SOME DISTRIBUTIONS USED TO DESCRIBE

HELMINTH EGG COUNTS

by

RICHARD ORRIE GILBERT

B. S., Ferris State College, 1963

———————————————

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1965

Approved by:

TABLE OF CONTENTS

A problem of major concern to parasitologists and others interested in the various aspects of host-parasite relationships has been the determination of the intestinal worm burden of various types of animals. Worm burden can be ascertained with a high degree of accuracy if the host animals are sacrificed and the intestinal contents examined for the numbers and species of worms present. If several hosts are killed and the number of worms counted in each, a frequency distribution of the counts, i.e., the number of animals with 0,1,2,... worms of a particular species, or of all species combined, can be calculated. It may then be possible to determine if this observed frequency distribution follows some theoretical statistical distribution. If it does, then the probability of a particular worm count occurring in any sample taken from this population can be determined. The sample estimates of the parameters of the theoretical distribution can also be calculated from sample counts adding to existing knowledge of the observed distribution.

If sacrificing the host to obtain worm counts is undesirable, then parasite egg counts, obtained by taking subsamples of faecal samples collected from the infected host, may be used to estimate the true worm burden of the host. Just as in the case of worm counts, a frequency distribution can then be constructed based on the observed number of faecal subsamples with 0,1,2,... eggs present.

This report is concerned with the most popular theoretical distributions used by various workers in the past to fit the observed egg counts under specified conditions.

Peters and Leiper (1940), using two different egg counting techniques, concluded that the distribution of successive counts of Haemonchus contortus

eggs, when these counts were all from the same suspension of eggs, was approximately the Poisson. Peters (1941) presents an interesting discussion of dilution egg counts and gives three examples showing that the variance or mean square of these counts approximates their respective means. He concludes that the Poisson series is applicable. Emik (1947) concluded (using a dilution counting technique) that nematode egg counts from 12 heterogeneous lambs were Poisson distributed. Brambell (1963) confirms Peters (1941) position by taking egg counts from housed sheep. Hunter and Quenouille (1952) took four faecal samples from each of 132 sheep and determined by a chi-square test (discussed in the next section) that the distribution of the replicate worm-egg counts in each sheep fitted the well-known Poisson distribution. They performed other trials that gave similar results, i.e., Poisson distributed counts; and in addition, the over-all chi-square test (i.e., the chi-square value over all trials) also indicated the Poisson distribution. When Hunter and Quenouille (1952) investigated 13 series of egg counts from different sheep however, they found that the negative binomial distribution (also known as the binomial waiting-time distribution or the Pascal distribution, (Wilks, 1962)) was appropriate.

If the assumption is made that egg counts are accurate estimates of the worm population, then it becomes important to investigate what distributions have been found to describe worm counts themselves. As regards this assumption it is interesting to note that Willmott and Pester (1952) performed an experiment in which they tried to ascertain within what limits egg counts were reliable criteria on which to base estimates of the number of flukes (paramphistomes) in the host. They admit that the number of observations was too few to permit definite conclusions, but based on their observations of

9 infected cattle (both egg and worm counts), they concluded that there appears to be a correlation between the number of paramphistome eggs present in a rectal sample and the number of paramphistome parasites in the intestinal tract.

Northam and Rocha (1958) confirmed the findings of Egerton (1953), who showed that the frequency distribution of worm counts in experimentally infected chickens followed the negative binomial distribution. Egerton and Hansen (1955) found that the negative binomial fitted the counts of Ascaridia galli which they obtained while working on the problem of immunity and tolerance of chickens to A. galli. Li and Hsu (1951) studied 15 species of parasitic nematodes, 3 species of cestodes, and 5 species of trematodes, and found that the frequency distribution of the counts of parasites found in their natural host (of various types) had the characteristics of being skewed in the positive direction and similar in appearance. They hypothesized that I_j of the Pearsonian frequency curves would probably fit most of the curves. Although Northam and Rocha (1958) do not discount this conclusion, they point out that the negative binomial distribution would probably fit the data as well, and it would have the advantage of being discrete just as the data are discrete.

From the conclusions of these various investigators, it appears that both egg counts and the underlying worm population can be described by the Poisson or the negative binomial distributions, depending on the sampling procedure. This report is therefore centered around these two distributions; their derivations, properties, and applicability to egg count studies.

THE POISSON DISTRIBUTION

Derivation

The Poisson distribution was discovered by Poisson in 1837. Bortkewetsch later expanded Poisson's work by illustration. This series was independently discovered in 1907 (Whitaker, 1914) by "Student" (1907) in his paper entitled "On The Error of Counting With a Haemacytometer", in which he showed that the distribution of small particles in a liquid followed the Poisson law

$$e^{-m} [1 + m + \frac{m^2}{2!} + \ldots + \frac{m^r}{r!} + \ldots] = \frac{e^{-m} m^r}{r!} \tag{1}$$

where m is the arithmetic mean number of particles per unit volume, e is the base of the natural logarithm and is approximately equal to 2.718, and the successive terms in the series gives the probability that a given unit volume contains $0, 1, 2, \ldots, r, \ldots$ particles.

The Poisson distribution can be derived in the following manner (Student, 1907; Ostle, 1963). Suppose that a liquid suspension of helminth eggs obtained from a faecal sample is thoroughly mixed and spread evenly over a surface marked off into N equal units of area. Suppose further that each area has an average of m eggs contained in it, resulting in a total of Nm eggs throughout the whole suspension. If the suspension has been thoroughly mixed, a given egg will have an equal probability ($\frac{1}{N}$) of falling on any one of the unit areas, and an equal probability ($1 - \frac{1}{N}$) of not falling on a particular unit area. Similarly, the probability that a given unit area will contain an egg is $\frac{m}{N}$, and not contain an egg is $1 - \frac{m}{N}$.

It must be assumed that each unit area has the capacity to hold any number of eggs without affecting the probability of still more eggs falling

on that area. That is, the effect of crowding is negligible.

If $P(X = x)$ is the probability of x eggs falling on a given unit area, then

$$P(X = x) = \binom{mN}{x}\left(\frac{1}{N}\right)^x \left(1 - \frac{1}{N}\right)^{mN-x}, \qquad x = 0,1,2,\ldots \tag{2}$$

which is simply the expansion of

$$\left(\frac{1}{N} + \left(1 - \frac{1}{N}\right)\right)^{mN}. \tag{3}$$

Equation (2) is usually written as

$$\binom{n}{x} p^x q^{n-x}, \qquad x = 0,1,2,\ldots \tag{4}$$

with $n = mN$, $p = \frac{1}{N}$, and $q = 1 - \frac{1}{N}$

where it should be noted that as $N \to \infty$, the quantity $\frac{1}{N} \to 0$.

The $(x + 1)^{th}$ term in the expansion of (2) is

$$\left[\frac{1}{x!}\right] mN[mN-1][mN-2] \ldots [mN-x+1]\left[\frac{1}{N}\right]^x \left[1-\frac{1}{N}\right]^{mN-x} \tag{5}$$

Letting $Nm = n$, it is evident that $\frac{1}{N} = \frac{m}{n}$ and $1 - \frac{1}{N} = 1 - \frac{m}{n}$.

Therefore, (5) becomes

$$\frac{n(n-1)(n-2)\ldots(n-x+1)}{x!} \left(\frac{m}{n}\right)^x \left(1 - \frac{1}{N}\right)^{Nm-x}. \tag{6}$$

Dividing both the numerator and denominator by n^x, it is seen that (6) reduces to

$$1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \ldots \left(1 - \frac{x+1}{n}\right) \frac{m^x}{x!} \left(1 - \frac{1}{N}\right)^{Nm} \left(1 - \frac{1}{N}\right)^{-x}. \tag{7}$$

Taking the limit of this expression as n → ∞ (which implies that mN → ∞ which in turn implies that N → ∞, since m is considered fixed) it is seen that (7) becomes

$$\frac{m^x}{x!} e^{-m}, \qquad\qquad x = 0,1,2,\ldots \quad (8)$$

which is the probability function of the Poisson distribution.

From this derivation, it is apparent that the Poisson distribution is a limiting form of the binomial distribution given by equation (4) where N → ∞ , resulting in $\frac{1}{N}$ → 0 while the mean number of particles per unit area (m) remains constant; or as Peters (1941) puts it, "The Poisson series is simply the binomial series pushed to the limit where p is indefinitely small, q is near unity, and n is so large that the mean, np, is an appreciable quantity."

It is not difficult to justify the premise that dilution egg counts follow the Poisson distribution (Peters, 1941). Dilution egg counts are usually performed using the McMaster slide technique in which the eggs are counted that lie under a centimeter square engraved on the fixed coverglass which is supported 0.15 cm. above a slide by 0.15 cc. of egg suspension. Because the faecal suspension is a half-saturated solution of salt, the eggs rise to the coverglass making them easy to count.

Peters (1941) estimates the average volume of 7 common sheep nematode eggs to be 0.0002 mm^3, where he is considering the average egg to be a square prism of dimensions 90μ by 45μ. Since 150 mm^3 is the volume of the suspension under the centimeter square, Peters calculates that there is room for 750,000 eggs closely packed in the available space. If the mean number of eggs in each square centimeter is 100, then p, the probability that any unit

volume will be an egg is 100/750,000 = 0.00013, and q, that it will not be an

egg is (1 - p) = 0.99987. The counts from successive slides with the same

mean 100 are distributed as

$$(p + q)^n = (0.00013 + 0.99987)^{750,000} .$$

Note that p is very small and n is very large, indicating that successive

terms can be expressed in the form of the Poisson distribution.

