RP-393

Analysis of binomial data in quantitative genetics experiments

Walter A. Becker
WASHINGTON STATE UNIVERSITY, PULLMAN, WA 99163

INTRODUCTION

In many of the experiments on binomial data one of the principal assumptions involves the simple binomial sampling variance. This sampling variance is valid only under certain conditions and it is the intention of this paper to examine the effects on the variance when three of the conditions are not met.

CONDITIONS UNDERLYING BINOMIAL SAMPLING VARIANCE

Simple sampling means random sampling where the probability, p, of success is the same for each event in the sample, the success of each event is completely independent of the success of all other events in the sample, and all samples are drawn from the same population (Yule & Kendall, 1950).

Under these conditions the mean of the proportion of successes is $\mu = p$, and the variance $\sigma^2 = pq/n$ where $q = 1 - p$, and n = number of events or observations within a sample.

Samples from Different Populations

If we remove the condition that samples must come from the same population while maintaining the other two conditions, the

Proceedings of the International Conference on QUANTITATIVE GENETICS

mean is $\mu = (p_1 + p_2 + \ldots . p_k)/N = p_o$ where p_k = proportion of successes in kth sample, N = number of samples, and p_o is mean value of the varying samples, p, and $\sigma^2 = p_o q_o/n + [(n-1)/n]\ \sigma_p^2$.

$p_o q_o/n$ is the variance due to chance fluctuations and σ_p^2 is due to real differences amongst the samples. For derivations of these and succeeding formulas see Yule and Kendall (1950).

For example, the proportion of blood spots in chicken eggs was analyzed by Becker and Bearse (1973) with a nested analysis of variance with three levels for sires, dams/sires and progeny/dams. The expectation of MS_W (mean square for progeny/dams) is $pq/n + [(n-1)/n]\ \sigma_W^2$ where σ_W^2 estimates 1/2 additive genetic variance, 3/4 dominance variance, < 3/4 epistatic variance and all the environmental variance within families.

Heterogenous Events Within A Sample

Relaxing the condition that the probability of success is the same for each event in the sample results in (the other two conditions remain): $\mu = p_o$, $\sigma^2 = p_o q_o/n - \sigma_p^2/n$ where p_o is the mean chance of a sample and σ_p^2 is the variance due to the probability of each event within a sample being different. For example, if we mate a sire to a series of dams, each producing n fertile poultry eggs, the probability of each egg within a dam hatching differs because of genetic and environmental variability. With a number of sires the hatchability analysis involves a nested design analysis of variance with sires and dams within sires. The progeny response is measured as the proportion of hatch of fertile eggs for each dam. Each dam will have an effect both genotypic and maternal on the hatchability of her progeny. In the one-way layout analysis of variance, the expectation of MS_D contains the binomial sampling variance, pq/n, the reduction of the variance because of σ_p^2/n ($\sigma_p^2 = \sigma_W^2$ where σ_W^2 is the variance between full sibs), and the increase in the variance

due to the samples (or dams) being different.

Correlation Between Events in the Sample

When the condition of independence between events in a sample is removed with the other two conditions held intact, the variance is $\sigma^2 = \frac{pq}{n}[1 + r\ (n-1)]$, where r is the correlation between the success of events in the sample.

Referring to the chicken hatchability example, each embryo is correlated genetically within the sample because they are full sibs and possibly correlated environmentally because the eggs of a dam are usually set in close proximity to one another within the incubator. Similar problems arise when investigating the genetics of infectious diseases.

DISCUSSION

In most quantitative genetic analyses, the three conditions of simple sampling are not met. The interpretation of the genetic model for quantitative genetic experiments using binomial data must take into account the effects on the simple binomial variance by the removal of the three conditions, a circumstance not considered completely by Becker and Marsden (1972) in a genetic study of blister rust resistance of Western White Pine.

The arc sine square root transformation is based upon the simple binomial sampling variance. When the variance of the proportion is different than the simple binomial sampling variance, the use of this transformation is probably not warranted.

Further difficulties arise when using the analysis of variance because of unequal numbers (see Gabriel, 1963) and the estimation of the underlying normal distribution on the binomial scale (Dempster and Lerner, 1950; Van Vleck, 1972).

BIBLIOGRAPHY

[1] Becker, W. A. & Bearse, G. E. (1973). Selection for high and low percentages of chicken eggs with blood spots. Brit. Poul. Sci. 14, 31-47.

[2] Becker, W. A. & Marsden, M. A. (1972). Estimation of heritability and selection gain for blister rust resistance in Western White Pine. Biology of Rust Resistance in Forest Trees, USDA Misc. Publ. No. 1221, 397-409.

[3] Dempster, E. R. & Lerner, I. M. (1950). Heritability of threshold characters. Genetics 35, 212-236.

[4] Gabriel, K. K. (1963). Analysis of variance of proportions with unequal frequencies. J. Amer. Stat. Assoc. 58, 1133-1157.

[5] Van Vleck, L. D. (1972). Estimation of heritability of threshold characters. J. Dairy Sci. 55, 218-225.

[6] Yule, G. U. & Kendall, M. G. (1950). An Introduction to the Theory of Statistics. 14th ed. London: Griffin, 386-409.