observed variables Z from a truncated distribution, $Z_1, Z_2, \ldots, Z_N$, is called a **truncated sample**. A formal definition of a doubly truncated variable will be given now.

Let the observed random variable Z represent a latent random variable $\xi \sim N(\mu,\sigma^2)$. Variable Z is observed on an interval scale between a lower truncation point B and an upper truncation point A of ξ; between B and A it is assumed that Z reflects the latent variable ξ accurately. Below B the variable Z is not observed at all, above A it is neither observed. In principle, the values of the truncation points B and A are known: they equal the smallest and the largest values of the observed variable Z, respectively. Thus a variable Z that is **doubly truncated** is defined as

$$\begin{aligned} &Z \text{ is not observed if } \xi < B \text{ ,} \\ &Z = \xi \quad \text{if } B \leq \xi \leq A \text{ ,} \\ &Z \text{ is not observed if } \xi > A \text{ ,} \end{aligned} \tag{5}$$

where the lower and upper **truncation points** of the truncated normal distribution are B and A, respectively. If B is replaced by $-\infty$ in (5), the distribution is singly **truncated from above**. If A is replaced by ∞ the distribution is singly **truncated from below**. Recall that Z is assumed to be a continuous variable. Therefore the equality signs in (5) are of no crucial

importance (cf. Muthén, 1988, p. 2-6); didactically, definition (5) has our preference.

According to Maddala (1983, p. 5) in the econometric literature it is customary to use the term truncated normal distribution to describe both the censored and the truncated variable case; it is uncommon to employ the term censored distribution. This is justifiable, because in the analysis of models for both types of variables properties of the truncated normal distribution play a role. The (standard) normal distribution and its truncated counterpart are defined explicitly in the following sections.

From Johnson & Kotz (1970a, Section 6.4) it may be learned that probability calculations and maximum likelihood estimation methods for censored normal data are strongly based on order statistics.

5. The standard normal distribution

The standard normal **cumulative distribution function** (cdf) of a random variable Z is defined, for $-\infty < z < \infty$, as

$$\Phi(z) = \Pr\{Z \leq z\} = \int_{-\infty}^{z} \phi(t)\, dt = (2\pi)^{-1/2} \int_{-\infty}^{z} e^{-t^2/2}\, dt \quad , \qquad (6)$$

where $\phi(z)$, the standard normal **probability density function** (pdf) of a random variable Z, is thus defined as

$$\phi(z) = \frac{d\Phi(z)}{dz} = (2\pi)^{-1/2}\, e^{-z^2/2} \quad , \quad -\infty < z < \infty \quad . \qquad (7)$$

Notice that in general for a normal random variable with mean μ and variance σ^2, the distribution function $\Phi(z)$ is replaced by $\Phi[(z - \mu)/\sigma]$, and the density function $\phi(z)$ by $\sigma^{-1}\phi[(z - \mu)/\sigma]$.

The lower α-quantile of the standard normal distribution, z_α, is implicitly defined as

$$\Phi(z_\alpha) = \Pr\{Z \leq z_\alpha\} = (2\pi)^{-1/2} \int_{-\infty}^{z_\alpha} e^{-t^2/2}\, dt = \alpha \quad . \tag{8}$$

The inverse standard normal distribution function Φ^{-1} is the inverse function of (6). It follows that

$$\Phi^{-1}(\alpha) = z_\alpha \quad . \tag{9}$$

Because the standard normal density function is symmetric around $z = 0$, it holds that $z_\alpha = -z_{1-\alpha}$. Given a value of α the corresponding quantile (percentage point) z_α of the standard normal distribution function can be computed, and reversely given a value of z_α the corresponding tail area α can be calculated.

6. Truncated standard normal distributions

A random variable Z has a doubly truncated standard normal distribution, with lower truncation point B and upper truncation point A, if its probability density function is defined as

$$\begin{aligned} f(z) &= [(2\pi)^{-1/2} e^{-z^2/2}]/[(2\pi)^{-1/2} \int_B^A e^{-t^2/2}\, dt] \\ &= \phi(z)/[\Phi(A) - \Phi(B)] \quad , \quad B \leq z \leq A \quad ; \\ f(z) &= 0 \quad , \quad z < B \quad \text{and} \quad z > A \quad , \end{aligned} \tag{10}$$

where $\phi(z)$ and $\Phi(z)$ are defined by (7) and (6), respectively. Notice that the constant $[\Phi(A) - \Phi(B)]$ in the denominator of (10) is necessary to define a proper probability density function: the integral of the numerator $\phi(z)$ over the range $B \leq z \leq A$ equals $[\Phi(A) - \Phi(B)]$. As can readily be seen from

the cumulative distribution function of the doubly truncated standard normal distribution,

$$F(z|B \le Z \le A) = Pr\{Z \le z|B \le Z \le A\}$$

$$= [\Phi(A) - \Phi(B)]^{-1} \int_B^z \phi(t)\, dt \qquad (11)$$

$$= [\Phi(z) - \Phi(B)]/[\Phi(A) - \Phi(B)] \quad , \quad B \le z \le A \quad ,$$

the integral of (10) over the whole range $B \le z \le A$ equals one, as required.

