Environmental Health Perspectives
Vol. 63, pp. 187-194, 1985

Statistical Studies in Genetic Toxicology: A Perspective from the U.S. National Toxicology Program

by Barry H. Margolin*

This paper surveys recent, as yet unpublished, statistical studies arising from research in genetic toxicology within the U.S. National Toxicology Program (NTP). These studies all involve analyses of data from Ames Salmonella/microsome mutagenicity tests, but the statistical methodologies are broadly applicable. Three issues are addressed: First, what is a tenable sampling model for Ames test data, and how does one best test the adequacy of the Poisson sampling assumption? Second, given that nonmonotone dose–response curves are fairly common in the Salmonella assay, what new statistical techniques or modifications of existing ones seem appropriate to accommodate to this reality? Finally, an intriguing question: How can the extensive NTP Ames test data base be used to assess the characteristics of any mutagen–nonmutagen decision rule? The last issue is illustrated with the commonly used "two-times background" rule.

Introduction

During the last decade the science of genetic toxicology has experienced dramatic growth in its volume of experimentation, its variety of assays, and the level of public awareness of it. Even laymen are likely to have heard of some of the tests in this field or seen newspaper accounts of results from one. This growth, in all its dimensions, is attributable to the ability of these test systems to detect, rapidly and relatively inexpensively, environmental agents that are genotoxic; these agents are thought to be implicated in such diverse human health problems as cancer, aging, and birth defects (*1*). It is reputed that over 2000 laboratories worldwide, in industry, academia, and government, currently perform the Ames Salmonella/microsome test (*2*), the best known and most widely employed of the short-term tests for genetic toxicity. In many parts of the industrialized world, regulatory decisions regarding the registration of pesticides or pharmaceuticals are based in part on results from tests for genetic toxicity. In some countries, such as Japan and the United States, these tests are used in national programs to screen agents already in the environment. It is worth remembering, however, that this area of scientific research is far from mature; much remains to be achieved in terms of understanding the precise implications of results from such tests for the assessment of risks to human health.

To date, man-made industrial agents have been the primary focus of research interest in this area; there is, however, an increasing emphasis on naturally occurring potential sources of genetic toxicity, such as common dietary components. The term "genetic toxicity" is applied to the induction of genetic damage by any agent, whether the damage be DNA point mutations at a particular locus, induction of DNA repair, binding to DNA, or chromosomal aberrations, such as fragments or aneuploidy. The chronic rodent carcinogenicity bioassay, technically speaking, is not a test for genetic toxicity because tumor development has not yet been demonstrated to result directly from genetic damage. The somatic-mutation theory of cancer (*3*), however, is seemingly reinforced weekly by new experimental findings.

Unlike the chronic rodent carcinogenicity bioassay, for which there is a rich statistical literature, the tests for genetic toxicity have only recently begun to attract the attention of research statisticians; witness the dearth of published papers containing new statistical methodology motivated wholly or in part by problems in genetic toxicology. Two exceptions are the works by Collings, Margolin and Oehlert (*4*) on the analysis of binomial data and by Tarone (*5*) on the use of historical control data. Although five years ago Hollstein et al. (*6*), in an excellent review of short-term tests for genetic toxicity, could cite over 100 assays that had at least a modicum of representation in the published literature, the U.S. National Toxicology Program (NTP) has fewer than 20 assays in use, undergoing validation or in de-

*Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709.

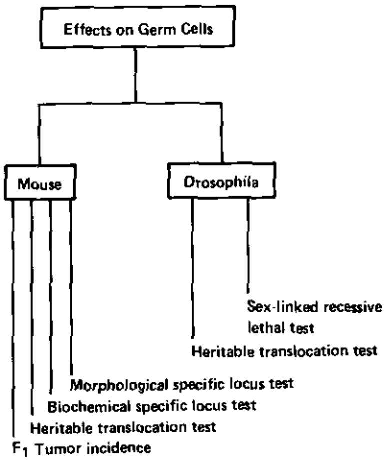
Effects on Germ Cells
Mouse
Drosophila
Sex-linked recessive lethal test
Heritable translocation test
Morphological specific locus test
Biochemical specific locus test
Heritable translocation test
F1 Tumor incidence

FIGURE 1. NTP tests for genetic toxicity in use, undergoing validation, or in development: germ cell targets. Listing courtesy R. Tennant, NTP.

velopment. Figures 1 and 2 present these assays, separated as to whether the target used to probe an agent's ability to induce genetic damage is a somatic or a germ cell. Fewer than half a dozen of these assays have been studied carefully by statisticians, and fewer still have methods of statistical analysis that are generally accepted.

