COMMUN. STATIST.—THEORY METH., 29(5&6), 1179–1195 (2000)

A COMPOSITE BINOMIAL MODEL DERIVED FROM CORRELATED RANDOM VARIABLES

Ruiguang Song[a], Paul C. Schlecht[b], and Jensen H. Groff[b]

[a]HGO/NIOSH, 4676 Columbia Parkway, Cincinnati, OH 45226, USA
[b]National Institute for Occupational Safety and Health
4676 Columbia Parkway, Cincinnati, OH 45226, USA

Key Words: Outlier; Bernoulli variable; Exchangeable; Link function

ABSTRACT

Motivated by a correlation problem in a power calculation of proficiency testing, a composite binomial model is developed to describe the distribution for the number of outliers from measurement results of multiple analytes contained in a single sampler. This model is different from other binomial models in that its component Bernoulli variables are derived from correlated random variables. By modeling the correlated random variables and selecting a binary link function to convert these correlated random variables to Bernoulli variables, various composite binomial distributions can be derived. Formulas for calculating the probabilities of the composite binomial distribution are provided. Although the composite binomial model is developed for a power calculation, it can be applied to other problems related to a sum of correlated Bernoulli variables.

1. INTRODUCTION

In environmental and industrial hygiene chemistry, various proficiency testing programs evaluate the performance of laboratories that conduct analyses of

toxic materials. In these proficiency testing programs, samples containing toxic materials are generated and distributed to participating laboratories for analysis. If it is feasible, similar analytes are usually loaded onto one sampler to reduce cost by maximizing sample usage. For example, two or three organics can be collected on a single charcoal or other absorbent tube and three heavy metals are put on a single filter in the American Industrial Hygiene Association (AIHA) administered Proficiency Analytical Testing (PAT) Program (see Esche, Groff, Schlecht, and Shulman, 1994). One result is reported for each analyte contained in a sample and each result is rated acceptable or not acceptable (called an outlier in PAT, see PAT outlier definition in section 5) according to the difference between the result and the reference value. A laboratory's performance is then rated as proficient or non-proficient based on the number (or percent) of acceptable results.

Note that analyzing a sample with multiple analytes may not be equivalent to analyzing multiple samples with a single analyte on each sample since analysis results of analytes collected on one sample are usually correlated. Although this correlation does not change a laboratory's chance to have an acceptable result, it may affect the distribution of the number of acceptable results, and hence, change the chance of the laboratory to be rated proficient. The higher the sample analysis correlation, the lower the power of the proficiency test. If the correlation is near perfect, then there is almost no gain on test power by loading multiple analytes onto one sample. Therefore, it is not cost effective to put multiple, highly correlated analytes on one sample in proficiency testing programs.

To design a proficiency test program in which the rating of a laboratory's performance is based on analyses of multiple analytes, a question that needs to be addressed is how to cost effectively make use of samples so that the two types of test errors are controlled within specified levels. To obtain an answer, one must know the power of the proficiency test. However, obtaining the power requires knowledge of the distribution for the number of outliers. In section 2, a calculation formula for the power is provided. The outlier distribution required in the formula is discussed in section 3. A composite binomial model is introduced to describe the distribution for the number of outliers from a sample that contains

multiple analytes. In this model, the correlated Bernoulli variables are defined by other correlated random variables through a binary link function. The correlation between these variables is due to a common random factor contained in these variables. Different correlation structures result in different composite binomial distributions. With a specified binary link function, formulas for calculating the probabilities of the composite binomial distribution are provided. In section 4, the composite binomial model is compared with other correlated binomial models. Actually, the composite binomial distribution can be seen as a concrete example of the abstract correlated binomial distribution introduced by Bahadur (1961) and the EXBERT distribution defined by Madsen (1993). The well known beta-binomial distribution is shown to be a special case of the composite binomial distribution. As an application, the distribution for the number of outliers from a sample with multiple analytes and how this distribution depends on the sample correlation are presented in section 5. Finally, an example of the correlation impact on proficiency rating is provided in section 6.

2. POWER CALCULATION

Suppose that a laboratory must analyze all n analytes from each of m samples and must have no more than λ percent of results not acceptable to be rated proficient. Let X_{ij} be the outlier indicator variable of the result for the j^{th} analyte in the i^{th} sample: $X_{ij} = 1$ if the result is an outlier and 0 otherwise. If each result has an equal chance to be an outlier and all results are independent, then the total number of outliers $S = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}$ has a binomial distribution $\text{Bin}(mn, p)$ here $p = \Pr(X_{ij} = 1)$. However, the assumption of independence may not apply. Analytes collected on the same sample are prepared for analysis with an identical procedure. Instrumental analysis usually involves an identical analyte separation and detection. Often all the analytes are then analyzed simultaneously with an instrument set (e.g., Gas Chromographic analysis for organics, Inductively Coupled Plasma - Atomic Emission Spectroscopy for heavy metals). Therefore, one would expect analysis results to be highly correlated. In fact, actual single

sample correlation coefficients in the PAT program range from 0.5 to 0.95 for heavy metals and 0 to 0.99 for organic solvents, see Song, Schlecht, and Groff, 1997. In contrast, results from different samples seem to be independent.