Notice that (10) and (11) are defined for any value of B and A on the real line. If B is replaced by $-\infty$ the standard normal distribution of Z is singly truncated from above; if A is replaced by ∞ it is singly truncated from below. The probabilities $\Phi(B)$ and $1 - \Phi(A)$ are called the **degrees of truncation** from below and above, respectively (Johnson & Kotz, 1970a, p. 81).

7. The expected value of truncated normal variables

Usually, the term **normal scores** refers to the expected values of the order statistics in a sample of size N from a standard normal distribution (cf. Kendall & Stuart, 1973, p. 504). Tables of those scores (for N=2,...,50) can be found in Owen (1962, p. 151ff.), for example.

In the present context the **normal scores** are by definition the expected values of normal truncated variables within a specific range $B \le Z \le A$. Therefore, those expectations shall now be defined explicitly for doubly, and singly truncated variables, respectively. Before doing so for the truncated standard normal random variable Z in (10), first notice that the first order derivative of the standard normal density function, as defined by (7), can be expressed as

$$\phi'(z) = \frac{d\phi(z)}{dz} = - z\, (2\pi)^{-1/2}\, e^{-z^2/2} = - z\, \phi(z) \quad . \qquad (12)$$

It then follows with (12) that the integral

$$(2\pi)^{-1/2} \int_B^A t\, e^{-t^2/2}\, dt = \int_B^A t\, \phi(t)\, dt = -\, \phi(t) \Big|_B^A \tag{13}$$

$$= \phi(B) - \phi(A) \quad .$$

a. Doubly truncated. Using (13) the expectation of the doubly truncated standard normal variable Z in (10), is by definition

$$E(Z|B \le Z \le A) = \int_B^A t\, \phi(t)\, [\Phi(A) - \Phi(B)]^{-1}\, dt$$

$$= [\Phi(A) - \Phi(B)]^{-1} \int_B^A t\, \phi(t)\, dt \tag{14}$$

$$= [\phi(B) - \phi(A)]/[\Phi(A) - \Phi(B)] \quad .$$

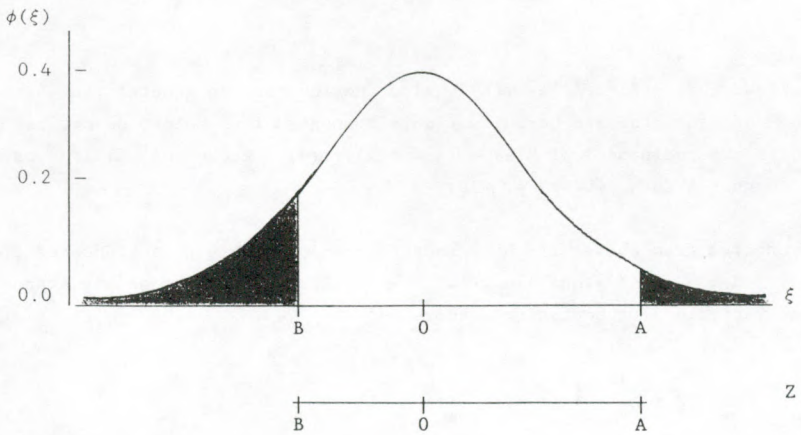

Figure 2. A latent, doubly truncated standard normal variable ξ, and its associated, observed continuous variable Z.

In Figure 2 a picture is given of a latent, doubly truncated standard normal variable $\xi \sim N(0,1)$. The values $\phi(B)$ and $\phi(A)$ are the values of the standard normal density function at the lower and upper truncation points, respectively. The values of the standard normal distribution function $\Phi(B)$ and $\Phi(A)$ quantify the proportional area under the density curve below the lower and upper truncation point, respectively.

The normal score $E\{\xi | B \leq \xi \leq A\}$ can be interpreted as a weighted average of ξ within the range between B and A, where (apart from a constant) the weight function is the probability density function $\phi(\xi)$.

b. Truncated from below. If the standard normal variable Z is truncated from below, then $B \leq Z$; thus in (14) $A = \infty$. The expectation of the standard normal variable Z truncated from below is

$$E\{Z|B \leq Z\} = \int_B^{\infty} t\, \phi(t)\, [\Phi(\infty) - \Phi(B)]^{-1}\, dt \tag{15}$$

$$= [\phi(B) - \phi(\infty)]/[\Phi(\infty) - \Phi(B)] = \phi(B)/[1 - \Phi(B)] \quad .$$

It follows that $\phi(B) > B[1 - \Phi(B)]$. Also, notice that in general the expectation of the standard normal variable truncated from below, as defined in (15), is the reciprocal of $R(x) = [1 - \Phi(x)]/\phi(x)$, known as Mill's ratio (cf. Johnson & Kotz, 1970b, p. 278).

