Comparison of the Observed Distribution of Aflatoxin in Shelled Peanuts to the Negative Binomial Distribution[1]

T.B. WHITAKER, J.W. DICKENS, Market Quality Research Division, ARS, USDA,
R.J. MONROE, Department of Experimental Statistics, and **E.H. WISER,**
Biological and Agricultural Engineering Department, North Carolina State University, Raleigh, North Carolina

ABSTRACT

Suitability of the negative binomial distribution for use in estimating the probabilities associated with sampling lots of shelled peanuts for aflatoxin analysis has been studied. Large samples, called "minilots," were drawn from 29 lots of shelled peanuts contaminated with aflatoxin. These minilots were subdivided into ca. 12 lb samples which were analyzed for aflatoxin. The mean and variance of these aflatoxin determinations for each minilot were determined. The shape parameter k and the mean aflatoxin concentration m were estimated for each minilot. A regression analysis indicated the functional relationship between k and m to be: $k = (2.0866 + 2.3898m) \times 10^{-6}$. The observed distribution of sample concentrations from each of the 29 minilots was compared to the negative binomial distribution by means of the Kolmogorov-Smirnov test. The null hypothesis that each of the true unknown distribution functions was negative binomial was not rejected at the 5% significance level for all 29 comparisons.

INTRODUCTION

The negative binomial distribution has been used by the authors to estimate the variation among analytical determinations of aflatoxin concentrations in replicate samples drawn from aflatoxin-contaminated lots of shelled peanuts (1,2). The probability function for the negative binomial distribution is

$$F(X) = (\Gamma[X + K]/[X!\Gamma(K)])\,(K/[M + K])^K\,(M/[M + K])^X \quad [1]$$

for $X = 0, 1, 2, \ldots$, where Γ is the gamma function, X is the quantity of aflatoxin per peanut kernel, M is the average quantity of aflatoxin in the total population of kernels and K is a shape parameter. When each peanut kernel in the total population is considered to weigh the same, X may be used to denote aflatoxin concentration in each kernel, and M will then denote the average concentration of aflatoxin in the total population of kernels. Since aflatoxin determinations are generally reported in concentrations of aflatoxin, X and M will designate aflatoxin concentration in the remainder of this paper.

Equation 1 can take the form $F(X) = (\Gamma[X + K]/[X!\Gamma(K)])\,p^K q^X$, where $p = K/(M + K)$, $q = M/(M + K)$ and $q = 1 - p$. The cumulative distribution may be written as

$$F'(X) = \sum_{r=0}^{X} F(r) = \sum_{r=0}^{X} (\Gamma[r + K]/[r!\Gamma(K)])\,p^K q^r \quad [2]$$

where r is a dummy variable for X.

[1]Journal Series Paper of the North Carolina State University Agricultural Experiment Station, Raleigh, N.C.

TABLE I

Aflatoxin Test Results ($\bar{x}$ Values) for Ten 12 lb Samples from Each of 29 Mini lots[a]

Mini lot	Observed $\bar{x}$ values, ppb										Average $\bar{x}$ values (m)
1	0	0	0	0	0	0	0	6	10	14	3.0
2	0	0	0	0	2	3	4	8	23	53	9.3
3	0	0	0	0	2	4	8	14	28	43	9.9
4	0	0	0	0	0	0	0	16	40	69	12.5
5	0	3	6	8	8	10	14	16	22	39	12.6
6	0	0	0	0	0	3	8	26	52	70	15.9
7	0	0	0	0	0	0	0	2	40	125	16.7
8	0	0	0	0	8	8	15	16	16	125	18.8
9	0	0	0	0	3	13	19	41	43	69	18.8
10	0	0	0	4	4	5	25	26	58	67	18.9
11	0	0	6	10	18	19	20	25	52	67	21.7
12	0	0	0	0	0	5	10	35	39	130	21.9
13	0	0	0	0	0	0	0	8	22	198	22.8
14	0	0	3	12	12	12	12	25	63	103	24.2
15	0	9	11	14	17	17	30	44	57	59	25.8
16	0	0	0	0	0	0	0	9	12	285	30.6
17	0	0	3	4	4	5	15	60	106	165	36.2
18	0	6	6	8	10	50	60	62	66	130	39.8
19	3	3	9	12	41	42	57	70	80	126	44.3
20	0	0	32	32	34	37	55	67	77	134	46.8
21	0	3	5	19	32	49	87	91	127	168	58.1
22	18	21	25	35	43	46	86	86	94	169	62.3
23	4	7	40	41	55	60	75	95	99	230	70.6
24	11	19	23	38	54	90	96	108	132	140	71.1
25	3	6	11	18	80	99	104	116	129	147	71.3
26	0	4	6	17	36	80	133	148	192	216	83.2
27	5	12	56	66	70	92	98	132	141	164	83.6
28	18	50	53	72	82	108	112	127	182	191	99.5
29	29	37	41	71	95	117	168	174	183	197	111.2

[a]Test results are given in ppb aflatoxin and are ordered according to aflatoxin concentration.