Let $Q_i = \sum_{j=1}^{n} X_{ij}$ be the number of outliers from the i^{th} sample and R_k the number of cases where $Q_i = k$, for $k = 0, 1, \ldots, n$. Then, $\sum_{k=0}^{n} R_k = m$ and $S = \sum_{k=0}^{n} kR_k$. Under the assumption that results from different samples are independent and $Q_1, \ldots, Q_m$ have an identical distribution given by $b_{nk} = \Pr(Q_i = k)$, the random variables $R_0, R_1, \ldots,$ and R_n have a multinomial distribution with parameters $(m; b_{n0}, b_{n1}, \ldots, b_{nn})$. The probability of a non-proficient rating is then given by

$$\begin{aligned} P_{NP} &= \Pr\left(\sum_{k=0}^{n} kR_k > \lambda mn\right) \\ &= \sum_{i_1+2i_2+\cdots+ni_n>\lambda mn} \Pr(R_0 = i_0, R_1 = i_1, \ldots, R_n = i_n) \\ &= \sum_{i_1+2i_2+\cdots+ni_n>\lambda mn} \frac{m!}{i_0!i_1!\cdots i_n!}(b_{n0})^{i_0}(b_{n1})^{i_1}\cdots(b_{nn})^{i_n} \end{aligned} \tag{1}$$

Thus, the distribution of Q_i is of critical importance in obtaining the power of the proficiency test. It is not simply a binomial distribution, since Q_i is not a sum of independent Bernoulli random variables.

3. COMPOSITE BINOMIAL MODEL

In many applications, observed variables are correlated because they contain a common random factor. This kind of correlation can be presented by the following model: For $i = 1, \ldots, n$, define $Z_i = J(U_i, V)$, where $U_1, \ldots, U_n$, and V are independent random variables, and $U_i \sim F(u), V \sim G(v)$, and $J(u,v)$ is called a correlation link function. $Z_1, \ldots, Z_n$ are independent if V is a constant. A simple example of the correlation link function is the linear function: $J(u,v) = au + bv + c$.

Let $I(z)$ be a binary link function from variable Z to a variable with only two possible values 0 and 1. Define $X_i = I(Z_i)$. Then, $X_1, \ldots, X_n$ are Bernoulli

random variables and they are correlated if $Z_1,\ldots,Z_n$ are correlated. The distribution of $S_n = \sum_{i=1}^{n} X_i$ is of interest when the Bernoulli variables are correlated. Two common binary link functions are the one-sided indicator function: $I(z) = I_h(z)$ and the two-sided indicator function: $I(z) = I_h(|z|)$, where $I_h(z) = 1$ if $z > h$ and 0 otherwise.

Note that $Z_1,\ldots,Z_n$ are equally correlated with an identical distribution. Therefore, $X_1,\ldots,X_n$ are exchangeable Bernoulli variables. That is

$$\Pr(X_1 = \delta_1,\ldots,X_n = \delta_n) = \Pr(X_1 = \delta_1',\ldots,X_n = \delta_n') \tag{2}$$

when $\sum_{i=1}^{n}\delta_i = \sum_{i=1}^{n}\delta_i'$ for $\delta_i,\delta_i' = 0$ or 1.

Defining

$$p_{nk} = \Pr(X_1 = 1,\ldots,X_k = 1, X_{k+1} = 0,\ldots,X_n = 0), \tag{3}$$

we have

$$b_{nk} = \Pr(S_n = k) = C_n^k p_{nk}. \tag{4}$$

Also, note that $X_1,\ldots,X_n$ are conditionally independent given V; therefore,

$$\begin{aligned} p_{nk} &= \int_{-\infty}^{\infty} \Pr(X_1 = 1,\ldots,X_k = 1, X_{k+1} = 0,\ldots,X_n = 0|V = v)dG(v) \\ &= \int_{-\infty}^{\infty} [\Pr(X_1 = 1|V = v)]^k [\Pr(X_1 = 0|V = v)]^{n-k} dG(v) \\ &= \int_{-\infty}^{\infty} [\Pr(I(Z_1) = 1|V = v)]^k [1 - \Pr(I(Z_1) = 1|V = v)]^{n-k} dG(v) \\ &= \int_{-\infty}^{\infty} [\Pr(I(J(U_1,v)) = 1|V = v)]^k [1 - \Pr(I(J(U_1,v)) = 1|V = v)]^{n-k} dG(v) \end{aligned} \tag{5}$$

Let $p(v|I,J,F) = \Pr(I(J(U_1,v)) = 1|V = v)$ and $\mathrm{Bin}(k,n,p) = C_n^k p^k (1-p)^{n-k}$.