c. Truncated from above. If the standard normal variable Z is truncated from above, then $Z \leq A$; thus in (14) $B = -\infty$. The expectation of the standard normal variable Z truncated from above is

$$E\{Z|Z \leq A\} = \int_{-\infty}^{A} t\, \phi(t)\, [\Phi(A) - \Phi(-\infty)]^{-1}\, dt \tag{16}$$

$$= [\phi(-\infty) - \phi(A)]/[\Phi(A) - \Phi(-\infty)] = -\phi(A)/\Phi(A) \quad .$$

It follows that $\phi(A) > -A\Phi(A)$.

d. The general case of truncated $Z \sim N(\mu,\sigma^2)$

If the truncated variable Z has a normal distribution with a mean $\mu \neq 0$ and/or a variance $\sigma^2 \neq 1$, in equations (14) through (16) Z, B, and A have to be substituted by $(Z - \mu)/\sigma$, $(B - \mu)/\sigma$, and $(A - \mu)/\sigma$, respectively.

In general for $Z \sim N(\mu,\sigma^2)$, it then follows that

$$E\{Z|B \leq Z \leq A\} = \mu + \frac{\phi\{(B - \mu)/\sigma\} - \phi\{(A - \mu)/\sigma\}}{\Phi\{(A - \mu)/\sigma\} - \Phi\{(B - \mu)/\sigma\}} \sigma \quad , \tag{17}$$

$$E\{Z|B \leq Z\} = \mu + \frac{\phi\{(B - \mu)/\sigma\}}{1 - \Phi\{(B - \mu)/\sigma\}} \sigma \quad , \tag{18}$$

and

$$E\{Z|Z \leq A\} = \mu - \frac{\phi\{(A - \mu)/\sigma\}}{\Phi\{(A - \mu)/\sigma\}} \sigma \quad . \tag{19}$$

These expressions for the expectation of Z can also be derived directly, using the probability density function of the doubly truncated normal random variable Z, which is defined as

$$g(z) = \sigma^{-1}\phi\{(z - \mu)/\sigma\}/[\Phi\{(A - \mu)/\sigma\} - \Phi\{(B - \mu)/\sigma\}] \quad , \; B \leq z \leq A \; ;$$

$$g(z) = 0 \quad , \quad z < B \quad \text{and} \quad z > A \quad . \tag{20}$$

This is left as an exercise to the interested reader.

In PRELIS the normal scores are of central importance in how to deal with ordinal variables and continuous censored variables. First the ordinal variable case, next the continuous censored variable case will be treated, each followed by an example using the PRELIS program.

8. Normal scores for ordinal variables

Let N be the total number of observations of an ordinal variable X, and let $n_j \leq N$ be the number of observations in category j of that variable, j=1,2,...,k. The **cumulative proportion of observations** in category 1 through category i is denoted as

$$P_i = \sum_{j=1}^{i} n_j/N \quad , \qquad i=0,1,2,\ldots,k \quad . \tag{21}$$

Obviously, $P_0 = 0$ and $P_k = 1$. In order to compute normal scores, the unknown threshold values c_i, i=1,2,...,k-1, in equation (1) have to be estimated first. This is done by taking the assumption into account that the underlying variable ξ has a standard normal distribution. The thresholds c_i are estimated from the marginal, discrete distribution of the observed ordinal variable as

$$\hat{c}_i = \Phi^{-1}(P_i) = \Phi^{-1}\left(\sum_{j=1}^{i} n_j/N\right) \quad , \qquad i=0,1,2,\ldots,k \quad , \tag{22}$$

where the inverse standard normal distribution function Φ^{-1} is defined by (9). Under the assumption of standard normality of ξ the estimated thresholds are thus determined by the proportions of observations in the k categories of the ordinal variable. Obviously, the thresholds c_0 and c_k do not have to be estimated from the sample of observations, because by definition $c_0 = -\infty$ and $c_k = \infty$. (These values coincide with those given by (22) for i = 0, and i = k, respectively.) From (22) it follows that

$$\Phi(\hat{c}_i) = P_i = \sum_{j=1}^{i} n_j/N \quad , \qquad i=0,1,2,\ldots,k \quad . \tag{23}$$

The probability that a standard normal variable takes on a value smaller than or equal the estimated threshold $\hat{c}_i$ is the cumulative proportion of observations in categories 1 through i.

Recall that by definition (1) an observation X belongs to category i (i.e. X = i) if $c_{i-1} < \xi \leq c_i$. The **normal score z_i** corresponding to an observation X = i is now defined as the expected value of ξ in the very same interval $c_{i-1} < \xi \leq c_i$. It thus follows from (14) that

$$z_i = E(\xi | c_{i-1} < \xi \leq c_i) = [\phi(c_{i-1}) - \phi(c_i)]/[\Phi(c_i) - \Phi(c_{i-1})] \quad , \quad i=1,2,\ldots,k \quad . \tag{24}$$

By using normal scores for ordinal variables all observations within a specific category are replaced by the expected value of an underlying variable ξ within a normal, estimated area. Thus, each observation X = i is replaced by **the same normal score z_i**.

