Published in final edited form as:

J Stat Plan Inference. 2012 November ; 142(11): 2965–2975. doi:10.1016/j.jspi.2012.04.017.

Two-way model with random cell sizes

Steven F. Arnold^a and Panagis G. Moschopoulos^{b,*}

^aPennsylvania State University, State College, PA, United States

^bUniversity of Texas at El Paso, El Paso, TX, United States

Abstract

We consider inference for row effects in the presence of possible interactions in a two-way fixed effects model when the numbers of observations are themselves random variables. Let N_{ij} be the number of observations in the (i, j) cell, π_{ij} be the probability that a particular observation is in that cell and μ_{ij} be the expected value of an observation in that cell. We assume that the $\{N_{ij}\}$ have a joint multinomial distribution with parameters n and $\{\pi_{ij}\}$. Then $\bar{\mu}_{i.} = \sum_j \pi_{ij} \mu_{ij} / \sum_j \pi_{ij}$ is the expected value of a randomly chosen observation in the i th row. Hence, we consider testing that the $\bar{\mu}_{i.}$ are equal. With the $\{\pi_{ij}\}$ unknown, there is no obvious sum of squares and F-ratio computed by the widely available statistical packages for testing this hypothesis. Let $\bar{Y}_{i.}$ be the sample mean of the observations in the i th row. We show that $\bar{Y}_{i.}$ is an MLE of $\bar{\mu}_{i.}$, is consistent and is conditionally unbiased. We then find the asymptotic joint distribution of the $\bar{Y}_{i.}$ and use it to construct a sensible asymptotic size α test of the equality of the $\bar{\mu}_{i.}$ and asymptotic simultaneous $(1 - \alpha)$ confidence intervals for contrasts in the $\bar{\mu}_{i.}$.

Keywords

Two-way model; Main effects; Analysis of variance; Unbalanced data; Multinomial cell sizes

1. Introduction

The two-way model with fixed effects for unbalanced data has received considerable attention in the past 30 years. There has been much discussion of the definition of the effects, calculation of sums of squares and definitions of hypotheses. The model is the subject of many texts, prominent among those Searle (1971, 1982), Graybill (1976), Scheffé (1959), and Arnold (1981). Several decompositions of sums of squares are presented in ANOVA tables, and testing the parameters of the model is routinely accomplished by subtracting sums of squares to isolate the contribution of a certain parameter. Searle's (1971) R -notation, for example $R(\alpha/\mu, \beta)$ is usually involved in testing the contribution of including α after μ and β in the model. The R -notation implies orthogonalizations of the design matrix that lead to sums of squares that test a host of hypotheses that depend on the cell sizes. Many hypotheses have interpretation problems and several articles have been written on the hypotheses implied by using Searle's $R(\cdot/\cdot)$ notation, e.g. see, Speed et al. (1978), Kutner (1974); for an extensive list see Macnaughton (1998).

This paper considers the same two-way model under the assumption that the overall sample size n is fixed but the cell sizes are themselves random variables following the multinomial distribution with unknown parameters π_{ij} . The setting has been introduced in Moschopoulos

and Davidson (1985). Applications of the new setting arise mainly in survey studies. In these studies it is often not possible to sample from the individual cells. Instead, a sample of size n is taken from the overall population and is divided after the fact into the two-way cells according to *sample* characteristics. The following two examples will help in understanding this setting:

Example 1

A random sample size n of women is drawn from a metropolitan area for the purpose of studying fertility, measured by the number of children born to women (the response y -variable). It is assumed that fertility is dependent on religion, the row factor (Catholic (1) or Protestant (2)) and educational level, the column factor (Elementary (1), High School (2), College (3)). Sampling 300 women from the overall population results in $2 \times 3 = 6$ cell sizes n_{ij} , $i = 1, 2$; $j = 1, 2, 3$ that are random variables following the multinomial distribution (for large n). Note that sampling from each of the six classifications is not possible since religion and education are only known after the sample is drawn. In this example, let i represent the woman's religion and j represent the woman's educational level. In this situation, π_{21} represents the proportion of women who are Protestant with Elementary education; μ_{12} represents the expected number of children in Catholic women with High School education; $\bar{\mu}_2^\pi$ represents the expected number of children by Protestant women averaged across educational levels. In testing that the $\bar{\mu}_i^\pi$, $i = 1, 2$ are all equal in this setting, we are testing the equality of the mean number of children for the two religions averaged over educational levels. Testing for no column effects with these weights is testing the equality between educational levels averaged over religion. For a real situation related to this, see Groat and Neal (1967).

Example 2

Consider an academic class size n viewed as a random sample from a population of similar students with the response y being the student performance during an examination. The purpose of the investigation here is to relate class performance to student anxiety, the row factor (High (1), Moderate (2), Low (3)) and student attitude towards the subject the column factor (Positive (1), Negative (2)). Note that sampling from each of the six classifications here is not possible as the student classifications are known only after the sample is drawn. Thus, the cell sizes are random following the multinomial distribution (for large n). As in the first example, here testing the equality of $\bar{\mu}_i^\pi$, $i = 1, 2, 3$ is testing the equality of performances of High, Moderate and Low anxiety students averaged over student attitudes. For a real situation on this, see Galassi et al. (1981).

Random cell sizes but from a different perspective are considered in Psychology, consideration there given to random cell sizes arising as a result of the underlying treatment, see Weiss (1999, 2006).

