The Institute for Perception

For more technical reports from The Institute for Perception or for information about our short courses, please visit www.ifpress.com or email us at mail@ifpress.com

Reprinted from IFPress (1998) 1(4)2,3

Models for Replicated Ratings

Jian Bi and Daniel Ennis

Background: A competitor has introduced a new fragrance for young females into a market that your company dominates with an existing fragrance. Since these fragrances are quite different according to your perfumers, you decide to test them among separate consumer samples to avoid effects introduced by testing the different fragrances on the same individuals. Your main interest is in comparing the degree of liking for your fragrance and the new rival. However, you also suspect that there may be differences in individual hedonic response to the fragrances and in how individuals use rating scales. For these reasons you decide to use replicated tests in which each consumer evaluates only one type of fragrance. Replicated ratings data are collected from two young female groups with 50 and 54 consumers, respectively. The smaller sample is for your product. Six replications are obtained from each consumer. A 5-point liking scale is used where "1" means "dislike very much" and "5" means "like very much".

Replicated ratings data such as these arise in sensory and consumer acceptance research. The main advantage of replicated measurement is that it can improve the reproducibility of an experiment. Replicated measurements typically yield more precise estimates or more powerful tests with the same number of consumers. Conventional statistical models such as the binomial and multinomial usually fail to fit replicated measurement data. Alternative models are needed for the kind of data given in the above example.

Ratings are ordered categorical data. Researchers in the social sciences including those in sensory and consumer science have relied on statistical methods, such as the *t*-test and the analysis of variance, which were designed for applications where the outcomes are continuous. However, analysis options have been changing gradually over the last two decades and new tools, more appropriate to categorical data, are emerging. There are many useful procedures now available for analyzing categorical or replicated categorical data. One group of techniques transforms the data in order to treat them using existing methods for continuous data. Examples are the general linear model (GLM)[1] and generalized estimating equations (GEE)[2]. Another group of techniques deals with categorical or replicated categorical data in their original form without transformation and these are the models that we discuss here. Our interest in these models is two-fold. First, we would like to remain as faithful as possible to the original data without transformation. Second, we can connect the parameters of these models directly to Thurstonian models[3] for ratings and difference testing methods and this allows great flexibility in interconnecting methodologies.

The Multinomial Model: The multinomial distribution is a multivariate discrete distribution. It is a natural extension of the binomial distribution when the number of response categories is more than two. In the following two extreme situations, the multinomial model could be used for each of the consumer samples in the fragrance example:

1) Assume that each consumer always gives the same rating score for a product. In this situation, replication is not necessary. So, the vectors of frequencies for the liking rating categories follow the multinomial distributions with sample size $N_1 = 50$ and $N_2 = 54$ and parameter vectors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, each summing to 1.

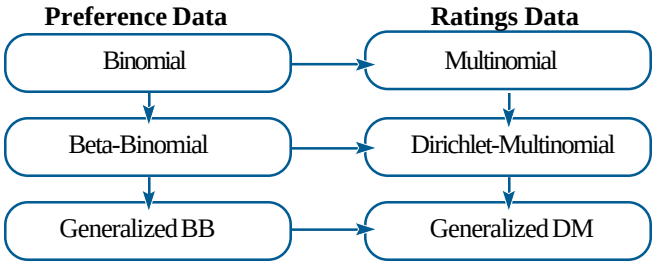
2) Usually a subject gives different rating scores for the same product at different times. These rating scores have some probability of occurring. If we assume that the responses for all consumers are independent of one another and that the consumers have the same response pattern, then the vector of rating frequencies for the pooled data across subjects follows a multinomial distribution with sample size N_j and parameter vector $\boldsymbol{\pi}_j = (p_{j1}, p_{j2}, \ldots p_{j5})$, $j = 1, 2$.

These two assumptions are quite naïve because each consumer may not always give the same rating for a product. In addition, consumers in the same group may not have identical response patterns because these patterns may depend on how a particular consumer interpreted the rating instrument used. If the two assumptions described above are not justified, pooled data for consumers in a group will not follow a multinomial distribution. The usual formulae for Pearson's chi-square and likelihood ratio tests may not provide reliable statistics, even for very large samples. This phenomenon is referred to as *overdispersion*. Ignoring the inter-consumer or inter-trial variation can result in an inflated Type I error level for tests on the mean response probabilities for each rating category. In the presence of overdispersion, where then does the multinomial assumption apply? For replicated ratings for each consumer in the fragrance example, the vector of liking frequencies *within* each consumer may be assumed to follow a multinomial distribution with sample size $n = 6$ (the number of replications) and parameter vector $\boldsymbol{p}_{ji}$, where j and i represent consumer samples and consumers within samples, respectively. The key issue to be resolved in developing a model to handle overdispersion is how to model the distribution of $\boldsymbol{p}_{ji}$ over consumers or trials.

