Modelling the distribution of aflatoxin amount by kernel
Analyzing Tom Whitaker's data
Jason Osborne, February, 2019

In each of 20 lots of almonds, there were 16 samples, each of 100g of crushed up almonds. Each sample consisted of $N = 7730$ almonds. (*is this right?*) A measurement for sample j from lot i, $\bar{X}_{ij}$ is obtained by dividing the observed ppb of aflatoxin, X_{ij} by the number of kernels in the sample: $\bar{X}_{ij} = (X_{ij}/N)$. So, X_{ij} is really a sum of N kernels,

$$X_{ij} = \sum_{l=1}^{N} X_{ijl}.$$

This analysis aims to quantify the distribution of X_{ijl} for a given lot, given observed summary statistics as described below.

The averages of the measured concentrations from the 16 samples from each lot i are denoted

$$\bar{X}_i = \frac{1}{16}\sum_{1}^{16}(X_{ij}/N) = \frac{1}{16}\sum_{j}\bar{X}_{ij} \qquad \text{for} \quad i = 1,\ldots,20$$

The sample variance among the 16 aflatoxin concentration sample averages from lot i is denoted S_i^2:

$$S_i^2 = \frac{1}{16-1}\sum_{j=1}^{j=16}(\bar{X}_{ij} - \bar{X}_i)^2 \qquad \text{for} \quad i = 1,\ldots,20$$

The negative binomial distribution is proposed to model ppb among kernels. The general form of this distribution, with mean parameter M and shape parameter K, and its moments are given below. Let X_{ijl} denote the ppb of a kernel sampled from sample j, lot i (even though sampling of individual kernels is not possible). Then

$$\begin{array}{rcccc} P(X_{ijl} = x) & = & \left(\begin{array}{c} x + K_i - 1 \\ K_i - 1 \end{array}\right)\left(\frac{K_i}{M_i+K_i}\right)^{K_i}\left(\frac{M_i}{M_i+K_i}\right)^{x} & \text{for} & x = 0,1,2,\ldots \\ E(X_{ijl}) & = & M_i & = & \mu \\ V(X_{ijl}) & = & M_i + M_i^2/K_i & = & \sigma^2 \end{array}$$

If the distribution among kernels is negative binomial with parameters M and K, then the sum of ppb in a sample (j) and lot (i) is also negative binomial, under the simplifying assumptions that the kernels are a random sample from the lot. The mean parameter is NM_i and the shape parameter is NK_i. The mean and variance are given by

$$\begin{array}{rcl} E(X_{ij}) & = & E(N\bar{X}_{ij}) = NM_i \\ V(X_{ij}) & = & V(N\bar{X}_{ij}) = NM_i + (NM_i)^2/(NK_i) \end{array}$$

For an average (or measurement on a sample j), $\bar{X}_{ij} = X_{ij}/N$, the moments are

$$\begin{array}{rcl} E(\bar{X}_{ij}) & = & M_i \\ V(\bar{X}_{ij}) & = & \frac{1}{N^2}(NM_i + (NM_i)^2/(NK_i)) \\ & = & \frac{M_i}{N} + M_i^2/(NK_i) \end{array}$$

For averages of sample averages, and sample variances like those appearing in the spreadsheet:

$$\begin{array}{rcl} \bar{X}_i & = & \frac{1}{16}\sum_{j=1}^{16}\bar{X}_{ij} \\ E(\bar{X}_i) & = & E(\frac{1}{16}\sum_{j=1}^{16}\bar{X}_{ij}) = M_i \\ E(S_i^2) & = & V(\bar{X}_{ij}) = \frac{M_i}{N} + M_i^2/(NK_i) \end{array}$$