As stated above, the two-way model with fixed effects and unbalanced data is not free of controversies. The main problems are the following: (1) the definition of main effects, i.e. row, column and interaction effects, and (2) hypotheses tested by typical ANOVA decompositions and interpretations of such hypotheses. Concerning (1) above, as it turns out, in the case of the two-way model with multinomially distributed cell sizes we can provide a natural definition of main effects. This definition is not arbitrary but is a natural consequence of the sampling scheme and it is expressed in terms of parameters (i.e. cell population proportions and the cell population means), see the next section. As for (2) above, we are concerned here with one particular hypothesis, namely the equality of row (column) means. Let μ_{ij} , $i = 1, \dots, r$; $j = 1, \dots, c$ be the mean of cell (i, j) , and let π_{ij} be the probability of obtaining an observation in cell (i, j) . We show that the i -th row mean is

$$\mu_i = \sum_{j=1}^c \frac{\pi_{ij}}{\pi_{i.}} \mu_{ij}, \quad (1.1)$$

where $\pi_{i.} = \sum_{j=1}^c \pi_{ij}$. The hypothesis of equality of row means is

$$H_0: \mu_i \text{ equal}, \quad i=1, \dots, r. \quad (1.2)$$

The hypothesis (1.2) above is equivalent to the hypothesis that the α -main effects (defined in Section 2) are zero. This hypothesis is well known in the literature but the means are defined with cell size weights, i.e. (see Searle, 1971)

$$\tilde{\mu}_i = \sum_{j=1}^c \frac{N_{ij}}{N_{i.}} \mu_{ij}. \quad (1.3)$$

Obviously, when the cell sizes are random variables, testing the equality of means in (1.3) is not proper. Testing of the right hypothesis (1.2) is the second goal of this paper. However, *there is no test derived from standard ANOVA sums of squares decompositions that is proper for testing* (1.2). We develop an asymptotic ($n \rightarrow \infty$) test for (1.2) and asymptotic $100 \times (1 - \alpha)\%$ confidence intervals for contrasts of row means. In addition, we provide some numerical evaluations of the performance of the proposed test under the null hypothesis. Under the multinomial assumption for the cell sizes, it is possible that a cell size may be zero. The paper is rigorously considering the case of zero cells.

2. The two-way model with random cell sizes

The two-way model is often modeled in the following over-parameterized form that includes terms for main effects and interactions. We observe Y_{ijk} independent,

$$Y_{ijk} \sim N(\theta + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2), \quad i=1, \dots, r; \quad j=1, \dots, c; \quad k=1, \dots, N_{ij}, \quad (2.1)$$

where θ , $\{\alpha_i\}$, $\{\beta_j\}$, $\{\gamma_{ij}\}$ and σ^2 are unknown parameters. In order to make these parameters identifiable, weights $w_{ij} > 0$ are often chosen and it is assumed in addition that

$$\sum_i \sum_j w_{ij} \alpha_i = 0, \quad \sum_i \sum_j w_{ij} \beta_j = 0, \quad \sum_i w_{ij} \gamma_{ij} = 0, \quad \sum_j w_{ij} \gamma_{ij} = 0. \quad (2.2)$$

It is well known that the problem of testing that the interactions $\gamma_{ij} = 0$ does not depend on the weights, nor does the problem of testing that the main effects $\alpha_i = 0$ when it is assumed that the $\gamma_{ij} = 0$ (e.g. see Arnold, 1981, pp. 93–96). Unfortunately, however, the problem of testing that the $\alpha_i = 0$ when the γ_{ij} may be non-zero depends on the weights chosen. Different weights lead to different hypotheses. Let

$$\mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij}, \quad \bar{\mu}_i^w = \sum_j w_{ij} \mu_{ij} / \sum_j w_{ij}. \quad (2.3)$$

Then testing that $\alpha_i = 0$ with the weights w_{ij} is the same as testing the equality of the $\bar{\mu}_i^W$. Therefore, the weights w_{ij} represent the relative importance of the observations in the i th row, j th cell out of all the observations in the i th row.

Now, suppose that the individuals observed are themselves a sample of size $n (= \sum_i \sum_j N_{ij})$ from a very large population. In that case, we may assume that the cell sizes $\{N_{ij}\}$ have a joint multinomial distribution with parameters n , and $\{\pi_{ij}\}$ written as

$$\{N_{ij}\} \sim M_{rc}(n, \{\pi_{ij}\}),$$

where π_{ij} is the probability that a randomly chosen individual is in the i th row and the j th column. We assume that the π_{ij} are unknown parameters for this model. In this case, naturally, we use the weights

$$w_{ij} = \pi_{ij}. \quad (2.4)$$

Using these weights, we see that the conditional expectation of an observation in row i is

$$\begin{aligned} E(Y_{ijk} | \text{row } i) &= \sum_{j=1}^c Pr(\text{column } j | \text{row } i) E(Y_{ijk} | \text{row } i, \text{column } j) \\ &= \sum_{j=1}^c \frac{\pi_{ij}}{\pi_i} \mu_{ij} = \bar{\mu}_i \end{aligned} \quad (2.5)$$

since

$$P(\text{column } j | \text{row } i) = \pi_{ij} / \sum_j \pi_{ij} = \frac{\pi_{ij}}{\pi_i}. \quad (2.6)$$

Therefore, in testing the equality of the $\bar{\mu}_i$, we are testing that the expected value of the response variable is the same in all the rows.

