STATISTICAL MODELS FOR THE ANALYSIS OF CERTAIN TOXICOLOGICAL EXPERIMENTS

by

Lawrence L. Kupper

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1169

May 1978

ABSTRACT

This paper provides a review of the current state-of-the-art with regard to the analysis of data generated via laboratory experiments designed to investigate the teratogenic or toxicological effect of certain compounds. In such experiments, the response of interest is typically binary in nature, namely, the occurrence or not of death (or of some particular abnormality) in each fetus in a litter of animals.

This paper discusses the advantages and disadvantages of the various models and methods of analysis which have been proposed for dealing with this type of data, and recommendations are made regarding which techniques seem to be most appropriate.

STATISTICAL MODELS FOR THE ANALYSIS OF CERTAIN TOXICOLOGICAL EXPERIMENTS

by

Lawrence L. Kupper

Department of Biostatistics
University of North Carolina at Chapel Hill

1. INTRODUCTION

In laboratory experiments designed to investigate the teratogenic or toxicological effect of certain compounds, the response of interest is frequently binary in nature, namely, the occurrence or not of "affected" fetuses or implantations in a litter. The "effect" under consideration is generally fetal death or the occurrence of some particular malformation.

The statistical treatment of such data generally requires that the variations in response be described by some underlying probabilistic model, and hence the quality of any subsequent statistical inferences will necessarily depend on how well such a model describes the phenomenon under study.

It is the purpose of this paper to discuss various models and methods of analysis which have been proposed for dealing with this type of data, and to make some recommendations regarding which techniques seem to be most appropriate.

To establish notation, let us suppose that there are ℓ_i litters in the i-th group ($i = 0$ for *control* group and $i = 1$ for *treatment* group), the j-th litter in the i-th group being of size n_{ij}, $j = 1,2,\ldots,\ell_i$. Let

$$x_{ij} = \sum_{k=1}^{n_{ij}} x_{ijk} ,$$

where x_{ijk} is a dichotomous variable taking the value 1 if the k-th fetus in the j-th litter of the i-th group possesses the attribute of interest and taking the value 0 if the attribute is not present. Thus, x_{ij} is the observed total number of affected fetuses out of the n_{ij} under consideration. Further, let

$$\hat{p}_{ij} = x_{ij}/n_{ij} ,$$

$$\hat{p}_{i} = \textstyle\sum_j x_{ij}/\sum_j n_{ij} = \sum_j n_{ij}\hat{p}_{ij}/\sum_j n_{ij} ,$$

$$\bar{p}_{i} = \textstyle\sum_j \hat{p}_{ij}/\ell_i .$$

Note that $\bar{p}_i$ involves no weighting of the $\{\hat{p}_{ij}\}$ by litter size, which can be a problem with small samples.

2. TYPES OF MODELS

Consider the following two probabilistic models.

MODEL A: For the j-th litter in the i-th group, the set $\{n_{ij}, x_{ij}, p_{ij}\}$ is a realization of the trivariate set of random variables $\{N_i, X_i, P_i\}$, where

(i) the marginal distribution of N_i is defined on the non-negative integers, with, say,

$$\Pr(N_i = n) = \pi_{in} , \quad n = 0,1,2,\ldots ;$$

(ii) the marginal density of P_i on the interval $[0,1]$ is, say, $f_i(p)$;

(iii) the conditional distribution of X_i *given* $N_i = n_{ij}$ and $P_i = p_{ij}$ is (typically taken to be) *binomial*, namely

$$\Pr(X_i = x_{ij}/n_{ij}, p_{ij}) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}} (1-p_{ij})^{n_{ij}-x_{ij}} , \quad x_{ij} = 0,1,\ldots,n_{ij} . \qquad (1)$$

Here, n_{ij} and x_{ij} are observable, but p_{ij} is *not* observable.