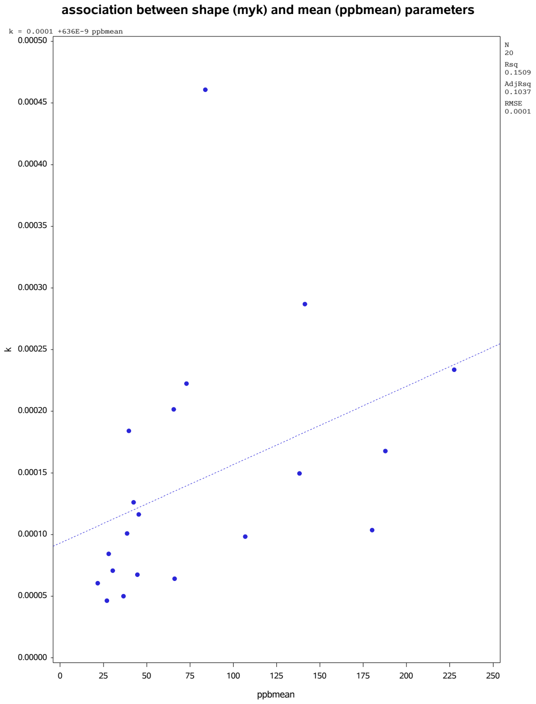
The *method of moments* sets the first two sample moments equal to their expectation, obtaining a system of two equations and two unknowns:

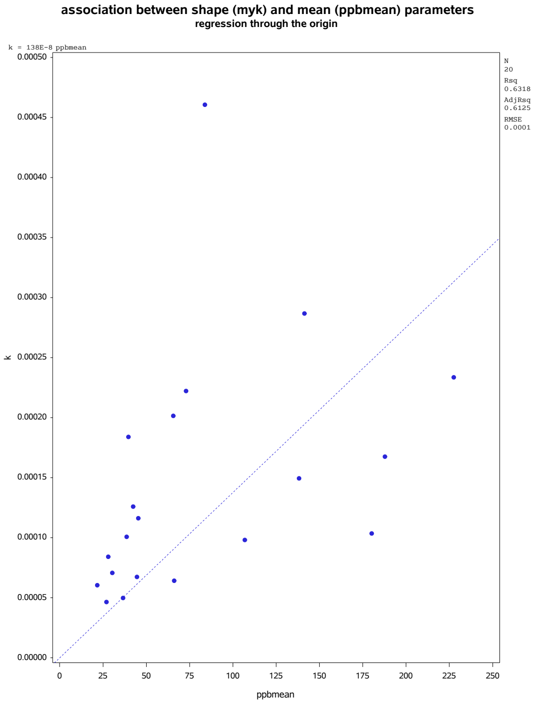
$$\begin{aligned} \bar{X}_i &\overset{\text{set}}{=} M_i \\ S_i^2 &\overset{\text{set}}{=} \left(\frac{M_i}{N} + M_i^2/(NK_i)\right) \end{aligned}$$

The solution $(\hat{M}_i, \hat{K}_i)$ is given by

$$\begin{aligned} \widehat{M}_i &= \bar{X}_i \\ \widehat{K}_i &= \frac{\widehat{M}_i^{\,2}}{NS_i^2 - \widehat{M}_i} \end{aligned}$$

The estimated parameters for all 20 lots appear in the plots below, with shape parameters (k) on the vertical axis and lot mean on the horizontal axis. The plot on the left allows for a non-zero intercept, the plot on the right is through the origin. Both provide some indication of a positive association between estimated shape and scale. Lots with more aflatoxin also tend to have a larger estimated shape parameter. The lot with the largest shape parameter may be an outlier of sorts.





