

# Test for Uniform Stochastic Ordering with Correlated Binary Data

Aniko Szabo<sup>1</sup>

*Huntsman Cancer Institute and Department of Oncological Sciences,  
University of Utah, 2000 Circle of Hope,  
Salt Lake City, UT 84112-5550*

Olusegun George

*Department of Mathematical Sciences,  
The University of Memphis, Memphis, TN 38152-6429*

## Abstract

We construct a trend test with exchangeable clustered binary data. The definition of the trend is based on uniform stochastic ordering. As application of the procedure, we analyze a developmental toxicity dataset on the effect of diethylhexyl phthalate (DEHP).

**Keywords:** risk assessment; trend; cluster; teratology.

---

<sup>1</sup>Corresponding author. Ph. (801) 585-5182, Fax: (801) 585-5357, email: aniko.szabo@hci.utah.edu

# 1 Introduction

Correlated data are usually obtained in longitudinal studies, experiments with several measurements on each subject, group randomization trials, animal teratology and developmental toxicity experiments. In these studies, when several dose levels of a treatment are to be compared, the definitions of treatment homogeneity or dose-related trend are ambiguous. In particular, when the data points are clusters of observations, a choice has to be made between defining treatment effect in terms of the response rate of the cluster elements, or the response rate of the cluster as a whole, or in terms of the probability of no adverse response within the cluster. Zucker and Wittes (1992), among others, have noted that the choice of a particular definition is directed by the intended application. For example, in the treatment of glaucoma it may be more appropriate to conduct tests of treatment effect based on the per-eye response rate, whereas in the treatment of eye cancer, the probability of no cancerous eye would seem to be a better choice.

In this paper we are concerned with the construction of tests for testing for trend when data points consist of clusters of exchangeable binary random variables. Ophthalmological and otolaryngological studies, as well as developmental toxicity experiments, are a few examples of applications where such data are obtained.

Developmental toxicity experiments are usually performed on pregnant laboratory rodents. In such experiments groups of pregnant female animals are randomly assigned to one of up to 6 dose groups of an agent in the period after implantation of the fetuses. Just prior to term, the uterine contents of each dam are examined. The data collected usually include the number of implantation sites, the number of live fetuses, and the state of each fetus – normal, malformed, dead or resorbed. In the evaluation of the toxicity of the agent the researcher may want to know whether morbidity and mortality increased with increasing dose levels.

As an example we will analyze data from a teratology study in which five groups of pregnant CD-1 mice were randomly assigned to dose levels 0, 0.025, 0.050, 0.100 and 0.150 grams per kilogram body weight of diethylhexyl phthalate (DEHP) (Tyl et al., 1988). For this analysis we combined the observed endpoints: malformation and fetal death. The data is available at <http://www.hci.utah.edu/groups/biostat/szabo.html>. The proportion of responding fetuses are (0.189, 0.103, 0.251, 0.696, 0.981), so empirically the response probability seems first to decrease, but then to steadily increase as dose increases.

For such a study let  $X_{ijk}$  represent the binary response of the  $k^{th}$  fetus in the  $j^{th}$  litter of the  $i^{th}$  dose group,  $i = 0, 1, \dots, g$ ,  $j = 1, \dots, M_i$ ,  $k = 1, \dots, n_{ij}$ , where

$$X_{ijk} = \begin{cases} 1, & \text{if the fetus is malformed resorbed or dead} \\ 0, & \text{if the fetus is normal} \end{cases}$$

The literature on the test for treatment homogeneity or trend when correlated binary data are involved has commonly been based on comparing the marginal response probability  $p_i$  across dose groups, or requiring  $p_i$  to be monotone with increasing dose levels, where  $p_i = E(X_{ijk})$ ,  $i = 0, 1, \dots, g$ . This approach corresponds to evaluating treatment based on the response rate of observations, rather than the sampling unit. Typically,

$p_i$  is modeled by dose-response functions such as logistic or Weibull, with a linear link for  $p_i$ . Thus, the test for trend in such analysis represents a test for linear trend.

Examples of such tests include those of Lefkopoulou et al. (1989), Ryan (1993) and Chen (1993), all of who propose generalized score tests based on a quasi-likelihood approach with a linear logistic model for  $p_i$ . In general, these procedures result in Cochran-Armitage type statistics.

An alternative procedure was proposed by Rao and Scott (1992), who suggested using the Cochran-Armitage statistic with a sampling survey adjustment for extrabinomial variation due to clustering. Rao and Scott's procedure may be considered non-parametric, since no model was imposed on  $p_i$ 's. However by stating the hypothesis of treatment homogeneity in terms of  $p_i$ 's only, their procedure is again based on per observational unit response, rather than the response of the cluster as a whole.

In the context of clustered discrete data the need for new methods which incorporate the effect of the cluster as a whole into the testing procedure is motivated by documented evidence of the effect of the treatment on correlations and higher moments (Kupper et al., 1986; Bowman et al., 1995; George and Kodell, 1996). Based on the assumption that within clusters  $X_{ijk}$ 's are exchangeable, George and Kodell (1996) proposed a non-parametric procedure for testing for trend based on the per-cluster response, not just the marginal response probability of each observational unit. We discuss this method further in the next section.

