TESTS OF HOMOGENEITY AND GOODNESS-OF-FIT TO A TRUNCATED GEOMETRIC DISTRIBUTION: AN M.S. THESIS PROBLEM

by

D. S. Robson

BU-602-M January, 1977

Abstract

Populations not available to direct counting methods can sometimes be estimated by the so-called "removal method". The number of fish in a closed off section of stream, for example, is not available to direct observation but can be estimated by repeatedly seining or electrofishing to obtain a cumulative but incomplete count for the section. In such a sequence of trials the number of newly captured fish per trial typically forms an approximately geometric sequence with the approximately constant ratio $q \doteq X_{n+1}/X_n$ of successive catch sizes then representing the constant rate of escapement. When this relationship holds exactly in expectation, $q = E(X_{n+1}/X_n)$, then the rate of escapement q and the initial population size can be estimated from the observed sequence $X_1, \cdots, X_k$ of counts of new captures, using only the sufficient summary statistics $n_1 = X_1 + \cdots + X_k$ and $t_2 = X_2 + 2X_3 + \cdots + (k - 1)X_k$. Computing algorithms are given for constructing exact tests of this model, conditional upon the sufficient statistic (n_1, t_2).

Paper No. BU-602-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

TESTS OF HOMOGENEITY AND GOODNESS-OF-FIT TO A TRUNCATED GEOMETRIC DISTRIBUTION: AN M.S. THESIS PROBLEM

by

D. S. Robson

BU-602-M January, 1977

The truncated geometric distribution is utilized as a statistical model in several classes of biometric problems, including "catch curve analysis" in fishery statistics [1], and the so-called "removal method" of estimating animal abundance by k successive trapping and removal experiments [2]. Some laboratory techniques for removing micro-organisms from an aliquot of diluted medium by successively mixing and centrifuging the sample k times [3] have also been modeled by the truncated geometric distribution. In such contexts the geometric model is the result of a great many simplifying assumptions concerning independence, homogeneity and stationarity in space and time, and each new application of the model requires preliminary testing of these assumptions. We consider here the statistical problem of developing exact and approximate conditional tests of homogeneity and tests of fit to this specific model

Goodness-of-fit

Fine structure and parameterization of the model does depend somewhat upon the context, and fine structure of tests of the model should vary accordingly to direct power against alternatives befitting the circumstances. In the removal method of sampling, each of the units present in the population at the time of the i^{th} sampling is assumed to have equal and independent chances p_i of being removed. If N_i elements are present then the number X_i removed on this occasion will be a

---

Paper No. BU-602-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

binomial random variable,

$$P(X_i = x_i | N_i) = \binom{N_i}{x_i} p_i^{x_i} (1 - p_i)^{N_i - x_i},$$

and if the population is closed to other types of loss or recruitment then

$$N_i = N_1 - X_1 - \cdots - X_{i-1}$$

and

$$P(X_1 = x_1, \cdots, X_k = x_k | N_1) = \prod_{i=1}^{k} \binom{N_i}{x_i} p_i^{x_i} q_i^{N_i - x_i}$$

$$= \frac{N_1!}{\prod_1^k x_i! (N_1 - \sum_1^k x_i)!} (p_1)^{x_1} (q_1 p_2)^{x_2} \cdots (q_1 q_2 \cdots q_{k-1} p_k)^{x_k} (q_1 q_2 \cdots q_k)^{N_1 - \sum_1^k x_i}. \quad (1)$$

In the stationary case where p_i does not depend upon i, $p_i = p$, then the marginal probability of removal in the i^{th} sample becomes the geometric probability pq^{i-1} and the above $(k + 1)$-class multinomial distribution becomes a two-parameter distribution

$$P(X_1 = x_1, \cdots, X_k = x_k | N_1) = \frac{N_1!}{\prod_1^k x_i! (N_1 - \sum_1^k x_i)!} p^{\sum_1^k x_i} q^{\sum_2^k (i-1)x_i + k(N_1 - \sum_1^k x_i)}$$

$$= \left[\binom{N_1}{n_1} (1 - q^k)^{n_1} (q^k)^{N_1 - n_1}\right] \left[\frac{N_1!}{\prod_1^k x_i!} \frac{q^{t_2}}{(1 + q + q^2 + \cdots + q^{k-1})^{n_1}}\right] \quad (2)$$

admitting a minimal sufficient statistic $\left(n_1 = \sum_1^k x_i,\ t_2 = \sum_2^k (i - 1)x_i\right)$ of the same dimension as (N_1, p).

