CZESŁAW STĘPNIAK (Rzeszów and Lublin)

CHARACTERIZING EXPERIMENTAL DESIGNS BY PROPERTIES OF THE STANDARD QUADRATIC FORMS OF OBSERVATIONS

Abstract. For any orthogonal multi-way classification, the sums of squares appearing in the analysis of variance may be expressed by the standard quadratic forms involving only squares of the marginal and total sums of observations. In this case the forms are independent and nonnegative definite. We characterize all two-way classifications preserving these properties for some and for all of the standard quadratic forms.

1. Background. Consider an experimental design with two-way classification of observations. Let y_{ijk} be the kth observation in the (i, j)th cell, and let

$$\mathbf{y} = (y_{111}, \dots, y_{11n_{11}}; \dots; y_{tv1}, \dots, y_{tvn_{tv}})^T$$

be the vector composed of the observations. A statistical model of \mathbf{y} is usually derived from the following framework:

(1.1)
$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

 $i = 1, \ldots, t; j = 1, \ldots, v; k = 1, \ldots, n_{ij}$, where μ , α_i , β_j and γ_{ij} are deterministic real values, while e_{ijk} is a random quantity representing the experimental error. The following assumptions about the error are standard:

- (a) Vanishing expectation, i.e. $E(e_{ijk}) = 0$ for all i, j and k,
- (b) Homogeneity of variances, say $var(e_{ijk}) = \sigma^2$ for all i, j and k,
- (c) Vanishing covariances, i.e. $\operatorname{cov}(e_{ijk}, e_{i'j'k'}) = 0$ for $(ijk) \neq (i'j'k')$.

For testing problems an additional assumption about the distribution of the vector \mathbf{y} is required. Usually it is assumed that e_{ijk} is normally distributed.

²⁰⁰⁰ Mathematics Subject Classification: Primary 05B06, 62K10; Secondary 51E05, 15A27.

Key words and phrases: 2-way classification, standard quadratic forms, independence, nonnegative definiteness, characterization of designs.

Consequently, (1.1) leads to the following model:

(1.2)
$$y_{ijk} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2)$$
 and are independent.

We mention that the quantities μ , α_i , β_j and γ_{ij} are not parameters in the proper sense, because they are not identifiable by distributions. The standard likelihood ratio tests for hypotheses concerning α 's, β 's and γ 's are based on the following quadratic forms of **y**:

(1.3)
$$S_A = \min_{\mu} \sum_{i,j,k} (y_{ijk} - \mu)^2 - \min_{\mu,\alpha_i} \sum_{i,j,k} (y_{ijk} - \mu - \alpha_i)^2,$$

(1.4)
$$S_B = \min_{\mu} \sum_{i,j,k} (y_{ijk} - \mu)^2 - \min_{\mu,\beta_j} \sum_{i,j,k} (y_{ijk} - \mu - \beta_j)^2,$$

(1.5)
$$S_{A,B} = \min_{\mu,\alpha_{i,\beta_{j}}} \sum_{i,j,k} (y_{ijk} - \mu - \alpha_{i} - \beta_{j})^{2} - \min_{\mu,\alpha_{i,\beta_{j}},\gamma_{ij}} \sum_{i,j,k} (y_{ijk} - \mu - \alpha_{i} - \beta_{j} - \gamma_{ij})^{2},$$

(1.6)
$$S_{E} = \min_{\mu,\alpha_{i,\beta_{j}},\gamma_{ij}} \sum_{i,j,k} (y_{ijk} - \mu - \alpha_{i} - \beta_{j} - \gamma_{ij})^{2}$$

(cf. Scheffé, 1959, or Lehmann, 1986). These quadratics represent the sums of squares for α , β , interaction $\alpha \times \beta$ and error, respectively.

It will be shown that the sums of squares (1.3), (1.4) and (1.6) reduce, respectively, to the quadratic forms

(1.7)
$$Q_A(\mathbf{y}) = \sum_i \frac{y_{i.}^2}{n_{i.}} - \frac{y^2}{n},$$

(1.8)
$$Q_B(\mathbf{y}) = \sum_j \frac{y_{.j}^2}{n_{.j}} - \frac{y^2}{n},$$

(1.9)
$$Q_E(\mathbf{y}) = \sum_{i,j,k} y_{ijk}^2 - \sum_{i,j} \frac{y_{ij}^2}{n_{ij}},$$

where $y_{ij} = \sum_k y_{ijk}$, $y_{i.} = \sum_j y_{ij}$, $y_{.j} = \sum_i y_{ij}$ and $y = \sum_i y_{i.}$, but computing the sum of squares $S_{A,B}$ may not be easy in general. Fortunately, as we will show in the next sections, this computation simplifies considerably in the case of proportional frequencies in cells, i.e. when