Properties

Normal Approximation to the Poisson Distribution. The Poisson distri-

bution is known to approach the normal distribution as m becomes very large.

This can be shown to be true by investigating the distribution statistics α_3

and α_4, where

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

and

$$\alpha_4 = \frac{\mu_4}{\sigma^4} \quad ,$$

where μ_3 and μ_4 are the third and fourth moments about the mean of the dis-

tribution, respectively, and σ^3 and σ^4 are the third and fourth power of the

standard deviation of the distribution. The statistic α_3 is a measure of

skewness, and the normal distribution is always equal to zero due to the

symmetry of the normal curve. A measure of kurtosis, i.e., a measure of

whether the distribution is more peaked or flat-topped than the normal curve,

is given by the statistic α_4 and is always equal to three in the normal dis-

tribution. It can be shown that for any point binomial distribution found

by evaluating $(q + p)^n$, the values for α_3 and α_4 are

$$\alpha_3 = \frac{q-p}{\sqrt{npq}} \quad , \qquad \alpha_4 = \frac{1}{npq} - \frac{6}{n} + 3 \; .$$

Now if p is assumed to be very, very, small (which results in q approaching 1), and n to be very large so that np is an appreciable quantity, then α_3 and α_4 become appropriate statistics for the Poisson distribution. In such a distribution we have

$$\alpha_3 = \frac{1}{\sqrt{m}} \; , \text{ and } \; \alpha_4 = 3 + \frac{1}{m} \quad .$$

Thus, if m (the arithmetic mean of a Poisson distribution) is large, $\alpha_3 \to 0$ and $\alpha_4 \to 3$, which implies the Poisson distribution approaches the normal distribution (Waugh, 1943).

Equality of the Mean and Variance. One of the most useful properties of the Poisson distribution is the equality of the mean and variance, which can be shown in the following manner:

The moment generating function of a random variable X with a Poisson distribution is $\phi(\theta)$ where

$$\phi(\theta) = e^{-m} \sum_{x=0}^{\infty} \frac{e^{\theta x} m^x}{x!}$$

$$= e^{-m} \sum_{x=0}^{\infty} \frac{[me^{\theta}]^x}{x!}$$

$$= e^{-m} \cdot e^{me^{\theta}}$$

$$= e^{m(e^{\theta} - 1)} \quad .$$

$$E(X) = \frac{\partial \phi}{\partial \theta} \Big|_{\theta=0} = m \, e^{\theta} \, e^{m(e^{\theta}-1)} \Big|_{\theta=0}$$

$$= m$$

$$E(X^2) = \frac{\partial^2 \phi}{\partial \theta^2} \Big|_{\theta=0}$$

$$= m^2 \, e^{2\theta} \, e^{m(e^{\theta}-1)} + e^{m(e^{\theta}-1)} \, m \, e^{\theta} \Big|_{\theta=0}$$

$$= m^2 + m$$

It is concluded that

$$E(X) = m = \text{mean}$$

and

$$E(X^2) - (E(X))^2 = m^2 + m - m^2$$

$$= m .$$

The Poisson distribution is therefore completely specified by the one

parameter m.

Transformations. To make the variance and mean independent (where the

observations are Poisson distributed), Bartlett (1947) recommends the square

root transformation, $\sqrt{x}$, or for very small numbers the transformation

$\sqrt{x + 1/2}$.

Emik (1947) used the square root transformation on each egg count ob-

tained on each of twelve sheep where the counts were Poisson distributed.

For each sheep he calculated the mean of the transformed counts and found that

the mean and variance were no longer significantly correlated. He also per-

formed an analysis of variance on these transformed counts and made several

tests for significance. He was able to make these tests only because the

transformation was made, causing the means to become independent of the variances and the counts somewhat more normal.

Testing Hypotheses

Given that an experimenter has taken faecal samples and made numerous dilution counts from these samples where the number of eggs per count ranged from zero to n, he is usually interested in testing some hypothesis about these counts. A common hypothesis tested, is that the counts follow the Poisson distribution. If the experimenter assumes this hypothesis to be true he can find the expected number of counts with $0,1,2,\ldots,n$ eggs per count by evaluating the probability function of the Poisson distribution for $x = 0,1,2,\ldots,n$ and for m calculated using his sample counts. The probability function can be evaluated by first calculating equation (8) for $x = 0$, then use the relationship

$$\frac{e^{-m} m^{x+1}}{(x+1)!} = \frac{e^{-m} m^{x}}{x!} \cdot \frac{m}{x+1}$$

for $x = 1,2,\ldots,n$. The probability function for each value of x is then multiplied by N, the total number of counts taken, to obtain the expected frequency of counts for $x = 0,1,2,\ldots,n$. Now the experimenter is in need of some statistical test that will help him make a decision as to whether his observed counts differ significantly from the calculated expected counts, resulting in a decision to either accept or reject the hypothesis that the counts are distributed according to the Poisson distribution.

Probably the most commonly used test is the chi-square test developed by Karl Pearson in 1899. He developed the χ^2 statistic, where

$$\chi^2 = \sum_{i=1}^{n} (O_i - E_i)^2 / E_i .$$ (9)

In an egg count study as outlined above

O_i = the number of egg counts observed in the ith class where each

class is one unit in length

and

E_i = the number of egg counts expected in the ith class under the

hypothesis of Poisson distributed counts.

Under these conditions χ^2 has approximately the chi-square distribution

with $n - p - 1$ degrees of freedom, where the number of class intervals used

in fitting the distribution is n, and p is the number of parameters in the

distribution. Since the Poisson distribution is being fitted, the appropriate

degrees of freedom are $(n - 2)$. Equation (9) can thus be used to test the

hypothesis that the observed egg counts follow the Poisson distribution

(Ostle, 1963).

Instead of calculating the expected number of counts, it is a common

procedure to use the fact that for all Poisson series, the variance is

numerically equal to the mean and to use a slightly different form of the

chi-square statistic than given in equation (9). Since

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 / \sigma^2$$

has the chi-square distribution with $(n - 1)$ degrees of freedom (Fisher,

1954), for true samples from a Poisson distribution,

$$\chi^2 = \sum_{i=1}^{n} (x_i - m)^2 / m$$ (10)

is approximately distributed as a chi-square with (n - 1) degrees of freedom, where

x_i = the number of eggs observed in the ith count,

n = the total number of counts taken,

and

m = the mean number of eggs per count.

The reader should note that equation (9) is used when working with frequencies of egg counts, whereas equation (10) is applicable when working with the number of eggs per count. When using equation (10) the hypothesis being tested is identical to the hypothesis underlying equation (9).

It is also of importance to note that the sum of k independent chi-squares is a chi-square with degrees of freedom equal to the sum of the degrees of freedom of each individual χ^2. The resulting χ^2 test is sensitive and will often show discrepancies that were not apparent in the separate χ^2 values.

If many values of χ^2 are available for testing (all with the same degrees of freedom), it is often advisable to distribute the various χ^2 values into classes bounded by values given in the chi-square table (depending, of course, on the degrees of freedom with which the table was entered) as Brambell (1963) does with chi-squares calculated from egg counts. The expected frequency of these classes can be obtained directly from the chi-square table. Thus a χ^2 test may be performed using equation (9) where

0_i = the observed frequency of occurrence of χ^2 values in the ith class interval,

and

E_i = the expected frequency of occurrence of χ^2 values in the ith class as taken from the chi-square table.

Now that various methods of calculating the χ^2 statistic have been discussed, it is logical to ask how this statistic is used to test the hypothesis that a given series of counts follows the Poisson distribution. It was established in 1900 that the χ^2 statistic was distributed differently for different degrees of freedom. The chi-square tables mentioned above are a consequence of this discovery. These tables, for specified degrees of freedom give values of P where

P = the probability that a χ^2 value will be observed that is larger than the calculated χ^2 value from the sample, if the hypothesis of Poisson distribution counts is true.

To every value of χ^2 there thus corresponds some value of P. The question now arises: What values of P would indicate the hypothesis should be accepted? Fisher (1954) states that a value of P between 0.1 and 0.9 would certainly indicate that the hypothesis should not be rejected.

If an experimenter desires to test

$$H_1: m \leq m_o$$
$$H_2: m > m_o$$

for a particular series of counts, this may be accomplished by using a table of the cumulative Poisson distribution. In this table are values of F(x) where

$$F(x_o) = P(x \leq x_o)$$

$$= \sum_{x=0}^{x_o} \frac{e^{-m} m^x}{x!} \tag{11}$$

for various values of m and x. The procedure for a sample of size one, is to calculate

$$P = 1 - F(x - 1) \tag{12}$$

where $F(x)$ is read from the above mentioned table assuming $m = m_o$. If $P \leq \alpha$, where α is a preassigned level of significance, H_1 is rejected and H_2 accepted (Ostle, 1963). For samples of size two or larger, the sample sum z, replaces x in equation (12) were z has the Poisson distribution do to the following theorem from Wilks (1962), page 206:

If $(x_1, x_2, \ldots, x_n)$ is a sample from the Poisson distribution Po(m),
then the samplying distribution of the sample sum, say z, is distributed as a Po(nm).