TABLE II

Average of $\bar{x}$ Values m, Variance of $\bar{x}$ Values $s^2_{\bar{x}}$, Number of Kernels in the Sample n and Shape Parameter for the Negative Binomial Equation k

Mini lot	m, ppb	$s^2_{\bar{x}}$	n	$\hat{k}$ (x 10^4)
1	3.0	26.9	11,960	0.2799
2	9.3	285.1	10,800	0.2809
3	9.9	214.8	10,800	0.4226
4	12.5	561.6	11,960	0.2326
5	12.6	126.9	10,800	1.1581
6	15.9	647.2	10,800	0.3617
7	16.7	1604.5	10,800	0.1609
8	18.8	1439.5	10,800	0.2273
9	18.8	588.4	10,800	0.5562
10	18.9	625.9	9,680	0.5908
11	21.7	481.1	10,800	0.9062
12	21.9	1663.9	11,960	0.2410
13	22.8	3838.3	10,800	0.1254
14	24.2	1093.7	9,990	0.5726
15	25.8	431.7	10,800	1.4276
16	30.6	8009.6	10,800	0.1082
17	36.2	3249.7	10,800	0.3734
18	39.8	1732.8	10,800	0.8464
19	44.3	1619.1	9,990	1.2142
20	46.8	1563.3	10,800	1.2973
21	58.1	3353.4	9,890	1.0233
22	62.3	2229.4	9,950	1.7474
23	70.6	4177.2	9,940	1.2025
24	71.1	2313.6	10,800	2.0231
25	71.3	3164.6	10,620	1.5120
26	83.2	6871.7	10,800	0.9687
27	83.6	2773.6	10,320	2.4421
28	99.5	3168.8	10,020	3.1200
29	111.2	4315.1	9,760	2.9385

If the random variable X is described by the negative binomial distribution, then the distribution of the sum of N independent observations is negative binomial with mean NM and shape parameter NK (3). The sum, $\sum_{i=1}^{N} X_i$ is equivalent to $N\bar{X}$, where $\bar{X}$ is the concentration of aflatoxin in the sample and N is the number of kernels in the sample. Therefore the cumulative distribution of the sum of N observations can be expressed as

$$F^*(N\bar{X}) = \sum_{r=o}^{N\bar{X}} (\Gamma[r + NK]/[r!\Gamma(NK)])\, p^{NK} q^r. \qquad [3]$$

The cumulative distribution of the aflatoxin concentrations in the samples $F^*(\bar{X})$ can be determined by a scale transformation of equation 3.

The negative binomial distribution is completely defined by two parameters M and K. By assigning values to these parameters the distribution of $\bar{X}$ values for replicated samples of N kernels from lots with a concentration of M ppb aflatoxin can be predicted by equation 3. The accuracy of this prediction is dependent upon a correct choice of K.

Rationale for the application of equation 3 to predict the distribution of aflatoxin test results as a function of sample size N and true lot mean M has been based upon theoretical considerations as discussed previously by the authors (1,2). Due to the limited amount of information concerning the model parameters, past predictions based on equation 3 were made with assumed values of K (2). Therefore the primary objective of this study was to obtain a more accurate estimate of K and determine if a functional relationship exists between K and M. A secondary objective was to compare the observed distribution of aflatoxin test results to the negative binomial distribution.

EXPERIMENTAL PROCEDURES

For this study 29 "minilots" weighing ca. 120 lb each were drawn from 29 commercial lots of shelled peanuts