The NTP statistical effort in the area of genetic toxicology has emphasized the development of objective analyses of individual test results, methods for meaningful assay validation regarding operating characteristics, and large data bases, which can be exploited for a variety of purposes, such as devising screening strategies, measuring interlaboratory and interassay concordance, and attempting to ascertain the degree of predictivity of short-term tests for the chronic rodent carcinogenicity bioassay.

This paper surveys the methodological components of a series of statistical projects, largely unpublished as of this date, that were conducted under the direction of the author in response to perceived needs in genetic toxicology within the NTP. The principal issues addressed are: tenable sampling models and goodness of fit; nonmonotone dose–response relationships and tests of significance; and external validation of tests of hypothesis when there is sufficient replication. All three issues will be illustrated with the Ames test, but their importance transcends this one assay.

Tenable Sampling Models and Goodness of Fit

The development of a parametric statistical analysis for data arising from short-term tests is best achieved after scrutiny of a variety of data, preferably generated by different technicians and at least two laboratories. One main component in the developmental process is the creation of a tenable sampling model. Significant departures of reality from assumption with regard to the sampling distribution can impact substantially on false positive and false negative rates, and on the efficiency of estimators (7,8).

For short-term tests, the response of interest is frequently a count, which may be bounded by definition or

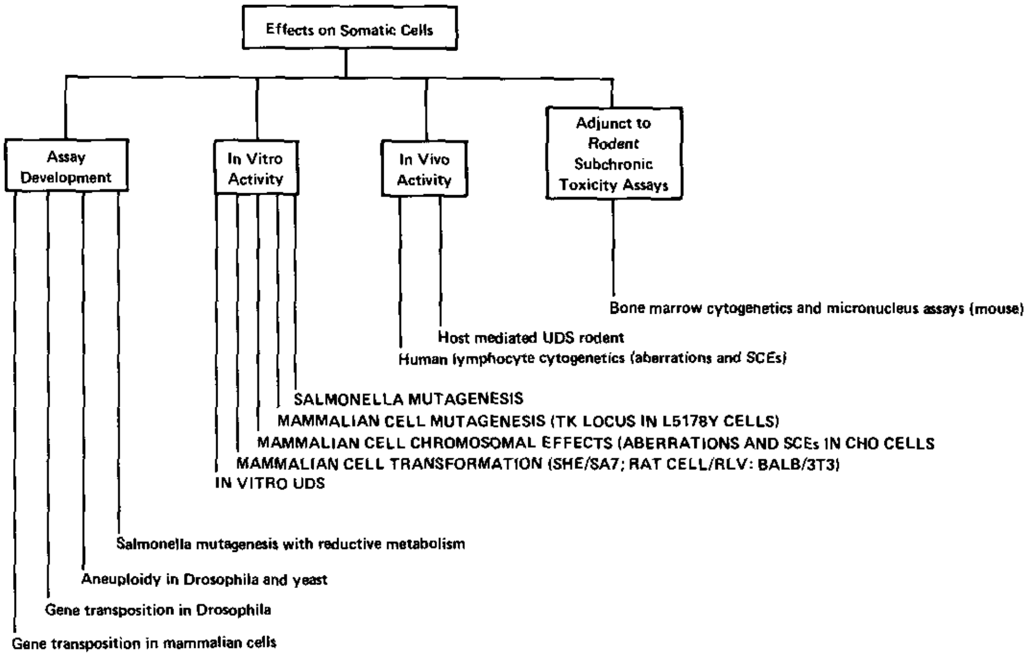
Effects on Somatic Cells
Assay Development
In Vitro Activity
In Vivo Activity
Adjunct to Rodent Subchronic Toxicity Assays
Bone marrow cytogenetics and micronucleus assays (mouse)
Host mediated UDS rodent
Human lymphocyte cytogenetics (aberrations and SCEs)
SALMONELLA MUTAGENESIS
MAMMALIAN CELL MUTAGENESIS (TK LOCUS IN L5178Y CELLS)
MAMMALIAN CELL CHROMOSOMAL EFFECTS (ABERRATIONS AND SCEs IN CHO CELLS
MAMMALIAN CELL TRANSFORMATION (SHE/SA7; RAT CELL/RLV: BALB/3T3)
IN VITRO UDS
Salmonella mutagenesis with reductive metabolism
Aneuploidy in Drosophila and yeast
Gene transposition in Drosophila
Gene transposition in mammalian cells

FIGURE 2. NTP tests for genetic toxicity in use, undergoing validation, or in development: somatic cell targets. Listing courtesy R. Tennant, NTP.

not. For the Ames Salmonella/microsome test the response is unbounded, being the number of visible bacterial colonies per plate after plating and incubation of approximately 10^8 histidine-dependent bacteria, together with minimal medium and a dose of test compound. All pipettings included in the protocol presumably contribute stochastic variability.