The quantity $P = 1 - F(z-1)$ is then determined from the cumulative Poisson tables and H_1 is accepted or rejected on the basis of the size of P in relation to α.

As an example, suppose 4 McMaster slide counts have been made, resulting in the following egg counts; 3, 0, 5, and 8. Let the null and alternative hypothesis be

$$H_1: m \leq 2$$

$$H_2: m > 2,$$

which is testing the hypothesis that the average number of eggs per count is 2, and let α be 0.01. Since n = 4 and $m_o = 2$, from the above theorem it is seen that z is distributed as Po(8). Using the cumulative Poisson tables in Ostle (1963) and equation (12) with z in place of x, it is seen that

$$P = 1 - F(16 - 1) = 1 - F(15) = 1 - 0.992$$

$$= 0.008$$

Since $P < 0.01$, the H_1 is rejected and it is concluded that $m > 2$.

Often an experimenter wishes to test for significant differences between means of two or more Poisson series. Suppose 4 series of counts are known

to be Poisson distributed with 4 means m_1, m_2, m_3, and m_4. Let S_1, S_2, S_3, and S_4 be the total number of eggs counted in series 1, 2, 3, and 4 respectively, and let the mean over all 4 series of counts be $\bar{x}$. Then

$$\chi^2 = \frac{1}{\bar{x}} \sum_{i=1}^{4} (S_i - \bar{x})^2 \tag{13}$$

is approximately distributed as a chi-square with $(n - 1)$ degrees of freedom, where n is the number of distributions being compared. If χ^2 is significant it is concluded that the means, m_i, are significantly different and a single Poisson distribution cannot adequately fit all 4 series of counts (Snedecor, page 234, 1956; Ostle, page 125, 1963).

If the number of counts taken is large so that the total number of eggs counted is large, it is appropriate to test for differences between 2 Poisson distributions by means of "Student's" t test, where

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2_{\bar{x}_1 - \bar{x}_2}}} \tag{14}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}} \tag{15}$$

Under the hypothesis that $\mu_1 = \mu_2$, (15) reduces to

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}} \tag{16}$$

As an example, suppose 100 counts were taken resulting in 300 eggs being counted over all counts. If the entire sample of 100 counts is considered as a unit with 300 eggs, it may be thought of as a single sample drawn from a population whose mean is estimated at 300 and whose variance is therefore 300 (assuming the counts are Poisson distributed). Therefore, since Poisson populations with such large means are approximately normal in distribution, this sample may be considered as drawn from a normal population whose mean is estimated by $\bar{x}$ = 300 with standard error $s_{\bar{x}}$ = $\sqrt{300}$. If another series of counts is taken, say 75 counts resulting in 275 eggs, the difference between these two distributions may be tested by equation (16). That is

$$t = \frac{300 - 275}{\sqrt{300 + 275}}$$

$$\doteq \frac{25}{24}$$

$$\doteq 1.042$$

which would indicate that the two series of counts probably are from the same population, i.e., the same Poisson distribution (Snedecor, p. 487, 1956).

Egg Count Studies

As was mentioned in the introduction, Peters and Leiper (1940) did a study using sheep in which they investigated the variation among successive dilution egg counts from the same suspension of eggs. They proposed two questions:

1) What is the form of such a distribution?

2) What is the relationship between the mean and variance of this distribution?

The first part of the paper is concerned with obtaining and analysing
faecal samples taken from a single sheep infected with _Haemonchus contortus_.
The authors were concerned with the distribution of egg counts in samples of
0.15 ml. drawn from a 200 ml. suspension of eggs in water.

The suspension was agitated before the 0.15 ml. samples were withdrawn
using a McDonald pipette. Eleven series of counts were made (where a count
is understood to be all of the eggs present on a slide containing 0.15 ml. of
the suspension) where each series consists of 25 counts. All counts in a
series were made from a single volume of 200 ml. of a suspension of eggs,
where the concentration of eggs was different from series to series.

Since each series had its own mean and variance, and since the authors
were interested in the overall distribution of the 275 counts, the trans-
formation

$$y = \frac{x - \bar{x}}{s}$$

was applied to each series, where $\bar{x}$ and s varied according to the series of
counts in which the x variate appeared. The mean and standard deviation
of the new compounded distribution of 275 transformed counts were 0.0055 and
0.9894 respectively. To show that the transformed counts were in approximate
agreement with the normal curve, first it was shown that the distribution
was roughly normal in shape. This was done by simply superimposing a normal
curve over the histogram made from the transformed counts being grouped into
10 intervals of length $y = 0.50$ and observing the visual goodness of fit.
Secondly, the chi-square test was performed, i.e., the observed frequencies
in each interval were compared with the frequencies expected under the
hypothesis that the observed values were normally distributed. In the

calculation of the χ^2 statistic (using equation (9)), 8 intervals were used instead of 10 due to the interval on each end of the distribution having an expected frequency that was too small. If 8 intervals had been used, this would have resulted in an artificially large chi-square value. Thus the most extreme intervals were combined with their nearest neighbor. The total χ^2 was equal to 2.3762 with 7 degrees of freedom, corresponding to a P of about 0.9. Therefore Peters and Leiper (1940) concluded that the egg counts were close to the normal distribution in form.

The authors point out the mean and variance of each of the 11 original series of counts are roughly equal. They also note that as the mean of a Poisson series becomes large, the distribution approaches the normal. This result has been shown to be true previously in this report. With these thoughts in mind, they continue on in an attempt to show that the counts are Poisson in nature.

The mean and variance of each of the original 11 series of counts was plotted with the mean on the abscissa and the variance on the ordinate. A linear regression analysis was conducted resulting in a graph showing the linear regression line along with its 95% confidence bound lines. The expected line based on the Poisson distribution for which the mean and variance are equal, just intercepted the lower limit line, resulting in Peters and Leiper drawing the conclusion that the series of counts was just barely consistent with the Poisson distribution.

These authors also performed a regression analysis on the logarithms of the counts due to the fact that low counts were anticipated. Plotting the log standard deviation against the log mean and then calculating the regression line with its 95% confidence limits, the expected Poisson-theory line was out

of the confidence belt only at low values of the log mean. For means up to
about 15, they concluded that the variance may be assumed to be equal to the
mean using their sampling technique; but the variance is greater than the
mean if the mean is greater than 15. Their overall conclusion was that the
distribution of counts of eggs, following their technique, is not far removed
from the form given by the Poisson distribution.

Hunter and Quenouille (1952) were interested in determining the optimum
number of samples to count for each sheep (where a sample in their case was
one McMaster slide) when egg counts of naturally infected sheep were being
compared. They were concerned with this problem because the accuracy of the
count increases as more eggs are counted when dilution counting techniques
are used. A related problem of interest concerned the optimum time lapse
before repeating the whole samplying procedure again. In order to inves-
tigate these two problems they first investigated the distribution of egg
counts both between and within sheep. They found these to be the negative
binomial distribution and the Poisson distribution, respectively. The
portion of their paper dealing with the negative binomial distribution will
be delayed for discussion until later.

The sheep used were ewes and gimmers (yearling female sheep) from
various flocks in various parts of Scotland. Faecal samples were taken from
the rectum of these sheep and were counted using the McMaster slide. The
authors made 9 series of counts in all, where each series consisted of ob-
taining several counts on each sheep.

The first series consisted of 4 counts on each of 132 sheep. The chi-
square statistic was calculated using equation (10) on each group of four
counts and these values were added together yielding a total chi-square of

415.8 with 3x132 = 396 degrees of freedom. P was greater than 0.2 resulting in the conclusion that chance variation could result in this outcome when the Poisson distribution was assumed to be applicable. The other 8 series of counts yielded various chi-square values with their appropriate degrees of freedom. The total chi-square over all 9 series was 1,322.0 with 1,339 degrees of freedom, indicating good agreement with the expected value (P > 0.50) and very good agreement with the expected values in each series of counts.

Brambell (1963) considers the extent of variation of egg counts taken from the same sample, and taken from several samples drawn from the same sheep whose host-parasite physiological relationship is unchanged (i.e., given a stable host-parasite relationship, the effects of the time of day, amount of faeces passed, and water content of the faeces, can be investigated with less error). In order to gain insight into these two problems, Brambell makes observations of egg counts obtained using the McMaster slide technique on housed sheep under experimental conditions that are infected with Haemonchus contortus.

He used two groups of sheep. Group I consisted of seven sheep infected with Haemonchus contortus, and group II contained 4 sheep reared indoors under worm-free conditions. Sixteen faecal samples were collected from each sheep in group I, where each sample consisted of four chamber counts. Equation (10) was used to compute a χ^2 value for each example, each with three degrees of freedom. A chi-square goodness-of-fit test was then performed using a slightly different technique than previous workers cited in this report had used.

Brambell's technique involves comparing the per cent of sample chi-square values falling within a given range of χ^2 values in the chi-square table against the expected frequency of χ^2 values falling within these intervals. The range of tabular χ^2 values has a known expected frequency.