In some applications, p is regarded as only a nuisance parameter which, nevertheless, must be estimated in order to obtain an estimate of the parameter N_1 of interest. In other applications such as catch curve analysis where p is identified as an annual survival rate, the parameter N_1 does not even appear in the model, and the likelihood function for a single sample is then given by the second factor of (2) representing the p.d.f. of the vector $\underset{\sim}{X} = (X_1, \cdots, X_k)$ conditional on the sum $n_1 = X_1 + \cdots + X_k$. In either case a goodness-of-fit test of size α is defined by a characteristic function $\phi(\underset{\sim}{X})$ having the property $E\left(\phi(\underset{\sim}{X})\middle| n_1, t_2\right) \leq \alpha$, and a solution of the testing problem thus requires calculation of the exact or approximate p.d.f. of t_2 conditional on n_1 in order to obtain the p.d.f. of $\underset{\sim}{X}$ conditional on (n_1, t_2):

$$P_k(T_2 = t_2 | n_1) = C_k(t_2; n_1) \frac{q^{t_2}}{(1 + q + \cdots + q^{k-1})^{n_1}}$$

$$P_k(X_1 = x_1, \cdots, X_k = x_k | t_2, n_2) = \frac{n_1!}{C_k(t_2; n_1) \prod_1^k x_i!} \tag{3}$$

$$C_k(t_2; n_1) = \sum_{x_1, \cdots, x_k \left| \begin{array}{l} \sum_1^k x_i = n_1 \\ \sum_2^k (i-1)x_i = t_2 \end{array} \right.} \frac{n_1!}{\prod_1^k x_i!} .$$

If $k > 3$ then the statistician has some latitude in the choice of a critical region. Lack of fit may in some circumstances be limited to the first one or few stages, as when clumping occurs and several stages of mixing are required to break up the initial clumps. If there is reason for suspecting the initial stages then the available latitude in test structure may be exploited by constructing tests

which are particularly sensitive against departures from expectation in the beginning stages of the k-stage sampling process.

Such sensitivity is achieved by factoring (3) into

$$P_k(T_3 = t_3|n_1,t_2)P_k(T_4 = t_4|n_1,t_2,t_3) \cdots P_k(T_k = t_k|n_1,t_2, \cdots, t_{k-1})$$

$$= \left[\frac{\binom{n_1}{x_1}C_{k-1}(t_3;t_2 - t_3)}{C_k(t_2;n_1)}\right] \cdot \left[\frac{\binom{n_1 - x_1}{x_2}C_{k-2}(t_4;t_3 - t_4)}{C_{k-1}(t_3;t_2 - t_3)}\right] \cdots \qquad (4)$$

$$\left[\frac{\binom{n - x_1 - \cdots - x_{k-3}}{x_{k-2}}C_2(t_k;t_{k-1} - t_k)}{C_3(t_{k-1};t_{k-2} - t_{k-1})}\right]$$

where $T_{\nu+1} = \sum_{i=\nu+1}^{k} (i - \nu)X_i$ is a conditionally sufficient summary of $X_\nu, \cdots, X_k$ if the truncated geometric model applies from the ν^{th} stage onward. Note that since $t_3 = x_1 - n_1 + t_2$, the first factor on the right-hand side is simply the p.d.f. of x_1 conditional upon n_1 and t_2, the second factor is the p.d.f. of x_2 conditional upon n_1, t_2, and x_1, and so on, so that (4) is indeed a factorization of (3). A critical region defined by the tails of these k - 2 p.d.f.'s should be particularly sensitive against alternatives where the lack of fit is limited to the first few stages.

The statistics T_ν are linear functions of H_0-multihypergeometrically distributed random variables and hence are themselves approximately multivariate normally distributed. The conditional p.d.f.'s appearing as factors in (4) are therefore approximately univariate normal, and the calculation of mean and variance formulas for these approximating normal distributions as well as a numerical evaluation of the adequacy of the normal approximation would be an essential part of this research and development.

Tests of homogeneity

When r independent samples are available, each distributed according to (1) or (2), there may be reason to test for homogeneity among these samples. If k is the same for all samples then a conditional test of whether the probabilities $p_1, \cdots, p_k$ in the k trials are the same for all samples is given by the conventional chi-square test for an r × k table. If homogeneity of the p-vectors is accepted then a chi-square test of homogeneity of the r marginal totals of this r × k table also provides a test of the hypothesis that all r samples share a common N_1-parameter.