Combining (4) and (5) gives

$$\begin{aligned} b_{nk} &= C_n^k \int_{-\infty}^{\infty} [p(v|I,J,F)]^k [1 - p(v|I,J,F)]^{n-k} dG(v) \\ &= \int_{-\infty}^{\infty} \mathrm{Bin}(k,n,p(v|I,J,F))dG(v) \end{aligned} \tag{6}$$

If V is a discrete random variable, the integral in (6) is converted to a summation:

$$b_{nk} = \sum_{v} \mathrm{Bin}(k,n,p(v|I,J,F)) \Pr(V = v). \tag{7}$$

The distribution of S_n given by (6) or (7) is called a composite binomial distribution since it has the form of a binomial distribution but with a random

parameter $p(v|I,J,F)$. The distribution of the parameter is determined by two link functions $I(z)$ and $J(u,v)$ and two distribution functions $F(u)$ and $G(v)$. Different choices of these functions derive different composite binomial distributions.

Given a value of V, S_n has a binomial distribution. If $V = c$ is a constant, then $X_1,\ldots,X_n$ are independent, and hence, S_n has a binomial distribution $\text{Bin}(n,p)$ with $p = p(c|I,J,F)$. On the other hand, if all $U_i = c$ are constant, then all Z_i, and hence, all X_i are the same. In this case, $S_n = nX_1$ and X_1 has a Bernoulli distribution $\text{Bin}(1,p)$ with $p = \Pr(X_1 = 1)$.

The mean and variance of S_n are given by

$$E(S_n) = n\Pr(X_1 = 1) = nE(p(V|I,J,F)) \tag{8}$$

and

$$\begin{aligned} Var(S_n) &= nVar(X_1) + n(n-1)Cov(X_1,X_2) \\ &= nVar(X_1)[1+(n-1)Corr(X_1,X_2)]. \end{aligned} \tag{9}$$

Note that given V, X_1 and X_2 are independent. Therefore,

$$\begin{aligned} Cov(X_1,X_2) &= E(Cov(X_1,X_2|V)) + Cov(E(X_1|V),E(X_2|V)) \\ &= 0 + Cov(p(V|I,J,F),p(V|I,J,F)) \\ &= Var(p(V|I,J,F)) \geq 0. \end{aligned}$$

Hence, $X_1,\ldots,X_n$ are positively correlated. This correlation does not change the mean of S_n, but increases the variance of S_n compared to the case where $X_1,\ldots,X_n$ are independent.

If the correlation link function is defined as $J(u,v) = u - v$ and the standard deviations for U_i and V are σ_u and σ_v, respectively, then the correlation coefficient between Z_is is $\rho(Z) = 1/[1+(\sigma_u/\sigma_v)^2]$. Further if the binary link function is the one-sided indicator function $I_h(z)$, the probability parameter in the composite binomial distribution becomes $p(V|I,J,F) = F(h+V)$. If $h = 0$ and U_i has a uniform distribution on the interval [0,1], then $p(V|I,J,F) = V$, for $0 \leq V \leq 1$, 0 for $V<0$, and 1 for $V>1$. In this case, the distribution of S_n is a mixture of a binomial distribution and the distribution of V. If the mixing

distribution is a uniform distribution over the interval [0,1], then S_n has a discrete uniform distribution with $\Pr(S_n = k) = 1/(n+1)$ (see Feller (1968)). If the mixing distribution is a beta-distribution, then the distribution of S_n is the well known beta-binomial distribution (see Skellam, 1948).

4. BINOMIAL MODELS WITH EXCHANGABLE BERNOULLI VARIABLES

There are many different ways to model the sum of correlated Bernoulli variables. See Rudolfer (1990) and Madsen (1993) for a review. In this paper, we are interested in the models with exchangeable Bernoulli variables. The distribution for the sum of exchangeable Bernoulli variables has been considered by Madsen (1993) and is referred to as an EXBERT distribution. The EXBERT distribution is developed by generalizing the beta-binomial type distributions. This generalization is broad and covers many correlated binomial distributions. However, it does not include the Markov chain model developed by Klotz (1973). The Bernoulli trials with Markov dependence are not equally correlated, and hence they are not exchangeable Bernoulli variables.

