THE NEGATIVE BINOMIAL DISTRIBUTION: ITS VALIDITY AS A MODEL IN WHITEFISH SAMPLING

A Thesis

Presented to

the Department of

Actuarial Mathematics and Statistics

University of Manitoba

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Clayton Alfred Mills

April 1966

TABLE OF CONTENTS

LIST OF TABLES

LIST OF GRAPHS

CHAPTER I

INTRODUCTION

Sampling and inspection of shipments of Whitefish drawn from Canadian Lakes, prior to shipping to the United States market, is necessary in order to maintain quality in the exported fish, as well as avoid rejection of shipments that reach the United States, with the concomitant wastage and loss to the Canadian Fisherman.

Federal inspection services are made available to Fisheries in Canada for the purpose of detecting shipments with an unusually high cyst, Triaenophorus crassus, content and preventing shipment of these parcels to the United States.

An important part of the sampling inspection operation is the statistical procedure that is followed in arriving at a satisfactory sampling plan: For example, the size of the sample, the equating as nearly as possible of the probability of rejecting a satisfactory parcel with the probability of accepting an unsatisfactory parcel and the underlying theoretical distribution and its related assumptions must be considered in the implementation of a sampling plan.

Recent research on the sampling aspect of this problem was based on the assumption that the distribution of fish with cysts followed a Negative Binomial Distribution (N.B.D.) (Paul-1961; Goldsmith-1963).

The object of the present investigation is:

(1) to examine the suitability of the N.B.D. as a model in the Fisheries sampling work, and

(2) to calculate the size of sample necessary to ensure an acceptable probability (producers' and consumers') when Whitefish shipments are sampled to determine the cyst content.

HISTORY

Miller (1952) said, "Triaenophorus crassus is a tapeworm which, in one of its immature stages, is very common in the flesh of Whitefish in Canadian lakes. Here it appears as a yellowish cyst about one-half inch long, filled with viscous yellow fluid and a long, coiled, thin worm. These cysts, while harmless to man and animal, are objectional in appearance, and when numerous, render the fish unmarketable."

Since infested fish are unmarketable, the Canadian Government had to pass, under the Fish Inspection Act, a set of Whitefish Inspection Regulations (1954) and its amendments (1958) to control the infestation rate in exported fish.

Kennedy (1948) recommended to the Fisheries Department that a sampling plan, consisting of samples which were sufficiently large enough to give valid computations, should be used to determine whether or not a shipment of Whitefish was acceptable.

Oakland (1950) suggested a sequential sampling plan which had the advantage of requiring a smaller sample size on the average, than did the non-sequential plans. However, this plan proved impractical because more than one sample was sometimes required from a shipment and inspecting officers, as well as Fishermen, objected to this procedure.

PARAMETRIC PROCEDURE

The procedure used in calculating the parameters p and k in the Negative Binomial Distribution is the one set forth by Williamson & Bretherton (1963) in the introduction to their Negative Binomial Tables. These tables are used in the fitting of the Whitefish data as well as in the construction of the tables in determining sample size.

INTERPOLATION

Linear interpolation is used because of its simplicity, and because the difference between it and non-linear interpolation is negligible.

This negligible difference is due to the fact that the numbers involved are quite small (< 1) and are negligibly affected by the two means of interpolation.

CHAPTER II

THE NEGATIVE BINOMIAL DISTRIBUTION

The N.B.D., as described by Wilks (1962), is a distribution of a discrete variable x where the variable (the number of cysts) runs from zero to infinity. Since previous workers have described this distribution in detail, only a brief discussion of it is presented here.

From the work of Goldsmith (1963) the N.B.D. has the form $(q - p)^{-k}$ where $p > 0$, $q = 1 + p$ and $k > 0$ is not necessarily an integer. From this expression the general term of the N.B.D. is

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

If one lets x equal the number of cysts ($x = 0, 1, 2, \ldots n$) then the general term of the N.B.D. yields the expected frequencies for x cysts.

FITTING k and p FROM POPULATION MOMENTS

The population variance σ^2, of the N.B.D. is given by $\sigma^2 = kpq$, and the population mean μ by $\mu = kp$.