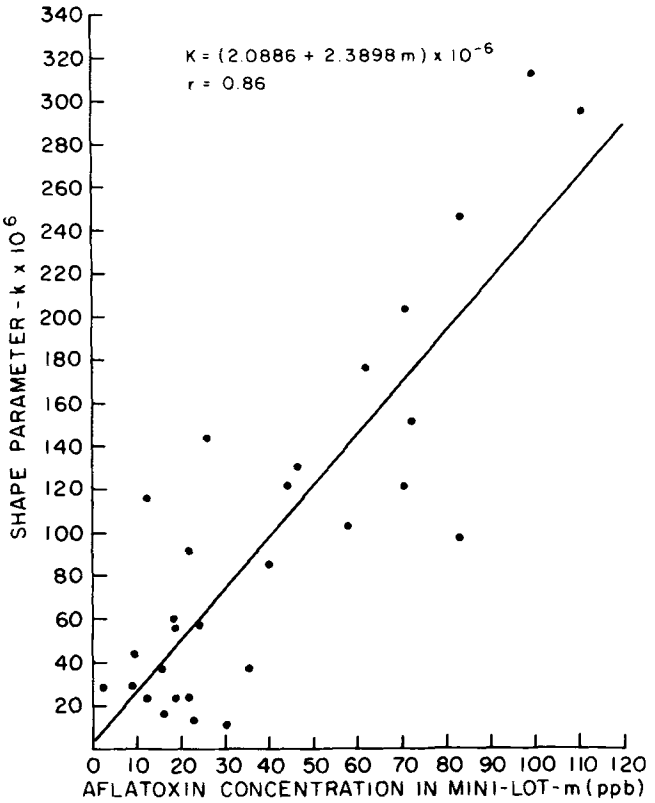

FIG. 1. Relationship between the negative binomial shape parameter k and the aflatoxin concentration in the minilot m. Correlation coefficient $r = 0.86$.

contaminated with aflatoxin. It was assumed that the distribution of aflatoxin among the kernels in the minilots was representative of the distribution found in typical commercial lots. Using a riffle divider, each minilot was divided into 10 samples of ca. 12 lb each. An estimate of the number of kernels N in each sample was based upon the weight of the sample and a kernel per pound determination for each minilot. Every sample was passed through a subsampling mill (4), and a single subsample of ca. 250 g of comminuted material from each sample was analyzed for aflatoxin with the Waltking method (5). As a result, ten 250 g subsamples (each representing a 12 lb sample of kernels) were analyzed for aflatoxin from each minilot. Aflatoxin test results are considered to be estimates of sample concentrations $\bar{X}$ and are denoted by $\bar{x}$. Estimates of M, N and K, based upon experimental values, are also denoted by m, n and k, respectively.

Parameter Estimation

Anscombe (3) discussed five methods to estimate the parameters M and K of the negative binomial distribution. The procedure listed by Anscombe as Method 1, often called "the method of moments," was used in this study. The method of moments was modified to use $\bar{x}$ values to estimate the parameters M and K.

The first moment of equation 1 is

$$\mu_1 = Kq/p = M \qquad [4]$$

and the second moment about the mean is

$$\mu_2 = Kq/p^2 = M + (M^2/K) = \sigma^2 \qquad [5]$$

where σ^2 is the variance of the kernel population in the minilot. Equation 5 shows that $\sigma^2 \geqslant M$ for the negative binomial distribution. As K goes to infinity, $\sigma^2 = M$, which is characteristic of the Poisson distribution.

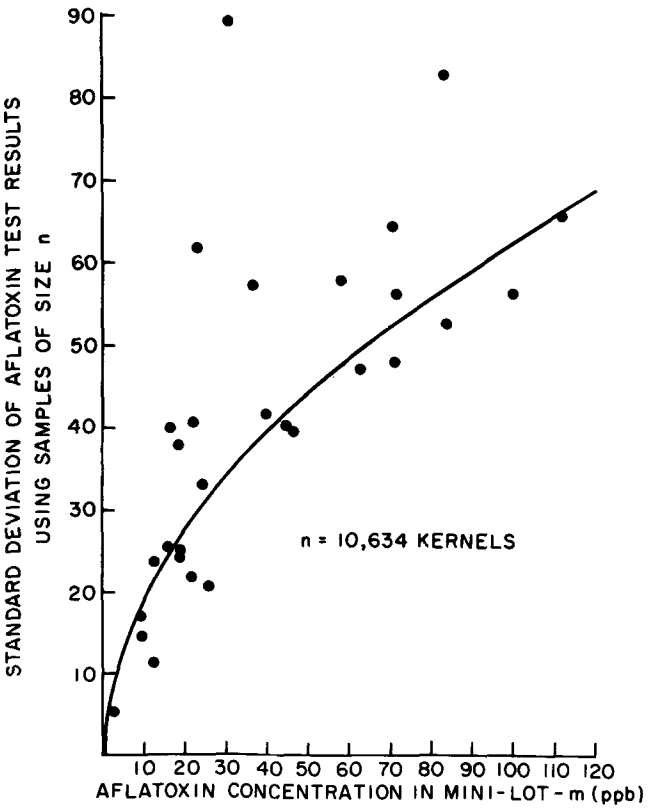

FIG. 2. Comparison of the standard deviation predicted by the fitted model (equation 10) and observed values of the standard deviation (Table II).