Given (23), the denominator in (24) can be estimated as

$$\Phi(\hat{c}_i) - \Phi(\hat{c}_{i-1}) = P_i - P_{i-1} = [\sum_{j=1}^{i} n_j - \sum_{j=1}^{i-1} n_j]/N = n_i/N \quad , \quad i=1,2,\ldots,k \quad . \tag{25}$$

The normal score z_i can thus be estimated as

$$\hat{z}_i = \hat{E}(\xi | c_{i-1} < \xi \leq c_i) = [\phi(\hat{c}_{i-1}) - \phi(\hat{c}_i)]\, N/n_i \quad , \quad i=1,2,\ldots,k \quad . \tag{26}$$

9. Normal scores for ordinal variables: An example

An example of how to calculate normal scores with ordinal variables is taken from Jöreskog & Sörbom (1988). The data, reproduced from page 1-20 of the PRELIS manual, are listed in Table 1. Missing data have the value -9.

Table 1. Twelve observations on four variables: DATA.EX9.

CASENR	VAR1	VAR2	VAR3	VAR4
1	1	3	-0.7	-0.4
2	2	4	2.3	1.6
3	3	3	1.2	1.7
4	1	-9	-0.4	-0.3
5	3	2	-1.2	-0.7
6	2	1	-9	1.2
7	2	1	0.8	0.3
8	3	3	1.6	1.5
9	1	2	-0.9	-9
10	1	4	-0.8	-0.8
11	1	1	0.7	0.8
12	2	2	1.1	1.3

Suppose normal scores are needed for variable 2 of these sample data. The ordinal variable 2 has k = 4 categories; its number of non-missing observations equals N = 11. The marginal frequencies of observations in each category and cumulative proportions P_i are given in Table 2.

Table 2. Estimated thresholds c_i and normal scores z_i.

category i	marginal frequency n_i	cumulative proportion P_i	upper threshold $\hat{c}_i$	normal score $\hat{z}_i$
1	3	3/11 = 0.273	-0.605	-1.218
2	3	6/11 = 0.545	0.114	-0.235
3	3	9/11 = 0.818	0.908	0.485
4	2	11/11 = 1.000	∞	1.452
k = 4	N = 11			

Given the cumulative proportions P_i, first the thresholds c_1, c_2, and c_3 are estimated using (22). Recall that $c_0 = -\infty$ and $c_4 = \infty$, by definition.

Given the estimated thresholds $\hat{c}_i$, the normal scores z_i can be estimated using (26). These estimates, presented in Table 2, are obtained as follows:

$$\hat{z}_1 = [\phi(\hat{c}_0) - \phi(\hat{c}_1)](11/3) = [\phi(-\infty) - \phi(-0.605)](11/3)$$
$$= [0 - 0.332](11/3) = -1.218;$$

$$\hat{z}_2 = [\phi(\hat{c}_1) - \phi(\hat{c}_2)](11/3) = [\phi(-0.605) - \phi(0.114)](11/3)$$
$$= [0.332 - 0.396](11/3) = -0.235;$$

$$\hat{z}_3 = [\phi(\hat{c}_2) - \phi(\hat{c}_3)](11/3) = [\phi(0.114) - \phi(0.908)](11/3)$$
$$= [0.396 - 0.264](11/3) = 0.485;$$

$$\hat{z}_4 = [\phi(\hat{c}_3) - \phi(\hat{c}_4)](11/2) = [\phi(0.908 - \phi(\infty)](11/2)$$
$$= [0.264 - 0](11/2) = 1.452 \quad .$$

From this example it can be seen that the weighted mean of the normal scores equals 0. This always holds, as can be derived with (26). Thus for ordinal variables the weighted average of the normal scores equals

$$\hat{\mu} = \sum_{i=1}^{k} n_i \hat{z}_i / N = 0 \quad . \tag{27}$$

The weighted estimated variance of the normal scores of variable 2 in the example equals 0.954 (these quantities are found in regular PRELIS output, as will be seen in Section 10). Because each observation in category i is replaced by the estimated normal score $\hat{z}_i$, this weighted variance,

$$\hat{\sigma}_b^2 = \sum_{i=1}^{k} n_i \hat{z}_i^2 / (N-1) \quad , \tag{28}$$

is interpreted by Jöreskog & Sörbom (1988, p. 1-22) as the between-category variance of the latent variable ξ. (Recall that the categories of the latent variable are determined by the threshold estimates $\hat{c}_i$.) By assumption the

total variance σ^2 of ξ equals 1. Therefore, the within-category variance of ξ is estimated as $\hat{\sigma}^2_w = 1 - \hat{\sigma}^2_b$. In our example $\hat{\sigma}^2_w = 1 - 0.954 = 0.046$.

In the PRELIS manual Jöreskog & Sörbom (1988, p. 1-7) suggest that "if required, the normal scores can be scaled so that the weighted variance is 1". This can be done, for example, after a PRELIS analysis, simply by dividing each of the estimated normal scores $\hat{z}_i$ by $\hat{\sigma}_b$, defined through (28).