In this paper we describe a method for defining trend in correlated binary data that considers the effect of the treatment on the cluster as a whole and has a clear intuitive meaning. It reduces the problem of testing for dose-related trend to testing for ordering of multinomial vectors. We discuss the idea in greater detail in Section 2, and we specify formally the hypothesis to be tested using uniform stochastic ordering. The next two sections are devoted to deriving a likelihood ratio test and finding its asymptotic distribution. In Section 4 we apply our method to a developmental toxicology data set on the effect of diethylhexyl phthalate (DEHP) in mice, and to compare our procedure with those in the statistical literature in Section 5 we analyze a developmental toxicology data set from the Shell Toxicology Laboratory.

## 2 The exchangeable model for correlated binary data

### 2.1 Definitions and basic results

A sequence of random variables  $X_1, X_2, \dots, X_n$  is called exchangeable if for vector  $(x_1, \dots, x_n)$ ,

$$P(X_{\pi(1)} = x_1, \dots, X_{\pi(n)} = x_n) = P(X_1 = x_1, \dots, X_n = x_n), \quad (1)$$

for any permutation  $\pi(1), \dots, \pi(n)$  of  $1, \dots, n$ . In many applications, especially in teratological and developmental toxicity studies, data from the same litter or cluster can be assumed exchangeable. Actually, any approach based only on the cluster sum  $R = \sum_{k=1}^n X_k$ , implicitly assumes exchangeability among littermates, hence all the previously mentioned models are built on the assumption of exchangeability. Among the

advantages of the exchangeable model is that correlations of all orders can be handled simply and efficiently (George and Bowman, 1995). Theoretically the existing models can account for higher order correlations, but they are either determined by the first two or become increasingly complicated as higher order correlations are introduced.

A convenient parameterization of the joint distribution of  $X_1, X_2, \dots, X_n$  can be given using the parameters  $\lambda_k = P(X_1 = 1, \dots, X_k = 1)$ ,  $k = 1, \dots, n$  and  $\lambda_0 = 1$ . Using an inclusion-exclusion argument, it can be shown (see George and Bowman 1995) that

$$P(X_1 = x_1, \dots, X_n = x_n) = \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \lambda_{r+j}, \quad (2)$$

where  $r = \sum_{i=1}^n x_i$ . Hence, if  $R = \sum_{i=1}^n X_i$  denotes the number of responses, then

$$P(R = r) = \binom{n}{r} p_r = \binom{n}{r} \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \lambda_{r+j}. \quad (3)$$

As a special case, if the  $X_i$ 's are independent, then  $\lambda_k = \lambda_1^k$  and

$$P(R = r) = \binom{n}{r} \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \lambda_1^{r+j} = \binom{n}{r} \lambda_1^r (1 - \lambda_1)^{n-r},$$

so  $R$  is a binomial random variable with parameters  $n$  and  $\lambda_1$ .

## 2.2 Defining trend

Suppose there is one control group and  $g$  treatment groups with  $M_0, M_1, \dots, M_g$  clusters respectively. The primary unit of the dataset is a cluster of binary points that are assumed to be exchangeable within each cluster. We also assume that the clusters are independent from each other. Let  $S$  denote the cluster size, which we will assume to be random, but independent from the treatment assignment. This assumption is often reasonable in teratology applications, since usually the administration of the drug begins only after conception, so it does not influence litter size. Also in practice the litter sizes rarely exceed 30, so we will assume that  $S$  has a maximum possible value denoted by  $K$ . Let  $m_n^{(i)}$  denote the number of clusters of size  $n$ ,  $i = 0, \dots, g$ ;  $n = 1, \dots, K$  at the  $i^{\text{th}}$  dose level. Then  $(m_1^{(i)}, \dots, m_K^{(i)})$  has a multinomial distribution with parameters  $(M_i; q_1, \dots, q_K)$ , where  $q_n = P(S = n)$ .

Recall that  $X_{ijk}$  denotes the  $k^{\text{th}}$  observation in the  $j^{\text{th}}$  cluster of the  $i^{\text{th}}$  dose group,  $i = 1, \dots, g$ ;  $j = 1, \dots, M_i$ ;  $k = 1, \dots, S$ . Among the clusters of size  $n$ , let  $A_{rn}^{(i)}$  denote the number of clusters with cluster size  $r$ , i.e. with exactly  $r$  responses. Note that  $(A_{0n}^{(i)}, \dots, A_{nn}^{(i)})$  has a multinomial distribution with parameters  $(m_n^{(i)}; \boldsymbol{\pi}_n^{(i)}) = (m_n^{(i)}; \pi_{0n}^{(i)}, \dots, \pi_{nn}^{(i)})$ , where  $\pi_{rn}^{(i)} = \binom{n}{r} p_{rn}^{(i)}$ .