Since the Bernoulli variables defined in the previous section are exchangeable, the composite binomial distribution is an EXBERT distribution. As an EXBERT distribution, the distribution is determined by the sequence $\{b_{00}, b_{11}, \ldots, b_{nn}, \ldots\}$ with $b_{00} = 1$ and the probability can be expressed as

$$b_{nk} = C_n^k \sum_{i=0}^{n-k} C_{n-k}^i (-1)^i b_{(k+i)(k+i)} \quad \text{for any} \quad 0 \le k \le n. \tag{10}$$

Note that for $0 \le k \le n-1$,

$$\begin{aligned} \sum_{i=0}^{n-k} C_{n-k}^i (-1)^i b_{(k+i)(k+i)} &\equiv b_{kk} + \sum_{i=1}^{n-k-1} (C_{n-k-1}^i + C_{n-k-1}^{i-1})(-1)^i b_{(k+i)(k+i)} + (-1)^{n-k} b_{nn} \\ &\equiv \sum_{i=0}^{n-1-k} C_{n-1-k}^i (-1)^i b_{(k+i)(k+i)} - \sum_{i=0}^{n-k-1} C_{n-k-1}^i (-1)^i b_{(k+1+i)(k+1+i)} \end{aligned} \tag{11}$$

Based on (10), this is equivalent to

$$b_{nk} / C_n^k = b_{(n-1)k} / C_{n-1}^k - b_{n(k+1)} / C_n^{k+1} \quad \text{for any} \quad 0 \le k \le n-1. \tag{12}$$

If we denote

$$p_{nk} = b_{nk} / C_n^k \quad \text{for} \quad 0 \le k \le n, \tag{13}$$

then we have

$$p_{nk} = p_{(n-1)k} - p_{n(k+1)} \quad \text{for any} \quad 0 \le k \le n-1. \tag{14}$$

On the other hand, the equation (10) can be derived from (13) and (14). This can be proved using mathematical induction. First of all, (10) is true for any n and k such that $n - k = 1$ by assuming that (14) is true. Suppose that (10) is true for any n and k such that $n - k = d - 1$. Then by mathematical induction, we only need to prove that (10) is true for any n and k such that $n - k = d$.

From (13) and (14), we get (12). By induction assumption, the two terms on the right hand side of (12) can be expressed in the form of (10). Then applying the identity (11) to the equation (12), we have just obtained the equality (10). Therefore, to determine whether a distribution with known probability expression is an EXBERT distribution, we need only to verify the equality (14). Based on this result, it is not difficult to verify that the correlated binomial distribution introduced by Kupper and Haseman (1978) is an EXBERT distribution. This binomial model is equivalent to the additive model developed by Altham (1978) except that they have different expressions for b_{nk}. Although the additive model is attractive, it has unpleasant restrictions on the correlation between X_is. This is because it is only an approximation of the general correlated binomial model introduced by Bahadur (1961). Figure 1 shows the upper and lower limits of the correlation coefficient, denoted by $\rho(X)$, for sample size $2 \le n \le 20$ and $p = p_{11} = 0.1, 0.3, 0.5$ (the limits for $p > 0.5$ are the same as that for $1 - p$). This figure shows that the correlation coefficient is more restricted on the lower (negative) side. The lower limit is greater than or equal to $-2/[n(n-1)]$ and the upper limit is close to $2/n$. These limits are needed to ensure that the resulting probability values are not negative. Within these limits, the probability distribution may become irregular when the correlation coefficient is close to the limit. For instance, when $n = 6$ and $p = 0.3$, the upper limit for $\rho(X)$ is 0.323. The distributions corresponding to $\rho(X) = 0(0.05)0.3$ are shown in Figure 2.