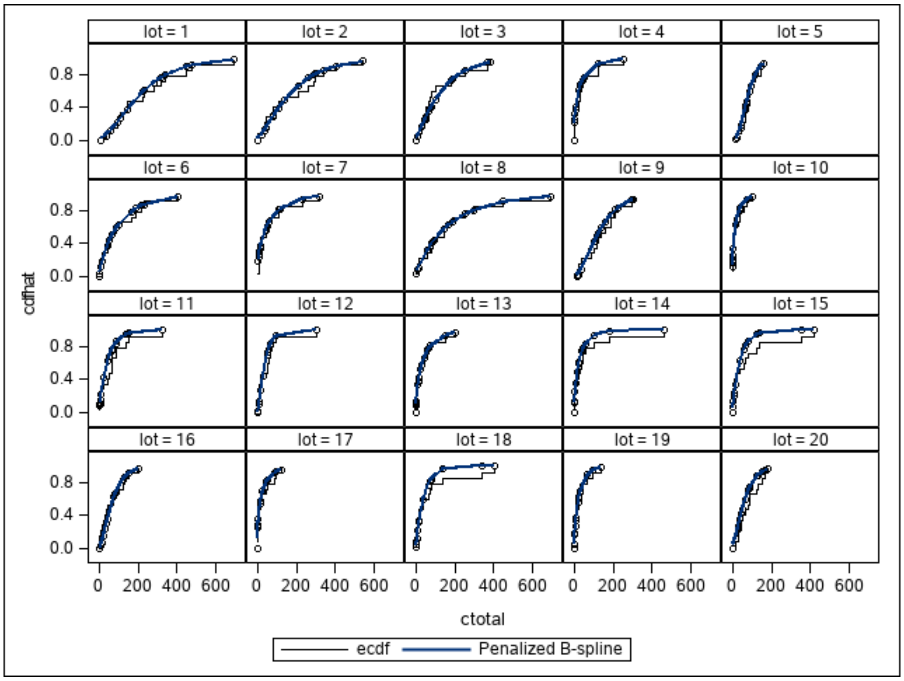
For each of the points (lots) in these plots, there were 16 samples, and from these samples, one or two subsamples were taken and from these subsamples, one or two aliquots were taken. This nested measurement is useful for estimating sources of variation in the assessment of aflatoxin. To assess the fit of the negative binomial model, the first aliquots from the first subsamples from each sample of each lot were used in empirical distribution function (ECDF) plots. The ECDF for a given lot at any concentration is simply the proportion of samples from that lot (out of 16) with an average that was less than or equal to the concentration. This is naturally a step-function with jump discontinuities at the observed concentration amounts for each sample. At each such concentration, the function increases by 1/16 since there are 16 sample averages per lot. This function is a nonparametric and unbiased estimate of the real cumulative distribution function for averages of N kernels. The distribution of aflatoxin in a sum of N kernels from lot i, $X_{ij} = \sum_l X_{ijl}$ is also negative binomial, with mean NM_i and NK_i. This fact can be used to obtain an expression for an estimate of the theoretical distribution function for the averages from lot i. Using the probability mass function for kernels, X_{ijl} from the previous page, we have

$$\begin{aligned} P(\bar{X}_{ij} \le x) &= P(\sum_l X_{ijl} \le Nx) \\ &= P(X_{ij} \le Nx) \\ &= \sum_{t=0}^{[Nx]} P(X_{ij} = t; N\widehat{M}_i, N\widehat{K}_i) \end{aligned}$$

Here, $[Nx]$ is the largest integer less than or equal to Nx. All statistical software packages have this function built-in. In SAS, the generic `CDF` function may be used along with the 'NEGB' parameter:

```
cdfhat = CDF('NEGB',N*xbar,phat,N*khat);
```

Plots in which the estimated negative binomial cumulative distribution function are overlaid with the ECDF for the 20 lots are given below:



The step function makes no assumptions about the distribution of aflatoxin among kernels, and the blue line is a smoothed version of the function which assumes aflatoxin among kernels follows the negative binomial distribution. There appears to be reasonable agreement between these two estimate of the distribution functions associated with the sample averages of all $N = 20$ lots. To formally assess the goodness-of-fit of the Negative Binomial model, one could obtain the Kolmogorov-Smirnov test statistic, which is the maximum distance between the estimate based on the Negative Binomial assumption and the ECDF across all values of the aflatoxin sum for each lot. To see whether or not these differences exceed the $\alpha = .05$ level critical value, D_{nn} for this test using $n = 16$ samples per lot, the ECDF was plotted with limits of the form $\pm D_{nn} = .3273$. In none of the lots did the Negative Binomial estimate fall outside of this interval about the ECDF. That is, the observed differences between the two estimates of the distribution function for aflatoxin were not significant at level $\alpha = .05$.