An important feature of  $A_{rn}^{(i)}$ ,  $i = 1, \dots, g$ ;  $n = 1, \dots, M_i$ ;  $r = 0, 1, \dots, n$  is that they are jointly sufficient statistics for the introduced parameters. This can be seen from the likelihood function derived as discussed by George and Kodell (1996):

$$\begin{aligned}
L &= \prod_{i=0}^g \prod_{j=0}^{M_i} P(X_{ij1} = x_{ij1}, \dots, X_{ijn_{ij}} | S_{ij} = n_{ij}) P(S_{ij} = n_{ij}) = \\
&= \prod_{i=0}^g \prod_{j=0}^{M_i} p_{r_{ij}, n_{ij}}^{(i)} q_{n_{ij}}^{(i)} = \prod_{i=0}^g \prod_{n=1}^K \prod_{r=0}^n [p_{rn}^{(i)} q_n]^{A_{rn}^{(i)}},
\end{aligned} \tag{4}$$

where  $r_{ij} = \sum_{k=1}^{n_{ij}} x_{ijk}$  is the observed number of responses in the  $j^{\text{th}}$  cluster of the  $i^{\text{th}}$  treatment group.

If the studied substance is toxic, then at a higher dose level there will tend to be more clusters with a large number of responses, so the distribution of  $R_n^{(i+1)}$  will tend to be stochastically larger than  $R_n^{(i)}$ ,  $i = 0, 1, \dots, g$ , where  $R_n^{(i)} = \sum_{j=1}^{m_n^{(i)}} X_{ijk}$ . Now  $P(R_n^{(i)} = r) = \pi_{rn}^{(i)}$ . Hence the problem of testing dose-related trend can be rewritten as testing for ordering of multinomial random vectors  $\boldsymbol{\pi}_n^{(i)}$ .

There are several advantages of this approach over the existing methods. First of all the above method is based on the entire vector  $\boldsymbol{\pi}_n^{(i)}$  and not only the marginal means. The resulting procedures are non-parametric, i.e. they do not test for a specific kind of trend, while most other methods test only for a linear trend. Also, as discussed in Section 1, different applications might require different definitions of trend. The desired flexibility can be achieved by using various orderings of the multinomial vectors (stochastic, uniform stochastic, likelihood-ratio, etc.). Finally, results of order restricted statistical inference can be also applied in this situation.

A special case of this approach was suggested by George and Kodell (1996). The authors define dose-related trend as having  $\lambda_{rn}^{(i)} \leq \lambda_{rn}^{(i+1)}$ ;  $r = 1, \dots, n$ ;  $n = 1, \dots, K$ ;  $i = 0, \dots, g - 1$  with at least one strict inequality. This definition follows naturally from the parameterization by the  $\lambda_{jn}^{(i)}$ 's described in section 2.1. Since

$$\lambda_{rn}^{(i)} = \sum_{j=0}^{n-r} \binom{n-r}{j} p_{n-j,n}^{(i)} = \sum_{j=0}^{n-r} \frac{\binom{n-r}{j}}{\binom{n}{n-j}} \pi_{n-j,n}^{(i)},$$

$i = 0, \dots, g$ , the above definition defines a particular ordering of the multinomial vectors  $\boldsymbol{\pi}_n^{(i)}$ . George and Kodell (1996) construct a likelihood ratio test, but are unable to find the exact distribution of the test statistic and have to use an upper bound for the p-value. Since they find the maximum likelihood estimates by non-linear optimization, simulation or resampling are essentially excluded.

### 2.3 Choice of ordering

To decide on the appropriate ordering for analyzing the DEHP data set described in Section 1, we examined the data graphically (see Fig. 1).

Fig. 1 here

For assessing stochastic ordering, we plotted the values of  $P(R > r)$ ,  $r = 0, 1, \dots, 18$  for each dose level (Fig. 1a). By taking sections of this plot parallel to the dose axis, we have clear empirical evidence of stochastic ordering. To examine the possibility of a more stringent trend, we plotted the values of  $P(R > r + 1 | R > r)$ ,  $r = 0, 1, \dots, 18$  – the probability of having 1 additional malformation (Fig. 1b),  $P(R > r + 2 | R > r)$ ,  $r = 0, 1, \dots, 17$  – the probability of having 2 additional malformations (Fig. 1c), etc.

While these graphs still indicate increasing trend, the evidence is much weaker and testing for such (uniform stochastic) trend might be of interest.

Intuitively, the absence of uniform stochastic ordering, that is having  $P^{(i+1)}(R > r_0 + s | P > r_0) < P^{(i)}(R > r_0 + s | P > r_0)$ , would indicate that the occurrence of  $r_0$  responses has a protective effect on the other fetuses. This could occur, for example, if a certain number of the littermates absorb a large amount of the chemical, thus reducing the exposure for the rest. However, given the setup of the experiments and the biological mechanism of exposure of the fetuses through maternal blood, such a situation is unplausible. So in most experiments with a real toxic effect we would expect the responses to be uniformly stochastically ordered.

Another reason for considering uniform stochastic ordering instead of stochastic ordering is the relative simplicity of parameter estimation and inference. In the area of order restricted inference it has been long recognized that estimation under stochastic ordering appears to be the most computationally difficult (see, for example, Robertson et al. (1988)). Hence in this paper we concentrate on using uniform stochastic ordering for defining trend with correlated binary data.