Early authors discussing analyses of Ames test data assumed Poisson sampling without producing any empirical supporting evidence. From a theoretical standpoint, the usual Poisson assumptions seem credible for a given plate, but to extrapolate from one plate and claim that a set of plate counts behaves like a random sample from a Poisson distribution requires an additional assumption of homogeneity of environments across plates. In some laboratories that condition may obtain, but the key point is that this issue is open to empirical study. The concept of uniformity trials (*9*) from agricultural research deserves renewed consideration by experimenters and statisticians; it suggests the desirability of running assays early in their development as one would to test a compound for genetic toxicity, but with no test compound added. Ideally, data from such negative or solvent control trials can then support or refute a particular sampling model, and can be used to assess the possibility of hidden components of variability.

Margolin et al. (*8*) reported results from 20 replicated control plates for Ames tests conducted by each of three laboratories. If $Y_1, \ldots, Y_n$ represent the control plate counts observed by a laboratory on a given day, and if $\bar{Y}$ denotes their mean, then a standard test of the Poisson sampling assumption is based on the statistic

$$T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2/\bar{Y} \tag{1}$$

When the data are a random sample from a Poisson distribution, the statistic T is well approximated by a chi-square random variable with n - 1 degrees of freedom. Using this fact, Margolin et al. (*8*) demonstrated that the Poisson model is inadequate to describe Ames test data; sample variance to mean ratios of 4 or larger were reported by them. In place of the Poisson, those authors adopted a negative binomial sampling model, which they motivated as a stochastic mixture of Poisson distributions created by pipetting errors. Additional evidence, both empirical and theoretical, in support of the negative binomial (NB) model for Ames test data is given by Collings and Margolin (*10*), who employed the following form of the negative binomial distribution:

$$P\{Y = y\} = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})} \left(\frac{cm}{1 + cm}\right)^{y} \left(\frac{1}{1 + cm}\right)^{c^{-1}} \tag{2}$$

for $y = 0, 1, 2, \ldots, 0 < m < \infty$, and $0 < c < \infty$. As a shorthand, Y will be said to be distributed NB (m,c) if Eq. (2) obtains. Here m is the mean of Y and the limit of Eq. (2) as $c \rightarrow 0$ is the Poisson distribution with mean m. Thus, Eq. (2) extends naturally to incorporate the Poisson distribution at $c = 0$. With this formulation, one can speak of the distribution of the maximum likelihood estimate (MLE) for c, which now has finite moments of all positive orders (*11*). Contrast this with the more common parametrization in terms of $k = c^{-1}$, where the MLE of k does not possess a proper distribution (*12*).

Although control trials are highly useful, they are rarely available. In general, even a good-sized random sample of control plates is hard to come by. For example, the data of Margolin et al. (*8,10*) are unique in the literature on the Ames test.

For the general short-term test in which unbounded count data are observed, the test results that would be available for assessing the goodness of fit of the Poisson assumption are from experiments with varying doses of true test compounds. These data are not identically distributed, but rather have a one-way layout structure indexed by dose. An extension to the one-way layout of the goodness of fit test for Poisson sampling based on Eq. (1) is studied by Collings and Margolin (*10*), who obtain the following result.

THEOREM: If Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, r$, are independent and Y_{ij} is distributed NB (m_i,c), then the locally most powerful unbiased test of H_0: $c = 0$ (Poisson) versus H_1: c > 0 (overdispersion) is conditional on

$$\{\bar{Y}_{i+} = \sum_{j=1}^{n_i} Y_{ij}/n_i\}_{i=1}^{r}$$

and rejects H_0 for large values of

$$T_C = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2/\bar{Y}_{++} \tag{3}$$

where

$$\bar{Y}_{++} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} \Big/ \sum_{k=1}^{r} n_k$$

The null sampling distribution of T_C and the power of the test based upon it were also studied (*10*).