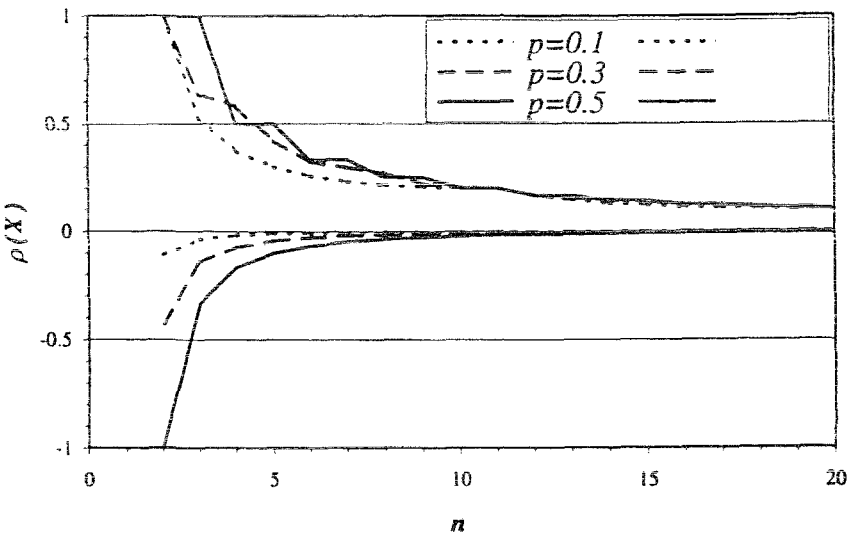

FIG. 1. Upper and lower limits of the correlation coefficient between Bernoulli random variables in the additive binomial model.

When $\rho(X)$ changes from 0 to 0.3, the distribution changes from an unimodal (binomial) distribution to a bimodal distribution.

There is no upper limit for the correlation coefficient between X_is in the composite binomial model. The correlation coefficient $\rho(X)$ depends on the correlation coefficient between Z_i s. $\rho(X) = 0$ if $\rho(Z) = 0$ and $\rho(X) = 1$ if $\rho(Z) = 1$. The relationship between the two correlation coefficients depends on the binary and correlation link functions and the two distribution functions $F(u)$ and $G(v)$. Two examples are given in the next section.

5. DISTRIBUTION FOR THE NUMBER OF OUTLIERS FROM A SAMPLE

In this section, we derive an appropriate composite binomial model for the number of outliers from a sample with multiple analytes in a proficiency test program. In the AIHA PAT program, each result (denoted by x) is converted to a

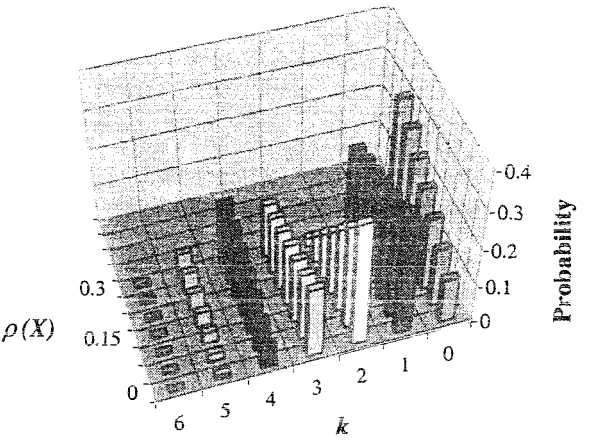

FIG. 2. Probability distributions based on the additive binomial model with $n = 6$ and $p = 0.3$.

z-score which is defined as $z = (x - \bar{y}) / s_y$, where $\bar{y}$ and s_y are the sample mean and standard deviation of results from pre-selected reference laboratories. Note that the z-score has a Student t distribution with degrees of freedom equal to the number of reference laboratories minus one. Since the number of reference laboratories in the PAT program is greater than 50, the z-score has an approximate normal distribution. Let μ and σ be the mean and standard deviation of the z-score, respectively. Then the mean μ is directly related to the laboratory's bias and the standard deviation σ reflects the laboratory's precision.

Suppose that there are n analytes on a sample and $Z_1,...,Z_n$ are the corresponding z-scores. Assume that they have a normal distribution with mean μ, standard deviation σ, and a common correlation coefficient $\rho(Z)$. These correlated normal random variables can be expressed as

$$Z_i = \mu + \sigma(\sqrt{1-\rho}U_i - \sqrt{\rho}V),$$

where $U_1,...,U_n$, V are independent standard normal variables.

A result is not acceptable if the absolute value of its corresponding z-score is greater than a pre-specified value, say h (in the PAT program, $h = 3$). Let Z_i be the z-score of the result for the i^{th} analyte on a sample, X_i the outlier indicator variable: $X_i = I_h(|Z_i|)$.