Therefore, for each range of values of χ^2 both the expected and observed frequencies of occurrence were available so that Brambell could calculate the χ^2 statistic for each interval. A pooled χ^2 value was calculated by adding together each χ^2 statistic computed for each interval resulting in a total chi-square value of 5.63 with 6 degrees of freedom (since he had 6 intervals). P was close to 0.30 which supported the hypothesis of Poisson distributed counts. Accordingly, Brambell concludes that the hypothesis of Peters (1941) that McMaster slide egg counts follow the Poisson distribution is confirmed.

Brambell (1963) considers his technique as outlined above to be superior to methods used that compare only one estimate of variance in a series of counts with the mean of the series. He bases his argument on the fact that other distributions than the Poisson have the characteristic that at certain values of the mean, the variance approximates to the mean. He reasons that to compare only one estimate of variance in a series of counts with the mean of the series is thus not sufficient to distinguish the Poisson series.

Of the four sheep in group II (labeled A, B, C, and D) sheep A, B, and C, had 120 samples counted (2 chamber counts per sample), and sheep D had 108 two chamber counts. The chi-square statistic was calculated on each sheep using the same procedure as was used on group I sheep. Of the four series, two, (A and C), deviated from the Poisson distribution with chi-square values of 23.38 and 27.02 respectively, with 6 degrees of freedom. In the series from sheep A, 13% of the counts showed abnormally high variability. With sheep C, the mean was so low and the number of chambers counted so few that the chi-square test became unreliable.

Many years earlier Peters (1941) hypothesized that a discrepancy such as occurred in sheep A could be explained by personal error on the part of the counter if the mean count is in the neighborhood of 50. The easiest way to overcome the difficulties in the series associated with sheep C is to count more chambers.

Brambell (1963) gives a table derived from the Poisson distribution as an aid in the determination of the range of populations from which an egg count could have been drawn. For counts from 0 to 170 the table gives the range of means of populations from which the given counts could be drawn more often than once in twenty times, assuming the counts are from a Poisson distribution.

Emik (1947) performed a chi-square test on 240 egg counts from 12 heterogeneous sheep in a manner similar to Brambell, but instead of using a range of chi-square values for each interval, he used probability limits in the χ^2 tables with one degree of freedom under the assumption of Poisson distributed counts. He then counted the number of observed χ^2 values that fell in each interval of P. The expected number of χ^2 values falling in each P interval was calculated directly from the probability intervals themselves. The χ^2 calculated was equal to 6.250 with 7 degrees of freedom, yielding a P of about 0.5. This size P resulted in acceptance of the hypothesis that the counts are Poisson distributed.

THE NEGATIVE BINOMIAL DISTRIBUTION

Derivation

The negative binomial distribution is one of several distributions proposed to describe a series of counts in which the variance is significantly larger than the mean. The negative binomial,

$$(q - p)^{-k}$$

where $p = \frac{m}{k}$, and $q = 1 + p$, (where m and k are defined below) is so called because of its analogy to the positive binomial $(q + p)^n$.

The first known derivation and publication of the distribution was due to Montomort in 1714. Pascal and Fermat are also recognized as having discussed the distribution (Bartko, 1961). Student (1907) obtained the negative binomial when he observed, while deriving the Poisson series from the binomial, that two of his series gave negative values for p and n yet fitted the data very well.

The distribution is completely specified by two parameters. The first is the mean, namely m. The second is an index of over-dispersion, denoted by k. The nature of k can be better understood if it is recalled that in the Poisson distribution the mean is equal to the variance. In the negative binomial distribution, however, the variance is given by $m + \frac{m^2}{k}$. Note that as k becomes large, the second term in the variance equation tends to zero, i.e., the variance will approach the mean in value. Thus the negative binomial distribution with parameters m and k, becomes very much like the Poisson with parameter m. As k becomes very small, the variance becomes very large, a property called "over-dispersion" (Northam and Rocha, 1958).

Wilks (1943) formally proved that the negative binomial approaches the Poisson as $k \rightarrow \infty$. In addition, it is pointed out by Bliss and Fisher (1953) that as $k \rightarrow \infty$ and the number of units containing no individuals is disregarded, the distribution approaches Fisher's logarithmic series (Williams, 1947).

If $(q - p)^{-k}$ is expanded, the probability P_x that an observational unit will contain $x = 0, 1, 2, \ldots$ individuals is

$$P_x = \frac{(k + x - 1)!}{x! \, (k-1)!} \, \frac{R^x}{q^k}$$

where $R = p/q = m/(k + m)$. Therefore

$$P_x = \frac{(k + x - 1)!}{x! \, (k-1)!} \, (\frac{m}{k+m})^x \, q^{-k}$$

$$= \frac{(k + x - 1)!}{x! \, (k-1)!} \, p^x \, q^{-k-x} \tag{17}$$

where k need not be an integer. To find the expected frequency of units with x individuals simply multiply equation (17) by N, the total number of counts.

If $1/q$ is taken to be the probability of a "success" and p/q as the probability of a "failure" in a trial, then equation (17) can also be interpreted to be the probability that $x + k$ trials will be required to obtain k successes (Wilks, 1943).

Equation (17) is not the only form of the negative binomial in use. Bartko (1961) lists the following two forms in addition to the form given in equation (17):

$$p(x) = \binom{x + r - 1}{x} p^r q^x, \qquad x = 0,1,2,\ldots \tag{18}$$

where r and p are parameters, $p + q = 1$, and r is an integer; and

$$p(x) = \binom{n-1}{r-1} p^r q^x, \qquad\qquad x = 0,1,2,\ldots \qquad (19)$$

where r and p are parameters with $n = r + x$. Wilks (1962) gives still another form:

$$p(x) = \binom{x-1}{k-1} p^k q^{x-k}, \qquad\qquad x = k,k+1,\ldots,$$

where x is a random variable denoting the number of trials performed in order to obtain exactly k "successes". Although these alternative forms have many useful applications, the negative binomial in the form of equation (17) will be used in this report due to its frequent use in egg count studies.

Wilks (1943) shows that the mean and variance of the negative binomial are kp and kpq respectively, which Fisher (1941) points out are identical with the first and second moments of the positive binomial except that k corresponds to $-n$ and $q = 1 + p$, i.e., the sign of p is changed. Wilks (1943) also formally proves that the negative binomial is an extension of the Poisson series as the variance of the negative binomial approaches the mean (as was noted above).

Models

The negative binomial distribution can arise from a variety of bio-logical situations. In fact it has generally been held that in some cases, one can start from two or more mutually incompatable biological hypotheses and arrive at the same over-dispersed distribution using deductive reasoning. As an example of this, Bliss (1958) states that one could assume that unit areas are unequally exposed to infestation and individuals completely in-dependent of each other, or that contagion was present and the initial in-festation uniform. With the proper definitions, both hypotheses could lead to the same over-dispersed distribution.

Anscombe (1950) and Bliss and Fisher (1953) discuss the several ways in which the negative binomial may arise. Only two of these will be presented here due to their applicability to egg count studies. The reader interested in the other models is referred to the above authors.

When the mean, m, of a Poisson distribution is not constant from trial to trial or sample to sample, then the counts may be a mixture of several homogeneous Poisson distributions. In such a mixture of Poissons, the mean represents a positive continuous variate. If m is distributed according to the Eulerian distribution (also known as the Pearson Type III distribution) then the counts conform to the negative binomial distribution. Also, if the mean degree of infestation in different sampling units follows the log-arithmic distribution the negative binomial is known to arise (Hunter and Quenouille, 1952).

<div align="center">Estimation of Parameters</div>

The parameters m and k are estimated from the frequency distribution of a sample by the statistics $\hat{m}$ and $\hat{k}$. The mean number of eggs per count is estimated efficiently from the frequency of counts, f, at each level of x, the number of eggs per count. That is,

$$\bar{x} = \hat{m} = \frac{1}{N} \sum_{x=0}^{n} xf_x, \qquad x = 0,1,2,\ldots,n \qquad (20)$$

where

N = total number of counts taken

and

n = the largest number of eggs observed for any count.

The statistic $\hat{k}$ is not as easily estimated however, there being at least three techniques commonly used; the method of moments, the method based on the ratio of the total number of counts to the number of counts with no eggs present, and the method of maximum likelihood.

The Method of Moments. The simplest and oldest method of estimation is based on the first and second moments, m and s^2 respectively. Since

$$s^2 = kpq$$

$$= m + \frac{m^2}{k} , \tag{21}$$

if equation (21) is solved for k we have the result:

$$k_1 = \frac{m^2}{s^2 - m} = \frac{m}{p} \tag{22}$$

where

$$s^2 = \frac{1}{N(N-1)} [N \sum_{x=0}^{n} (xf_x)^2 - (\sum_{x=0}^{n} xf_x)^2]. \tag{23}$$

In practice, $\hat{m}$ is calculated using equation (20) and replaces m in equation (22) to give $\hat{k}_1$, which is the estimate of k_1 derived by the method of moments, i.e., in this instance, by equating the variance of the sample to the variance of the distribution (Anscombe, 1949).

Fisher (1941) showed that the estimate of p by the method of moments is given by the equation

$$\hat{p} = \frac{s^2 - m}{m} . \tag{24}$$

This result is easily seen by replacing k by $\frac{m}{p}$ in equation (22) and solving for p.