### 3 Likelihood ratio test

#### 3.1 Formulation of the hypotheses

*Definition:* Let  $X$  and  $Y$  be two discrete random variables taking values from the space  $\Gamma = \{1, 2, \dots, k\}$  with outcome probabilities defined by the probability vectors  $\mathbf{P} = (p_1, p_2, \dots, p_k)$  and  $\mathbf{Q} = (q_1, q_2, \dots, q_k)$ , where  $p_i = P(X = i)$  and  $q_i = P(Y = i)$ . We say that  $X$  is larger than  $Y$  with respect to the *uniform stochastic ordering* (notation:  $X >^{ust} Y$ ), if

$$P(X \geq j + m | X \geq j) \geq P(Y \geq j + m | Y \geq j)$$

for any  $j$  and  $m$  with at least one strict inequality. In terms of the probability vectors  $\mathbf{P} >^{ust} \mathbf{Q}$ , if  $\frac{p_k}{q_k} \geq \frac{p_{k-1} + p_k}{q_{k-1} + q_k} \geq \dots \geq \frac{p_1 + \dots + p_k}{q_1 + \dots + q_k}$  with at least one strict inequality.

Let as above  $R_n^{(i)}$  denote the number of adversely affected fetuses from a litter of size  $n$  in the  $i^{th}$  dose group,  $i = 0, 1, \dots, g$ ;  $n = 1, 2, \dots, K$ . If the administered treatment is toxic, then it is reasonable to assume that  $P(R_n^{(i)} \geq r | R_n^{(i)} \geq j)$  does not decrease with increasing dose level for any values of  $r \geq j$ , i.e. the risk of having additional affected fetuses increases with dose level. Conversely, if  $P(R_n^{(i)} \geq r | R_n^{(i)} \geq j)$  increases with dose level, then the studied substance should be declared toxic.

Hence we define trend as uniform stochastic ordering between the random variables  $R_n^{(i)}$  or, equivalently, between the vector of sufficient statistics  $(A_{0,n}^{(i)}, A_{1,n}^{(i)}, \dots, A_{n,n}^{(i)})$ . More formally, define the sample space as

$$\Omega = \{(\boldsymbol{\pi}_n^{(0)}, \boldsymbol{\pi}_n^{(1)}, \dots, \boldsymbol{\pi}_n^{(g)}) \mid \frac{\sum_{j=r}^n \pi_{jn}^{(i)}}{\sum_{j=r+1}^n \pi_{jn}^{(i)}} \geq \frac{\sum_{j=r}^n \pi_{jn}^{(i+1)}}{\sum_{j=r}^n \pi_{jn}^{(i+1)}}; r = 0, 1, \dots, n-1; i = 0, 1, \dots, g-1\}$$

$$\begin{aligned} & \sum_{j=0}^n \pi_{jn}^{(i)} = 1; i = 0, 1, \dots, g; n = 1, 2, \dots, K \} = \\ & = \{(\boldsymbol{\pi}_n^{(0)}, \boldsymbol{\pi}_n^{(1)}, \dots, \boldsymbol{\pi}_n^{(g)}) | \boldsymbol{\pi}_n^{(i)} <^{ust} \boldsymbol{\pi}_n^{(i+1)}; i = 0, 1, \dots, g-1; n = 1, 2, \dots, K\} \end{aligned}$$

and let the space for the null hypothesis be

$$\Omega_0 = \{(\boldsymbol{\pi}_n^{(0)}, \boldsymbol{\pi}_n^{(1)}, \dots, \boldsymbol{\pi}_n^{(g)}) | \boldsymbol{\pi}_{rn}^{(0)} = \dots = \boldsymbol{\pi}_{rn}^{(g)}; r = 0, 1, \dots, n; n = 1, 2, \dots, K\}.$$

Then we define a uniform stochastic trend test to be the test of the hypothesis

$$H_0 : (\boldsymbol{\pi}_n^{(0)}, \boldsymbol{\pi}_n^{(1)}, \dots, \boldsymbol{\pi}_n^{(g)}) \in \Omega_0$$

versus

$$H_A : (\boldsymbol{\pi}_n^{(0)}, \boldsymbol{\pi}_n^{(1)}, \dots, \boldsymbol{\pi}_n^{(g)}) \in \Omega - \Omega_0$$