Thus, it follows that

$$\sigma^2 = \mu + \mu p$$

Therefore

$$\mu p = \sigma^2 - \mu$$

and hence

$$p = \frac{\sigma^2 - \mu}{\mu}$$

Using this form of p, it follows that

$$k = \frac{\mu^2}{\sigma^2 - \mu}$$

Similarly from sample moments, on replacing σ^2 by s^2 and μ by $\bar{x}$

$$p = \frac{s^2 - \bar{x}}{\bar{x}} \qquad k = \frac{\bar{x}^2}{s^2 - \bar{x}}$$

EFFICIENCY OF MOMENT METHOD

The reciprocal of the efficiency (Fisher 1941) is given by

$$E^{-1} = 1+\frac{4}{3}\frac{p}{q(k+2)} + \frac{3p^2}{q^2(k+2)(k+3)} + \ldots\ldots$$

If $p < 1/9$ for any k, or $k > 18$ for any p, then high efficiency is assured. Also the moment method has a high efficiency for small values of μ when $k/\mu > 6$, for large values of μ when $k > 13$, and for μ in the intermediate zone when $\frac{(k+\mu)(k+2)}{\mu} \geq 15$ (Bliss & Fisher 1953)

Williamson & Bretherton (1963) wrote the N.B.D. in the form $p^k(1-q)^{-k}$, with the general term

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

where $0 \leq p \leq 1$, $k > 0$, k not necessarily an integer and $q = 1-p$.

Upon letting x equal the number of cysts ($x = 0, 1, 2, \ldots\ldots n$), the general term of the N.B.D. then yields the expected frequencies for x cysts.

METHOD OF MOMENTS AND CUMULANTS

From the general term of the N.B.D. of Williamson & Bretherton, the moment generating function is

$$M(t) = \sum_{x=0}^{\infty} e^{tx} P(x)$$

$$= \sum_{x=0}^{\infty} \frac{e^{tx}(k+x-1)!p^k q^x}{x!(k-1)!}$$

$$= \sum_{x=0}^{\infty} \frac{(k+x-1)!p^k}{x!(k-1)!} (qe^t)^x$$

$$= p^k(1-qe^t)^{-k}$$

The moments of the distribution, however, are more easily obtained from the cumulant generating function, K(t), which is the logarithm of the moment generating function; i.e.,

$$K(t) = \log M(t)$$

$$= \log \left(p^k(1-qe^t)^{-k} \right)$$

$$= -k \log \left(\frac{1-qe^t}{p} \right)$$

$$= -k \log \left(\frac{p+q-qe^t}{p} \right)$$

$$= -k \log \left(1-\frac{q}{p}(e^t-1) \right)$$

$$= -k\left(-\frac{p}{q}(e^t-1) - \frac{p^2}{q^2}(e^t-1)^2 - \ldots\ldots\right)$$

Since $e^t = 1 + \frac{t}{1!} + \frac{t^2}{2!} + \ldots\ldots$

$$K(t) = k\left[\frac{q}{p}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right) + \frac{q^2}{2p^2}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^2 + \frac{q^3}{3p^3}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^3 + \ldots\right]$$

The general term of the moment generating function is

$$\sum_{x=0}^{\infty} \mu_r^1 \frac{t^r}{r!}$$

and consequently on equating the coefficients of $\frac{t^r}{r!}$ we obtain

$$\mu_1^1 = \frac{kq}{p} \text{ , the mean } \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(0)$$

$$\mu_2 = \sigma^2 = \frac{kq}{p} + \frac{k2q^2}{2p^2} = \frac{kqp + kq^2}{p^2} = \frac{kq}{p^2}, \text{ the variance.}$$

If the first two moments are estimated from sample moments, then p is estimated by $\frac{\bar{x}}{s^2}$

Setting $q = 1-p$ in (0), k is found to be estimated by $\frac{\bar{x}p}{1-p}$

APPROXIMATION OF p and k BY THE METHOD OF MAXIMUM LIKELIHOOD - HALDANE (1941)

The maximum likelihood method of estimating the parameters p and k has been discussed by many writers, e.g. Haldane (1941), Anscombe (1950), and Sichel (1951). The following approximation makes use of the method described by Haldane (1941) using the (x+1)th term of the N.B.D. which is given by

$$P(x) = \frac{(k + x - 1)!}{x!(k-1)!} p^k q^x \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Let n_x be the observed frequency of x events, R be the maximum value of x observed, the total number of observations $N = \sum_{x=0}^{R} n_x$, and the mean of x, $\mu = \frac{1}{N} \sum_{x=0}^{R} x n_x$.