10. Checking the example of normal scores for ordinal variables with PRELIS

The results from the previous example could be checked by using the PRELIS program (here 386-PRELIS, version 1.20). The data from Table 1 were stored on file DATA.EX9. First, for comparative reasons, all variables are treated as continuous variables (Case a). Next, both variables 1 and 2 are considered to be of ordinal type (Case b).

From the output of Case b it can be seen what the estimated normal scores are that replace the original ordinal scores of variables 1 and 2. Effects on the estimated covariance matrix **S** can be observed clearly; for convenience of comparison the corresponding correlation matrix **R** is also shown. (The product-moment correlations were obtained by the output instruction OU MA=KM.) Notice carefully that the pairwise deletion option was used. Therefore, the estimated correlation coefficients and covariances are based on n_{ij} pairwise non-missing data, whereas the estimated variances and standard deviations are based on n_{ii} non-missing data! (See the effective sample sizes in the output from Case a, and Jöreskog & Sörbom (1988, p. 1-22).)

Case a: Continuous variables VAR1-VAR4

INPUT

```
Case a. Normal scores in PRELIS
        Continuous variables VAR1-VAR4
DA NI=5 NOBS=12 MISSING=-9 TREATMENT=PAIRWISE
RAW_DATA_FROM FILE = DATA.EX9
LABELS
CASENR VAR1 VAR2 VAR3 VAR4
CONTINUOUS VAR1-VAR4
SD CASENR
OUTPUT MA=CM
```

OUTPUT

DISTRIBUTION OF MISSING VALUES
TOTAL SAMPLE SIZE = 12

NUMBER OF MISSING VALUES	0	1
NUMBER OF CASES	9	3

EFFECTIVE SAMPLE SIZES
UNIVARIATE (IN DIAGONAL) AND PAIRWISE BIVARIATE (OFF DIAGONAL)
TOTAL SAMPLE SIZE = 12

	VAR1	VAR2	VAR3	VAR4
VAR1	12			
VAR2	11	11		
VAR3	11	10	11	
VAR4	11	10	10	11

PERCENTAGE OF MISSING VALUES
UNIVARIATE (IN DIAGONAL) AND PAIRWISE BIVARIATE (OFF DIAGONAL)
TOTAL SAMPLE SIZE = 12

	VAR1	VAR2	VAR3	VAR4
VAR1	0.00			
VAR2	8.33	8.33		
VAR3	8.33	16.67	8.33	
VAR4	8.33	16.67	16.67	8.33

UNIVARIATE SUMMARY STATISTICS FOR CONTINUOUS VARIABLES

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS	MINIMUM	FREQ.	MAXIMUM	FREQ.
VAR1	1.833	0.835	0.354	-0.994	**1.000**	5	**3.000**	3
VAR2	2.364	1.120	0.155	-0.771	**1.000**	3	**4.000**	2
VAR3	0.336	1.180	0.173	-0.894	-1.200	1	2.300	1
VAR4	0.564	0.972	-0.271	-1.213	-0.800	1	1.700	1

ESTIMATED COVARIANCE (LOWER TRIANGULAR)
AND CORRELATION MATRIX (UPPER TRIANGULAR)

	VAR1	VAR2	VAR3	VAR4
VAR1	0.697	0.039	0.424	0.466
VAR2	0.036	1.255	0.131	-0.048
VAR3	0.437	0.172	1.393	0.946
VAR4	0.376	-0.056	1.103	0.945

Case b: VAR1-VAR2 ordinal, VAR3-VAR4 continuous

INPUT

```
Case b. Normal scores in PRELIS
        VAR1-VAR2 ordinal, VAR3-VAR4 continuous
```

```
DA NI=5 NOBS=12 MISSING=-9 TREATMENT=PAIRWISE
RAW_DATA_FROM FILE = DATA.EX9
LABELS
CASENR VAR1 VAR2 VAR3 VAR4
ORDINAL VAR1-VAR2
CONTINUOUS VAR3-VAR4
SD CASENR
OUTPUT MA=CM
```

OUTPUT

CONVERSION OF ORIGINAL VALUES TO CATEGORIES

VARIABLE	CATEGORY 1	2	3	4
VAR1	1.00	2.00	3.00	
VAR2	1.00	2.00	3.00	4.00

UNIVARIATE FREQUENCY DISTRIBUTIONS FOR ORDINAL VARIABLES

VARIABLE	CATEGORY 1	2	3	4
VAR1	5	4	3	
VAR2	3	3	3	2

NORMAL SCORES FOR ORDINAL VARIABLES

VARIABLE	CATEGORY 1	2	3	4
VAR1	**-0.936**	**0.217**	**1.271**	
VAR2	**-1.218**	**-0.235**	**0.485**	**1.452**

UNIVARIATE SUMMARY STATISTICS FOR CONTINUOUS VARIABLES

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS	MINIMUM	FREQ.	MAXIMUM	FREQ.
VAR3	0.336	1.180	0.173	-0.894	-1.200	1	2.300	1
VAR4	0.564	0.972	-0.271	-1.213	-0.800	1	1.700	1