### 3.2 Maximum likelihood estimates

As in Dykstra et al. (1991), we simplify the expression (4) for the likelihood function using the following reparametrization: denote  $\theta_{rn}^{(i)} = \sum_{j=r+1}^n \pi_{jn}^{(i)} / \sum_{j=r}^n \pi_{jn}^{(i)}$ , if  $r = 0, 1, \dots, n-1$ . Then the log-likelihood function can be rewritten as

$$\ell = \sum_{n=1}^K \sum_{i=0}^g \sum_{r=0}^{n-1} [A_r^{(i)} \ln(1 - \theta_{rn}^{(i)}) + (s_{rn}^{(i)} - A_{rn}^{(i)}) \ln \theta_{rn}^{(i)}], \quad (5)$$

where  $s_{rn}^{(i)} = \sum_{l=r}^n A_{ln}^{(i)}$ , and the constraints become  $\theta_{rn}^{(0)} \leq \theta_{rn}^{(1)} \leq \dots \leq \theta_{rn}^{(g)}$ ,  $r = 0, 1, \dots, n$ ,  $n = 1, 2, \dots, K$ . Hence the maximization of  $\ell$  can be done separately for each  $n$  and  $r$ , and using an isotonic regression procedure for maximization (Robertson et al., 1988), the maximum likelihood solution is obtained as the isotonic regression of the unrestricted MLE's  $\hat{\theta}_{rn}^{(i)} = (s_{rn}^{(i)} - A_{rn}^{(i)}) / s_{rn}^{(i)}$  with weights  $(s_{rn}^{(0)}, s_{rn}^{(1)}, \dots, s_{rn}^{(g)})$ . Here we use the convention that  $0/0 = 1$ . The solution  $(\bar{\theta}_{rn}^{(0)}, \dots, \bar{\theta}_{rn}^{(g)})$  can be obtained by the pool-adjacent-violators algorithm or the minimum lower sets algorithm, that are described in detail in the same book.

In data reporting fetuses that are malformed, only live animals are considered. Malformed fetuses that are lost due to maternal death or some other independent event are not counted. But these ‘‘censored’’ observations can be easily included in the above model. In such a case the number of fetal responses observed at the time of death can be used as a lower bound for the possible number of responses that could have occurred if the animal have gone to full gestation (we make the realistic assumption that the malformations are not reversible). In this case  $A_{rn}^{(i)}$  is still defined as the number of clusters with exactly  $r$  responses, and  $s_{rn}^{(i)}$  is the number of clusters with at least  $r$  responses, but the latter will now include the ‘‘censored’’ observations as well.

Under the null hypothesis of no treatment effect,  $\theta_{rn}^{(i)} = \theta_{rn}$  for all  $i$ , and the maximum likelihood estimates are easily found to be

$$\theta_{rn}^* = \left( \sum_{i=0}^g s_{rn}^{(i)} - \sum_{i=0}^g A_{rn}^{(i)} \right) / \sum_{i=0}^g s_{rn}^{(i)}. \quad (6)$$

Then the likelihood ratio test statistic is

$$T = \sum_{n=1}^K \sum_{r=0}^{n-1} T_{rn} = 2 \sum_{n=1}^K \sum_{r=0}^{n-1} \sum_{i=0}^g A_{rn}^{(i)} \ln \left( \frac{1 - \bar{\theta}_{rn}^{(i)}}{1 - \theta_{rn}^*} \right) + (s_{rn}^{(i)} - A_{rn}^{(i)}) \ln \left( \frac{\bar{\theta}_{rn}^{(i)}}{\theta_{rn}^*} \right). \quad (7)$$

### 3.3 The limiting distribution of the likelihood ratio test statistic

We use Taylor series expansion of the likelihood ratio statistic to obtain the limiting distribution of  $T_{rn}$ , when the number of clusters increases to infinity so that  $\alpha_{in} = \lim_{m_n \rightarrow \infty} m_n^{(i)} / m_n$  is finite, where  $m_n = \sum_{i=0}^g m_n^{(i)}$ . Similar to Dykstra et al. (1991) it can be shown that the limiting distribution has a chi-bar-square form, but it depends on the values of  $\alpha_{in}$ :

$$P(T_{rn} \geq t) = \sum_{l=0}^g P(l+1, g+1, \boldsymbol{\alpha}_n) P(\chi_l^2 \geq t),$$

where  $\chi_0^2 \equiv 0$  and  $P(l+1, g+1, \boldsymbol{\alpha}_n)$  is the so called level probability, i.e. the probability that  $E_{\alpha_n}(\mathbf{Z}|\mathcal{C})$  has exactly  $l+1$  distinct values, if  $Z_i \sim \text{iid } N(0, \alpha_{in}^{-1})$ ,  $i = 0, 1, \dots, g$ . Unfortunately the results are not easily computable unless one assumes that  $m_n^{(0)} \sim m_n^{(1)} \sim \dots \sim m_n^{(g)}$  and so we can take  $\alpha_{0n} = \alpha_{1n} = \dots = \alpha_{gn}$ . This assumption is reasonable, for example, in the case of completely random treatment allocation. Actually, from the work of Dykstra and Robertson (1983) it is known, that the limiting distribution is not sensitive to different values of  $\alpha$  as long as their ratios stay between 1/4 and 4.