(1.10)
$$n_{ij} = \frac{n_{i.}n_{.j}}{n}$$
 for $i = 1, \dots, t; j = 1, \dots, v$,

where $n_{i.} = \sum_{j} n_{ij}, n_{.j} = \sum_{i} n_{ij}$ and $n = \sum_{i,j} n_{ij}$.

The condition (1.10) may be written in the concise form rank(\mathbf{N}) = 1, where $\mathbf{N} = (n_{ij})$ is the incidence matrix. It will be shown that in this regular

case $S_{A,B}$ may be represented as the quadratic form

(1.11)
$$Q_{A,B}(\mathbf{y}) = \sum_{i,j} \frac{y_{ij}^2}{n_{ij}} - \sum_i \frac{y_{i.}^2}{n_{i.}} - \sum_j \frac{y_{.j}^2}{n_{.j}} + \frac{y^2}{n_{.j}}$$

Of course the expression (1.11) makes sense for arbitrary **N** with positive entries, although it may no longer coincide with the sum of squares $S_{A,B}$. We shall refer to (1.11) as the *standard quadratic form*. The aim of this paper is to characterize the experimental designs (i.e. matrices **N**) for which the quadratic form $Q_{A,B}$:

- (i) coincides with the corresponding sum of squares $S_{A,B}$,
- (ii) is nonnegative definite,
- (iii) is independent of the other quadratic forms Q_A , Q_B and Q_E .

Throughout this paper we shall assume that the incidence matrix $\mathbf{N} = (n_{ij})$ is *complete* in the sense that

(1.12)
$$n_{ij} > 0$$
 for all $i = 1, \dots, t; j = 1, \dots, v$,

i.e. all cells in the two-way classification are non-empty.

2. Sums of squares in a two-way classification with proportional frequencies. Most of the statistical literature on this subject focuses on experiments with equal frequencies (cf. Scheffé, 1959, or Lehmann, 1986), or on general experiments with arbitrary frequencies in cells (see Searle, 1971). It is understandable that in the latter case the formulae derived are rather far from being explicit. Some more complete results concerning proportional frequencies can be found in Oktaba and Mikos (1970) in the context of models with restraints (weighted and nonweighted). They noted that the two kinds of restraints may lead to different tests. To avoid this disadvantage, we shall use a coordinate-free approach (cf. Lehmann, 1986, Sect. 7) based on the formulae (1.3)-(1.6).

Let us rewrite model (1.2) in vector-matrix form

(2.1)
$$\mathbf{y} \sim \mathcal{N}(\mu \mathbf{1}_n + \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\gamma}, \ \sigma^2 \mathbf{I}_n),$$

where **y** is the observation *n*-vector, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_t)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_v)^T$ and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{tv})^T$ are vectors of parameters, $\mathbf{A}_{n \times t} = [\mathbf{a}_1, \ldots, \mathbf{a}_t]$ and $\mathbf{B}_{n \times v} = [\mathbf{b}_1, \ldots, \mathbf{b}_v]$ are matrices of zeros and ones satisfying the conditions

(2.2)
$$\sum_{i=1}^{t} \mathbf{a}_i = \sum_{j=1}^{v} \mathbf{b}_j = \mathbf{1}_n,$$

while

(2.3)
$$\mathbf{C}_{n \times tv} = [\mathbf{c}_{11}, \dots, \mathbf{c}_{tv}]$$

C. Stępniak

is the matrix with columns $\mathbf{c}_{ij} = \mathbf{a}_i * \mathbf{b}_j$, $i = 1, \dots, t$; $j = 1, \dots, v$, where * denotes the Hadamard product of vectors (see e.g. Stępniak, 1983).

We shall prove

THEOREM 1. Given a complete incidence matrix $\mathbf{N} = (n_{ij})$ let y_{ijk} , $i = 1, \ldots, t; \ j = 1, \ldots, v; \ k = 1, \ldots, n_{ij}$, be observations represented by (1.1), S_A , S_B , $S_{A,B}$ and S_E be the sums of squares defined by (1.3)–(1.6), and Q_A , Q_B , Q_E and $Q_{A,B}$ be the quadratic forms of the observations defined by (1.7)–(1.9) and (1.11). Then $S_A = Q_A$, $S_B = Q_B$ and $S_E = Q_E$, while $S_{A,B} = Q_{A,B}$ providing rank(\mathbf{N}) = 1.