Comments. Condition (iii) implicitely assumes that, conditional on $N_i = n_{ij}$ and $P_i = p_{ij}$, the within-litter Bernoulli responses are mutually independent, which is often not a realistic assumption. Also, note that $0 \le X_i \le N_i$, that N_i and P_i may or may not be assumed to be independent, and that $\Pr(N_i = 0) = \pi_{i0}$ may or may not be assumed to be equal to zero.

MODEL B: For the j-th litter in the i-th group, the set $\{x_{ij}, \lambda_{ij}\}$ is a realization of the bivariate set of random variables $\{X_i, \Lambda_i\}$, where

(i) the marginal density of Λ_i on the interval $[0,\infty)$ is, say, $g_i(\lambda)$;

(ii) the conditional distribution of X_i given $\Lambda_i = \lambda_{ij}$ is (typically taken to be) *Poisson*, namely

$$\Pr(X_i = x_{ij}/\lambda_{ij}) = \frac{\lambda_{ij}^{x_{ij}} e^{-\lambda_{ij}}}{x_{ij}!}, \quad x_{ij} = 0,1,\ldots,\infty .$$

Here, x_{ij} is observable, but λ_{ij} is *not* observable.

Comments. Model B (in contrast to Model A) completely ignores litter size and considers only the *number* (and not the *proportion*) of dead or malformed fetuses.

In the next two sections, we will consider some particular methodologies associated with Models A and B.

2.1. Methods Associated With Model A.

For

$$\mu_i = \int_0^1 p\, f_i(p)\,dp$$

and for $\pi_{i0} = 0$, VanRyzin [9] has shown that $\bar{p}_i$ is the minimum-variance linear (in the $\hat{p}_{ij}$) unbiased, strongly consistent ($\bar{p}_i \to \mu_i$ with probability 1 as $\ell_i \to \infty$) estimator of μ_i, and that $\sqrt{\ell_i}(\bar{p}_i - \mu_i)/s_i$ converges

in distribution to $N(0,1)$ as $\ell_i \to \infty$, where

$$s_i^2 = \frac{1}{(\ell_i - 1)} \sum_j (\hat{p}_{ij} - \bar{p}_i)^2 .$$

When $0 < \pi_{i0} < 1$, we do not have an identifiable estimation problem if the joint distribution of N_i and P_i is unrestricted; however, the above properties will hold for an estimator based only on litters with $n_{ij} > 0$ as long as N_i and P_i are independent. When P_i is related to N_i, which is likely to be the case in many biological and reproductive systems, VanRyzin develops a moment estimator for μ_i.

A statistical test of $H_0: \mu_1 = \mu_0$ can be constructed based on the asymptotic properties of $\bar{p}_1$ and $\bar{p}_0$. Hoel [4] has discussed an application of these results to the problem of estimating the probability of death and the probability of a malformation given that the fetus is alive.

Results Conditional on Fixed $\{n_{ij}\}$

Southward and VanRyzin [8] compared the variances of $\hat{p}_i$ and $\bar{p}_i$, which are functions of the unknown quantities $\sigma_i^2 = \int_0^1 (p-\mu_i)^2 f_i(p)dp$ and $\tau_i = \int_0^1 p(1-p) f_i(p)dp$; they concluded that neither $\hat{p}_i$ nor $\bar{p}_i$ is for all choices of $f_i(p)$ relatively more efficient than the other. They derived the minimum-variance linear unbiased estimator of μ_i and developed an asymptotically optimal confidence interval for μ_i which involved estimates of σ_i^2 and τ_i.

Williams [10] assumed that model (1) held, took $f_i(p)$ to be a beta distribution with parameters α_i and β_i, and worked with the *beta-binomial* model

$$\Pr(X_i = x_{ij}/n_{ij}) = \binom{n_{ij}}{x_{ij}} \frac{B(\alpha_i + x_{ij}, n_{ij} + \beta_i - x_{ij})}{B(\alpha_i, \beta_i)} ,$$

where $B(\alpha_i,\beta_i) = \Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i+\beta_i)$. Under this model, the within-litter responses are conditionally independent, but are unconditionally dependent. The parameters $\mu_i = \alpha_i(\alpha_i+\beta_i)^{-1}$ and $\sigma_i^2 = \mu_i(1-\mu_i)(\alpha_i+\beta_i+1)^{-1}$ are estimated by maximum likelihood, and treatment and control groups are compared with respect to these parameter values by using asymptotic likelihood ratio tests. Haseman and Soares [3] have shown by simulation that Williams' procedure tends to yield inflated Type I error rates.