Under this assumptions the moment generating function of the limiting distribution of  $T_{rn}$  is

$$m_{T_{rn}}(t) = \Theta_g(s) = E(e^{tT_{rn}}) = \int_0^\infty e^{ty} dF_{T_{rn}}(y) = \sum_{l=0}^g \int_0^\infty e^{ty} P(l+1, g+1) dF_{\chi_l^2}(y) = \sum_{l=0}^g P(l+1, g+1) s^l, \quad (8)$$

where  $s = (1 - 2t)^{-1/2}$  for  $t < 1/2$  and  $P(l+1, g+1)$  is the equal weight level probability. While the general unequal weight level probabilities are difficult to compute,  $P(l, g)$  can be easily generated from the recurrence by Miles (1959):

$$P(l+1, g+1) = \frac{1}{g+1} P(l, g) + \frac{g}{g+1} P(l+1, g), \quad (9)$$

$$P(0, g) = P(g+1, g) = 0.$$

This recurrence implies that

$$\Theta_g(s) = \frac{s+g}{g+1} \Theta_{g-1}(s), \quad g = 1, 2, \dots,$$

so the mgf can be explicitly calculated as

$$\Theta_g(s) = \frac{(s+1) \dots (s+g)}{2 \dots (g+1)}. \quad (10)$$

When calculating the distribution of  $T$ , we have to take into account that because of the sparseness of the data for some combinations of  $r, n$  and  $i$  the isotonic weights  $s_{rn}^{(i)}$  take the value 0. For these values of

$r$  and  $n$  the isotonic regression is actually performed on less than  $g + 1$  variables and the distribution of  $T_{rn}$  will reflect this fact. Let  $g_{rn} + 1$  denote the number of non-zero weights among  $s_{rn}^{(i)}$ , i.e. the number of variables on which the isotonic regression is actually performed. Then the moment generating function corresponding to the term  $T_{rn}$  is  $\Theta_{g_{rn}}(s)$ , where  $s = (1 - 2t)^{-1/2}$ .

Since the  $T_{rn}$ 's are independent, the moment generating function of  $T$  is given by

$$\Psi(s) = \prod_{n=1}^K \prod_{r=0}^{n-1} \Theta_{g_{rn}}(s) = \prod_{\gamma=1}^g [\Theta_{\gamma}(s)]^{K_{\gamma}} = \prod_{\gamma=1}^g \left( \frac{s + \gamma}{\gamma + 1} \right)^{N_{\gamma}}, \quad (11)$$

where  $K_{\gamma} = \#\{g_{rn} = \gamma\}$  and  $N_{\gamma} = \sum_{x=\gamma}^g K_x = \#\{g_{rn} \geq \gamma\}$ .

By expanding the product, it is easily shown that  $\Psi(s)$  is a polynomial in  $s = (1 - 2t)^{-1/2}$ . Hence we have proved the following

**Theorem 1** *If  $H_0$  is true, then the asymptotic distribution of  $T$  is chi-bar-squared, that is for any real  $t > 0$*

$$\lim_{M_n \rightarrow \infty} P(T \geq t) = \sum_{l=0}^g a_l P(\chi_l^2 \geq t),$$

where  $a_l$  is the coefficient of  $s^l$  in the expansion of  $\Psi(s)$ .

A computationally simpler, but less accurate method for computing the p-value for the test is to use normal approximation of the null-distribution of  $T$ . Since  $\ln \Psi(s) = \sum_{\gamma=1}^g N_{\gamma} \ln((s + \gamma)/(\gamma + 1))$ , we can find the parameters of the normal approximation:

$$E[T] = (\ln \Psi(s(t)))'|_{t=0} = \sum_{\gamma=1}^g \frac{N_{\gamma}}{1 + \gamma} \quad (12)$$

$$\text{Var}[T] = (\ln \Psi(s(t)))''|_{t=0} = \sum_{\gamma=1}^g N_{\gamma} \left[ \frac{3}{1 + \gamma} - \frac{1}{(1 + \gamma)^2} \right].$$

## 4 Application: DEHP data

We illustrate the methods described above by using the DEHP data from Section 1.

The computed value of the test statistic based on uniform stochastic order is  $T = 212.29$ . The moment generating function of the asymptotic distribution of  $T$  is  $\Psi(s) = \left(\frac{s+1}{2}\right)^{89} \left(\frac{s+2}{3}\right)^{49} \left(\frac{s+3}{4}\right)^{28} \left(\frac{s+4}{5}\right)^{15}$ . By expanding this polynomial, its coefficients can be obtained. The asymptotic p-value obtained as described in the previous section, is 0.000013. For a crude, but quick computation we use the normal approximation.  $E[T] = 70.83$  and  $\text{Var}[T] = 182.45$ , resulting a  $Z$ -score 10.47 with an approximate p-value 0.

The asymptotic results were derived under the condition that the number of clusters for each litter size tends to infinity. In the DEHP data these numbers are fairly small, they never exceed 30. To check the validity of the asymptotics, we estimated the exact p-value of the test statistic using bootstrap resampling. The p-value of the actual test statistic is the probability of observing an even more extreme value if the null-hypothesis of no treatment effect is true. If the null-hypothesis is true, then the outcomes should not

depend on the treatment assignment. In that case we can pool the litters from all the treatment groups. We resample from the pooled data with replacement. Then we randomly assign each selected cluster to a treatment group, being careful to maintain the number of clusters in each group to be equal to those of the actual experiment. For each set of data generated by the resampling described above, we calculate the corresponding test statistic. By repeating this procedure a large number of times (e.g. 10,000 times), we can estimate the distribution of the test statistic, and in particular the p-value. Since for a likelihood ratio test large values of  $T$  indicate deviation from the null-hypothesis, we estimate the p-value as the proportion of the bootstrap samples whose test statistic exceed the actually observed value.