The theorem above generalizes a result obtained by Potthoff and Whittinghill (*13*) for the case of the goodness-of-fit test for a Poisson random sample based upon Eq. (1). A test statistic competitive to that in Eq. (3) is to aggregate the value of Eq. (1) obtained for each group separately, i.e.,

$$S_C = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2/\bar{Y}_{i+} \tag{4}$$

and reference a chi-square distribution with $(\Sigma n_i) - r$

degrees of freedom. Collings and Margolin (*10*) prove that if $n_i/\Sigma n_j \rightarrow p_i$, a constant for each $i = 1, \ldots, r$, such that $0 < p_i < 1$ and $\Sigma p_i = 1$, then the Pitman asymptotic relative efficiency of S_C to T_C is given by

$$e_C = \left(\sum_i m_i p_i\right)^2 \Big/ \sum_i m_i^2 p_i \qquad (5)$$

Moreover

$$(\min_i p_i) \leqslant e_C \leqslant 1$$

with equality obtaining on the right if and only if the $\{m_i\}$ are all equal. Loosely speaking, the gain in sensitivity of T_C over S_C increases as the $\{m_i\}$ become more disparate. Those authors also discuss goodness of fit testing for the Poisson sampling assumption when the data are such that Y_i is distributed NB (m_i, c), with $m_i = \beta_i m$ for β_i a known positive constant, $i = 1, \ldots, n$; they obtain a test for this regression-through-the-origin case that is suitable when there is reason to believe the linearity, and when the $\{\beta_i\}$ are quite disparate, as in dosing studies with doses spaced logarithmically.

Nonmonotone Dose–Response and Tests of Significance

Were the possibility of hyper-Poisson sampling variability for Ames test data their only distinguishing feature, one could readily modify inference procedures intended for Poisson data so that these procedures were appropriate for negative binomial data, thereby accommodating the overdispersion. To illustrate, a commonly used procedure to test a quantitative factor d, such as dose, for its effect on Poisson means is to compute the Cochran-Armitage test (*14,15*) for trend in the means. If for each i, X_i is distributed as a Poisson random variable with mean λ_i and this observation is associated with a level d_i of a quantitative factor, then the trend test of H_0: $\lambda_i = \lambda$ for all i, versus H_1: λ_i ordered by d_i, is based on the statistic

$$Z = \sum_{i=1}^{n} X_i(d_i - \bar{d})/s_x \left[\sum_{j=1}^{n} (d_j - \bar{d})^2\right]^{1/2} \qquad (6)$$

where $s_x^2 = \bar{X} = \Sigma X_i/n$ and $\bar{d} = \Sigma d_i/n$. Z in Eq. (6) can easily be seen to be the regression coefficient for X regressed on d, normalized by its estimated standard deviation. Under H_0, Z is distributed approximately as a standard normal random variable. Tarone (*5*) has shown that the test based upon Eq. (6) is asymptotically locally optimal against any smooth monotone function expressing λ in terms of d.

The modification of the Cochran-Armitage trend test needed to permit its use for negative binomial data is to define $s_x^2 = \bar{X}(1 + \hat{c}\bar{X})$, where $\hat{c}$ is the MLE of c in Eq. (2) when the data are considered as a random sample (H_0). Again, the reference distribution for Z is the standard normal. The Appendix contains a demonstration paralleling that of Tarone (*5*), which establishes that the test for trend among negative binomial means is asymptotically locally optimal against any smooth monotone function that expresses m in terms of d. As Collings and Margolin (*10*) note, the negative binomial distribution in Eq. (2) can be extended to include the binomial as well as the Poisson distribution. The Appendix then contains a proof that holds for all three models.

Table 1 presents results from a small Monte Carlo study of the size of the one-tailed test for trend in negative binomial means. To mimic typical experimentation, the Monte Carlo included six dose groups, with either three or five replicate observations per dose. The dosing was either linear (specified by $d = 0, 1, \ldots, 5$) or logarithmic (specified by $d = 0, 1, 10, \ldots, 10^4$). Note that these specifications entail no loss in generality because Eq. (6) is invariant to scale transformations of dose. The values for m were set at 15 and 150, whereas c was either 0 (Poisson) or $3/m$ (highly overdispersed). Each of the 1000 data sets randomly generated for a given set of conditions was analyzed two ways, once with the true c used in s_x in Eq. (6) and once with the MLE of c, as would be the case with real data. The results indicate that the size of the trend test is well approximated by the standard normal tail area whether c is known or estimated from the data.

A more interesting characteristic of Ames test data that separates them from most other dose response data treated in the statistics literature is that the dose–response for Ames test data is frequently not monotone (*8*). There are other *in vitro* assays for genetic toxicity that exhibit similar behavior, e.g., the fluctuation test (*4*) and the mouse lymphoma assay (personal communication from W. Caspary, NTP). The common decrease in mean response at high doses, sometimes to levels below that for the control, is usually attributed to toxicity that prevents an experimental unit from exhibiting phenotypic evidence of mutagenicity. Decreases in the mean response at high doses, especially to or below control levels, impact heavily on the power of trend tests (*4*), which place their greatest weight on the responses to the control and maximum dose.

