AN APPLICATION OF GOODNESS OF FIT TO A MULTINOMIAL MODEL WITH A FIXED AND KNOWN NUMBER OF EQUIPROBABLE BUT UNLABELED CELLS

Douglas S. Robson
Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-907-M June 1986

ABSTRACT

Data from a replicated experiment were used to test for avoidance of superparasitism on the part of a parasitoid encaged with a fixed number (N) of unlabeled virgin host larvae for a 24-hour period. The null hypothesis specifies that the parasitoid randomly and independently selects a larva for each successive oviposition so that, conditional on the total number of eggs laid, the vector of egg counts per larva is multinomial with $p_i = 1/N$ for $i=1,\cdots,N$. Each replicate experiment utilized a new one-day-old bred female and a new batch of N same-age larvae. The number (r) of replicates was sufficiently large to enable asymptotic normal theory in testing for an excess of singly parasitized larvae, thus requiring only the calculation of conditional H_0-mean vectors and covariance matrices for constructing the goodness of fit test statistic.

In a laboratory test for superparasitism, a one-day-old bred female parasitoid fly was encaged for 24 hours in a petri dish containing N=40 sessile host larvae arranged in a spatially regular pattern over the floor of the dish. During the 24-hour period she oviposited E eggs, one at a time, in the larval hosts. The larvae were later dissected to count these eggs in order to test the null hypothesis that her selections of host were random and independent for each successive oviposition, the alternative hypothesis being that she tended to avoid repeated oviposition in the same host (i.e., avoided "superparasitism").

Spatial location of the larva was not recorded in making these egg counts so the data consisted of an *unordered* set of counts $\{X_1, X_2, \cdots, X_N\}$ of eggs in each of the N=40 larvae. The sample frequency distribution of eggs per larva,

$$F_i = \text{number of larvae containing exactly i eggs}$$

for i=0,1,2, etc., thus provides a sufficient summary of these data, with

$$F_1 + 2F_2 + 3F_3 + \cdots + EF_E = E = \text{total egg count}$$

and

$$F_0 + F_1 + F_2 + \cdots + F_E = N = 40 = \text{total number of larvae} \quad .$$

In replicates of this experiment the total egg count E_j, $j=1,2,\cdots,r$, for the 24-hour period fluctuated in an uncontrolled manner but was always substantially less than the number (N=40) of available hosts, so superparasitism could have been completely avoided.

Since the alternative hypothesis specifically implies an excess of singly parasitized larvae, we are justified in using this specific residual as a basis for constructing our test statistic; namely, the residual R_{1j} provided by the j'th female parasitoid,

$$R_{1j} = F_{1j} - NP_{E_j}(1)$$

where

E_j	is the total number of eggs laid by this female during the 24-hour period
F_{1j}	is the number of larvae in which she deposited a single egg

$$P_{E_j}(1) = E_j\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right)^{E_j-1}$$

is the conditional probability of a particular one of the N larvae receiving a single egg, given that she randomly and independently selected a larva for each of the E_j oviposits.

Conditional upon this fact (i.e., this H_0-assumption) that E_j eggs were randomly distributed among the N larvae, the expected value of R_{1j} is zero for all $j=1,2,\cdots,r$ and the R_{1j} are independent random variables with variance

$$V(R_{1j}) = NP_{E_j}(1)\left[1 - P_{E_j}(1)\right] + N(N-1)\left[P_{E_j}(1,1) - P_{E_j}(1)P_{E_j}(1)\right]$$

where

$$P_{E_j}(1,1) = E_j(E_j-1)\left(\frac{1}{N}\right)^2\left(1 - \frac{2}{N}\right)^{E_j-2}$$

is the conditional probability that two particular larvae will have received a single egg each.