The parameter M is estimated by taking an average of the 10 $\bar{x}$ values for each minilot. The variance σ^2 is equal to the variance of the sample means $\sigma^2_{\bar{x}}$ times the sample size n. For each minilot the variance of the 10 $\bar{x}$ values, $s^2_{\bar{x}}$ was computed as an estimate of $\sigma^2_{\bar{x}}$. Therefore

$$s^2 = n\, s^2_{\bar{x}}. \qquad [6]$$

From equation 5,

$$k = m^2/(s^2 - m). \qquad [7]$$

Substituting equation 6 into 7, the moment estimate of k is

$$k = m^2/(n\, s^2_{\bar{x}} - m). \qquad [8]$$

In the range of k and m values expected, the statistical efficiency of the method of moments described above is not as high as the method of measuring the proportion of kernels not contaminated with aflatoxin (3). However present aflatoxin assay techniques for individual kernels are insensitive and too costly for measuring the large population of kernels necessary to accurately estimate the per cent of kernels with zero aflatoxin.

Comparison of the Observed Distribution of $\bar{x}$ Values to the Theoretical Distribution

The theoretical distribution of $\bar{x}$ values $F^*(\bar{x})$ defined by equation 3 can be generated using values of k and m calculated by the procedure outlined above. The Kolmogorov-Smirnov (K-S) test (6-8) was used to determine the probability that the observed cumulative distribution of $\bar{x}$ values $C(\bar{x})$ came from a population having a true but unknown distribution function $F(\bar{x})$ that can be specified by the negative binomial equation $F^*(\bar{x})$. The test is based upon the greatest absolute differences D_{max} between $C(\bar{x})$ and $F^*(\bar{x})$. If D_{max} is greater than some critical value D_{nn}, then the null hypothesis H_o that $F(\bar{x})$ is equal to $F^*(\bar{x})$ is

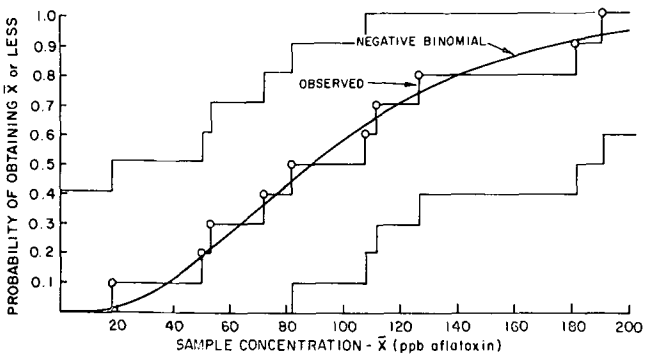

FIG. 3. Comparison of the cumulative distribution of aflatoxin concentrations in samples of peanuts as predicted by the negative binomial equation and the cumulative distribution of aflatoxin determinations ($\bar{x}$ values) on samples from minilot number 28. Values for the aflatoxin concentration in the minilot, m = 99.5 ppb; sample size, n = 10,020 kernels; and shape parameter k = 3.1200 x 10^{-4} were determined from the minilot and used in the solution of the negative binomial equation. Upper and lower bounds around the observed distribution are shown for the 5% significance level.

rejected with significance α. Values of D_{nn} for various significance levels α and number of samples nn are presented in several texts (6-8).

The K-S test is exact when the hypothesized distribution function $F^*(\bar{x})$ is continuous; otherwise the test is conservative (8). Also the K-S test is valid only when the parameters of $F^*(\bar{x})$ are evaluated independent of the observed data (9). Little is known about the D_{nn} statistic when the parameters are evaluated from the observed data except the critical value of D_{nn} listed in the tables should be reduced slightly in magnitude (9,10). However Kendall and Stuart (9) indicate that, when the parameters are determined from the observed data the Kolmogorov two-sided test statistic, D_{nn} may be used to form a confidence band for the true unknown distribution function $F(\bar{x})$ for any significance level 1-α. The confidence band is a band of width $\pm D_{nn}$ around the observed cumulative distribution function $C(\bar{x})$, and the probability that the true unknown distribution function $F(\bar{x})$ lies entirely in the band is 1-α. Therefore, if $F^*(\bar{x})$ lies completely with the band $C(\bar{x}) \pm D_{nn}$, then the null hypothesis H_o that $F(\bar{x}) = F^*(\bar{x})$ cannot be rejected with significance α.