R. A. Fisher (1941) also derived the equation of $\frac{1}{E}$, where E is the efficiency of calculating k by the method of moments compared with k calculated using the fully efficient maximum likelihood technique. Anscombe (1950) plots this efficiency for various values of m and k. In general, the method of moments has an efficiency of 90% or more for small values of m when k/m > 6, for large values of m when k > 13, and for intermediate values of m when $(k + m)(k + 2)/m \geq 15$ (Bliss and Fisher 1953). Fisher (1941) states that if p is less than 1/9 for any value of k, or if k exceeds 18 for any value of p, then high efficiency is assurred. If the efficiency in any particular situation turns out to be low, then a more exact fitting may be acquired by the maximum likelihood method which is presented following the ratio method.

The Ratio Method. To estimate k from the ratio of the total number of observational units, (N), to the number of units with no eggs present, (f_o), it is necessary to note from equation (17) that the probability for x = 0 is $P_o = 1/q^k$. Replacing P_o with the proportion of empty units to total units, it is seen that

$$P_o = f_o/N = 1/q^k = 1/(1 + m/k)^k.$$

Since

$$\log(f_o/N) = \log 1 - \hat{k}_2 \log(1 + m/\hat{k}_2)$$

or

$$\hat{k}_2 \log(1 + m/\hat{k}_2) = -\log f_o + \log N$$

or

$$\hat{k}_2 \log(1 + m/\hat{k}_2) = \log(N/f_o), \tag{25}$$

the value of $\hat{k}_2$ desired is that which balances equation (25). The procedure used is to calculate the left side of equation (25) twice, using two different values of k whose large and smaller values (denoted by k_1' and k_2', respectively) give a product both larger and smaller than the right side of equation (25), which is calculated from the sample values and is therefore a constant. The first approximation of $\hat{k}$ (denoted by k_3') is obtained by interpolation between these two products. The left side of equation (25) is then calculated using k_3', and interpolation between k_1' and k_3' or k_2' and k_3' is executed depending on whether the new product (using k_3') is larger or smaller than the right side of equation (25). This final interpolation gives the desired estimate of k.

If k is to be estimated with an efficiency of 90% or more, at least 1/3 of the observation units must be empty. If m is less than 10, enough empty units must exist such that

$(m + 0.17)(P_o - 0.32) > 0.20$ (Bliss and Fisher 1953).

The Method of Maximum Likelihood. It sometimes happens that k cannot be efficiently estimated by the above techniques. If this is the case, the method of maximum likelihood may be used, resulting in a fully efficient estimate of k.

Haldane (1941) derives the following maximum likelihood equation:

$$N[\log(k+m)-\log k] = \frac{f_1+f_2+\ldots+f_n}{k} + \frac{f_2+f_3+\ldots+f_n}{k+1} + \ldots + \frac{f_n}{k+n-1} \qquad (26)$$

where

f_i = observed frequency in the ith class,

n = maximum number of eggs counted in a single count,

N = total number of counts, and

m = mean number of eggs per count.

The maximum likelihood estimate of k, is that value of k which satisfies equation (26) (Edgerton, 1953). Haldane (1941) points out that interpolation becomes easier if both sides of equation (26) are multiplied by k, since this causes one side to increase with k while the other side decreases.

Bliss and Fisher (1953) present another maximum likelihood technique for estimating k but their method is fully efficient only if the largest frequency does not exceed 30. The equation

$$Z_i = \sum_{x=0}^{n} [\frac{A_x}{k_i' + x}] - N \ln[1 + \frac{\bar{x}}{k_i'}], \quad (27)$$

where A_x is the number of observations exceeding x, is calculated using trial values of k_i', selected so that they bracket the required estimate $\hat{k}$ such that Z_i = zero. Equation (27) is computed using k_1', which is usually obtained by the method of moments. If Z_1 is positive, the value of k_1' is increased slightly, yielding the value k_2' such that $k_2' > k_1'$. If Z_1 is negative, k_2' is taken as less than k_1'. Then k_2' is used in the calculation of Z_2 and the new value k_3' is obtained by interpolation between k_1' and k_2' for Z = 0. To increase the precision of $\hat{k}$, a Z_4 may be computed that is opposite in sign from Z_3 by selecting k_4' at about the same distance as k_3' beyond a newly in-terpolated k' for Z = 0. Interpolation between k_3' and k_4' gives the final maximum likelihood estimate of $\hat{k}$. Anscombe (1950) presents a good dis-cussion of the preceding estimation techniques and Bliss and Fisher (1953) give examples of all three methods.

The Transformation Method. An alternative method of estimating k given by Anscombe (1949 and 1950) is to guess a value of k and then make a transformation (depending on k) of the counts to a new variable y whose variance depends only on k. An improved estimate of k is then obtained by calculating the sample variance of y and equating it to the expected variance. The process is then repeated if the new value of $\hat{k}$ is much different than the old one. Anscombe (1949) suggests using the transformation

$$y = \log_{10}[r + \frac{k}{2}] \tag{28}$$

where r is an actual count, if m $\geq$ 15 and if 2 $\leq$ k $\leq$ 5. If 2 > k > 5, then equation (28) may still be used but only if m is sufficiently large. Under these conditions, the expected variance of y is approximately independent of m and is equal to $0.1886\psi'(k)$, where $\psi'(k)$ is the second derivative of $\ln\Gamma(k)$ with respect to k.

The equation

$$y = \sinh^{-1}\sqrt{\frac{r + c}{k - 2c}} \tag{29}$$

may be used if k $\geq$ 2. The constant c is equal to 0.375 if k is large and 0.2 when k = 2. The expected variance of y using this transformation is $0.25\psi'(k)$. The mean, m, may be as small as 4 or 5 (Anscombe, 1949).

One important characteristic of the negative binomial must be noted. Since the distribution may be very skewed, confidence limits and standard errors should be calculated initially in terms of 1/k, which has a relatively symmetrical distribution.

<div align="center">Variance of $\bar{x}$ and $\hat{k}$</div>

The variance of $\bar{x}$ is

$$V(\bar{x}) = \frac{1}{N}[m + \frac{m^2}{k}] \tag{30}$$

where m and k are necessarily replaced by $\bar{x}$ and $\hat{k}$.

The variance of $\hat{k}$ depends on the method by which $\hat{k}$ is estimated. If equation (22) is used to estimate k, its large sample variance is

$$V(\hat{k}_1) \doteq \frac{2\hat{k}[\hat{k}+1]}{NR^2} \tag{31}$$

where

$$R = \frac{\bar{x}}{\hat{k} + \bar{x}} \quad . \tag{32}$$

If $\hat{k}$ is calculated using equation (25) then the appropriate large sample variance is given by

$$V(\hat{k}_2) \doteq \frac{[1-R]^{-k} - 1 - \hat{k}R}{N[-\ln_e (1-R) - R]^2} \tag{33}$$

where R is as defined in equation (32). The variance of $\hat{k}$, where $\hat{k}$ is calculated using the method of maximum likelihood (equation (27)), is given by the ratio

$$V(\hat{k}) = \frac{k_3' - k_4'}{Z_4 - Z_3} \tag{34}$$

where Z_3 and Z_4 are the two values of Z_i just above and below zero and calculated using k_3' and k_4'.

Anscombe (1950) derives the large sample variance of $\hat{k}$ calculated
using the transformation method. The variance of $\hat{k}$ is

$$V(\hat{k}) \doteq \frac{\psi'''(k) + 2[\psi'(k)]^2}{m\left[\psi'''(k) + \dfrac{k}{(k+1)^2 n}\right]^2}$$ (35)

where $\psi(k)$, $\psi'(k)$... denote the successive derivatives of $\ln\Gamma(k)$. Equation
(35) is accurate for practical purposes according to Anscombe if $m > 50$
and $k > 5$ and assuming the appropriate hyperbolic sine transformation is used.

Goodness of Fit Tests

The chi-square test is used to test the adequacy of the negative bi-
nomial distribution in fitting a series of counts much the same way as the
Poisson distribution is tested for goodness of fit.

Test 1. Bliss and Fisher (1953) presented the following procedure:

1). Compute the expected frequencies using equation (17). Start with
the number expected at $x = 0$, which is $\phi_o = N/q^k$.

2). Find the expected frequencies for $x = 1, 2, 3, \ldots$ by using the
relationship

$$\phi_x = \frac{[k+x-1]}{x} R \cdot \phi_{x-1} .$$ (36)

3). Avoid accumulating rounding errors by retaining more decimals
in the calculator than necessary.

4). Pool the frequencies with small expectations so that no expectation
is less than 5.

5). Compute

$$\chi^2 = \sum_{x=0}^{n} \frac{(f_x - \phi_x)^2}{\phi_x} \tag{37}$$

where f_x is the observed frequency for each x, and the degrees of freedom are (n - 3), i.e., three less than the number of ratios summed. If χ^2 is small and m and k are efficient estimates, probably no other test is needed.

Other tests have been proposed by Anscombe (1950) and have been discussed and illustrated by Bliss and Fisher (1953). Two tests in particular are described that are applicable when N is large, each being based on the difference between the observed and the expected second and third moments of the negative binomial. These tests are not influenced by chance irregularities in the observed frequencies and furthermore, the few large values in the tails of the observed distribution are not ignored.