Note—The expectation in (2.5) is approximate if some cells are empty in row i ; the case of empty cells is treated in this paper, see Basic results. It is now interesting to note that if the artificial parameters θ , α_i , β_j and γ_{ij} above are defined using the weights $w_{ij} = \pi_{ij}$, then

$$\theta = \sum_i \sum_j \pi_{ij} \mu_{ij}, \quad \alpha_i = \bar{\mu}_i - \theta, \quad \beta_j = \bar{\mu}_{.j} - \theta, \quad \gamma_{ij} = \mu_{ij} - \bar{\mu}_i - \bar{\mu}_{.j} + \theta. \quad (2.7)$$

3. Basic results

For simplicity we consider here the cell means μ_{ij} instead of the over-parameterized model. Let Y_{ijk} be independent,

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2), \quad i=1, \dots, r; \quad j=1, \dots, c; \quad k=1, \dots, N_{ij},$$

where

$$\{N_{ij}\} \sim M_{rc}(n, \{\pi_{ij}\},)$$

i.e., the $\{N_{ij}\}$ are (jointly) multinomially distributed with parameters n and $\{\pi_{ij}\}$. We assume that

$$\pi_{ij} > 0 \quad \text{for all } i \text{ and } j$$

and that the π_{ij} are unknown and must be estimated from the data, but that n is known and fixed in advance. It is important to remember that the N_{ij} are random variables, and may in fact be 0.

Let q be the number of cells (i,j) in which $N_{ij} > 0$. (Note that q is a random variable dependent on the N_{ij}). Let

$$\bar{Y}_{ij.} = \sum_k Y_{ijk} / N_{ij} \text{ if } N_{ij} > 0, \quad \bar{Y}_{ij.} = 0 \text{ if } N_{ij} = 0, \quad (3.1)$$

$$S^2 = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 / (n - q). \quad (3.2)$$

Then it is easily shown that the sets $\{\bar{Y}_{ij.}\}$, $\{N_{ij}\}$ together with S^2 form a sufficient statistic for this model and that $\bar{Y}_{ij.}$, N_{ij}/n and $(n - q)S^2/n$ are the MLE's of μ_{ij} , π_{ij} and σ^2 respectively. (Note that the MLE for μ_{ij} is only unique if $N_{ij} > 0$.) Furthermore, N_{ij}/n is an unbiased estimator of π_{ij} . Unbiased estimators for the μ_{ij} do not appear possible, due to the possibility that $N_{ij} = 0$. Finally, note that conditionally on q ,

$$(n - q)S^2 / \sigma^2 | q \sim \chi^2_{n - q} \quad (3.3)$$

so that S^2 is consistent and unbiased.

Now, with $\bar{\mu}_{i.}$ as in (2.5) and $\pi_{i.}$ and η_{ij} as in (2.6), let $N_i = \sum_j N_{ij}$ and

$$\hat{\pi}_{i.} = N_i / n, \quad \hat{\eta}_{ij} = N_{ij} / N_i \text{ if } N_i > 0 \quad \text{and} \quad \hat{\eta}_{ij} = 0 \text{ if } N_i = 0$$

be MLE's of $\pi_{i.}$ and η_{ij} respectively. Further, let

$$\bar{Y}_{i..} = \sum_j \hat{\eta}_{ij} \bar{Y}_{ij.}.$$

Note that $\bar{Y}_{i..}$ is the MLE for $\bar{\mu}_{i.}$ by the invariance principle, and is unique as long as $N_i > 0$. Note also that as long as $N_i > 0$,

$$\bar{Y}_{i..} = \sum_j \sum_k Y_{ijk} / N_i. \quad (3.4)$$

Finally, note that $\bar{Y}_{j..}$ has a point mass at 0, so is not a continuous random variable.

$$P(\bar{Y}_{i..}=0)=P(N_i=0)=(1-\pi_i)^n. \quad (3.5)$$

The Distribution of $\bar{Y}_{j..}$ —We are primarily interested in inference about the $\bar{\mu}_{j..}$. In particular, we are interested in testing that the $\bar{\mu}_{j..}$ are all equal. In order to draw inference about the $\bar{\mu}_{j..}$, we need to learn about the distribution of $\bar{Y}_{j..}$. For notational convenience we let

$$\vec{\eta}_i' = (\eta_{i1}, \dots, \eta_{ic})', \quad \widehat{\vec{\eta}}_i' = (\widehat{\eta}_{i1}, \dots, \widehat{\eta}_{ic})', \quad \vec{m}_i' = (\mu_{i1}, \dots, \mu_{ic}).$$

The following random variables are useful in our development here and represent row-means weighted by cell sizes. For $i = 1, \dots, r$, let

$$\tilde{\mu}_i = \sum_j \frac{N_{ij}}{N_i} \mu_{ij} = \sum_j \widehat{\eta}_{ij} \mu_{ij} = \vec{m}_i' \widehat{\vec{\eta}}_i. \quad (3.6)$$

We note here that testing for main effects in the presence of interaction and using the usual conditional F -test entails testing the hypothesis that the $\tilde{\mu}_i$'s are equal, see Searle, 1971, pp. 292–231, Arnold, 1981, pp. 93–96. Obviously this hypothesis makes no sense especially if the cell sizes are random variables. Define

$$e_{ijk} = Y_{ijk} - \mu_{ij}, \quad \bar{e}_{i..} = \sum_{jk} e_{ijk} / N_i \text{ if } N_i > 0, \quad \bar{e}_{i..} = 0 \text{ if } N_i = 0. \quad (3.7)$$

If $N_i > 0$, then

$$\bar{Y}_{i..} - \tilde{\mu}_i = \sum_{jk} (Y_{ijk} - \mu_{ij}) / N_i = \bar{e}_{i..}. \quad (3.8)$$

When $N_i = 0$,

$$\bar{Y}_{i..} = \tilde{\mu}_i = \bar{e}_{i..} = 0.$$

Therefore, for all N_i ,

$$\bar{Y}_{i..} = \tilde{\mu}_i + \bar{e}_{i..} = \vec{m}_i' \widehat{\vec{\eta}}_i. \quad (3.9)$$

Lemma 1

Conditionally on $N_{j\cdot}$ and $\hat{\eta}_i$ are independent.