The Dirichlet-Multinomial Model: There is a close parallel between the generalization of the binomial to the beta-binomial (BB)[4] and the generalization of the multinomial to the Dirichlet-multinomial (DM)[5]. See Figure 1 to see how these models are interconnected. The Dirichlet distribution is the multivariate beta distribution and allows us to treat the within-trial multinomial probabilities as random variables very much the way that the beta distribution treats binomial probabilities. This means that we have a formal way of accounting for differences among consumers within

groups in how they use rating scales. For a particular group, we estimate two parameters: $\boldsymbol{\pi}_j$, a vector of response probabilities for the scales used and C_j, an overdispersion parameter for the group. The number of elements in $\boldsymbol{\pi}_j$ is the number of rating categories and C_j provides information on the extent of overdispersion present.

Figure 1. Relationships among categorical data models. All of the models are special cases of the generalized Dirichlet-multinomial model.



In the figure above, it can be seen that the binomial is a special case of the beta-binomial when there is no overdispersion. If there is more than one source of overdispersion, then the appropriate model is the generalized BB where a different beta distribution applies to each source of overdispersion.

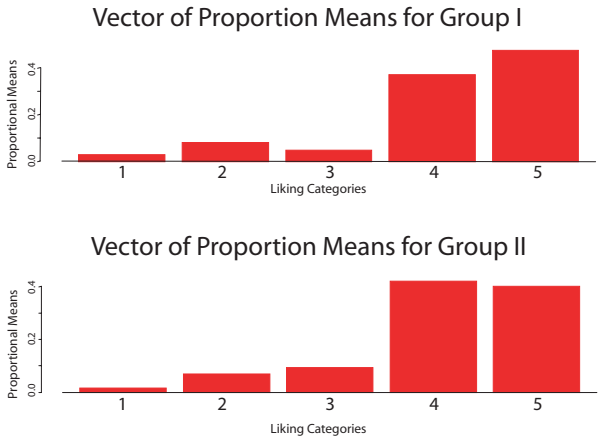
The right side of Figure 1 parallels the left side for choice responses involving more than two categories. The multinomial replaces the binomial, a special case, and the DM models replace the BB models.

Fitting the Fragrance Data to the DM Model: Your first goal is to test for overdispersion by comparing the DM model to the multinomial. The generalized Tarone's Z statistic[5] is used and the *p*-values for the goodness of fit tests are smaller than 0.01. This means that the variation among consumers within both of the samples cannot be ignored and the DM model fits the data better than the multinomial.

Estimating and Testing the Parameters of the DM Model: Using methods for fitting and testing the DM models, your estimates of π_j are: $\hat{\boldsymbol{\pi}}_1 = (.03, .08, .05, .37, .48)$ and $\hat{\boldsymbol{\pi}}_2 = (.02, .07, .09, .43, .39)$ as shown in Figure 2. Your estimates of C_j are: $\hat{C}_1 = 2.45$ and $\hat{C}_2 = 2.33$. The C_j values are significantly larger than one but not different from each other. You conclude that individual consumers differ within both groups in how they rate the fragrances, but that there is no difference betweens groups in rating heterogeneity. This result once again confirms that overdispersion is present in both groups because of differences among consumers in rating scale use.

Your main interest in evaluating the fragrance data is to test for a difference in degree of liking for the fragrances. This test shows that the two vectors of proportions *are not* significantly different (*p*-value = .38), while Pearson's χ^2 statistic shows that the two vectors of proportions *are* significantly different (*p*-value = .04). These results are different because variation among consumer is ignored in Pearson' χ^2 test. This leads to an inflated Type I error. The importance and reliability of the DM model is that it accounts for both inter-trial and intra-trial variation in replicated ratings.

Figure 2: Proportion means for the five point liking scale among two consumer samples who evaluated fragrances.



Other Applications: Overdispersion can arise in numerous ways. In the fragrance example, differences among consumers was the source. In highly trained panels, differences among subjects may not be the main cause of overdispersion, but it may be associated with other aspects of the experiment. For instance, if a product is produced at different factories and evaluated by a panel of experts whose rating response pattern is uniform across panelists, overdispersion may be due to factories. The number of replications in this case would be the number of panelists. Hierarchical or generalized DM models can be constructed to account for multiple sources of overdispersion such as that due to consumers, differences in experimental material and time.

Conclusion: The DM model is an extension of the multinomial model and is an appropriate model for replicated ratings data when inter-trial variation cannot be ignored. Critical applications of these models are in product development, claims support and quality assurance. When overdispersion occurs, Pearson's χ^2 test may result in a higher Type I error level than planned because of an underestimate of variance. The DM test is more reliable than Pearson's χ^2 test in these applications because more sources of variation are accounted for. When overdispersion does not exist, the DM model reduces to the multinomial.

References:

1. McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models.* London: Chapman and Hall.
2. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* **73**, 13-22.
3. Bi, J. and Ennis, D.M. (1998). A Thurstonian variant of the beta-binomial model for replicated difference tests. *Journal of Sensory Studies*, **13,** 461-466.
4. Ennis, D.M. and Bi, J. (1998). The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, **13**, 389-412.
5. Ennis, D. M. and Bi, J. (1998). The Dirichlet-multinomial model: Accounting for inter-trial variation in replicated ratings. *Journal of Sensory Studies. In Press*.