ESTIMATED COVARIANCE (LOWER TRIANGULAR)
AND CORRELATION MATRIX (UPPER TRIANGULAR)

	VAR1	VAR2	VAR3	VAR4
VAR1	0.856	0.024	0.437	0.477
VAR2	0.022	0.954	0.108	-0.064
VAR3	0.499	0.123	1.393	0.946
VAR4	0.426	-0.064	1.103	0.945

11. Normal scores for censored continuous variables

For variables that are censored below a threshold B the PRELIS program uses the normal score associated with the interval $-\infty < \xi \leq B$, which means that the expected value of the censored variable in this interval is calculated. From (17) or (19) it follows that this normal score z_B can be estimated as

$$\begin{aligned} \hat{z}_B &= \hat{\mu} + \frac{\phi\{(-\infty) - \phi\{(B - \hat{\mu})/\hat{\sigma}\}}{\Phi\{(B - \hat{\mu})/\hat{\sigma}\} - \Phi(-\infty)} \hat{\sigma} \\ &= \hat{\mu} - \frac{\phi\{(B - \hat{\mu})/\hat{\sigma}\}}{\Phi\{(B - \hat{\mu})/\hat{\sigma}\}} \hat{\sigma} \quad . \end{aligned} \tag{29}$$

For variables that are censored above a threshold A the PRELIS program uses the normal score associated with the interval $A \leq \xi < \infty$. From (17) or (18) it follows that this normal score z_A can be estimated as

$$\begin{aligned} \hat{z}_A &= \hat{\mu} + \frac{\phi\{(A - \hat{\mu})/\hat{\sigma}\} - \phi(\infty)}{\Phi(\infty) - \Phi\{(A - \hat{\mu})/\hat{\sigma}\}} \hat{\sigma} \\ &= \hat{\mu} + \frac{\phi\{(A - \hat{\mu})/\hat{\sigma}\}}{1 - \Phi\{(A - \hat{\mu})/\hat{\sigma}\}} \hat{\sigma} \quad . \end{aligned} \tag{30}$$

In the PRELIS manual (Jöreskog & Sörbom, 1988, p. 1-7) it is stated that the mean μ and the variance σ^2 in (29) and (30) are estimated using maximum likelihood (ML) estimators. In fact, however, according to Jöreskog (1991a), μ and σ^2 are not estimated by ML. Rather, the last sentence on p. 3-16 of the manual ("After the maximum or minimum value of the censored variables has been replaced by its corresponding normal score, these variables are treated as continuous variables".) tells how it is actually done.

Maximum likelihood estimators for μ and σ^2 have to be based on the assumption that the observed variable is a censored normal variable. Calculation of such estimates is not an easy job to do. For more details, see for example Johnson & Kotz (1970a, p. 77ff.), Cohen (1950), and Gupta (1952).

It should be noticed by comparison of (29) that the equation for $\hat{z}_A$ in the PRELIS manual, i.e. the last equation on page 1-7, is wrong. However, we have no reason to believe that the actual calculations in PRELIS are incorrect.

12. Censored continuous variables in PRELIS: An example

In order to illustrate the material from the previous section, the effects of treating continuous variables as censored variables were studied for the example used earlier in Section 9. Variables 1 and 2 are treated as censored continuous variables, variables 3 and 4 as ordinary, uncensored continuous variables. In Case c through Case e, variables 1 and 2 are censored from below, from above, and doubly censored, respectively.

From the output as collected from PRELIS, notice the estimates of the normal scores, which are printed in boldface here as minimum and/or maximum values under the univariate summary statistics. It is slightly confusing that the headings MINIMUM and MAXIMUM are used, where actually estimates of expected values are presented.

Unfortunately, the PRELIS program does not give separate estimates for the mean μ and the variance σ^2 of the underlying, latent variable ξ of censored variables. (These estimates can be obtained with the forthcoming PRELIS 2 program; see Jöreskog & Sörbom, 1991, p. 7). Since the latter estimates differ whether a variable is singly or doubly censored, clearly this affects the normal score estimates (see (29) and (30)), even if we are dealing with the same sample of observations $Z_1, Z_2, \ldots, Z_N$.

Case c: Continuous variables; VAR1-VAR2 censored from below

INPUT

```
Case c. Normal scores in PRELIS
        Continuous variables; VAR1-VAR2 censored from below
DA NI=5 NOBS=12 MISSING=-9 TREATMENT=PAIRWISE
RAW_DATA_FROM FILE = DATA.EX9
LABELS
CASENR VAR1 VAR2 VAR3 VAR4
CONTINUOUS VAR1-VAR4
CB VAR1-VAR2
SD CASENR
OUTPUT MA=CM
```