Proof

If $N_j = 0$, then both $j\cdot$ and $\hat{\eta}_i$ are degenerate and hence independent. If $N_j > 0$, then $j\cdot$ is the average of all N_j of the errors in the j th row, irrespective of which cell they are in. Since the errors are independently identically distributed, $j\cdot$ does not depend on the relative proportions in the columns, and hence on $\hat{\eta}_i$.

This result implies that conditionally on $\vec{N} = (N_1, N_2, \dots, N_r)$, the $\bar{Y}_{i\cdot}$, $i = 1, \dots, r$ are not normally distributed; by (3.9), $\bar{Y}_{i\cdot}$ is the sum of two independent random variables and if it were normal then both components should be normal, which obviously is not the case. This result is of importance in our development for the following reasons: The definition of $\bar{\mu}_{i\cdot}$ depends only on μ_{ij} and η_{ij} and the distribution of $\vec{N}$ depends only on the $\pi_{i\cdot}$. Therefore, $\vec{N}$ is an ancillary statistic for this problem and it makes sense to work conditionally on $\vec{N}$. Note also that the N_{ij} are not ancillary for the $\mu_{i\cdot}$, so that we should not work conditionally on the N_{ij} .

Comment—Working conditionally on $\vec{N} = (N_1, N_2, \dots, N_r)$, is of course equivalent to making inferences about r means of non-normal sub-populations with the overall population being a mixture of these r sub-populations, each with mixing probability $\pi_{i\cdot}$. This will lead to an approximate test statistic for the problem for large n and essentially reduces the two-way problem to a ‘one-way problem’. However, preliminary numerical evaluations showed that the ‘heuristic’ one-way ANOVA test of equality of row means in this case is totally inappropriate, because the distribution of the row-sample means is NOT normal. Obviously, any test statistic for testing the hypothesis that the $\mu_{i\cdot}$ ’s in (2.5) are equal would have to rely on the distribution of the row-sample mean. We first study the conditional distribution of the $\bar{Y}_{i\cdot}$, given $\vec{N}$.

Theorem 1

- In the conditional distribution given $\vec{N}$, the $\bar{Y}_{i\cdot}$, $i = 1, \dots, r$ are independent.
- If $N_j = 0$, then $E(\bar{Y}_{i\cdot} | \vec{N}) = \text{Var}(\bar{Y}_{i\cdot} | \vec{N}) = 0$.
- If $N_j > 0$, then

$$E(\bar{Y}_{i\cdot} | \vec{N}) = \bar{\mu}_{i\cdot}, \quad \text{Var}(\bar{Y}_{i\cdot} | \vec{N}) = \delta_i \sigma^2 / N_i,$$

where

$$\delta_i = 1 + \sum_j \eta_{ij} (\mu_{ij} - \bar{\mu}_{i\cdot})^2 / \sigma^2. \quad (3.10)$$

Proof

- Conditionally on the $\vec{N}$, the $j\cdot$ are independent as are the $\hat{\eta}_i$, $i = 1, \dots, r$. Hence, using the lemma, the $\bar{Y}_{i\cdot} = \bar{e}_{i\cdot} + \hat{m}_i' \hat{\eta}_i$ are also independent.
- When $N_j = 0$, then $\bar{Y}_{i\cdot}$ is degenerate at 0, so these results follow.
- If $N_j > 0$, then $N_j \hat{\eta}_i$ conditional on $\vec{N}$ has the following multinomial distribution:

$$N_i \widehat{\vec{\eta}} | \vec{N} \sim M_c(N_i, \vec{\eta}_i). \quad (3.11)$$

Now,

$$\begin{aligned} E(\vec{e}_{i..} | \vec{N}) &= 0, \\ E(\tilde{\mu}_i | \vec{N}) &= \vec{m}_i' E(\widehat{\vec{\eta}}_i | \vec{N}) = \vec{m}_i' \vec{\eta}_i = \bar{\mu}_i. \end{aligned} \quad (3.12)$$

and therefore,

$$E(\bar{Y}_{i..} | \vec{N}) = E(\vec{e}_{i..} | \vec{N}) + E(\tilde{\mu}_i | \vec{N}) = 0 + \bar{\mu}_i. \quad (3.13)$$

To see the formula for the variance, note that

$$\text{Var}(\vec{e}_{i..} | \vec{N}) = \sigma^2 / N_i$$

and using the variance–covariance matrix of the multinomial proportions

$$\widehat{\vec{\eta}}_i' = (\widehat{\eta}_{i1}, \dots, \widehat{\eta}_{ic})'$$

we obtain

$$\text{Var}(\tilde{\mu}_i | \vec{N}) = \vec{m}_i' (\text{Cov}(\widehat{\vec{\eta}}_i | \vec{N})) \vec{m}_i = \vec{m}_i' V_i \vec{m}_i / N_i,$$

where V_i has (j, g) element V_{ijg} given by

$$V_{ijj} = \eta_{ij}(1 - \eta_{ij}), \quad V_{ijg} = -\eta_{ij}\eta_{ig}, \quad j \neq g. \quad (3.14)$$

Since $\tilde{\mu}_i$ and $\bar{Y}_{i..}$ are independent (conditionally on $\vec{N}$)

$$\text{Var}(\bar{Y}_{i..} | \vec{N}) = \text{Var}(\vec{e}_{i..} | \vec{N}) + \text{Var}(\tilde{\mu}_i | \vec{N}) = (\sigma^2 / N_i)(1 + \vec{m}_i' V_i \vec{m}_i / \sigma^2) = \sigma^2 \delta_i / N_i$$

and the theorem is proved.