In this paper, the likelihood is defined by

$$\prod_{x=0}^{R} \left[P(x) \right]^{n_x}$$

and hence the logarithm of the likelihood, L, is given by

$$L = \sum_{x=0}^{R} n_x \log P(x)$$

$$= \sum_{x=0}^{R} n_x \left(k \log p + x \log q + \sum_{s=0}^{x-1} \log(k+s) - \log x! \right)$$

To obtain the value of p and k which gives maximum likelihood, the first partial derivatives of the logarithm of the likelihood with respect to p and k are taken and set equal to zero; i.e.,

$$\text{Maximum L when } \frac{\partial L}{\partial p} = 0$$

$$\text{and } \frac{\partial L}{\partial k} = 0$$

Since $\frac{\partial L}{\partial p} = \sum_{x=0}^{R} n_x \left(\frac{k}{p} - \frac{x}{1-p} \right) = \sum_{x=0}^{R} n_x \left(\frac{k(1-p)-xp}{p(1-p)} \right)$

Then L is maximum when

$$\frac{1}{pq} \sum_{x=0}^{R} n_x \left(kq-xp \right) = 0$$

i.e., when

$$\sum_{x=0}^{R} n_x kq = \sum_{x=0}^{R} n_x xp$$

and upon substituting N for $\sum_{x=0}^{R} n_x$ and μN for $\sum_{x=0}^{R} xn_x$ we obtain

$$Nkq = \mu Np$$

from whence

$$p = \frac{kq}{\mu} \quad \text{.......................................(2)}$$

Therefore the best estimate of p makes use of the arithmetic mean.

Similarly, L is maximum when

$$\frac{\partial L}{\partial k} = \sum_{x=0}^{R} n_x \left(\log p + \sum_{s=0}^{x-1} \frac{1}{k+s} \right)$$

$$= N \log p + \sum_{x=0}^{R} \frac{1}{k+x} \sum_{s=x+1}^{R} n_s = 0$$

i.e., when

$$\sum_{x=0}^{R} \frac{1}{k+x} \sum_{s=x+1}^{R} n_s = -N \log p$$

$$= N \log \frac{1}{p} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

From (2)

$$\mu = \frac{k(1-p)}{p}$$

Thus $$p = \frac{k}{\mu+k}$$

Therefore equation (3) becomes

$$N \log \left(1+\frac{\mu}{k} \right) - \frac{n_1+n_2+\ldots+n_R}{k} - \frac{n_2+n_3+\ldots+n_R}{(k+1)} - \ldots\ldots - \frac{n_R}{(k+R-1)} = 0 \quad \ldots (4)$$

The solution of this equation, other than $k = \infty$, provides an estimate of k. Using the iterative process of Newton-Raphson (Whittaker and Robinson - 1944), the best estimate of k is obtained.

CHAPTER III

FITTING THE NEGATIVE BINOMIAL DISTRIBUTION TO WHITEFISH DATA

In the following investigation the N.B.D. is fitted to Whitefish data from five Lakes in Manitoba. In each case the goodness of fit is measured by a Chi-Squared test.

Chi-Squared is given by

$$\chi^2_{0.05\ n-3} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where

O_i is the observed frequency in the ith class,

E_i is the corresponding expected frequency for the ith class,

n-3 is the degree of freedom, and

0.05 is the level of significance of the test

(Probability of rejecting the Null Hypothesis when it is true.)