Let $Q = \sum_{i=1}^{n} X_i$ be the number of outliers from the sample. Then, Q has a composite binomial distribution

$$b_{nk} = C_n^k \int_{-\infty}^{\infty} [p(v|I,J,F)]^k [1 - p(v|I,J,F)]^{n-k} d\Phi(v) \tag{15}$$

and

$$p(v|I,J,F) = 1 + \Phi\left(\frac{(-h-\mu)/\sigma + \sqrt{\rho}v}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{(h-\mu)/\sigma + \sqrt{\rho}v}{\sqrt{1-\rho}}\right). \tag{16}$$

The chance for getting an outlier depends on the value of h as well as the laboratory's performance parameters μ and σ:

$$p_{11} = \Pr(|Z_i| > h) = 1 + \Phi((-h-\mu)/\sigma) - \Phi((h-\mu)/\sigma). \tag{17}$$

To see how the correlation between outliers relates to the correlation between the z-scores with the two-sided indicator link function $I_h(|z|)$, the relation functions for $p = p_{11} = 0.1(0.2)0.9$ are shown in Figure 3. We see that the correlation between outliers is less than the correlation between the z-scores and the difference increases as the value of p_{11} increases. The distributions of S_n for $n = 6$, $p_{11} = 0.3$, and $\rho(X) = 0(0.1)1$ are shown in Figure 4.

As a comparison, the relation functions for the one-sided indicator link function $I_h(z)$ are shown in Figures 5. In this case,

$$p(v|I,J,F) = 1 - \Phi\left(\frac{(h-\mu)/\sigma + \sqrt{\rho}v}{\sqrt{1-\rho}}\right) \tag{18}$$

and $p_{11} = \Pr(Z_i > h) = 1 - \Phi((h-\mu)/\sigma)$. The correlation between outliers is less than the correlation between the z-scores, but the difference is not as great as that associated with the two-sided indicator link function $I_h(|z|)$. The distributions of S_n for $n = 6$, $p_{11} = 0.3$, and $\rho(X) = 0(0.1)1$ are shown in Figure 6.

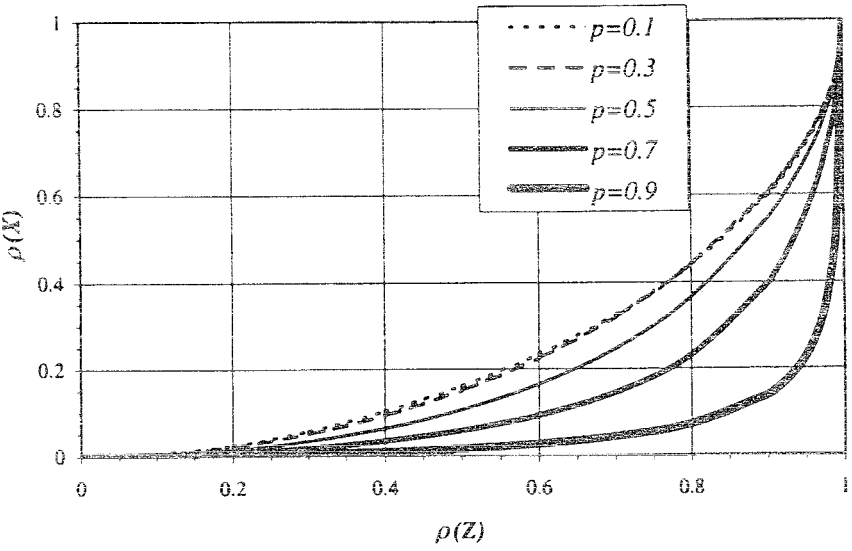

FIG. 3. The correlation coefficient between Bernoulli random variables as a function of the correlation coefficient between normal random variables when the binary link function is the two-sided indicator function.

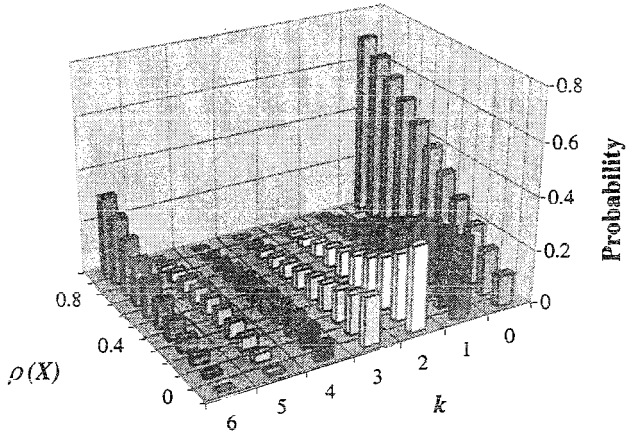

FIG. 4. Probability distributions based on the composite binomial model with $n = 6, p_{11} = 0.3$, and the two-sided indicator link function.