Note from part (c) that $\bar{Y}_{i..}$ is a conditionally unbiased estimator of $\bar{\mu}_i > 0$ but is not unbiased. In fact,

$$E\bar{Y}_{i..} = \bar{\mu}_i P(N_i > 0) = \bar{\mu}_i (1 - (1 - \pi_i)^n).$$

The asymptotic distribution of $\bar{Y}_{i..}$ is given in the following.

Theorem 2

- a. $\bar{Y}_{j..}$ is a consistent estimator of $\bar{\mu}_j$
- b. $N_i^{1/2}(\bar{Y}_{i..} - \bar{\mu}_{i.})|\vec{N} \xrightarrow{d} U_i \sim N(0, \delta_i \sigma^2)$ as $N_i \rightarrow \infty$.

Proof

(a) Note first that $\pi_{i.} > 0$ so that

$$P(N_i > 0) = 1 - (1 - \pi_{i.})^n \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Therefore,

$$E\bar{Y}_{i..} = E(E\bar{Y}_{i..}|\vec{N}) = \bar{\mu}_{i.} P(N_i > 0) \rightarrow \bar{\mu}_{i.} \quad \text{as } n \rightarrow \infty.$$

To show consistency we need to show that $\text{Var}(\bar{Y}_{j..}) \rightarrow 0$. Recall that

$$\text{Var}(\bar{Y}_{i..}) = \text{Var}(E(\bar{Y}_{i..}|\vec{N})) + E(\text{Var}(\bar{Y}_{i..}|\vec{N})). \quad (3.15)$$

Now,

$$\text{Var}(E(\bar{Y}_{i..}|\vec{N})) = \bar{\mu}_{i.}^2 (1 - (P(N_i > 0))^2) \rightarrow \bar{\mu}_{i.}^2 (1 - 1^2) = 0.$$

Finally,

$$\begin{aligned} E(\text{Var}(\bar{Y}_{i..}|\vec{N})) &= E\left(\delta_i \sigma^2 \frac{1}{N_i} P(N_i > 0)\right) \\ &= P(N_i > 0) \delta_i \sigma^2 \sum_{m=1}^n m^{-1} P(N_i = m) \\ &= P(N_i > 0) \delta_i \sigma^2 \sum_{m=1}^n m^{-1} \binom{n}{m} \pi_{i.}^m (1 - \pi_{i.})^{n-m} \\ &= P(N_i > 0) \delta_i \sigma^2 \frac{1}{(n+1)\pi_{i.}} \sum_{m=1}^n \left(1 + \frac{1}{m}\right) \binom{n+1}{m+1} \pi_{i.}^{m+1} (1 - \pi_{i.})^{n-m} \\ &\leq P(N_i > 0) \delta_i \sigma^2 \frac{2}{(n+1)\pi_{i.}} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Therefore, the $\text{Var}(\bar{Y}_{j..})$ and the bias of $\bar{Y}_{j..}$ converge to 0 and $\bar{Y}_{j..}$ is consistent. (b). By the asymptotic approximation of the multinomial proportions,

$$N_i^{1/2}(\widehat{\vec{\eta}}_i - \vec{\eta}_i)|\vec{N} \xrightarrow{d} \vec{W}_i \sim N_c(\vec{0}, V_i),$$

where V_j is given in (3.14). Therefore,

$$N_i^{1/2}(\tilde{\mu}_i - \bar{\mu}_i) | \vec{N} = N_i^{1/2} \vec{m}_i' (\widehat{\eta}_i - \vec{\eta}_i) | \vec{N} \xrightarrow{d} N(0, \vec{m}_i' V_i \vec{m}_i).$$

Now, $\bar{\mu}_{i\cdot}$ is the average of all the errors in the i th row, so that

$$N_i^{1/2} \bar{e}_{i\cdot} | \vec{N} \sim N(0, \sigma^2).$$

By (3.9) and the lemma above, conditionally on N , $N_i^{1/2}(\tilde{\mu}_i - \bar{\mu}_i)$ and $N^{1/2} \bar{e}_{i\cdot}$ are independent. Therefore,

$$N_i^{1/2}(\bar{Y}_{i\cdot} - \bar{\mu}_i) | \vec{N} = (N_i^{1/2} \bar{e}_{i\cdot} + N_i^{1/2}(\tilde{\mu}_i - \bar{\mu}_i)) | \vec{N} \xrightarrow{d} N(0, \sigma^2 + \vec{m}_i' V_i \vec{m}_i).$$

However,

$$\sigma^2 + \vec{m}_i' V_i \vec{m}_i = \sigma^2 \delta_i$$

and this proves the theorem.

Corollary

$\vec{Q}_n = n^{1/2}(\bar{Y}_{1\cdot} - \bar{\mu}_{1\cdot}, \dots, \bar{Y}_{r\cdot} - \bar{\mu}_{r\cdot})' \xrightarrow{d} \vec{Q} \sim N_r(\vec{0}, \sigma^2 T)$, as $n \rightarrow \infty$, where T is a diagonal matrix whose i th element is $\delta_i / \pi_{i\cdot}$.