Three published significance tests for various short-term tests attempt to cope with a nonmonotone dose response. First, Collings et al. (*4*) proposed the use of an isotonic test for fluctuation test data; this test, while not tailored to the situation under discussion, exhibits a greater degree of power robustness against downturns than does the binomial trend test. Second, Bernstein et al. (*16*) proposed a recursive analysis for Ames test data in which the response at highest dose is sub-

Table 1. A Monte Carlo study of the true size of a one-tailed test for trend in negative binomial means (1000 replications).

				Treatment of c in the analysis							
				True c used				c estimated from data			
m	c	r[a]	Scale[b]	0.100[c]	0.050	0.025	0.010	0.100	0.050	00.025	0.010
15	0	3	LIN	0.090	0.041	0.012	0.007	0.076	0.029	0.009	0.002
15	0	3	LOG	0.089	0.057	0.026	0.012	0.084	0.045	0.016	0.004
15	0	5	LIN	0.082	0.044	0.016	0.006	0.074	0.034	0.010	0.005
15	0	5	LOG	0.088	0.040	0.019	0.010	0.081	0.033	0.015	0.008
15	0.2	3	LIN	0.088	0.038	0.012	0.005	0.090	0.041	0.013	0.005
15	0.2	3	LOG	0.100	0.059	0.035	0.015	0.113	0.064	0.034	0.011
15	0.2	5	LIN	0.084	0.039	0.022	0.005	0.089	0.043	0.017	0.006
15	0.2	5	LOG	0.092	0.045	0.026	0.014	0.096	0.050	0.028	0.010
150	0	3	LIN	0.109	0.056	0.029	0.009	0.095	0.039	0.018	0.006
150	0	3	LOG	0.112	0.059	0.029	0.010	0.094	0.048	0.018	0.006
150	0	5	LIN	0.103	0.061	0.029	0.012	0.094	0.048	0.022	0.008
150	0	5	LOG	0.108	0.057	0.030	0.007	0.096	0.040	0.018	0.004
150	0.02	3	LIN	0.102	0.052	0.026	0.011	0.116	0.050	0.022	0.010
150	0.02	3	LOG	0.113	0.061	0.033	0.013	0.117	0.059	0.030	0.013
150	0.02	5	LIN	0.104	0.058	0.032	0.014	0.111	0.056	0.029	0.011
150	0.02	5	LOG	0.111	0.060	0.031	0.008	0.108	0.059	0.028	0.007

[a] r is the number of replicates per dose at each of six doses.
[b] For LIN, the doses were 0, 1, 2, 3, 4, and 5; for LOG, the doses were 0, 1, 10, 100, 1000, and 10000.
[c] Nominal size.

jected to a pretest for downward departure from linearity. If the pretest supports such a downturn, then the highest dose is excluded from the analysis and the next highest dose is similarly scrutinized. When this "point-rejection" procedure terminates, the remaining doses are subjected to a trend test modified for unequal variances. Finally, Margolin et al. (*8*) developed mechanistic biomathematical models that reflect a somewhat simplified view of the underlying biology of an Ames test. They proposed a test of significance based on the MLE of a parameter in their model that represents a mutagenic index. The last two analyses are clearly in need of further study to understand better their operating characteristics. Work on the latter is nearing completion and will be reported elsewhere.

The use of nonparametric procedures, especially Jonckheere's test (*17*), has been advocated for analyzing data from short-term tests for genetic toxicity (*18,19*). Simpson and Margolin (unpublished manuscript) have shown that nonparametric tests that are tailored to detect ordered alternatives, such as Jonckheere's, can have their power functions substantially depressed by a downturn in the underlying dose response function. Consequently, they devised a recursive strategy that excludes data obtained at the highest dose if there is evidence of a substantial downturn in response at that dose, i.e., a departure from monotonicity.

This check for a downturn is performed recursively with a Wilcoxon test, and when it terminates, the remaining doses are subjected to Jonckheere's test. The key consideration in doing this analysis recursively is to retain control of the size of the test. Simpson and Margolin present both empirical and analytic evidence for proper size behavior of their test. They also show that their procedure is consistent for the cases of interest and offers substantial improvement in power over Jonckheere's test when there is a sizeable downturn in dose response at high doses. This gain is achieved at a cost of a modest loss of power when the underlying response is, in fact, monotone in dose.

