Overdispersion or Extra Binomial Variation

Count data often do not conform to simple variance assumptions implied in using the binomial or multinomial distribution. In the binomial model it is assumed that each individual has the same underlying probability of a “tail" (homogeneity assumption) and that outcomes are independent (independence assumption). Under this set of assumptions,

$$\text{var}(y) = np(1-p) \text{ and } \text{var}(\hat{p}) = \frac{p(1-p)}{n}\ .$$

These models of count data carry an *assumed* variance (and covariance) structure. If these ideal assumptions are violated to some degree, the MLEs are typically consistent (some small sample bias might be present), but the estimated sampling variances and covariances are *underestimated* (i.e., the actual variability in the estimates will exceed the estimated sampling variance). If the empirical sampling variance > the theoretical variance, the situation is called “overdispersion" or “extra-binomial variation." This reflects a lack of independence or heterogeneity among individuals.

There are approaches to coping with this real world issue (see Burnham et al. 1987:243-246; the theory for this is due to Wedderburn 1974; also see Anderson et al. 1994 dealing with the effect of overdispersion in model selection in open capture-recapture models).

Extra-binomial variation might be expected when brood mates or male-female pairs are banded. If these behave as a unit, then independence is likely to be violated. In the example of a paired male and female bird, their fates (outcomes) may be linked; thus rather than a sample of 2, in some sense they behave as a sample of 1. Then, in the estimator of sampling variance $[p(1-p)]/n$, the sample size in the denominator is too large, thus the estimate is too small (the sampling variance is underestimated).

Wedderburn suggests some sophisticated procedures for actually *modeling* the residuals to obtain improved estimates of the variance-covariance matrix. However, a simple approach is often very effective and easy to implement. Here, one starts with the goodness-of-fit (GOF) test statistic for the global model (after any needed pooling of cells to avoid expected values that are small) and its degrees of freedom (again after any required pooling). Then, a **variance inflation factor** is computed simply as

$$\hat{c} = \chi^2/df\ .$$

The estimate of a variance inflation factor should come from the global model (the most highly parameterized model in the set of candidate models). In the case of the band recovery models, the deviance or the Pearson GOF test statistic could be used. In the case of the open population capture-recapture models, the estimation of *c* should come from the summation of TEST2 and TEST3 (see Burnham et al. 1987)

In the case of independence and homogeneity $c \equiv 1$. In cases of some dependence or heterogeneity, $c > 1$. We recommend 2 considerations before using the variance inflation factor: (1) knowledge of the biology leading to a suspicion of overdispersion, and (2) a value of $\hat{c}$ that is, say, 1.3 or greater (perhaps consider the "significance" of the GOF test if the p-value was < 0.2). If the degrees of freedom are small, (say less than about 5) then a smaller p-value should probably be used (say 0.05 or even 0.01).

If, based on these 2 criteria (a biological basis and an estimated variance inflation factor > 1), overdispersion is though to be present, then both the estimated sampling variances and covariances should be "inflated" by multiplying them by $\hat{c}$. Thus, the sampling variance would be $\hat{c} \cdot \text{var}(\hat{\theta})$ (then the $\text{se}(\hat{\theta})$ would be multiplied by the square root of $\hat{c}$). One can see that if the estimated variance inflation factor is near 1, the effect on the estimated standard errors is quite small. In cases where overdispersion is more marked, then the inflation of the variances and covariances becomes important. One might often expect $1 \leq c \leq 4$. A simple example of the variance inflation factor for the open C-R models is given by Burnham et al. (1987:252-254).

If overdispersion has been identified from using GOF test and its degrees of freedom from the global model, then *model selection* must be based on QAIC or QAIC$_c$ as

$$\text{QAIC} = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K,$$

and

$$\text{QAIC}_c = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K + \frac{2K(K+1)}{n-K-1},$$

$$= \text{QAIC} + \frac{2K(K+1)}{n-K-1}.$$

When no overdispersion exists, $c = 1$, the formulae for QAIC and QAIC$_c$ reduce to AIC and AIC$_c$, respectively (see Lebreton et al. 1992 for originally suggesting this approach; further details are given in Anderson et al. 1994). Program *MARK* assumes c of 1; if overdispersion is a consideration, then the value of $\hat{c}$ should be input into *MARK* (under ADJUSTMENTS).

Of course, $\hat{c}$ might be > 1 because the global model is structurally inadequate and no overdispersion, in fact, is present. In this case, careful biological considerations should often allow the investigator to conclude that the GOF test is detecting lack of model fit – not overdispersion. However, if one does not understand that the inadequate model structure is being detected by the GOF test and proceeds to inflate the estimated sampling variances and covariances, at least this might lead to a conservative approach to inference.

As a technical comment, Wedderburn (1974) provides theory to justify the usual MLEs as (assymptotically) optimal point estimators of the model parameters, even when there is overdipserion in the data.