If the alternative hypothesis is true then each of the R_{1j}'s has positive rather than zero expectation, and averaging the R_{1j}'s should therefore reinforce positivity under the alternative hypothesis while negatives would tend to cancel positives under the null hypothesis to then give an improved estimate of zero. Since the individual R_{1j}'s do have different variances, due to the unequal E_j's, however, a weighted average provides a better H_0-estimate of zero than a simple unweighted average. The optimal weights are the reciprocals of the variance, $1/V(R_{1j})$, so the test statistic based on this weighted average is

$$Z = \frac{\sum_{j=1}^{r} \frac{R_{1j}}{V(R_{1j})}}{\sqrt{\sum_{1}^{r} \frac{1}{V(R_{1j})}}}$$

which is H_0-distributed as a normal random variable with mean zero and variance unity. Z's from different treatments may be combined by summing and rescaling:

$$Z = \frac{\sum_{1}^{k} Z_i}{\sqrt{k}} \quad .$$

APPENDIX

The H_0-probability model for any replicate receiving E eggs is best derived by assigning labels to the N larvae, say $1,2,\cdots,N$ and noting that when she selects a larva to receive an egg, the larva labeled i has probability $p_i = 1/N$ of being selected. This is most apparently true for the first egg, but the H_0-model insists that each egg is so laid - independently of all preceding egg laying events. The outcome of these E independent and identically distributed trials is a multinomially distributed vector random variable $\mathbf{X} = (X_1,X_2,\cdots,X_N)$ where X_i is the number of eggs deposited in the larva labeled i; thus, since $X_1 + X_2 + \cdots + X_N \equiv E$,

$$P(\mathbf{X}=\mathbf{x}) = \frac{E!}{x_1!x_2!\cdots x_N!} p_1^{x_1} p_2^{x_2}\cdots p_N^{x_N} = \frac{E!}{x_1!x_2!\cdots x_N!}\left(\frac{1}{N}\right)^E \qquad (1)$$

Under the H_0-model the labels (and the spatial configuration) of the egg counts $X_1,X_2,\cdots,X_N$ convey no information; i.e., the N X's are exchangeable, so the information in the *vector* **X** is no greater than the information in the *set* $\{X_1,\cdots,X_N\}$. The latter information may be expressed in the form of a frequency *vector* $\mathbf{F} = (F_0,F_1,F_2,\cdots,F_E)$ where F_i is the number of X's equal to i,

$$F_i = \#\{j\,|\,X_j=i\} \quad .$$

The number of distinct vectors **X** producing the same frequency vector **F** is

$$\frac{N!}{F_0!F_1!F_2!\cdots F_E!} = \frac{N!}{\prod\limits_{i=0}^{E} F_i!} \tag{2}$$

where

$$F_0 + F_1 + F_2 + \cdots + F_E \equiv \sum_{i=0}^{E} F_i \equiv N$$

$$F_1 + 2F_2 + \cdots + EF_E \equiv \sum_{i=0}^{E} iF_i \equiv E \quad .$$

The product of (1) and (2) expressed in the form

$$\frac{E!}{\prod\limits_{i=0}^{E} (i!)^{F_i}} \cdot \frac{1}{N^E} \cdot \frac{N!}{\prod\limits_{0}^{E} F_i!} \tag{3}$$

then represents the joint (conditional on E) H_0-probability distribution of the F_i's,

$$P\{\mathbf{F}=\mathbf{f}\} = \frac{E!N!}{N^E \prod\limits_{i=0}^{E} f_i!(i!)^{f_i}} \tag{4}$$