OUTPUT

UNIVARIATE SUMMARY STATISTICS FOR CONTINUOUS VARIABLES

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS	MINIMUM	FREQ.	MAXIMUM	FREQ.
VAR1	1.494	1.221	-0.026	-1.329	**0.185**	5	3.000	3
VAR2	2.140	1.567	-0.487	-1.052	**0.181**	3	4.000	2
VAR3	0.336	1.180	0.173	-0.894	-1.200	1	2.300	1
VAR4	0.564	0.972	-0.271	-1.213	-0.800	1	1.700	1

ESTIMATED COVARIANCE (LOWER TRIANGULAR)
AND CORRELATION MATRIX (UPPER TRIANGULAR)

	VAR1	VAR2	VAR3	VAR4
VAR1	1.490	0.151	0.498	0.527
VAR2	0.274	2.455	0.152	0.061
VAR3	0.745	0.258	1.393	0.946
VAR4	0.617	0.092	1.103	0.945

Case d: Continuous variables; VAR1-VAR2 censored from above

INPUT

```
Case d. Normal scores in PRELIS
        Continuous variables; VAR1-VAR2 censored from above
DA NI=5 NOBS=12 MISSING=-9 TREATMENT=PAIRWISE
RAW_DATA_FROM FILE = DATA.EX9
LABELS
CASENR VAR1 VAR2 VAR3 VAR4
CONTINUOUS VAR1-VAR4
CA VAR1-VAR2
SD CASENR
OUTPUT MA=CM
```

OUTPUT

UNIVARIATE SUMMARY STATISTICS FOR CONTINUOUS VARIABLES

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS	MINIMUM	FREQ.	MAXIMUM	FREQ.
VAR1	1.974	1.058	0.682	-0.699	1.000	5	**3.562**	3
VAR2	2.484	1.406	0.335	-0.706	1.000	3	**4.660**	2
VAR3	0.336	1.180	0.173	-0.894	-1.200	1	2.300	1
VAR4	0.564	0.972	-0.271	-1.213	-0.800	1	1.700	1

ESTIMATED COVARIANCE (LOWER TRIANGULAR)
AND CORRELATION MATRIX (UPPER TRIANGULAR)

	VAR1	VAR2	VAR3	VAR4
VAR1	1.119	0.102	0.359	0.409
VAR2	0.143	1.978	0.205	0.036
VAR3	0.470	0.320	1.393	0.946
VAR4	0.422	0.049	1.103	0.945

Case e: Continuous variables; VAR1-VAR2 doubly censored

INPUT

```
Case e. Normal scores in PRELIS
        Continuous variables; VAR1-VAR2 doubly censored
DA NI=5 NOBS=12 MISSING=-9 TREATMENT=PAIRWISE
RAW_DATA_FROM FILE = DATA.EX9
LABELS
CASENR VAR1 VAR2 VAR3 VAR4
CONTINUOUS VAR1-VAR4
CE VAR1-VAR2
SD CASENR
OUTPUT MA=CM
```

OUTPUT

UNIVARIATE SUMMARY STATISTICS FOR CONTINUOUS VARIABLES

VARIABLE	MEAN	ST. DEV.	SKEWNESS	KURTOSIS	MINIMUM	FREQ.	MAXIMUM	FREQ.
VAR1	1.573	1.777	0.278	-1.062	**-0.228**	5	**4.007**	3
VAR2	2.245	1.954	-0.141	-0.848	**-0.075**	3	**4.963**	2
VAR3	0.336	1.180	0.173	-0.894	-1.200	1	2.300	1
VAR4	0.564	0.972	-0.271	-1.213	-0.800	1	1.700	1

ESTIMATED COVARIANCE (LOWER TRIANGULAR)
AND CORRELATION MATRIX (UPPER TRIANGULAR)

	VAR1	VAR2	VAR3	VAR4
VAR1	3.157	0.130	0.438	0.478
VAR2	0.427	3.819	0.157	0.035
VAR3	0.961	0.335	1.393	0.946
VAR4	0.821	0.067	1.103	0.945

---

13. Effects of scale type of variables on correlation estimates

Above, for all censored variables it was assumed that the threshold values B and A are known. In the PRELIS program these values are determined by the smallest and largest values in a sample of size N. It should be realized that if a constant b would be added to all observed scores of a variable, all normal scores, and thus the estimated mean of that variable, would increase by that same constant b; estimates of the standard deviation, the skewness, the kurtosis, and the correlations of the continuous variables are invariant under such an additive operation. In practice, it thus would not matter for the latter statistics if, for example, two items were added to a test that would not discriminate between the examinees; items that would be

answered either correctly, or wrongly, by each subject in the sample.

By comparing the univariate statistics from Case a with those from Cases c through e, it can be observed that considering variables as being censored may strongly affect the estimates of skewness and kurtosis, statistics so crucial in checking normality assumptions.

From the previous results from Cases a through e, each based on different types of variables, the effects of the scale (or measurement) assumptions on estimates of correlations ρ_{ij} are summarized. Clearly, it could make a difference for subsequent analyses based on these correlations, or covariances, which scale type the researcher is ready to assume for the variables under study.

Table 3. Estimated correlations $\hat{\rho}_{ij} = \hat{\rho}$(VARi,VARj) under different assumptions on the type of variables for both VAR1 and VAR2; VAR3 and VAR4 are continuous variables.