The structure of the conventional r × k contingency chi-square test is independent of the fine structure of the particular multinomial p.d.f.'s under consideration, however, and an improvement in power should be achievable through modification of the test structure to fit the circumstances at hand. In particular, if the p-vectors of the k-class multinomials were each specified by a truncated geometric p.d.f. as in (2) then the pairs (n_1, T_2) for each of the r samples would sufficiently summarize the samples, so the r - 1 d.f. chi-square statistic comparing the T_2's to their expected values conditioned on row and column totals of the r × k table should provide a more powerful test of homogeneity in this circumstance. Conditioned upon the T_2's as well as row and column totals, the p.d.f. of the T_3's could presumably also be approximated by a multinormal density of rank r - 1, the quadratic form of which then provides another chi-square test statistic on r - 1 d.f. In this manner, k - 1 H_0-independent chi-square statistics could be constructed, each on r - 1 d.f., and might be used to indicate at which stage ν of sampling that the r samples become homogeneous.

Numerical example of lungworm larvae counts

The following data are counts of lungworm larvae impinged on the cover slip by centrifuging a test tube full of a suspended aliquot of fecal material from dogs infected with this parasite. The 9 samples represent 3 aliquots each from fecal samples from each of 3 dogs, and each sample is centrifuged k = 3 times. If the larvae that are present in a sample are randomly dispersed throughout the suspension by stirring, shaking, and mixing in a standardized manner prior to each centrifuging then the truncated geometric model might be expected to hold with a constant $p_1 = p_2 = p_3 = p$ for all 9 samples. The numbers N_1 of larvae initially present should be homogeneous among aliquots from the same fecal sample, but heterogeneous among the 3 dogs.

Table 1. Numbers of lungworm larvae removed from fecal samples by centrifuging.

	Dog a aliquot				Dog b aliquot				Dog c aliquot				
Trial	1a	2a	3a	Total	1b	2b	3b	Total	1c	2c	3c	Total	Grand Total
1	20	19	19	58	13	48	37	98	2	2	3	7	163
2	3	2	2	7	3	6	11	20	3	0	0	3	30
3	0	1	0	1	1	0	1	2	0	0	0	0	3
n_1	23	22	21	66	17	54	49	120	5	2	3	10	196
t_2	3	4	2	9	5	6	13	24	3	0	0	3	36

The counts in some samples are too small to admit tests at the conventional size $\alpha = .05$, and the only test which I illustrate here is the exact test of goodness-of-fit to geometric model using the combined counts in the last column of Table 1. Noting that

$$P_3(T_3 = t_3 | n_1, t_2) = \frac{(t_2 - 2t_3 + 2)(t_2 - 2t_3 + 1)}{t_3(n_1 + t_3 - t_2)} P_3(T_3 = t_3 - 1 | n_1, t_2)$$

so that

$$P_3(1|196, 36) = \frac{36(35)}{1(161)} P_3(0|196, 36)$$

$$P_3(2|196, 36) = \frac{34(33)}{2(162)} P_3(1|196, 36)$$

$$P_3(3|196, 36) = \frac{32(31)}{3(163)} P_3(2|196, 36)$$

$$\vdots$$

$$P_3(18|196, 36) = \frac{2(1)}{18(178)} P_3(17|196, 36)$$

where

$$P_3(0|196, 36) + P_3(1|196, 36) + \cdots + P_3(18|196, 36) = 1$$

we find:

t_3	$P_3(T_3 = t_3 \mid n_1 = 196, t_2 = 36)$	$P(T_3 \leq t_3 \mid 196, 36)$
0	.00356	.00356
1	.02736	.03093
2	.09304	.12396
3	.18533	.30930
4	.24137	.55067
5	.21236	.76790
6	.13925	.90716
7	.06460	.97175
8	.00536	.99892
9	.00095	.99987
10	.00012	.99999
11	.00001	1
12	.00000	1

The observed outcome $t_3 = x_3 = 3$ is thus entirely compatible with the truncated geometric model.

References

[1] D. S. Robson and D. G. Chapman, "Catch curves and mortality rates". Trans. Amer. Fish. Soc. 90:181-189, 1961.

[2] G. A. F. Seber, The Estimation of Animal Abundance. Hafner, 1973.

[3] J. Georgi. Personal communication, 1977.