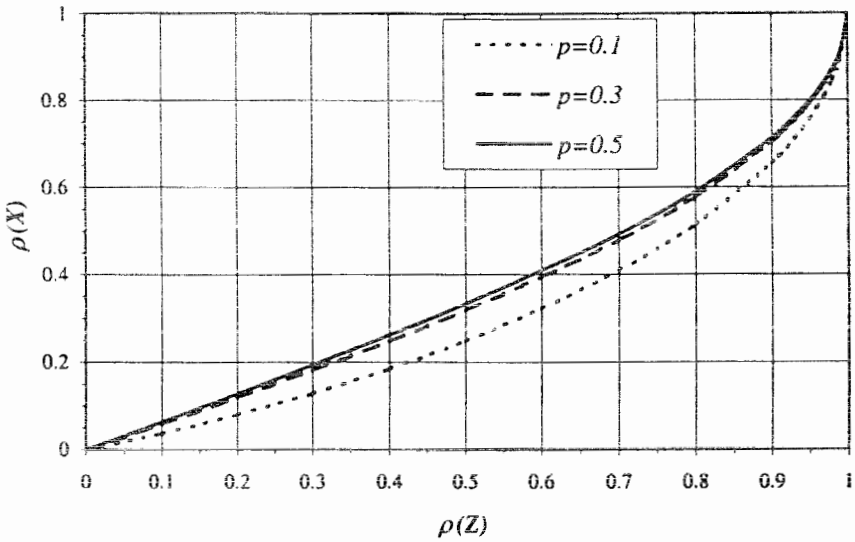

FIG. 5. The correlation coefficient between Bernoulli random variables as a function of the correlation coefficient between normal random variables when the binary link function is the one-sided indicator function.

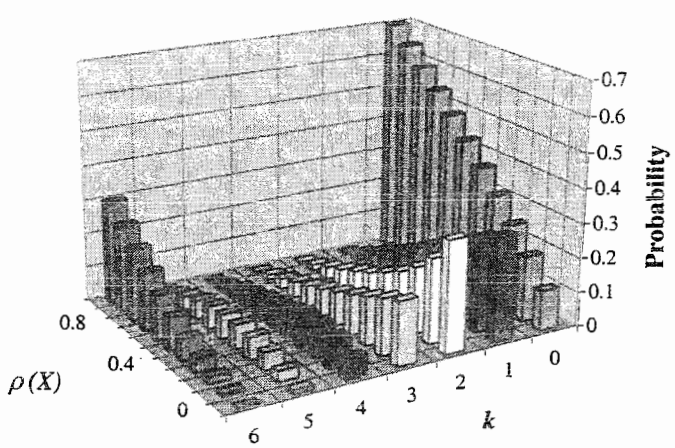

FIG. 6. Probability distributions based on the composite binomial model with $n = 6, p_{11} = 0.3$, and the one-sided indicator link function.

In both cases, the correlation coefficient does not affect the mean, but the standard deviation of the distribution. The higher the correlation, the greater the standard deviation. As the correlation increases, the probability in the middle $(0 < k < n)$ moves gradually to the two ends $(k = 0 \text{ or } n)$, and finally, when $\rho(X) = 1$, the probabilities in the middle are reduced to zero, and all probabilities are concentrated on the two ends.

6. CORRELATION IMPACT ON PROFICIENCY RATING

The correlation changes the chance for a laboratory to get a non-proficient rating because it has an impact on the distribution of outliers from one sample. To see this, let's look at an example that describes the case in the PAT program: $n = 3$, $h = 3$, and $\lambda = 1/4$. We show in Table I the probabilities of a non-proficient rating for $p = 0.3, 0.35, 0.4$ and $\rho(X) = 0, 0.5, 0.9, 1$. For each value of p, two combinations of mean and standard deviation of z-score are considered, one has mean zero and the other has standard deviation one: $(0, \sigma_Z)$ and $(\mu_Z, 1)$. In both cases, the corresponding values of p_{11} are the same and equal to one of the selected values: 0.3, 0.35, and 0.4. Probabilities of a non-proficient rating based on $m = (8, 12, 16)$ samples are provided, see columns under $\Pr(NP)$.

The combination of a laboratory's bias and precision determines the probability that a result is not acceptable. For a fixed value of p_{11}, different combinations of bias and precision result in different combinations of mean and standard deviation of the z-score and in turn result in different effects on the proficiency rating. The laboratory with a combination of a zero bias and a relatively poor precision has a slightly higher probability to be rated non-proficient than the laboratory with a combination of good precision and relative high bias. In other words, with a fixed value of p_{11}, the program has a slightly better chance to detect laboratories with poor precision than detect laboratories with large bias.