Case	Type of variable	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{14}$	$\hat{\rho}_{23}$	$\hat{\rho}_{24}$
a.	continuous	0.039	0.424	0.466	0.131	-0.048
b.	ordinal	0.024	0.437	0.477	0.108	-0.064
c.	censored below	0.151	0.498	0.527	0.152	0.061
d.	censored above	0.102	0.359	0.409	0.205	0.036
e.	doubly censored	0.130	0.438	0.478	0.157	0.035

14. Discussion

In this paper an expository overview was given of the use of normal scores in the PRELIS program, for ordered and continuous censored variables, respectively. Now, suppose a researcher wants to estimate proper correlation coefficients with PRELIS. If the variables are continuous censored variables, there is no alternative but to use normal score estimates. If, however, the variables are of ordinal scale type and declared ordinal, using normal scores is just one of several approaches in obtaining proper estimates of correlations with PRELIS. Three options are available then.

a. The KM option: normal scores are determined from the marginal frequencies, and product-moment correlation coefficients are subsequently based on

these normal scores (ρ_{NS}).

b. The PM option: polychoric correlations (ρ_{PC}) are estimated (or polyserial correlations if one of the pair of variables is declared continuous).

c. The OM option: ordinal scores are replaced by optimal scores, which are determined for each pair of variables separately; i.e. for each pair of variables canonical correlations (ρ_{OS}) are estimated.

Jöreskog & Sörbom (1988, p. 1-9ff.) did several Monte Carlo studies to compare the behaviour of various correlation estimates in the ordinal variable case. They concluded (o.c., p. 1-10) that among six correlation measures the polychoric correlation is "generally the best estimator" (in terms of bias, mean squared error, consistency, and robustness against non-normality). The robustness of the polychoric correlation coefficient against departures from bivariate normality was studied in further detail by Quiroga (1992).

In conclusion: given the advantageous behaviour of the polychoric correlation estimate it does not seem not to be recommended to use normal scores (the KM option) when ordinal variables are declared ordinal.

Finally, a potentially useful extension of polychoric correlation should be mentioned. Quiroga (1992) developed a new general underlying bivariate distribution for the observed ordinal variables, one that includes the bivariate normal. It is a mixture of the standard bivariate normal distribution and two independent univariate skew-normal distributions. (See Azzalini (1985) for a detailed description of the skew-normal distribution.) The correlation coefficient associated with this more general distribution was called the extended polychoric correlation. Parameters involved are estimated using maximum likelihood procedures. Here too, the hypothesized probability model can be tested by a likelihood ratio test. It can be expected that the extended polychoric correlation will be an option of the PRELIS program within the near future.

Acknowledgement

The author wishes to thank Herbert Hoijtink, Ivo Molenaar and Tom Snijders for their useful comments to an earlier version of this paper.

ontvangen 7-1-1992
geaccepteerd 8-7-1992

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. **Scandinavian Journal of Statistics**, **12**, 171-178.

Cohen, A.C. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. **Annals of Mathematical Statistics**, **21**, 557-569.

Cohen, A.C. (1991). **Truncated and censored samples: Theory and applications**. New York: Dekker.

Gupta, A.K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. **Biometrika**, **39**, 252-259.

Johnson, N.L., & Kotz, S. (1970a). **Distributions in statistics: Continuous univariate distributions - 1**. Boston: Houghton Mifflin.

Johnson, N.L., & Kotz, S. (1970b). **Distributions in statistics: Continuous univariate distributions - 2**. Boston: Houghton Mifflin.

Jöreskog, K.G. (1991a). Personal communication (July 4).

Jöreskog, K.G. (1991b). **Latent variable modeling with ordinal variables**. Unpublished manuscript, Uppsala University, Department of Statistics.

Jöreskog, K.G., & Sörbom, D. (1988). **PRELIS: A program for multivariate data screening and data summarization. A preprocessor for LISREL** (2nd ed.). Mooresville, IN: Scientific Software.

Jöreskog, K.G., & Sörbom, D. (1991). **New features in PRELIS 2**. Unpublished manuscript, Uppsala University, Department of Statistics.

Kendall, M.G., & Stuart, A. (1973). **The advanced theory of statistics** (Vol. 2, 3rd ed.). London: Griffin.

Maddala, G.S. (1983). **Limited-dependent and qualitative variables in econometrics**. Cambridge: Cambridge University Press.

Muthén, B.O. (1988). **LISCOMP: Analysis of linear structural equations with a comprehensive measurement model** (2nd ed.). Mooresville, IN: Scientific Software.

Owen, D.B. (1962). **Handbook of statistical tables**. Reading, MA: Addison-Wesley.

Quiroga, A.M. (1992). **Studies of the polychoric correlation and other correlation measures for ordinal variables** (Acta Universitatis Upsaliensis, 29). Doctoral dissertation, Uppsala University, Department of Statistics.

Schneider, H. (1986). **Truncated and censored samples from normal populations**. New York: Dekker.