Proof

Note first that $\hat{\pi}_{i\cdot} = N_i/n \rightarrow \pi_{i\cdot}$ a.s. (almost surely) Therefore, $N_i \xrightarrow{a.s.} \infty$, so that we can use the previous theorem. From that theorem and the conditional independence of the $\bar{Y}_{i\cdot}$,

$$\vec{P}_n = (N_1^{1/2}(\bar{Y}_{1\cdot} - \bar{\mu}_{1\cdot}), \dots, N_r^{1/2}(\bar{Y}_{r\cdot} - \bar{\mu}_{r\cdot}))' | \vec{N} \xrightarrow{d} \vec{P} \sim N_r(0, \sigma^2 M),$$

where M is a diagonal matrix whose i -th diagonal element is δ_i . Since this limiting distribution does not depend on $\vec{N}$, unconditionally we have

$$\vec{P}_n \xrightarrow{d} \vec{P}.$$

Now, $\hat{\pi}_{i\cdot}$ is a consistent estimator of $\pi_{i\cdot}$. Therefore, by Slutsky's theorem,

$$\vec{Q}_n = (\hat{\pi}_{1\cdot}^{-1/2} P_{n1}, \dots, \hat{\pi}_{r\cdot}^{-1/2} P_{nr})' \xrightarrow{d} (\pi_{1\cdot}^{-1/2} P_1, \dots, \pi_{r\cdot}^{-1/2} P_r)' \sim N(0, T).$$

Now consider testing the hypothesis

$$H_0: \bar{\mu}_i \text{ equal for all } i=1, \dots, r. \quad (3.16)$$

If the normal distribution in the corollary were exact and the δ_i/π_i were known, then the model would be a generalized linear model, and this hypothesis would be tested using the following (see (3.3) and recall that q is the number of non-empty cells):

$$F = n \sum_i \pi_i (\bar{Y}_{i..} - \tilde{Y})^2 / \delta_i (r-1) S^2, \quad (3.17)$$

where

$$\tilde{Y} = \sum_i (\pi_i \bar{Y}_{i..} / \delta_i) / \sum_i (\pi_i / \delta_i)$$

and under the null hypothesis

$$F \sim F_{r-1, n-q}. \quad (3.18)$$

In addition, Scheffe' simultaneous confidence intervals for the set of contrasts between the $\bar{\mu}_i$ (functions $\sum c_i \bar{\mu}_i$, $\sum c_i = 0$) are given by

$$\sum_i c_i \bar{\mu}_i \in \sum_i c_i \bar{Y}_{i..} \pm S [(r-1) F^\alpha]^{1/2} [\sum_i c_i^2 \delta_i / n \pi_i]^{1/2}, \quad (3.19)$$

where F^α is the upper $100(1 - \alpha)\%$ point of the F -distribution defined in (3.18). Of course, the δ_i and π_i would not be known in practice, so we estimate them in the obvious way. Let

$$\hat{\pi}_i = N_i / n, \quad \hat{\delta}_i = 1 + \sum_j \hat{\eta}_{ij} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 / S^2.$$

(By the invariance principle, $\hat{\pi}_i$ and $\hat{\delta}_i$ are the MLE's of π_i and δ_i). We suggest using the test statistic:

$$\hat{F}_n = n \sum_i \hat{\pi}_i (\bar{Y}_{i..} - \hat{Y})^2 / \hat{\delta}_i (r-1) S^2 = \sum_i N_i (\bar{Y}_{i..} - \hat{Y})^2 / \hat{\delta}_i (r-1) S^2, \quad (3.20)$$

where

$$\hat{Y} = \sum_i (\hat{\pi}_i \bar{Y}_{i..} / \hat{\delta}_i) / \sum_i (\hat{\pi}_i / \hat{\delta}_i). \quad (3.21)$$

Since $\hat{\pi}_i$ and $\hat{\delta}_i$ are consistent (see the proof below), $\hat{F}_n$ should have an approximate F -distribution $F_{r-1, n-q}$ under the null hypothesis. Similarly, we suggest using the following approximate confidence intervals for contrasts in the $\bar{\mu}_i$:

$$\sum c_i \bar{\mu}_i \in \sum c_i \bar{Y}_{i..} \pm S [(r-1)F^\alpha]^{1/2} \left[\sum c_i^2 \widehat{\delta}_i / N_i \right]^{1/2}. \quad (3.22)$$

We finish this section by showing that the test and confidence intervals given above are at least asymptotically correct.

Theorem 3

- a. Let $\widehat{F}_n$ be defined in Eq. (3.20). Then under the null hypothesis that $\mu_{i.}$, $i = 1, \dots, r$ are equal,

$$(r-1)\widehat{F}_n \xrightarrow{d} H \sim \chi_{r-1}^2,$$

and hence the test which rejects the null hypothesis when

$$F_n \geq F_{r-1, n-q}^\alpha$$

is an asymptotic size α test.

- b. The confidence intervals given in Eq. (3.19) are asymptotic $100(1 - \alpha)\%$ simultaneous confidence intervals for the set of contrasts given above.

Proof

- a. Note first that $\widehat{\pi}_{ij} = N_{ij}/n$ are known to be consistent estimators of the π_{ij} so that the $\widehat{\pi}_{i.}$ and $\widehat{\eta}_{ij}$ are consistent estimators of the $\pi_{i.}$ and η_{ij} . We have shown that the $\bar{Y}_{i.}$ and S_n^2 are consistent estimators of $\mu_{i.}$ and σ^2 . Therefore, the $\widehat{\delta}_i$ is consistent estimator of δ_i . Let $\widehat{Q}_n$ and $\widehat{Q}$ be defined as in the corollary above. Let

$$\begin{aligned} \vec{\pi} &= (\pi_1, \dots, \pi_r), & \vec{\pi}_n &= (\widehat{\pi}_1, \dots, \widehat{\pi}_r), & \vec{\delta} &= (\delta_1, \dots, \delta_r), \\ & & \vec{\delta}_n &= (\widehat{\delta}_1, \dots, \widehat{\delta}_r), \\ h(\vec{Q}, \vec{\pi}, \vec{\delta}, \sigma^2) &= \sum_i \pi_i (Q_i - \vec{Q})^2 / \delta_i \sigma^2, & \vec{Q} &= \sum_i (\pi_i Q_i / \delta_i) / \sum_i (\pi_i / \delta_i). \end{aligned}$$