Test 2. Using the method of moments (equation (22)), estimate the first two moments of the sample from a negative binomial distribution and find T, the difference between the third sample moment and the expected third sample moment predicted from the first two moments of the same sample. That is

$$T = \frac{(x^3)}{N} - pq[q + p]m$$

$$= \frac{(x^3)}{N} - s^2[\frac{2s^2}{\hat{m}} - 1]$$

$$= \frac{1}{N}\left[\sum_{x=0}^{n} [f_x(x - \hat{m})^3]\right] - s^2[\frac{2s^2}{\hat{m}} - 1]$$

$$= \frac{1}{N}\left[\sum_{x=0}^{n} [f_x x^3] - 3\hat{m}\sum_{x=0}^{n} [f_x x^2] + 2\hat{m}^2 \sum_{x=0}^{n} [f_x x]\right] - s^2[\frac{2s^2}{\hat{m}} - 1], \tag{38}$$

where $\hat{m}$ and s^2 are calculated using equations (20) and (23) respectively. Whether there is a significant difference between the observed and expected third moments is determined by comparing T with its standard error, the square root of its large sample variance, V(T), where

$$V(T) = \frac{1}{N}[2\hat{m}\ (\hat{k} + 1)\hat{p}^2\hat{q}^2 \left\{ 2(3 + 5\hat{p}) + 3\hat{k}\hat{q} \right\}], \qquad (39)$$

where $\hat{p} = \hat{m}/\hat{k}$, $\hat{q} = 1 + \hat{p}$, and $\hat{k}$ is the maximum likelihood estimate if available (Anscombe (1950) and Bliss (1953)).

Test 3. Compute the observed second moment and subtract from it its expectation. That is, calculate U, where

$$U = s^2 - (\hat{m} + \frac{\hat{m}^2}{\hat{k}_2}) \qquad (40)$$

where $\hat{k}_2$ is calculated using the ratio method (equation (25)). U has the large sample variance

$$V(U) = 2\hat{m}[k+1]\ \hat{p}\hat{q}^2 \left[1 - \frac{R^2}{-\ln[1-R]-R} \right] \Big/ [N + \hat{p}^4 V(\hat{k}_2)] \qquad (41)$$

where $V(\hat{k}_2)$ is defined by equation (33) but computed with the maximum likelihood estimate of $\hat{k}$ if it is available, as are the other terms in equation (41).

It should be noted that the expected moments in tests 2 and 3 are computed using other than maximum likelihood estimates. Bliss and Fisher (1953) state that this is the procedure because V(T) and V(U) are derived assuming T and U are calculated with a $\hat{k}$ estimated using some method other than maximum likelihood. Thus V(T) and V(U) are of doubtful applicability unless the procedure given above is followed.

Transformations

There have been two transformations proposed for stabilizing the variance of negative binomial distributions. One of these is the logarithmic transformation given by

$$y = \log \left[m + \frac{k}{2} \right].\tag{42}$$

Beall (1954) suggests using the more sophisticated transformation given by

$$y = \sqrt{k} \, \sinh^{-1} \sqrt{\frac{x}{k}} \ ,\tag{43}$$

which has been tabled for various values of x and 1/k ranging from 0 to 1. In both of these transformations, k is actually a common k as defined below if a common k is available (Northma and Rocha, 1958; Bliss and Owen, 1958). These transformations, in addition to making the variance independent of the mean, also tend to make the transformed scores normal in distribution, and the real effects additive. An analysis of variance can be computed on the transformed counts and F tests performed since the assumptions underlying such tests are more nearly met.

Calculation of a Common k

Several authors, for example, Anscombe (1949 and 1950), Bliss and Fisher (1953), Bliss and Owen (1953), and Bliss (1958), have discussed the calculation of a common k among several negative binomial distributions. Bliss and Owen (1958) state that since observed counts are often compared in terms of their means, a stable k would simplify such comparisons materially. Bliss and Fisher (1953) see an advantage in a common k in that

it increases both the utility of the negative binomial and the confidence one has in its suitability for a given problem.

Anscombe (1949) gives the following three methods for estimating a common k, where methods 1, 2, and 3, are based on equations (22), (25), and the transformation method for estimating k, respectively.

1). Guess a value of k and calculate T for each set of counts, i.e., for each negative binomial distribution, where

$$T = \frac{[N-1]s^2 - [N-1-\frac{1}{k}]m[1 + \frac{\hat{m}}{k}]}{[\hat{m} + k]^2} , \qquad (44)$$

where N is the number of observation units counted, and s^2 and $\hat{m}$ are as defined in equations (20) and (23) respectively. The common k is that value of k for which the sum of the expressions T, over all sets of counts, is zero. N should be at least 10.

2). Guess a value of k and calculate U for each set of counts, where

$$U = \log[1 + \frac{\hat{m}}{k}] \left[n_o - (1 + \frac{\hat{m}}{k})^{-k} \left[N - \frac{\hat{m}[k+1]}{2[\hat{m} + k]} \right] \right] , \qquad (45)$$

where n_o is the number of observation units that are empty. As in the first method, the object is to use a value of k such that the sum of all U over all series of counts is zero. Again, N is assumed to be at least 10. For details on the derivation of equation (44) and (45) see Anscombe (1950).

3). Calculate the variance of the transformed variate y (where y is defined by equation (28) or (29)) for each set of counts, pool the answers and equate to the theoretical variance. N may be as small as two in this case. The restrictions on m and k for an appropriate transformation to exist must be observed as mentioned following equations (28) and (29).

Bliss and Fisher (1953), by an expansion of the maximum likelihood method of estimating k, present a method of calculating a common k (which he denotes by k_c) from several series of distributions. Equation (27) is used to compute the score Z for each component distribution with the same trial values of k'. These values of Z are added over all component distributions for each k' trial value resulting in sums S(Z). Different trial values of k' are used until two of the sums are obtained which bracket zero very closely. By interpolation between these two sums, the required k_c is that value of k' for which S(Z) equals zero. Denoting these two sums by $S(Z_3)$ and $S(Z_4)$ from corresponding values of k_3' and k_4', the error variance of $\hat{k}_c$ can be calculated using equation (34) if Z_3 and Z_4 of that equation are replaced by $S(Z_3)$ and $S(Z_4)$.

Bliss and Fisher (1953) also show how to test for homogeneity of k values over the component distributions by use of χ^2. That is, do the values of k from the component distributions differ significantly. If they do not then a common k can be derived. For each negative binomial distribution Z_3 and Z_4 are calculated using k_3' and k_4'. Then for each distribution, the ratio

$$\frac{Z_3^2 \ [k_4' - k_3']}{Z_3 - Z_4} \tag{46}$$

is computed. From the sum of all of these ratios over all distributions for which the ratio was computed is subtracted the ratio (equation (46)) calculated using the sums $S(Z_3)$ and $S(Z_4)$ in place of Z_3 and Z_4 for k_3' and k_4' respectively. The difference is distributed as a chi-square with (g-1) degrees of freedom, where g is the number of distributions being considered. If the difference between the k's is not significant, then a

common k may be fitted and any differences between the component negative binomial distributions may be discovered in investigating only the means of the distributions.

Bliss and Owen (1958) developed another method of calculating a common k which is an extension of Anscombes' weighted moment estimate in terms of regression and small series. For each negative binomial distribution, the quantities x' and y' are computed, where

$$x' = \bar{x}^2 - \frac{s^2}{N} \tag{47}$$

and

$$y' = s^2 - m . \tag{48}$$

If y' is plotted against x', where x' is on the abscissa, the line fitting these points and passing through the origin has the slope $b = 1/k_c$. Since the increase in the variability of y' is often roughly proportional to the increase in x', a first estimate of $1/k_c$ is

$$\frac{1}{k_c} = \frac{\sum y'}{\sum x'} , \tag{49}$$

where the summation extends over all component distributions.

It is possible to determine whether a common k holds over all distributions by calculating

$$\frac{1}{k_1} = \frac{v'}{x'} \tag{50}$$

for each distribution and plotting it against its mean, $\bar{x}$. If $1/k_1$ does not consistently increase or decrease with $\bar{x}$, and if there is no distinct clustering of points, a common k may be fitted.

A more accurate method for fitting k_c has been developed (Bliss and Owen, 1958). For each negative binomial distribution in the set of distributions for which a common k is to be fitted, calculate the following quantities: $\bar{x}$, s^2, x', y', y'/x', $(m + k')^2$, wx', and $wx'y'$, where

$\bar{x}$ = observed mean given by equation (20),

s^2 = observed variance given by equation (23),

x' = a statistic given by equation (47),

y' = a statistic given by equation (48),

$k' = \sum x'/\sum y'$,   (51)

         where the summation extends over all distributions,

and

$$w = \frac{0.5[N - 1]k'^4}{k'[k'+1] - [2k'-1]/N - 3/N^2} \cdot \frac{1}{x'[\hat{m} + k']^2} \qquad (52)$$

$$= \frac{A}{x'[\hat{m}+k']^2} \quad ,$$

where A is a constant for each distribution if N is constant from distribution to distribution. The quantity k_c is then given by

$$k_c = \frac{\sum [wx'^2]}{\sum [wx'y']} \qquad (53)$$

where the summation is over all distributions. If k_c differs appreciably from k', the values of $(\hat{m} + k')^2$ and wx' are recalculated using k_c instead of k'. Equation (51) is a good first estimate of k_c if the quantities wx' are relatively stable over all distributions.