Derivation of the mean and covariance matrix of the vector $\mathbf{F}$ is facilitated by returning to the labeled case (1) and defining counting functions (characteristic functions)

$$\delta_i(j) = \begin{cases} 1 & \text{if} \quad X_j = i \\ 0 & \text{otherwise} \end{cases}$$

so that

$$F_i = \sum_{j=1}^{N} \delta_i(j)$$

and

$$1 = \sum_{i=0}^{E} \delta_i(j) \quad . \tag{5}$$

For a fixed i the $\delta_i(j)$ are identically but not independently distributed for $j=1,2,\cdots,N$; their common distribution is

$$P\{\delta_i(1)=1\} = P\{X_1=i\} = \frac{E!}{i!(E-i)!}\left(\frac{1}{N}\right)^i\left(1-\frac{1}{N}\right)^{E-i} = P_E(i) \qquad (6)$$

and for any two larvae j and j', say, the joint distribution of $\delta_i(j)$ and $\delta_i(j')$ gives

$$P_E(i,i) =$$
$$P\{\delta_i(1)=1,\delta_i(2)=1\} = P\{X_1=i,X_2=i\} = \frac{E!}{i!i!(E-2i)!}\left(\frac{1}{N}\right)^i\left(\frac{1}{N}\right)^i\left(1-\frac{2}{N}\right)^{E-2i} . \qquad (7)$$

The mean value of F_i is therefore

$$\sum_{j=1}^{N} P\{X_j=i\} = \sum_{j=1}^{N} P_E(i) = NP_E(i)$$

and

$$V(F_i) = V\left(\sum_{j=1}^{N}\delta_i(j)\right) = \sum_{j=1}^{N} V\left(\delta_i(j)\right) + \sum_{j\neq j'}^{N} \text{Cov}\left(\delta_i(j),\delta_i(j')\right)$$
$$= NV\left(\delta_i(1)\right) + N(N-1)\ \text{Cov}\left(\delta_i(1),\delta_i(2)\right)$$

where the variance of the Bernoulli variable $\delta_i(1)$ is

$$V\left(\delta_i(1)\right) = P_E(i)\left[1 - P_E(i)\right]$$

and the covariance of $\delta_i(1)$ and $\delta_i(2)$ is

$$\text{Cov}\left(\delta_i(1),\delta_i(2)\right) = P_E(i,i) - P_E(i)P_E(i) \quad .$$

In a similar manner we find, for $i\neq j$,

$$\text{Cov}(F_i,F_j) = -NP_E(i)P_E(j) + N(N-1)\left[P_E(i,j) - P_E(i)P_E(j)\right]$$

where

$$P_E(i,j) = P\{X_1=i,X_2=j\} = \frac{E!}{i!j!(E-i-j)!}\left(\frac{1}{N}\right)^i\left(\frac{1}{N}\right)^j\left(1 - \frac{2}{N}\right)^{E-i-j}$$

to completely define the covariance matrix of **F**.

REMARKS

- The normality of Z derives from the Central Limit Theorem for independent random variables. The asymptotics in this case appeal to the large number of replicates (r=20) contributing to Z. A simulation study could be undertaken to confirm the validity of the normal approximation in this setting.

- The H_0-probability model (4) would follow from an assumption that egg laying during the period in question is a Poisson process, though (4) does not imply that egg laying must be a Poisson process. While the Poisson model does seem untenable in requiring exponentially and independently identically distributed inter-egg laying times, a test of the Poisson model is readily available in the form of the Poisson variance test:

$$\chi^2_{r-1 d.f.} = \frac{\sum_{j=1}^{r} (E_j - \bar{E})^2}{\bar{E}} \quad .$$

- If the Poisson model (or some other well defined model) for the between-replication variation in E_j were available then it would not be necessary to condition on the E_j's. In the Poisson case where the E_j's are independent and identically distributed Poisson random variables the data from the r replicates could be composited; i.e., the rN larvae could be viewed as having all been placed in one large dish exposed to r identical, independently operating parasitoid females. Lacking such a model, however, we are obliged to condition on the E_j's (i.e., regard them as given, unequal sample sizes) and weight the individual replicates accordingly.