The Null (H_0) and Alternate (H_1) hypotheses

H_0: the N.B.D. is a good model for Whitefish cyst distribution,

H_1: the N.B.D. is not a good model for Whitefish cyst distribution,

must be formulated in order to complete the test.

A significant chi-squared value designates the rejection of the H_0 and the acceptance of H_1, while for a non-significant value the reverse is true.

Expected frequencies less than one are grouped with those immediately following until the expected frequency becomes one or greater. The corresponding observed frequencies are also grouped.

TABLE 3-1. Medium Whitefish Cysts From Knee Lake (1960)

Cysts	Observed Frequency	Expected Probability	Expected Frequency
0	61	.44373	52.80
1	21	.31115	37.03
2	22	.14969	17.81
3	14	.06102	7.26
4^+	1	.03436	4.09
Total	119	.99995	118.99

Mean = 0.93 Variance = 1.25

p = 0.74 k = 2.70

The calculated chi-squared goodness of fit, with two degrees of freedom, is equal to 17.752. This signifies a poor fit of the N.B.D. to the data (since the critical value is 5.991).

TABLE 3-2 Medium Whitefish Cysts
From Kisseynew Lake (1961)

Cysts	Observed Frequency	Expected Probability	Expected Frequency
0	29	.54452	43.02
1	19	.14440	11.41
2	16	.08140	6.43
3	6	.05394	4.26
4	4	.03841	3.03
5	1	.02849	2.25
6^+	4	.10875	8.59
Total	79	.99991	78.99

Mean = 1.87 Variance = 13.32

p = 0.14 k = 0.31

The calculated chi-squared goodness of fit, with four degrees of freedom, is equal to 28.120, thus signifying a poor fit (since the critical value is 9.488).

TABLE 3-3 Medium Whitefish Cysts From Dore Lake (1962)

Cysts	Observed Frequency	Expected Probability	Expected Frequency
0	490	.53121	510.49
1	219	.21024	202.04
2	121	.10905	104.80
3	59	.06099	58.61
4	28	.03535	33.97
5	14	.02092	20.10
6	9	.01255	12.06
7	8	.00760	7.30
8	3	.00464	4.46
9	6	.00284	2.73
10^{+}	4	.00452	4.34
Total	961	.99991	960.90

Mean = 1.10 Variance = 3.04

p = 0.36 k = 0.62

The calculated chi-squared goodness of fit, with eight degrees of freedom, is equal to 12.918, this signifying a good fit (since the critical value is 15.507).

TABLE 3-4 Medium Whitefish Cysts From South Indian Lake (1963)

Cysts	Observed Frequency	Expected Probability	Expected Frequency
0	994	.64650	1013.07
1	338	.20318	318.38
2	134	.08289	129.89
3	62	.03638	57.01
4	15	.01652	25.89
5	10	.00766	12.00
6	6	.00360	5.64
7^+	8	.00321	5.03
Total	1567	.99994	1566.91

Mean = 0.64 Variance = 1.29

p = 0.50 k = 0.63

The calculated chi-squared goodness of fit, with five degrees of freedom, is equal to 8.826, thus signifying a good fit (since the critical value is 11.070).

TABLE 3-5 Medium Whitefish Cysts
From Cormorant Lake (1963)

Cysts	Observed Frequency	Expected Probability	Expected Frequency
0	31	.20255	38.28
1	30	.17207	32.52
2	31	.13856	26.19
3	26	.10952	20.70
4	18	.08574	16.20
5	11	.06673	12.61
6	13	.05174	9.78
7	4	.04000	7.56
8	6	.03087	5.83
9	5	.02378	4.49
10	4	.01829	3.46
11	4	.01405	2.66
12	2	.01079	2.04
13^{+}	4	.03523	6.66
Total	189	.99992	188.98

Mean = 3.61 Variance = 15.20

p = 0.24 k = 1.12

The calculated chi-squared goodness of fit, with eleven degrees of freedom, is equal to 8.869, thus signifying a good fit (since the critical value is 19.675).

Thus in three out of the five lakes chosen, the N.B.D. is a good model for the Whitefish data, and the following table provides a summary of the statistics calculated for each of the lakes sampled, including those of the five lakes previously presented in detail.