External Validation of Tests of Hypothesis

One further important way in which the Ames Salmonella assay is unusual is in its sheer volume of usage; because the assay is fast and relatively inexpensive, it lends itself nicely to screening efforts. Since its creation in 1978, the NTP has had as one of its broad goals the extensive screening of environmental agents for evidence of genetic toxicity. To date, the data collected have come overwhelmingly from Ames tests on four strains of *Salmonella typhimurium* (TA98, TA100, TA1535, TA1537) tested separately at each of three levels of metabolic activation: rat liver, hamster liver, or none. The two mammalian liver (S9) preparations represent an attempt to recreate *in vitro* the metabolic processes that occur in humans. It is well known that apparently innocuous chemicals can be converted *in vivo* into noxious metabolites, so the use of an S9 activation attempts to provide for this possibility.

Chemicals are nominated in many different ways for NTP testing. If the scientific interest or evidence for concern is sufficient to justify the experimentation, the selected chemical proceeds through a 12 strain-activation battery of tests. The NTP Salmonella/microsome database currently consists of over 24,000 experiments, where an experiment refers to a test with a particular chemical, strain and activation in a given laboratory on

Table 2. Frequency of replication by strain and activation among the 941 chemicals.

		Number of replicates					
Strain	Activation	1	2	3	4	5	6
TA100	None	72	774	70	22	1	1
	Hamster	30	810	69	27	2	2
	Rat	33	815	68	21	3	1
TA98	None	114	775	42	9	1	0
	Hamster	80	784	62	10	3	2
	Rat	76	787	61	13	2	2
TA1537	None	140	742	52	5	1	0
	Hamster	121	765	47	6	2	0
	Rat	122	764	47	7	1	0
TA1535	None	116	742	65	17	0	1
	Hamster	99	756	73	8	4	1
	Rat	104	746	74	12	3	1

one day. The standard protocol requires two replicates of each experiment, usually two weeks apart. The emphasis, as in all good laboratory science, is on demonstrably reproducible results.

In some 20% of the experiments in the NTP database, the number of replicates is different from two. Table 2, from Margolin, Kim, and Risko (unpublished manuscript, hereafter referred to as MKR), indicates the frequency of replication by strain and activation among 941 chemicals tested; zero frequencies have been suppressed. Experimental loss due to contamination or extreme toxicity, together with *ad hoc* decisions by experimenters not to take a second replicate produced the singlets. Equally *ad hoc* decisions to obtain additional replicates beyond the two required by the protocol account for the replicates numbering greater than two. MKR report that the decision to proceed with additional replicates beyond two was apparently triggered on occasion by results observed for TA100 with either rat or hamster S9 activation. These two combinations were viewed by the experimenters as the two combinations with highest sensitivity to mutagens, and so clear resolution of these cases was frequently sought. The potential bias in the results for these two combinations suggests focusing attention on results for the other ten.

MKR note that if a given chemical is tested in n replicates of a given strain and activation, then the operating characteristics of any decision rule that assigns a "mutagenic" or "nonmutagenic" label to the individual experiments can be assessed by use of a finite mixture of binomials model. Specifically, in the notation of MKR, if Y of the n replicates are judged positive and labeled mutagenic by a decision rule, then the probability distribution function of Y can be written as:

$$f(Y_i;p,\tau_i) = z_i b(Y_i;n_i,p) + (1 - z_i)\, b(Y_i;n_i,\tau_i); \quad (7)$$

where $b(x;n,\phi)$ is the binomial probability distribution function for x successes out of n trials with success probability ϕ; z_i is an indicator variable with value 1 for nonmutagenicity and 0 for mutagenicity of chemical i in the particular strain/activation; p is the true probability that an experiment with a nonmutagen in the particular strain/activation will yield a result judged positive by the decision rule; τ_i is the probability that an experiment with the particular strain/activation for chemical i, given that chemical i is a mutagen in this combination, will yield a result judged positive by the decision rule; and, by assumption, $\tau_i > p$ for all i that correspond to mutagens.