Instead of calculating w for each distribution, the product wx' may be obtained by using the equation

$$wx' = \frac{A}{(m+k')^2} \qquad . \tag{54}$$

The weight w is derived by noting that $y' - x'/k$ has an expectation of zero (since $E(x') = m^2$, and $E(y') = m^2/k$). The reciprocal of the variance of this difference is the weight w (Bliss and Owen, 1958).

The χ^2 test may be used here also to test whether a series of negative binomials distributions has a common k. The procedure is to calculate the quantity wy' for each distribution. This is most easily done by first computing y'/x' and computing

$$(wx')(y'/x') = wy' \qquad . \tag{55}$$

The products of $(y')(wy')$ are summed over all distributions to obtain $\sum [wy'^2]$. The quantity

$$\sum [wy'^2] - [\sum (wx'y')]^2 \bigg/ \sum (wx'^2) \tag{56}$$

is distributed approximately as a χ^2 with $(g - 2)$ degrees of freedom, where g is the number of distributions. An additional degree of freedom is lost due to the second term of expression (56) being the slope of the line with a zero intercept. It is important that k_c and k' agree closely before column wy' is calculated because this chi-square test is sensitive to any discrepancy between the two.

Bliss and Owen (1958) also describe a method of calculating a common k that is useful when a k_c is necessary in transforming negative binomial counts so that an analysis of variance may be computed.

Egg Count Studies

There have not been many papers written describing egg count studies where the counts follow the negative binomial distribution. Most egg count studies are concerned with replicate counts from the same animal and these counts are usually Poisson distributed. Two papers will be discussed here that involve the negative binomial distribution.

Hunter and Quenouille (1952) took egg counts on two different breeds of sheep from three counties during the months of June, January, and July in Scotland, and found some sheep tend to give much higher counts than would arise in a random distribution, possibly due to different levels of resistence to the parasites. They believe that the negative binomial might be appropriate because the following two points were reasonable assumptions to make in light of the egg counts they observed:

1). Counts from the same sheep follow a Poisson distribution.

2). The quantity m in different animals follows the logarithmic distribution.

Hunter and Quenouille (1952) fitted 13 series of observations to the negative binomial by computing the expected number of counts for $x = 0,1,2,\ldots$ and comparing these with the observed frequencies. The number of sheep in these series ranged from 49 to 90. A χ^2 test was computed for each series, yielding a total χ^2 value over all 13 series of 38.35 with 34 degrees of freedom resulting in a $P > 0.25$. They conclude therefore that the negative binomial is applicable to these counts.

The authors point out that a knowledge of p and k provides a convenient summary of any series of counts. They state that p usually is more

sensitive to changes in the mean count than is k, and that p depends upon the dilution used in counting. They calculated p and k and their standard errors for each of the 13 series and found that p varied substantially (from 18.3 ± 4.19 to 0.89 ± 0.40) due chiefly to different levels of egg output, and k remained fairly uniform, varying from 1.45 ± 0.61 to 0.42 ± 0.12. The authors mention that the use of k allows the comparison of distributions of egg counts under differing conditions and at different times to determine how uniformly the worms are distributed under these varying conditions.

The relative efficiency of different size samples in terms of the variance is also investigated. The equation

$$\frac{kp[1 + p]}{kp[\frac{1}{r} + p]} \times 100\% = \frac{1 + p}{\frac{1}{r} + p} \times 100\% \qquad (57)$$

is derived for this purpose, where $kp(1 + p)$ is the variance of a negative binomial distribution for a particular sample size, and $kp(1/r + p)$ is the variance per unit sample of a sample r times as large. Thus, once p is known for a particular sample size, the relative efficiency of various multiples of this sample size may be computed.

Hunter and Quenouille (1952) also use the expression

$$\frac{100\%}{1+\frac{k}{m}} \qquad (58)$$

to calculate the relative efficiency in terms of the possible efficiency obtainable with a sample of unlimited size. By using various values of m and k in equation (58) they conclude that for large k (over 0.6) it is advisable to take a sample large enough to make the mean number of slides counted

per sheep about four. For values of k between 0.4 and 0.7 a mean of 2 slides is sufficient.

The problem of how much time should lapse between repetition of the whole samplying procedure is also discussed. The percentage gain in information from taking a second set of samples n days after the first is

$$100 \times \frac{1 + \frac{k}{m} - \rho_n^{\ n}}{1 + \frac{k}{m} + \rho_n} \tag{59}$$

where ρ_n is the correlation between the infestations n days apart. After examination of this expression for various values of k/m and ρ_n, Hunter and Quenouille conclude that weekly sampling is adequate for most practical purposes.

These two authors conclude by pointing out that a comparison of egg counts between two flocks of sheep should take into account the parameter k, since the dispersion of eggs throughout the population is just as important as the number of parasites. The question of whether a high or low k is most beneficial to the parasite is unknown. In more practical terms, this could be stated as Hunter and Quenouille do; "In other words whether eggs dropped on pasture would survive better if isolated in a few large groups (low k) than in more evenly dispersed smaller groups (high k)", is unknown.

Northam and Rocha (1958) did a statistical study on the worm burden of chickens and showed that the distribution of the worm counts followed the negative binomial law. They present 4 large frequency distributions of counts of _Ascarida galli_ in experimentally infected chickens. They assumed that

1). the frequency distribution of the number of worms per bird follows the Poisson series, provided the chickens have the same genetic resistence and that external factors are controlled, and

2). the mean number of worms per bird follows the Gamma distribution given that the genetic resistence varies from one chicken to another. This model was discussed earlier in this report as one that led to the negative binomial. Northam and Rocha (1958) show that assumption 1 holds in their case by citing three series of worm counts per bird where each series was subdivided into 3 parts; male, female, and combined worm counts. The chickens over all 3 series were taken from the same inbred line in an attempt to control genetic resistence variability. Each distribution for male, female, and combined, was shown to follow the Poisson law. They stress the importance of controlling external factors in arriving at a Poisson distribution by citing another experiment similar to the one above, except that several illnesses occurred, the result being over-dispersion and non-Poisson distributions of worm counts. The authors do not test assumption two directly.

For each of the 4 distributions, they calculate the mean, standard deviation, standard error of the mean, and 3 estimates of k by the method of maximum likelihood, method of moments, and the proportion of birds with no worms to the total number of birds investigated. They note that the 4 distributions do not have a common value of k. A χ^2 test was performed on each distribution with resulting values of p ranging from 0.10 to 0.25 thus indicating the negative binomial distribution adequately described the distributions.

The fact is noted that with the values of k and m they calculated, the relative efficiency of the method of moments as compared with the maximum likelihood approach is around 50%, while the ratio method is somewhat better, giving a relative efficiency between 75% and 90%. Northam and Rocha (1958) continue on and apply these statistical results to an experiment by Rocha in 1955 using the drug Phenothiazine.

CONCLUSIONS

The preceding discussion has centered around the Poisson and the negative binomial distributions with a consideration of many of their properties such as derivations, parameter estimation procedures, transformations, and applicability to egg and worm count studies.

It was shown that replicate worm and egg counts obtained by the use of a dilution technique (usually the McMaster slide) are distributed according to the Poisson distribution. This distribution was not too difficult to work with due to the equality of the mean and variance, and the property of convergence to the normal distribution with large m.

The negative binomial distribution is not as simple to work with as the Poisson. This is due to the fact that two parameters must be specified for each negative binomial distribution, namely m and k, where k is a measure of over-dispersion, i.e., a measure of how much variability is present byyond what we could expect in a Poisson series.

Perhaps two of the more important techniques presented under the negative binomial section were the calculation of a common k and the calculation of test statistics used to determine if the common k calculated is valid over all component negative binomial distributions. In sheep studies for example, an experimenter may be interested in determining whether two negative binomial distributions fitted to different egg counts are really significantly different from each other. A comparison of means between the two distributions in an attempt to answer this question is much more meaningful if it can be shown that a common k can be calculated.

As a concluding remark, it should be noted that the negative binomial is not the only over-dispersed distribution, although it is the easiest to

compute and the most widely applicable (Anscombe, 1950; Bliss and Fisher, 1953). Three other distributions in particular are Neyman's Contagious type A, the Thomas double Poisson, and the Polya (Bliss and Fisher, 1953; Feller, 1943). Each of these has been based on a mathematical model of biological interest.

Several distributions such as the Poisson Pascal (Katti and Gurland, 1961), the Conditional Poisson (Cohen Jr., 1960), and the truncated negative binomial (Bartko, 1961; Sampford, 1955) are also discussed in the literature, but are of only secondary interest to the topic of this report.

ACKNOWLEDGMENTS

I am grateful to be able to acknowledge the help and advise of Dr. Gary F. Krause in both the selection of the topic for this report and in its planning and execution. His contributions go far beyond these however. The friendship he has so freely given and the unending support he has shown this graduate student are most deeply appreciated. I would also like to thank Drs. Holly C. Fryer, Arlin M. Feyerherm, and M. F. Hansen for reading the manuscript and making helpful suggestions.

REFERENCES

Anscombe, F. J. "Sampling Theory of the Negative Binomial and Logarithmic Series Distributions." Biometrika, Dec. 1950, 37:358-382.

_____. "The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution." Biometrics, June 1949, 5:165-173.

Bartko, John J. "The Negative Binomial Distribution: A Review of Properties and Applications." The Virginia Jour. of Sci., January 1961, 12 (3).