Gladen [2] has proposed the use of a *jackknifed* estimator of μ_i of the form

$$\tilde{p}_i = \hat{p}_i + (\ell_i-1)\bar{y}_i ,$$

where

$$\bar{y}_i = \frac{1}{\ell_i}\sum_j y_{ij} = \frac{1}{\ell_i}\sum_j \frac{(x_{ij}-n_{ij}\hat{p}_i)}{(\ell_i-n_{ij})} .$$

It is easy to show that $\tilde{p}_i = \sum_j w_{ij}\hat{p}_{ij}$ and that $\tilde{p}_i = \bar{p}_i$ when $n_{ij} = n_i$. One can regard (e.g., see Miller [7]) either $(\tilde{p}_i-\mu_i)/\tilde{s}_i$ or $(\hat{p}_i-\mu_i)/\tilde{s}_i$ as having approximately a t-distribution with (ℓ_i-1) degrees of freedom, where

$$\tilde{s}_i^2 = \ell_i^{-1}(\ell_i-1)\sum_j (y_{ij}-\bar{y}_i)^2 .$$

To compare treatment and control groups, one can work with the statistic $(\tilde{p}_1-\tilde{p}_0)/(\tilde{s}_1^2+\tilde{s}_0^2)^{\frac{1}{2}}$, which has an approximate t-distribution with $(\ell_1+\ell_0-2)$df under H_0.

Gladen claims that the real advantage of the jackknife approach shows up when dealing with small samples and variable litter sizes; she argues that its use leads to more realistic standard error estimates than those based directly on the binomial distribution, e.g., like $[\hat{p}_i(1-\hat{p}_i)/\sum_j n_{ij}]^{\frac{1}{2}}$. However, note that $\tilde{p}_i$ can take a value outside the interval $[0,1]$.

Starting with model (1), Gladen shows that $(\tilde{p}_i-\mu_i)/\tilde{s}_i$ and $(\hat{p}_i-\mu_i)/\tilde{s}_i$ are both asymptotically $N(0,1)$ under the assumption that the litter sizes come from some bounded distribution. Asymptotic efficiency results based on the beta-binomial model using observed litter size distributions and estimated parameter values for three particular data sets (presented in Section 3) indicated that $\tilde{p}_i$ (or, equivalently, $\hat{p}_i$) was almost fully efficient, while $\bar{p}_i$ was somewhat less efficient but still respectable. Some of her simulation results will be mentioned in Section 3.

Results Conditional on Fixed $\{n_{ij}\}$ and $P_i = p_i$

Kupper and Haseman [5] considered a two-parameter generalization of model (1) of the form

$$\Pr(X_i = x_{ij}/n_{ij}, p_i) = \binom{n_{ij}}{x_{ij}} p_i^{x_{ij}} (1-p_i)^{n_{ij}-x_{ij}} \left\{1 + \frac{\theta_i}{2p_i^2(1-p_i)^2}\left[(x_{ij}-n_{ij}p_i)^2 + x_{ij}(2p_i-1) - n_{ij}p_i^2\right]\right\},$$

where θ_i is the covariance between any two responses within the same litter. This *correlated-binomial* model allows for the possiblity of negative intra-litter correlation, while Williams' beta-binomial model permits only supra-binomial variation. As with Williams' model, likelihood ratio tests are employed to assess the significance of treatment-control differences, and the use of either model necessitates consideration of boundary conditions with regard to maximization of the corresponding likelihood functions. A comparison between the correlated-binomial and beta-binomial models will be presented in Section 3.