ILLUSTRATION

Suppose we had only N=4 larvae instead of N=40 and we labeled them 1, 2, 3 and 4, and suppose E=8 eggs were deposited with the outcome $X_1=2$, $X_2=3$, $X_3=1$, $X_4=2$. The probability (under H_0) of this outcome is

$$\frac{8!}{2!3!1!2!}\left(\tfrac{1}{4}\right)^8 = \frac{8(7)(6)(5)(4)(3)(2)(1)}{2(3)(2)(1)(2)}\left(\tfrac{1}{4}\right)^8 = \frac{105}{4096} \quad .$$

This outcome gives the frequency vector $F_0=0$, $F_1=1$, $F_2=2$, $F_3=1$, $F_i=0$ for $i \geq 4$. There are, however, a total of 12 **X**-vectors producing this frequency vector **F**; namely:

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
3	2	2	1	2	3	2	1	2	2	3	1	2	2	1	3
3	2	1	2	2	3	1	2	2	1	3	2	2	1	2	3
3	1	2	2	1	3	2	2	1	2	3	2	1	2	2	3

and each has probability $\frac{105}{4096}$ of occurring; hence, the probability of this **F**-vector outcome is

$$\frac{12(105)}{4096} = \frac{315}{1024} \quad .$$

A complete list of all possible **F**-vector outcomes and their probabilities of occurrence when N=4 and E=8 are:

F_0	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	Probability (Formula (4))
2				2					$420/4^8$
1	1		1	1					$6720/4^8$
1		2		1					$5040/4^8$
	2	1		1					$10080/4^8$
1		1	2						$6720/4^8$
	2		2						$6720/4^8$
	1	2	1						$20160/4^8$
		4							$2520/4^8$
	3				1				$1344/4^8$
1	1	1			1				$4032/4^8$
2			1		1				$672/4^8$
1	2					1			$672/4^8$
2		1				1			$336/4^8$
2	1						1		$96/4^8$
3								1	$4/4^8$

Sum = $65536/4^8$

= 1

Note that we can now verify the mean and variance formulas (page 1) for F_1. From the above table we find

$$P_E(F_1=0) = (420+5040+6720+2520+672+336+4)/4^8 = 15712/4^8$$

and, similarly,

$$P_E(F_1=0) = 15712/4^8$$

$$P_E(F_1=1) = 31008/4^8$$

$$P_E(F_1=2) = 17472/4^8$$

$$P_E(F_1=3) = 1344/4^8$$

giving a mean value of

$$[0(15712) + 1(31008) + 2(17472) + 3(1344)]/4^8 = \frac{69984}{4^8}$$

in agreement with the formula

$$NP_E(1) = 4\frac{8!}{1!7!}(\tfrac{1}{4})^1(1 - \tfrac{1}{4})^7 = 4(8)(1)(3)^7/4^8 = \frac{69984}{4^8} \quad .$$

Similarly, the variance calculated from this table,

$$[0^2(15712) + 1^2(31008) + 2^2(17472) + 3^2(1344)]/4^8 - [69984]^2/4^{16} = \frac{2448519}{4^{11}}$$

agrees with the formula from page 1,

$$\frac{69984}{4^8}\left(1 - \frac{17496}{4^8}\right) + 4(3)\left[\frac{3584}{4^8} - \frac{4782969}{4^{13}}\right] = \frac{2448519}{4^{11}} \quad .$$

SIMULATION

The sampling distribution of the test statistic Z is simulated here for the case of r=20 replicates using N=40 larvae per replicate. A computer program was written to simulate this experiment M times and compare the resulting sampling distribution to a normal distribution by calculating the sample cumulative distribution at selected quantiles of the standard normal, such as $Q_{.01} = -2.326$, $Q_{.025} = -1.96$, $Q_{.05} = -1.645$, etc. The following results were obtained from M=5000 experiments in which E_j=15 for all j:

Normal fraction	.01	.025	.05	.10	.5	.9	.95	.975	.99
Observed fraction	.0104	.0268	.0554	.0988	.4970	.9008	.9502	.9702	.9896