RESULTS AND DISCUSSION

Observed values of $\bar{x}$ from all minilots are tabulated in Table I. Values of m, $s^2_{\bar{x}}$ and k computed from the data in Table I are listed in Table II, along with sample size n. The minilots are ranked according to their m values in both Table I and II. Inspection of Table II shows that for each minilot the variance of the sample means $s^2_{\bar{x}}$ is greater than the average of the 10 sample means m. This implies $s^2 \geqslant m$, which is a necessary condition for the negative binomial distribution to be applicable.

The 29 values of k are plotted vs. m in Figure 1. A multiple regression analysis performed on the 29 values of k and m gave the expression:

$$k = (2.0886 + 2.3898m) \times 10^{-6}. \qquad [9]$$

An analysis of variance indicated that the quadratic term was not significant at the 10% level, and higher order terms were negligible. The relationship between k and m is not determined for values greater than m = 111.2 ppb, but the critical level for aflatoxin determinations is usually 25 ppb or less, and a relationship for higher m values is of little practical value.

The variance $s^2_{\bar{x}}$ of aflatoxin test results $\bar{x}$, as predicted by the fitted model, can be calculated as a function of m

and n by substituting equation 9 into equation 8.

$$s^2_{\bar{x}} = (1/n)(m^2/[(2.0886 + 2.3898m) \times 10^{-6}] + m). \quad [10]$$

The coefficient of variation, expressed as a per cent is

$$CV = 100\, s_{\bar{x}}/m. \quad [11]$$

Figure 2 shows a plot of equation 10 for n = 10,634 kernels (average of the 29 n values in Table II) and m values up to 120 ppb. The measured values of $s^2_{\bar{x}}$ shown in Table II are plotted for comparison. It should be noted that the curve is derived from a least square fit of computed k values and not $s^2_{\bar{x}}$ values.

From aflatoxin test results listed in Table I, observed cumulative distributions of $\bar{x}$ were constructed for each of the 29 minilots. Using the Kolmogorov two-sided test statistic D_{nn} of ±0.409 for nn = 10 observations and a 5% significance level, an upper and lower bound was placed on each of the 29 observed distribution functions $C(\bar{x})$. For each minilot, the theoretical distribution $F^*(\bar{x})$ (equation 3) was generated for k and m values listed in Table II and compared to $C(\bar{x})$. Due to the insensitivity of the aflatoxin assay procedure used in this study, samples that truly had 0.5 ppb or less aflatoxin tested as zero ppb. Therefore all observed zero values were treated as 0.5 ppb or less when calculating the cumulative probability distribution of observed values.

Figure 3 shows one such comparison of $C(\bar{x})$ and $F^*(\bar{x})$ for minilot number 28. All 29 comparisons cannot be illustrated here, but in all cases the negative binomial equation $F^*(\bar{x})$ fell entirely within the upper and lower bounds around each $C(\bar{x})$. Therefore the null hypothesis H_o that the true unknown distribution function $F(\bar{x})$ is equal to the negative binomial distribution, $F(\bar{x}) = F^*(\bar{x})$, cannot be rejected at the 5% significance level for any of the 29 minilots.

The results of this study provide an estimate of the functional relationship between the negative binomial parameters K and M. Use of the relationship should provide a more accurate prediction of the risk levels associated with aflatoxin sampling plans based on the negative binomial equation. Variability of the observed data and comparisons between the model and the observed distribution indicate that the negative binomial distribution is a reasonable choice for the simulation model.

ACKNOWLEDGMENT

Skippy Laboratories, Best Foods, CPC Internationa, gave financial support.

REFERENCES

1. Whitaker, T.B., and E.H. Wiser, JAOCS 46:377 (1969).
2. Whitaker, T.B., E.H. Wiser and J.W. Dickens, Ibid. 47:501 (1970).
3. Anscombe, F.J., Biometrika 37:358 (1950).
4. Dickens, J.W., and J.B. Satterwhite, Food Technol. 23:90 (1969).
5. Waltking, A.E., G. Bleffert and M. Kiernan, JAOCS 45:880 (1968).
6. Siegel, S., "Nonparametric Statistics," McGraw-Hill, New York, 1956.
7. Ostle, B., "Statistics in Research," Iowa State University Press, Ames, Iowa, 1963.
8. Conover, W.J., "Practical Nonparametric Statistics," John Wiley & Sons, Inc., New York, 1971.
9. Kendall, M.G., and A. Stuart, "The Advanced Theory of Statistics," Vol. 2, Charles A. Griffen and Co., Ltd., London, 1961.
10. Benjamin, J.R., and C.A. Cornell, "Probability, Statistics and Decision for Civil Engineers," McGraw-Hill, New York, 1970.

[Received February 22, 1972]