Altham [1] considered another two-parameter generalization of model (1) of the form

$$\Pr(X_i = x_{ij}/n_{ij},p_i) = \binom{n_{ij}}{x_{ij}} p_i^{x_{ij}} (1-p_i)^{n_{ij}-x_{ij}} \delta_i^{x_{ij}(n_{ij}-x_{ij})} / f(p_i,\delta_i,n_{ij}) ,$$

where

$$f(p_i,\delta_i,n_{ij}) = \sum_j \binom{n_{ij}}{x_{ij}} p_i^{x_{ij}} (1-p_i)^{n_{ij}-x_{ij}} \delta_i^{x_{ij}(n_{ij}-x_{ij})} .$$

This model reduces to (1) when $\delta_i = 1$; if $\delta_i > 1$, the distribution is (strongly) unimodal and more sharply peaked than the binomial, which implies negative intra-litter association; for $0 < \delta_i < 1$, the distribution is more diffuse than the binomial and the intra-litter responses are positively correlated. Altham compares this model to the correlated-binomial model, and generally seems to prefer the latter.

Other Procedures

A commonly used approach is to employ ordinary χ^2 tests (CHI) for comparing $\hat{p}_1$ and $\hat{p}_0$. Haseman and Soares [3] have shown that such tests tend to operate at much above nominal significance levels (see Section 3), the main reason being that they involve the use of binomial standard errors.

Most other approaches involve the use of t-tests or Mann-Whitney U-tests (MWU) or the $\{\hat{p}_{ij}\}$ or on transformations thereof. Popular transformations are the Freeman-Tukey binomial (FTB) variance-stabilizing arc-sine transformation

$$\frac{1}{2}\left\{\sin^{-1}\sqrt{\frac{x_{ij}}{n_{ij}+1}} + \sin^{-1}\sqrt{\frac{x_{ij}+1}{n_{ij}+1}}\right\} ,$$

or the arc-sine transformation (ARC)

$$\begin{cases} \sin^{-1}\sqrt{x_{ij}/n_{ij}} & \text{if } x_{ij} \neq 0, n_{ij} \\ \sin^{-1}\sqrt{1/4n_{ij}} & \text{if } x_{ij} = 0 \\ \sin^{-1}(1) - \sin^{-1}\sqrt{1/4n_{ij}} & \text{if } x_{ij} = n_{ij} . \end{cases}$$

2.2. Methods Associated With Model B

McCaughran and Arnold [6] take $g_i(\lambda)$ to be a gamma distribution, so that the unconditional distribution of X_i is negative binomial. They employ the method of moments to estimate parameters of interest, then perform a variance-stabilizing transformation which depends on the parameter estimates, and then do t-tests on the transformed data to assess treatment-control differences.

The Freeman-Tukey Poisson (FTP) transformation

$$\sqrt{x_{ij}} + \sqrt{x_{ij}+1}$$

has also been utilized, with t-tests performed on the transformed data.

3. SOME SIMULATION RESULTS

Haseman and Soares [3] utilized the empirical distributions of fetal death presented in Table 1 to illustrate that such data rarely can be adequately modeled using a binomial or a Poisson distribution (see Table 2). They then constructed three infinite populations having the same relative frequencies of fetal death as the control groups shown in Table 1. From each population two random samples of 20 pregnant females were drawn, and comparisons of fetal death were made by chi-square (CHI) and by a Student's t-test based on the Freeman-Tukey transformation for Poisson counts (FTP), the Freeman-Tukey transformation for binomial

proportions (FTB), and the arc-sine transformation (ARC). The Mann-Whitney U-test based on the proportion of dead implants (MWU) was also considered. The process was repeated 5000 times for each population, and the frequency with which each test rejected the null hypothesis of no effect was recorded. It was found (see Table 3) that all procedures except chi-square appeared to be operating at approximately the correct level of significance. As far as power considerations are concerned, a number of computer simulations with various distributions of fetal death led to the following conclusions: (1) For treatments producing no pre-implantation loss, differences in power among the four procedures for detecting increases in fetal death were generally slight, with no one procedure being consistently superior to any other; (2) As pre-implantation losses increased in the treated group, FTP became progressively worse relative to the other three procedures for detecting increases in fetal death.