In Table I, we also listed the value L for each $\Pr(NP)$, such that $\Pr(NP)$ is the probability of getting a non-proficient rating based on L samples with a single

TABLE I. Probability of a non-proficient rating and equivalent sample size of independent samples.

p_{11}	μ_z	σ_z	$\rho(X)$	n=8 m=3		n=12 m=3		n=16 m=3	
				Pr(NP)	L	Pr(NP)	L	Pr(NP)	L
0.30	0.000	2.895	0.0	0.6114	24	0.6746	36	0.7204	48
			0.5	0.5909	22	0.6489	30	0.6915	42
			0.9	0.5600	18	0.6080	22	0.6431	30
			1.0	0.4482	8	0.5075	12	0.5501	16
	2.476	1.000	0.0	0.6114	24	0.6746	36	0.7204	48
			0.5	0.5771	18	0.6300	26	0.6692	34
			0.9	0.5518	18	0.6015	22	0.6363	30
			1.0	0.4482	8	0.5075	12	0.5501	16
0.35	0.000	3.210	0.0	0.7894	24	0.8615	36	0.9057	48
			0.5	0.7571	22	0.8293	30	0.8761	38
			0.9	0.6941	14	0.7602	22	0.8065	26
			1.0	0.5722	8	0.6533	12	0.7108	16
	2.615	1.000	0.0	0.7894	24	0.8615	36	0.9057	48
			0.5	0.7265	18	0.7961	26	0.8434	34
			0.9	0.6802	14	0.7478	18	0.7939	26
			1.0	0.5722	8	0.6533	12	0.7108	16
0.40	0.000	3.565	0.0	0.9040	24	0.9551	36	0.9781	48
			0.5	0.8759	21	0.9344	30	0.9639	38
			0.9	0.8036	14	0.8705	18	0.9116	26
			1.0	0.6846	8	0.7747	12	0.8334	16
	2.747	1.000	0.0	0.9040	24	0.9551	36	0.9781	48
			0.5	0.8393	18	0.9036	22	0.9400	30
			0.9	0.7859	14	0.8555	18	0.8983	22
			1.0	0.6846	8	0.7747	12	0.8334	16

analyte on each sample. We know that $L=nm$ when $\rho(X) = 0$ and $L=m$ when $\rho(X) = 1$. Apparently, when $0 < \rho(X) < 1$ and $p_{11} > \lambda$, the corresponding L falls in the interval (m, mn) and L increases as $\rho(X)$ decreases.

By looking at the L value, we can see the value of loading additional analytes on a sample. For example, when $p_{11} = 0.3$ ($\mu_Z = 0, \sigma_Z = 2.895$) and $\rho(X) = 0.9$, 48 results from 16 samples (3 analytes on each sample) have the same power as 30 results from 30 separate samples (1 analyte on each sample). Having two additional analytes on each of 16 samples is equivalent to having two additional samples without additional analytes on each sample.

7. COMMENTS

The composite binomial model introduced in this paper can generate various binomial distributions by selecting different binary and correlation link functions and the distributions of the random variables supporting the Bernoulli variables. Although the model is developed from a correlation problem in a power calculation, it could be applied to other problems related to correlated binomial distributions. For example in toxicology studies, the status that an animal is dead or alive at the end of an experiment is represented by a Bernoulli variable. This Bernoulli variable is actually a function of the variable that describes the life-time of the animal and the link function is the one-sided indicator function $I_h(z)$. If we have sufficient information on the supporting variables and the correlation between these variables, the distribution for the sum of Bernoulli variables derived from these variables is well determined.

ACKNOWLEDGEMENT

The authors thank the referees for their helpful comments that led to a significant improvement of the original manuscript.

BIBLIOGRAPHY

Althem, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics,* 27(2), 162-167.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In studies in Item Analysis and Prediction, H. Solomon (ed.), Stanford University Press, Stanford, California.

Esche, C. A., Groff, J. H., Schlecht, P. C. and Shulman, S. A. (1994). Laboratory Evaluations and Performance Reports for the Proficiency Analytical Testing (PAT) and Environmental Lead Proficiency Analytical Testing (ELPAT) Programs. DHHS (NIOSH 94 - 102).

Feller, W. (1966). *An Introduction to Probability and Its Applications*, Vol. 2. 2nd ed., New York: Wiley.

Klotz, J. (1973). Statistical Inference in Bernoulli Trials with Dependence. *The Annals of Statistics* 1(2), 373-379.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* 34, 69-76.

Madsen, R. W. (1993). Generalized binomial distribution. *Communications in Statistics - Theory and Methods* 22(11), 3065-3086.

Rudolfer, S. M. (1990). A Markov chain model of extrabinomial variation, *Biometrika* 77(2), 255-264.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society Series B* 10, 257-261.

Song, R., Schlecht, P. C., and Groff, J. H. (1997). Sample Analysis Correlation and Its Impact on Proficiency Rating. *The 1997 American Statistical Association Proceedings of the Section on Physical and Engineering Sciences, 118-123.*

Received June, 1999; Revised January, 2000.