Based on the DEHP dataset we generated 1,000,000 bootstrap samples, but none of them had a test statistics exceeding the observed one. Hence the estimated p-value is less than  $10^{-6}$ . From this analysis we conclude that DEHP exhibits fetal toxicity that is statistically highly significant, since with increasing dose level the probability of having additional affected fetuses increases.

A program calculating the value of the likelihood ratio test statistic and the three approximations of the p-value (chi-bar-squared, normal and resampled) is available upon request.

## 5 Application: Shell Toxicology data

For comparing our procedure with some of those in the literature, we analyze the Shell Toxicology Laboratory data given by Paul (1982). The data were obtained from an experiment in which pregnant banded Dutch rabbits were treated with one of the four experimental levels (Control, Low, Medium and High) of a chemical and the presence of skeletal and visceral abnormalities were recorded as malformations. This dataset can also be obtained at <http://www.hci.utah.edu/groups/biostat/szabo.html>.

We want to point out several interesting aspects of this dataset. First, as the actual dose levels are unknown, parametric analysis can not be performed, but our non-parametric procedure is applicable. We can also check the assumption of no treatment effect on the cluster sizes: there is a slight evidence that at the highest dose group the litters might be smaller, but this difference is not significant (especially because that dose level has the smallest number of clusters). An interesting phenomenon is that the number of responses does not seem to increase at each increase of the administered dose, for example at Medium dose level there seem to be more responses than at High. Empirical estimates of the marginal probabilities of response are estimated as (0.158, 0.135, 0.338, 0.235), so the observed dose- response curve is not monotone.

The test statistics for uniform stochastic trend test is  $T = 41.44$ . Based on the chi-bar-squared approximation, a p-value of 0.023 was obtained. Using bootstrap resampling with 10,000 samples the p-value was approximated to be 0.035. These indicate the existence of a trend defined by uniform stochastic ordering.

Fung et al. (1994) compare several tests, including the Rao-Scott test with design effect adjustment, the Lefkopoulu-Moore-Ryan and Ryan tests for linear trend based on a GEE approach and obtain p-values ranging from 0.003 to 0.009. These results demonstrate that a trend defined in terms of uniform stochastic

ordering is more stringent than the one defined by the above procedures.

## References

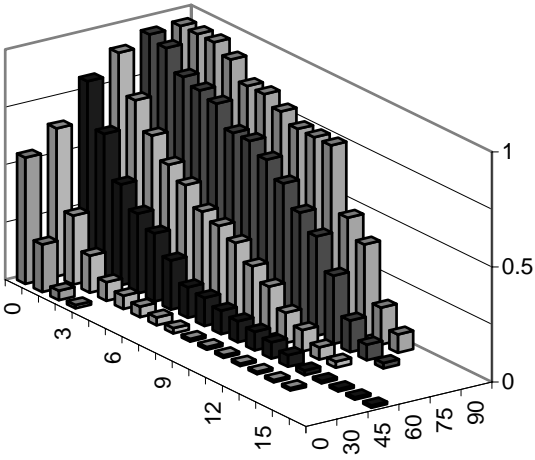
- D. Bowman, J. J. Chen, and E.O. George. Estimating variance functions in developmental toxicity studies. *Biometrics*, 51:1523–1528, 1995.
- J. J. Chen. Trend test for overdispersed proportions. *Biometrical Journal*, 35:949–958, 1993.
- Richard L. Dykstra and Tim Robertson. On testing monotone tendencies. *Journal of the American Statistical Association*, 78(382):342–350, 1983.
- R. L. Dykstra, S. Kochar, and T. Robertson. Statistical inference for uniform stochastic ordering in several populations. *The Annals of Statistics*, 10:870–888, 1991.
- K. Y. Fung, D. Krewski, J. N. K. Rao, and A. J. Scott. Tests for trend in developmental toxicity experiments with correlated binary data. *Risk Analysis*, 14:639–648, 1994.
- E.O. George and D. Bowman. A full likelihood procedure for analyzing exchangeable binary data. *Biometrics*, 51:512–523, 1995.
- E. O. George and R.L. Kodell. Tests of independence, treatment heterogeneity, and dose-related trend with exchangeable binary data. *Journal of the American Statistical association*, 91:1602–1610, 1996.
- L. L. Kupper, C. Portier, M.D. Hogan, and E. Yamamoto. The impact of litter effects on dose- response modeling in teratology. *Biometrics*, 42:85–98, 1986.
- M. Lefkopoulou, D. Moore, and L. Ryan. The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *Biometrics*, 34:69–76, 1989.
- R.E. Miles. The complete amalgamation into blocks by weighted means, of a finite set of real numbers. *Biometrika*, 46:317–327, 1959.
- S. R. Paul. Analysis of proportions of affected fetuses in teratological experiments. *Biometrics*, 38:361–370, 1982.
- J. N. K. Rao and A. J. Scott. A simple method for the analysis of clustered data. *Biometrics*, 48:577–586, 1992.
- T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. Wiley, London, 1988.
- L. M. Ryan. Using historical controls in the analysis of developmental toxicity data. *Biometrics*, 49:1126–1135, 1993.