Gladen [2] also conducted some simulation studies using these three control populations as well as three others. Her results supported the findings of Haseman and Soares, and she also found that her jackknife procedure performed about as well as the FTP, FTB, ARC and MWU approaches.

Finally, Table 4 presents a comparison of the beta-binomial and correlated-binomial model fits to the three data sets considered by Haseman and Soares. As can be seen, there is little difference between the fits of the two models, and the improvement in fit relative to the binomial distribution is quite impressive.

RECOMMENDATIONS

Based on all these results, what should we do? Motivated by the KISS philosophy ("keep it simple, stupid"), I would be inclined to use a Mann-Whitney U-test on the $\{\hat{p}_{ij}\}$ to compare treatment and control groups, since it seems to operate at about the right significance level, appears to have power comparable to other more complex procedures, and is simple to understand, explain and use. When the study design is of the multi-factor type (involving, for example, additional factors like time period or strain of mice), then an analysis of variance using the FTB or ARC transformation would be appropriate.

REFERENCES

[1] Altham, P.M.E. (1977). "Two Generalizations of the Binomial Distribution," unpublished manuscript.

[2] Gladen, B. (1977). "The Use of the Jackknife to Estimate Proportions from Toxicological Data in the Presence of Litter Effects," unpublished manuscript.

[3] Haseman, J.K. and Soares, E.R. (1976). "The Distribution of Fetal Death in Control Mice and Its Implications on Statistical Tests for Dominant Lethal Effects," *Mutation Research* 41, 277-288.

[4] Hoel, D.G. (1974). "Some Statistical Aspects of Experiments for Determining the Teratogenic Effects of Chemicals." In *Statistical and Mathematical Aspects of Pollution Problems*, Edited by John W. Pratt, Marcel Dekker, Inc., NY, 375-381.

[5] Kupper, L.L. and Haseman, J.K. (1978). "On the Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," to appear in the March 1978 issue of *Biometrics*.

[6] McCaughran, D.A. and Arnold, D.W. (1976). "Statistical Models for Numbers of Implantation Sites and Embryonic Deaths in Mice," *Toxicology and Applied Pharmacology* 38, 325-333.

[7] Miller, R.G. (1974). "The Jackknife-A Review," *Biometrika* 61, 1-15.

[8] Southward, G.M. and VanRyzin, J. (1972). "Estimating the Mean of a Random Binomial Parameter," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 4, 249-263.

[9] VanRyzin, J. (1975). "Estimating the Mean of a Random Binomial Parameter With Trial Size Random," *Sankhyā, Series B* 37, 10-27.

[10] Williams, D.A. (1975). The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity," *Biometrics* 31, 949-952.

TABLE 1

DISTRIBUTION OF FETAL MORTALITY IN 3 GROUPS OF CONTROL MICE

Number of implants	Data Set 1							Data Set 2						Data Set 3						
	Number of dead implants							Number of dead implants						Number of dead implants						
	0	1	2	3	4	5	>5[a]	0	1	2	3	4	>4[b]	0	1	2	3	4	5	>5[c]
1	2							15	1					7						
2	2							6	1	2				7						
3	3							6	6					6						
4	5	1	1					7	2	3		2		5	2	1				
5	2	2						16	9	3	3	1		8	2	1		1	1	
6	2	2						57	38	17	2	2		8						
7	2	2	2	1				119	81	45	6	1	1	4	4	2	1			
8	6	1		1	1			173	118	57	16	3	1	7	7	1				
9	2	3	1					136	103	50	13	6	2	8	9	7	1	1		
10	2	4	2		2			54	51	32	5	1	1	22	17	2		1		2
11	19	11	3	3				13	15	12	3	1		30	18	9	1	2		2
12	33	24	11	5	4	4	1		4	3	1			54	27	12	2	1		2
13	39	27	12	6	5	2	1			1			1	46	30	8	4	1	1	1
14	34	30	14	6	6		1							43	21	13	3	1		2
15	38	22	18	4	2	1								22	22	5	2	1		
16	13	16	14	4	3	1								6	6	3		1	1	
17	8	4	3	3	2	1	1													
18		4	2	1										3		2	1			
19	2	1																		
20							1													

[a]These five females had 12/12, 8/13, 7/14, 7/17 and 13/20 dead implants.