Bartlett, M. S. "The Use of Transformations." Biometrics, March 1947, 3 (1): 39-52.

Beall, Geoffrey. "Data in Binomial or Near-Binomial Distribution: With Particular Application to Problems in Entomological Research." Statistics and Mathematics in Biology, Ames, Iowa: Iowa State College Press, 1954.

Bliss, C. I. "The Analysis of Insect Counts as Negative Binomial Distributions." Proceedings Tenth International Congress of Entomology, Volume 2, 1956 (1958).

Bliss, C. I. and Fisher, R. A. "Fitting the Negative Binomial Distribution to Biological Data and Note on the Efficient Fitting of the Negative Binomial." Biometrics, June 1953, 9 (2): 176-196.

Bliss, C. I., and Owen, A. P. "Negative Binomial Distributions With a Common K." Biometrika, June 1958, 45: 37-58.

Blom, Gunnar. "Transformations of the Binomial, Negative Binomial, and Poisson Distributions." Biometrika, December 1954, 41 (3 and 4): 302-316.

Brambell, M. R. "Variation in Counts of Haemonchus contortus Eggs in the Faeces of Housed Sheep." Jour. of Helminthol., 1963, 37: 1-10.

Cohen, A. Clifford, Jr. "An Extension of a Truncated Poisson Distribution." Biometrics, September 1960, 16 (3): 446-450.

_____. "Estimating the Parameter in a Conditional Poisson Distribution." Biometrics, June 1960, 16 (2): 203-211.

Egerton, J. R. "Investigations on the Response of a Host Animal to a Parasitic Nematode Based Upon Passive Transfer of Immunity." Doctor of Philosophy Thesis, Kansas State University, 1953.

Egerton, J. R., and Hansen, M. F. "Immunity and Tolerance of Chickens to the Roundworm, Ascaridia galli (Schrank)." Experimental Parasitology, July 1955, 4 (4): 335-350.

Emik, Otis L., "Statistical Treatment of Counts of Trichostrongylid Eggs." Biometrics, June 1947, 3 (2): 89-93.

Feller, W. "On the General Class of Contagious Distributions." Ann. of Math. Stat., 1943, 14: 389-400.

Fisher, R. A. Statistical Methods for Research Workers. 12th edition. London: Oliver and Boyd, 1954.

_____. "The Negative Binomial Distribution." Ann. of Eugenics, Oct. 1941, 11 (2): 182-187.

Fisher, R. A., Corbet, A. S., and Williams, C. B. "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of Animal Populations." Jour. of Animal Ecology, 1943, 12: 42-58.

Haldane, J. B. S. "The Fitting of Binomial Distributions." Ann. of Eugenics, October 1941, 11 (2): 179-181.

Hunter, G. C., and Quenouille, M. H. "A Statistical Examination of the Worm Egg Count Sampling Technique for Sheep." Jour. of Helminth., 1952, 26 (4): 157-170.

Katti, S. K., and Gurland, John. "The Poisson Pascal Distribution." Biometrics, December 1961, 17 (4): 527-538.

Li, S. Y. and Hsu, H. F. "On the Frequency Distribution of Parasitic Helminths in Their Naturally Infected Hosts." Jour. of Parasit., February 1951, 37 (1): 32-41.

Morgan, D. O., Parnell, I. W., and Rayski, C. "The Seasonal Variations in the Worm Burden of Scottish Hill Sheep." Jour. of Helminth., 1951, 25: 177-212.

Northam, J. I., and Rocha, U. F. "On the Statistical Analysis of Worm Counts in Chickens." Exper. Parasit., July 1958, 7 (4): 428-438.

Ostle, Bernard. Statistics in Research. 2nd. edition. Ames, Iowa: Iowa State Univ. Press, 1963.

Patil, G. P. "On the Evaluation of the Negative Binomial Distribution with Examples." Technometrics, November 1960, 2 (4): 501-505.

Peters, B. G. "Dilution Egg-Counts and Poisson Series." Jour. of Helminth., 1941, 19: 59-62.

Peters, B. G., and Leiper, J. W. G. "Variation in Dilution counts of Helminth Eggs." Jour. of Helminth., 1940, 18: 117-142.

Quenouille, M. H. "A Relation Between the Logarithmic, Poisson, and Negative Binomial Series." Biometrics, June 1949, 5: 161-164.

Sampford, M. R.   "The Truncated Negative Binomial Distribution." _Biometrika_,
     June 1955, 42: 58-69.

Snedecor, George W. _Statistical Methods_. 5th edition. Ames, Iowa: Iowa
     State University Press, 1956.

Spedding, C. R. W. "Variation in the Egg Count of Sheep Faeces Within One
     Day." _Jour. of Helminth._, 1952, 26: 71-86.

_____. "Variation in the Nematode Egg Content of Sheep Faeces from
     Day to Day." _Jour. of Helminth._, 1953, 27: 9-16.

Student. "On the Error of Counting With a Haemacytometer." _Biometrika_,
     Oct. 1907, 5: 351-369.

Waugh, Albert E. _Elements of Statistical Method_. 2nd edition. New York
     and London: McGraw-Hill Book Company, 1943.

Whitaker, L. "On the Poisson Law of Small Numbers." _Biometrika_,
     1914, 10: 36-71.

Whitlock, J. H. "Parasitology, Biometry, and Ecology." _Brit. Veterinary
     Jour._, August 1961, 117: 337-348.

Wilks, Samuel S. _Mathematical Statistics_. New Jersey: Princeton Univ.
     Press, 1943.

_____. _Mathematical Statistics_. New York: John Wiley and Sons,
     1962.

Williams, C. B. "The Logarithmic Series and Its Application to
     Biological Problems." _Jour. of Ecology_, August 1947, 34 (2):
     253-272.

Willmott, S. and Pester, F. R. N. "Variations in Faecal Egg- counts in
     Paramphistome Infections as Determined by a New Technique."
     Jour. of Helminth, 1952, 26: 147-156.

SOME DISTRIBUTIONS USED TO DESCRIBE
HELMINTH EGG COUNTS

by

RICHARD ORRIE GILBERT

B. S., Ferris State College, 1963

————————————

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

The parasitologist is often interested in the frequency distribution of egg counts of internal parasites found in various animals. If such a distribution is known the parasitologist can predict more accurately the number of eggs expected in a future count from the same population. His interest in egg counts stems from his ability to obtain an accurate worm count at necropsy only.

Two statistical distributions have been shown to be of major importance in egg count studies. The first of these is the Poisson distribution which is applicable when a dilution technique in counting is used (such as the McMaster slide method) and when the counts are taken from a homogeneous population, such as repeated sampling from the host.

Various workers such as Peters, Leiper, Emik, Hunter and Quenouille, and Brambell have performed egg count studies on sheep and have found the Poisson distribution to adequately fit the egg counts obtained.

The mean and variance of a Poisson distribution are equal, resulting in only one parameter having to be estimated, namely the mean. In the event that an analysis of variance is going to be calculated using Poisson distributed counts a transformation is usually required. The transformation applied to Poisson distributed data is usually the square root, $\sqrt{x}$, or $\sqrt{x+1/2}$, depending on the size of the count.

The chi-square test may be used to test; (1) does a given series of counts actually follow the Poisson distribution, (2) is the mean of a Poisson greater than a particular value and (3) do several Poisson distributions have the same mean. For two Poisson distributions with large means, the t statistic may be used to test for significant differences between the means of the distributions.

The amount of variation in egg counts taken from several different sheep has been found to be much greater than that taken from one sheep. Due to this greater variability, the Poisson distribution is not adequate to describe such counts. Several distributions have been proposed to fit these counts but the easiest to work with and the most applicable is the negative binomial.

The negative binomial can arise from many different biological circumstances, but egg count distributions seem mainly to arise from two models; one where m is not constant from one Poisson distribution to the next and is, in fact, distributed according to the Pearson type III curve, and the other, where m follows the logarithmic distribution from one Poisson to another.

The negative binomial distribution is specified by two parameters, the mean, m, and an index of over-dispersion, k. The parameter k has been calculated using at least four different techniques; (1) the method of moments, (2) the ratio of the total number of counts to the number of counts with no eggs present, (3) the method of maximum likelihood and (4) the transformation method. The maximum likelihood approach is fully efficient while the other methods are not. Their efficiency can usually be calculated however. The variances of $\bar{x}$ and k can also be found.

The chi-square statistic may be used to test for goodness of fit of the negative binomial. Tests using the difference between the observed and expected second and third moments are applicable for this purpose also.

Two transformations are commonly used to stabilize the variance of a negative binomial distribution. One is the logarithmic transformation described by Anscombe and the other was suggested by Beall and involves $\sinh^{-1}$.

A stable k between several negative binomial distributions is often advantageous (especially in comparing means). Accordingly, several methods have been developed for finding a common k. Anscombe discusses three such

methods based on methods for calculating a single k. Bliss uses an extension
of the maximum likelihood technique, and Bliss and Owen present a method based
on a weighted moment estimate in terms of regression. Various χ^2 tests are
available for determining whether a common k exists between several distri-
butions.

Both egg and worm count studies have produced counts distributed ac-
cording to the negative binomial as shown by Hunter and Quenouille, Egerton,
Egerton and Hansen, and Northam and Rocha.