By the usual results on the generalized linear model,

$$H = h(\vec{Q}, \vec{\pi}, \vec{\delta}, \sigma^2) \sim \chi_{r-1}^2.$$

Therefore, by Slutsky's theorem,

$$h(\vec{Q}_n, \vec{\pi}_n, \vec{\delta}_n, S_n^2) \xrightarrow{d} H \sim \chi_{r-1}^2.$$

If the $\bar{\mu}_{i.}$ are all equal, then

$$(r-1)\widehat{F}_n = h(\vec{Q}_n, \vec{\pi}_n, \vec{\delta}_n, S_n^2) \xrightarrow{d} H \sim \chi_{r-1}^2.$$

Now, $q \rightarrow rc$, so that $n - q \rightarrow \infty$ and hence $(r-1)F_{r-1,n-q}^\alpha \rightarrow \chi_{r-1}^2$. Therefore, the probability of rejecting under the null hypothesis is

$$p(h(\vec{Q}_n, \widehat{\vec{\pi}}_n, \widehat{\vec{\delta}}_n, S_n^2) - (r-1)F_{r-1,n-q}^\alpha > 0) \rightarrow P(H - \chi_{r-1,\alpha}^2 > 0) = \alpha.$$

- b. The simultaneous confidence intervals given in Eq. (3.22) are all satisfied if and only if

$$h(\vec{Q}_n, \widehat{\vec{\pi}}_n, \widehat{\vec{\delta}}_n, S_n^2) \leq (r-1)F_{r-1,n-q}^\alpha.$$

The remainder of the argument is similar to part (a).

Some additional comments

1. For ease in defining the model, we have assumed that the $\pi_{ij} > 0$. A careful reading of the paper, however, will show that we have only used $\pi_{i\cdot} > 0$. If $\pi_{i\cdot} = 0$, we can just drop the i th row from the design, so that this is really no assumption.
2. Again, for simplicity, we have assumed that the Y_{ijk} are normally distributed. However, the results are true under greater generality. In particular, the proof of Theorem 1 has only used that

$$EY_{ijk} = \mu_{ij}, \quad \text{var}(Y_{ijk}) = \sigma^2.$$

For Theorems 2 and 3, all we need is that

$$Y_{ijk} = \mu_{ij} + e_{ijk},$$

where e_{ijk} are independently identically distributed with mean 0 and with finite variance σ^2 . We have only used the normality to establish that

$$N_i^{1/2} \bar{e}_{i\cdot} \xrightarrow{d} R \sim N(0, \sigma^2),$$

which can be established under these more general conditions using the central limit theorem.

3. For simplicity, we have assumed that we were sampling from a very large population so that the distribution of the N_{ij} could be assumed multinomial. If we sample from a smaller population we should replace the multinomial distribution with a hypergeometric distribution. Theorem 1 follows equally well for a multivariate hypergeometric (with population size m) if we replace the covariance matrix V for the multinomial with the covariance for the hypergeometric, getting

$$\delta_i^* = 1 + (m-n) \sum \eta_{ij} (\mu_{ij} - \bar{\mu}_{i\cdot})^2 / (m-1) \sigma^2.$$

Theorem 2 can be similarly modified.

4. As a final generalization, we mention a different sampling scheme. Suppose instead of getting a sample of size n from the whole population, we get independent samples of size N_i from each of the row classifications. As before, let N_{ij} be the number of observations in the (i, j) cell. We assume that

$$\vec{N}_i = (N_{i1}, \dots, N_{ic}) \sim M_c(N_i, \vec{\eta}_i), \quad \vec{\eta}_i = (\eta_{i1}, \dots, \eta_{ic}),$$

where η_{ij} is the proportion of the i th row class in column class j . Let $\bar{Y}_{i.}, \widehat{\eta}_{ij}, \bar{\mu}_{i.}, \tilde{\mu}_{i.}, e_{ijk}$ and $\hat{\mu}_{i.}$ be defined as in Section 2. The lemma is still true for this model, since it is conditional on the N_i . Similarly, Theorem 1 is conditional on the N_i , so that we see that the $\bar{Y}_{i.}$ are independent (since they are from independent samples) and

$$E(\bar{Y}_{i.}) = \mu_{i.}, \quad \text{var}(\bar{Y}_{i.}) = \delta_i \sigma^2 / N_i.$$

(Note that for this model, we do not have to worry about the case $N_i = 0$.) Hence $\bar{Y}_{i.}$ is unbiased and consistent for this model. Similarly Theorem 2 is derived conditionally on the N_i , so that we have

$$N_i^{1/2}(\bar{Y}_{i.} - \bar{\mu}_{i.}) \xrightarrow{d} U_i \sim N(0, \delta_i \sigma^2) \quad \text{as } N_i \rightarrow \infty.$$

Now suppose that

$$n = \sum_i N_i, \quad N_i/n \rightarrow \pi_i \quad \text{as } n \rightarrow \infty$$

for some constants π_i . Then $N_i/n \xrightarrow{P} \pi_i$ as $N_i \rightarrow \infty$, so that the corollary and Theorem 3 follow also. Therefore, the procedure derived in Section 2 with π_i replaced by $\lim_{n \rightarrow \infty} (n_i/n)$ applies equally well to the model considered in this paragraph.