TABLE 3-6.

A Summary of Statistics Calculated for each of The Lakes Sampled.

Lake	Year	Number of fish	Mean	Var.	p	k	χ^2
Molson	1961	457	2.18	29.51	0.07	0.17	152.158**
"	1962	388	0.98	6.69	0.15	0.17	107.158**
"	1963	793	0.82	5.78	0.14	0.14	128.009**
Knee	1960	119	0.93	1.25	0.74	2.70	17.752*
"	1961	90	0.96	1.48	0.65	1.77	1.992
"	1962	357	1.59	8.27	0.19	0.38	32.357*
"	1963	392	0.98	1.99	0.49	0.95	3.993
"	1964	206	1.19	2.36	0.50	1.21	3.036

TABLE 3-6 CONTINUED

Lake	Year	Number of fish	Mean	Var.	p	k	χ^2
Kisseynew	1961	79	1.87	13.32	0.14	0.31	28.120*
"	1962	94	0.35	0.40	0.88	2.45	0.091
"	1963	97	0.84	4.39	0.19	0.20	20.872*
Attawapiskat	1963	222	0.44	0.72	0.61	0.69	6.764
"	1964	126	0.38	0.53	0.72	0.98	0.840
Thompson	1961	68	0.34	0.62	0.55	0.69	0.605
"	1962	215	0.36	0.89	0.40	0.24	4.540
"	1963	114	0.33	0.77	0.43	0.25	2.935
"	1964	79	0.37	0.83	0.45	0.30	9.808*
Dore	1961	1493	1.11	2.86	0.39	0.70	16.953
"	1962	961	1.10	3.04	0.36	0.62	12.918
"	1963	608	0.86	1.65	0.52	0.94	3.639
Reed	1963	263	9.16	174.06	0.05	0.51	57.894*
" Large	1963	31	5.77	75.71	0.08	0.48	12.584

TABLE 3-6 CONTINUED

Lake	Year	Number of fish	Mean	Var.	p	k	χ^2
South Indian	1961	1594	0.76	3.84	0.20	0.19	500.288**
"	1962	1528	0.64	2.28	0.28	0.25	202.316**
"	1963	1567	0.64	1.29	0.50	0.63	8.826
"	1964	812	0.59	0.96	0.61	0.94	10.620*
" Large	1961	527	0.86	4.96	0.17	0.18	233.451**
" "	1962	81	0.86	1.62	0.53	0.97	0.915
" "	1963	246	0.57	2.85	0.20	0.14	632.499**
" Jumbo	1961	353	0.50	1.65	0.77	1.67	9.305*
" "	1962	35	0.49	0.73	0.67	1.00	4.254*
" "	1963	156	0.58	16.36	0.04	0.02	-------
Cormorant	1963	189	3.61	15.20	0.24	1.12	8.869
Paint	1963	137	3.49	17.06	0.20	0.90	9.478

From table 3-6, one can see that the N.B.D. is a suitable model for Whitefish data in just over one-half of the lakes sampled. This fit or lack of fit in the various sets of data may be due to many factors:

1) A favourable sample may have been drawn even from a highly infested lake.

2) An unfavourable sample may have been drawn from a previously favourable lake.

3) The grouping of the data in the table may have caused a few fits to be unfavourable.

The data for Molson Lake, for example, does not fit the N.B.D. The values of the parameters p and k for this lake are quite compatible with each other, while the mean and variance are larger than average. The same is basically true for South Indian Lake.

The data for Dore Lake, on the other hand, does fit the N.B.D., with the values of the parameters p and k being reasonably well spaced.

Thus, on the whole, the parameters p and k are quite important in determining goodness of fit of Whitefish data.

Even though the data from the lakes sampled fits the N.B.D. in only about fifty percent of the cases, and some variable factors enter into the fitting, the N.B.D. will be considered as a suitable model for Whitefish cyst data.

GRAPHICAL REPRESENTATION

The following is a graphical representation of the number of cysts per fish (observed and expected) for the five lakes which were previously dealt with in detail.