MKR reason that p is presumably constant for all nonmutagens tested with a given strain and activation, but that τ clearly depends upon a mutagen's potency and toxicity for a given strain and activation. Nevertheless, they argue that the paucity of information regarding the behavior of a given chemical for a specific strain and activation suggests as a first approximation assuming τ to be constant across all mutagens for a given strain and activation. Moreover, z_i may be viewed as a Bernoulli trial with $\mathrm{pr}\{z_i = 1\} = \pi$, where π represents the proportion of nonmutagens among the chemicals selected for testing. With this construct, the results of applying a decision rule to data from M test chemicals for a particular strain-activation are given by $\{Y_i;n_i\}_{i=1}^{M}$ with attendant log-likelihood l specified by:

$$l = \sum_{i=1}^{M} \log \{\Pi \cdot b(Y_i;n_i,p) + (1 - \Pi) \cdot b(Y_i;n_i,\tau)\} \quad (8)$$

This likelihood is for independent, but not identically distributed data, a case little treated in the literature. MKR prove that the model in Eq. (8) is identifiable if and only if

$$\max_i (n_i) \geqslant 3$$

They then construct a version of the EM algorithm (*20*) for the MLEs of (π, p, τ). Using results of Louis (*21*), MKR also obtain the observed information matrix for the parameters, and so produce estimates of the precision of the MLEs as well. MKR apply their analytic technique to two decision rules. The first is a modified statistical analysis based on the mechanistic models of Margolin et al. (*8*), while the second is really not a rule, but rather a set of decisions arrived at by a senior NTP toxicologist upon his review of the experimental data.

In the present paper, the same technique is applied to a decision rule that has been widely employed in toxicology, but poorly understood. Labeled the "two-times background" rule, this rule declares a chemical mutagenic if the average response for at least one dose of test chemical is greater than twice the observed concurrent control mean. This rule, which has a long history of application, is indifferent to the number of doses tested, the number of replicates observed per dose, any empirical measure of variability, and any consideration of level of significance. The results of applying the MKR technique to the decisions of the "two-times background" rule with regard to the NTP database are in

Table 3. Estimates ± one standard deviation of the proportion of mutagens among chemicals tested, the false positive probability, and the true positive probability for the "two-times background" rule by strain and activation.

Strain	Activation	Proportion of mutagens $1 - \pi$	Probability of false positive p	Probability of true positive τ
TA100	None	0.088 ± 0.010	0.008 ± 0.003	0.911 ± 0.024
	Rat	0.126 ± 0.012	0.010 ± 0.004	0.925 ± 0.029
	Hamster	0.154 ± 0.013	0.008 ± 0.003	0.863 ± 0.027
TA98	None	0.056 ± 0.011	0.020 ± 0.005	0.908 ± 0.067
	Rat	0.101 ± 0.013	0.019 ± 0.005	0.833 ± 0.053
	Hamster	0.095 ± 0.011	0.019 ± 0.004	0.932 ± 0.030
TA1535	None	0.076 ± 0.014	0.018 ± 0.006	0.808 ± 0.069
	Rat	0.212 ± 0.040	0.009 ± 0.014	0.623 ± 0.065
	Hamster	0.211 ± 0.045	0.013 ± 0.016	0.637 ± 0.073
TA1537	None	0.130 ± 0.089	0.070 ± 0.029	0.619 ±0.191
	Rat	0.089 ± 0.020	0.063 ± 0.010	0.863 ± 0.082
	Hamster	0.069 ± 0.016	0.067 ± 0.009	0.938 ± 0.079

Table 3. As one might well predict intuitively, this rule is moderately conservative, yielding false positive rates of approximately 0.01 for TA100 and 0.02 for TA98 and TA1535, irrespective of activation level. For TA1537, however, with its very low background rates, this rule has a false-positive rate of approximately 0.07. These estimates apply to the NTP protocol as executed by the NTP contractual laboratories, and to no other context. If one requires a repeated positive result for confirmation, then the probability of a falsely confirmed positive is p^2. For TA100, TA98, and TA1535, this probability is estimated to be 1×10^{-4} to 4×10^{-4}. For the NTP screening program, in which scientific judgment in chemical nomination and selection produces a population of test chemicals highly enriched with mutagens, decision rules with probabilities of confirmed false positives on the order of 10^{-4} are too conservative and counterproductive. The attendant loss in sensitivity to detect weak mutagens is a heavy price to pay in order to obtain a simple rule of thumb. In many instances, mutagens may not be able to achieve a doubling of background levels because of toxicity, solubility or other limitations, yet they may well exhibit highly reproducible patterns of mutation induction. An excellent example of this phenomenon is phenobarbital (*22*).

Concluding Remark

The statistical studies briefly surveyed here all had their origins in problems that arose from genetic toxicology. From this survey, one conclusion is clear: genetic toxicology represents a rapidly growing area of science that is rich with research opportunities for statisticians.