4. Numerical evaluations of the test in (3.20)

The proposed test F_n in (3.20) was evaluated numerically under the null hypothesis of equality of row means in (3.16) for several models of 2×2 , 2×3 and 3×4 designs with simulation size 10,000. Tables 1–3, give the models, the cell probabilities, the cell means and the equal row means. The assumed interactions were non-zero but non-significant and the means are plotted in the associated Figs. 1–3. Entries in the tables are the percent of rejections out of the 10,000 samples, for various sample sizes. The size of the entries should be compared to 5%. As seen from all three tables, although asymptotic, the proposed test performs well in maintaining the correct Type-I error.

Acknowledgments

Dr. Moschopoulos's research was partially supported by Grants from the National Center for Research Resources 5G12RR008124 and the National Institute for Minority Health and Health Disparities Grant G12MD007592 from the National Institutes of Health. We express our thanks to Dr. Julia Bader of the Statistical Consulting Laboratory of UTEP for help with the simulations. Finally, we thank the editors and two anonymous referees for comments and revisions that improved the final form of the paper.

References

- Arnold, SF. Theory of Linear Models and Multivariate Analysis. Wiley; New York: 1981.
- Galassi JP, Frierson HT Jr, Sharer R. Behavior of high, moderate and anxious students during an actual test situation. *Journal of Consulting and Clinical Psychology*. 1981; 49 (1):51–62. [PubMed: 7217475]
- Groat, HT.; Neal, AG. *American Sociological Review*. 1967. Social psychological correlates of urban fertility; p. 945-959.
- Graybill, FA. Theory and Application of linear Model. Duxbury; North Scituate, MA: 1976.
- Kutner MH. Hypothesis testing in linear models. *The American Statistician*. 1974; 28:98–100.
- Macnaughton, D. Which sums of squares are best in unbalanced analysis of variance?. Paper presented at the Joint Statistical Meetings; Boston. 1992; 1998. (<http://www.matstat.com/ss/easleao.pdf>)
- Moschopoulos PG, Davidson MI. Hypothesis testing in ANOVA under multinomial sampling. *Sankhyâ Series B*. 1985; 47(Pt. 3):301–309.
- Scheffè, H. The Analysis of Variance. Wiley; NY: 1959.
- Searle, SR. Linear Models. Wiley; NY: 1971.
- Searle, SR. Linear Models for Unbalanced Data. 1982.
- Speed FM, Hoching RR, Hackney OP. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*. 1978; 73:105–113.
- Weiss DJ. An analysis of variance test for random attrition. *Journal of Social Behavior and Personality*. 1999; 14:433–438.
- Weiss, DJ. Analysis of Variance and Functional Measurement: A Practical Guide. Oxford University Press; New York: 2006.

Fig. 1.
Graphs of means for 2×2 factorial designs.

Fig. 2.
Graphs of means for 2×3 factorial designs.

Fig. 3.
Graphs of means for 3×4 factorial designs.

Table 1

Type_I error rates of the new test for row effect of 2×2 factorial designs, with non-significant interaction.

Model	Cell probabilities	Cell means, SD	Row means	n = 30	n = 50	n = 70	n = 100
Model I	(.3,.4,.1,.2)	(50,15,35,27.5) SD = 20	(30,30)	4.35	4.93	4.35	4.83
Model II	(.3,.4,.1,.2)	(50,30,5.71,55) SD = 50	(38.571,38.571)	4.90	5.62	5.59	6.34
Model III	(.3,.4,.2,.1)	(50,15,40,10) SD = 4	(30,30)	5.23	5.92	6.35	6.52
Model IV	(.2,.4,.3,.1)	(32,20,30,6) SD = 10	(24,24)	4.95	5.84	6.24	6.71

Table 2

Type I error rates of the new test, for row effect of 2×3 factorial designs, with non-significant interaction.

Model	Cell probabilities	Cell means, SD	Row means	n = 50	n = 70	n = 100
Model I	(.2,.3,.1,.1,.1,.2)	(50,15,20,35,27.5,23.75) SD = 15	(27.5,27.5)	4.22	4.37	5.03
Model II	(.2,.3,.1,.1,.1,.2)	(40,35,20,45,45,23.33) SD = 10	(34.17,34.17)	4.83	4.97	5.49
Model III	(.2,.3,.1,.1,.1,.2)	(65,5,60,30,45,30.83) SD = 50	(34.17,34.17)	4.08	4.30	4.95
Model IV	(.3,.1,.05,.05,.1,.4)	(30,40,60,30,45,33.89) SD = 30	(35.56,35.56)	3.98	4.23	4.38

Table 3

Type I error rates of the new test, for row effect of 3×4 factorial designs, with non-significant interaction.

Model	Cell probabilities	Cell means, SD	Row means	$n = 70$	$n = 100$	$n = 200$
Model I	(.8,.2,.15,.05,.1,.02,.04,.03,.05,.1,.1,.08)	(50,15,20,10,25,27.5,10,23.54,15,15,40,12.11) SD = 40	(21.88,21.88,21.88)	2.77	3.53	4.13
Model II	(.08,.2,.15,.05,.1,.02,.04,.03,.05,.1,.1,.08)	(40,30,20,50,35,27.5,10,45.63,30,15,20,63.83) SD = 40	(30.63,30.63,30.63)	2.81	3.46	4.07
Model III	(.08,.05,.15,.05,.1,.16,.04,.03,.05,.1,.1,.08)	(40,30,20,50,35,30.45,3.33,20,15,30,60.74) SD = 50	(30.91,30.91,30.91)	3.17	3.65	4.11
Model IV	(.08,.02,.1,.05,.1,.33,.04,.03,.05,.02,.1,.08)	(60,15,20,10,25,35,15,18.33,30,5,20,50) SD = 50	(30.4,30.4,30.4)	3.38	3.72	4.94