[b]These six females had 7/7, 8/8, 5/9, 6/9, 10/10 and 7/13 dead implants.

[c]These nine females had 7/10, 8/10, 6/11, 8/11, 6/12, 6/12, 7/13, 7/14 and 9/14 dead implants.

TABLE 2

Poisson and Binomial Fit to Control Data[a]

Number of dead implants	Data Set 1			Data Set 2			Data Set 3		
	Observed frequency	Expected frequency		Observed frequency	Expected frequency		Observed frequency	Expected frequency	
		Poisson	binomial		Poisson	binomial		Poisson	binomial
0	214	162.3	160.0	602	564.1	548.5	286	241.1	241.2
1	154	190.2	191.0	429	483.0	500.1	165	200.6	199.5
2	83	111.5	114.2	225	206.8	213.5	66	83.5	84.8
3	34	43.5	43.8	49	59.0	55.1	15	23.1	23.3
>3	39	16.4	14.9	23	15.1	10.8	22	5.7	5.3
χ^2_3 (test of fit)		63.6[c]	75.1[c]		16.0[b]	30.5[c]		67.8[c]	74.0[c]

[a]See Haseman and Soares [3] for the complete distributions of fetal death.

[b]$P < 0.01$.

[c]$P < 0.001$.

TABLE 3

SIMULATED SIGNIFICANCE LEVELS FOR COMPARISONS OF FETAL DEATH IN TWO SAMPLES OF 20 PREGNANT MICE DRAWN FROM THE SAME POPULATION[a]

Population	Assumed level of significance	Actual level of significance				
		MWU	ARC	FTP	FTB	CHI
Data Set 1	0.10	0.101	0.102	0.100	0.103	0.187
	0.05	0.052	0.049	0.051	0.048	0.129
	0.01	0.010	0.008	0.010	0.011	0.051
Data Set 2	0.10	0.103	0.103	0.103	0.105	0.133
	0.05	0.052	0.056	0.054	0.055	0.082
	0.01	0.010	0.011	0.012	0.011	0.024
	0.10	0.106	0.106	0.106	0.107	0.181
Data Set 3	0.05	0.055	0.053	0.055	0.054	0.127
	0.01	0.012	0.010	0.013	0.011	0.052

[a]Based on 5000 iterations. Test procedures are defined in the text.

TABLE 4

Beta-Binomial and Correlated-Binomial Fits to Control Data in Table 1

Number of dead implants	Data Set 1			Data Set 2			Data Set 3		
	Observed frequency	Expected frequency		Observed frequency	Expected frequency		Observed frequency	Expected frequency	
		Beta-binomial	Correlated-binomial		Beta-binomial	Correlated-binomial		Beta-binomial	Correlated-binomial
0	214	222.26	202.45	602	612.18	594.72	286	299.86	272.14
1	154	139.87	149.11	429	421.47	440.43	165	134.74	153.30
2	83	79.16	81.85	225	194.15	200.90	66	64.03	74.12
3	34	42.34	51.23	49	71.22	70.33	15	30.41	36.63
>3	39	40.37	39.36	23	28.97	21.62	22	24.46	17.81
Parameter estimates		$\hat{\mu}$=.0900 $\hat{\sigma}^2$=.0056	$\hat{p}$=.0931 $\hat{\theta}$=.0037		$\hat{\mu}$=.1090 $\hat{\sigma}^2$=.0042	$\hat{p}$=.1086 $\hat{\theta}$=.0027		$\hat{\mu}$=.0735 $\hat{\sigma}^2$=.0051	$\hat{p}$=.0760 $\hat{\theta}$=.0027
χ^2_2 (test of fit)		3.61	6.63		13.37	9.83		15.66	16.25