Appendix

Assume that Y_j is distributed NB (m_j,c), that the $\{Y_j\}$ are independent and that $m_j = H(\alpha + \beta d_j)$ for H monotone and twice differentiable, and $j = 1, \ldots, n$. Without loss of generality, assume further that $\Sigma d_j = 0$. Then, from Eq. (2), the log-likelihood of the data is given by

$$l = \sum_j Y_j \ln H(\alpha + \beta d_j) - \\ (Y_j + c^{-1}) \ln [1 + cH(\alpha + \beta d_j)] + \text{terms in } c \text{ alone.}$$

It then follows that

$$\frac{\partial l}{\partial \beta} = \sum_j H'(\alpha + \beta d_j)\, d_j\, (Y_j - m_j)/m_j(1 + cm_j) \tag{A-1}$$

Straightforward computation yields the result that $E(\partial^2 l/\partial\beta\partial c) = 0$ and that $E(\partial^2 l/\partial\alpha\partial\beta)|_{\beta = 0} = 0$. Thus the normalized score statistic to test H_0: $\beta = 0$ is given by the ratio of Eq. (A-1) evaluated at ($\beta = 0$, $\alpha = \bar{X}$, $c = \hat{c}$) to its asymptotic standard deviation. The result is Eq. (6) with $s_x{}^2 = \bar{X}(1 + \hat{c}\bar{X})$.

REFERENCES

1. Ames, B. N. Identifying environmental chemicals causing mutations and cancer. Science 204: 587–593 (1979).
2. Ames, B. N., McCann, J., and Yamasaki, E. Methods for detecting carcinogens and mutagens with the *Salmonella*/microsome mutagenicity test. Mutat. Res. 31: 347–364 (1975).
3. Boveri, T. Zur Frage der Entstehung Maligner Tumoren. Gustov Fischer, Jena, 1914.
4. Collings, B. J., Margolin, B. H., and Oehlert, G. W. Analyses for binomial data, with application to the fluctuation test for mutagenicity. Biometrics 37: 775–794 (1981).
5. Tarone, R. E. The use of historical control information in testing for a trend in Poisson means. Biometrics 38: 457–462 (1982).
6. Hollstein, M., McCann, J., Angelosanto, F., and Nichols, W. Short-term tests for carcinogens and mutagens. Mutat. Res. 65: 133–226 (1979).
7. Paul, S. R., and Plackett, R. L. Inference sensitivity for Poisson mixtures. Biometrika 65: 591–602 (1978).
8. Margolin, B. H., Kaplan, N., and Zeiger, E. Statistical analysis of the Ames *Salmonella*/microsome test. Proc. Nat. Acad. Sci. (U.S.) 78: 3779–3783 (1981).
9. Cochran, W. G. C. Catalog of uniformity trial data. J. Roy. Statist. Soc. (Suppl.) 4: 233–253 (1937).
10. Collings, B. J., and Margolin, B. H. Testing goodness of fit for the Poisson assumption when observations are not identically distributed. J. Am. Statist. Assoc. 80: 411–418 (1985).
11. Collings, B. J. The negative binomial distribution: an alternative to the Poisson. Ph.D. Thesis, The University of North Carolina at Chapel Hill, NC, 1981.
12. Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37: 358–382 (1950).
13. Potthoff, R. F., and Whittinghill, M. Testing for homogeneity. II: The Poisson distribution. Biometrika 53: 183–190 (1966).
14. Cochran, W. G. Some methods for strengthening the common χ^2 tests. Biometrics 10: 417–451 (1954).
15. Armitage, P. Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386 (1955).
16. Bernstein, L., Kaldor, J., McCann, J. and Pike, M. C. An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test. Mutat. Res. 97: 267–281 (1982).
17. Jonckheere, A. R. A distribution-free k-sample test against ordered alternatives. Biometrika 41: 133–145 (1954).
18. Vollmar, J. Statistical problems in the Ames test. In: Progress in Mutation Research (A. Kappas, Ed.), Elsevier/North Holland, Amsterdam, 1981, pp. 179–186.
19. Boyd, M. N. Examples of testing against ordered alternatives in the analysis of mutagenicity data. Mutat. Res. 97: 147–153 (1982).
20. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum

likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B39: 1–38 (1977).

21. Louis, T. A. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. B44: 226–233 (1982).

22. Ashby, J., deSerres, F. J., Draper, M., Ishidate, M. Jr., Margolin, B. H., Matter, B., and Shelby, M. (Eds.). Evaluation of Short-Term Tests for Carcinogens: Report of the International Programme on Chemical Safety's Collaborative Study on In Vitro Assays. Elsevier/North Holland, Amsterdam, 1985.