- R. W. Tyl, C. J. Price, M. C. Marr, and C. A. Kimmel. Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, 10:395–412, 1988.
- D. Zucker and J. Wittes. Testing the effect of treatment in experiments with correlated binary outcomes. *Biometrics*, 48:695–710, 1992.

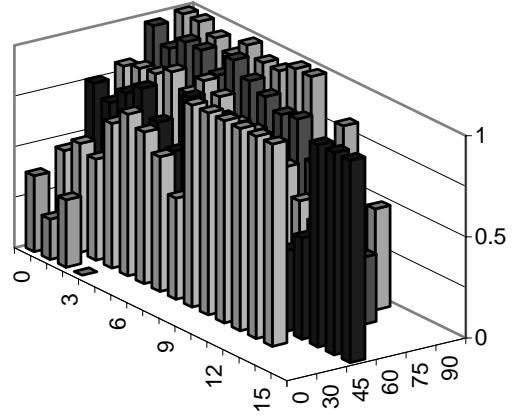
## List of Figures

Figure 1 Plots for assessing stochastic and uniform stochastic ordering for the DEHP data

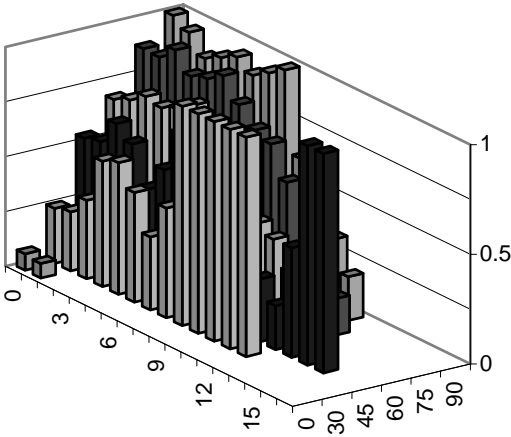
**Stochastic order**



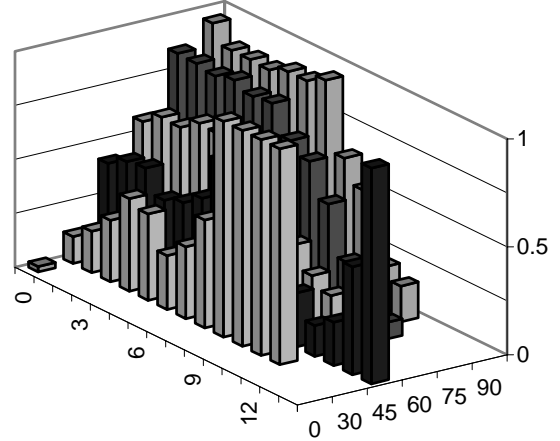
**Uniform stochastic order  
1 additional malformation**



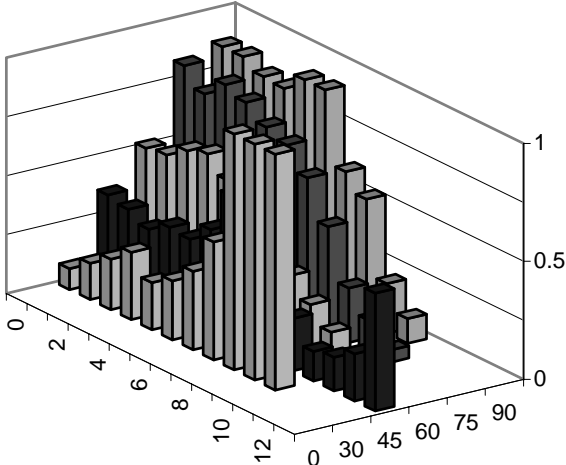
**Uniform stochastic order  
2 additional malformations**



**Uniform stochastic order  
3 additional malformations**



**Uniform stochastic order  
4 additional malformations**



**Uniform stochastic order  
5 additional malformations**

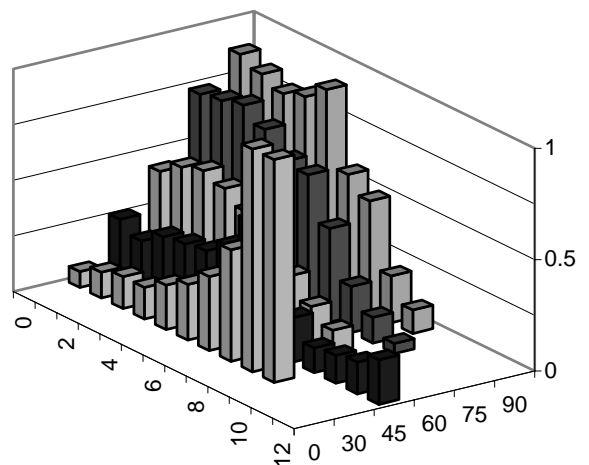


Figure 1  
13