The proof of this theorem is based on the following lemma:

LEMMA 1. Let $\mathbf{a}_1, \ldots, \mathbf{a}_t$ and $\mathbf{b}_1, \ldots, \mathbf{b}_v$ be arbitrary n-vectors of zeros and ones satisfying the conditions (2.2), \mathbf{c}_{ij} , $i = 1, \ldots, t$; $j = 1, \ldots, v$, be vectors defined by (2.3), $\mathbf{N} = (n_{ij})$ be the complete matrix defined by $n_{ij} =$ $\mathbf{a}_i^T \mathbf{b}_j$, with marginals n_i and $n_{.j}$ for $i = 1, \ldots, t$; $j = 1, \ldots, v$, and $\mathbf{1}_n$, \mathbf{A} , \mathbf{B} and \mathbf{C} be the matrices appearing in (2.1). Then the orthogonal projector onto the column space of $[\mathbf{A}, \mathbf{B}]$ is given by

$$\mathbf{P} = \sum_{i=1}^{t} \frac{1}{n_{i.}} \mathbf{a}_{i} \mathbf{a}_{i}^{T} + \sum_{j=1}^{v} \frac{1}{n_{.j}} \mathbf{b}_{j} \mathbf{b}_{j}^{T} - \frac{1}{n} \mathbf{1}_{n} \mathbf{1}_{n}^{T}$$

providing the condition (1.10) holds.

Proof. Let us rewrite \mathbf{P} in the form

$$\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3,$$

where

(2.5)
$$\mathbf{P}_1 = \sum_{i=1}^t \frac{1}{n_i} \mathbf{a}_i \mathbf{a}_i^T - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T,$$

(2.6)
$$\mathbf{P}_2 = \sum_{j=1}^v \frac{1}{n_{,j}} \mathbf{b}_j \mathbf{b}_j^T - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T,$$

(2.7)
$$\mathbf{P}_3 = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T.$$

First we will show that

(2.8)
$$\mathbf{P}_i \mathbf{P}_j = \begin{cases} \mathbf{P}_i & \text{if } i = j, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

It is easy to verify that $\mathbf{P}_i^2 = \mathbf{P}_i$ for i = 1, 2, 3 and $\mathbf{P}_1\mathbf{P}_3 = \mathbf{P}_2\mathbf{P}_3 = \mathbf{0}$. Moreover,

$$\mathbf{P}_1 \mathbf{P}_2 = \left(\sum_{i=1}^t \frac{1}{n_{i.}} \mathbf{a}_i \mathbf{a}_i^T - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\right) \left(\sum_{j=1}^v \frac{1}{n_{.j}} \mathbf{b}_j \mathbf{b}_j^T - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\right)$$
$$= \sum_{i,j} \frac{n_{ij}}{n_{i.}n_{.j}} \mathbf{a}_i \mathbf{b}_j^T - \frac{1}{n} \sum_i \mathbf{a}_i \mathbf{1}_n^T - \frac{1}{n} \sum_j \mathbf{1}_n \mathbf{b}_j^T + \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T.$$

Thus, by (1.10) and (2.2), we get

$$\mathbf{P}_1\mathbf{P}_2 = \frac{1}{n}\sum_i \mathbf{a}_i \left(\sum_j \mathbf{b}_j^T - \mathbf{1}_n^T\right) - \frac{1}{n}\mathbf{1}_n \left(\sum_j \mathbf{b}_j^T - \mathbf{1}_n^T\right) = \mathbf{0}.$$

In this way we have proved (2.8), which implies that **P** is idempotent.

Now it remains to show that range(\mathbf{P}) = range([$\mathbf{a}_1, \ldots, \mathbf{a}_t, \mathbf{b}_1, \ldots, \mathbf{b}_v$]). Indeed, $\mathbf{P}\mathbf{x} = \mathbf{0}$ for any vector \mathbf{x} such that $\mathbf{a}_i^T\mathbf{x} = \mathbf{b}_j^T\mathbf{x} = 0$ for all i and j. Moreover, by (2.2), rank([$\mathbf{a}_1, \ldots, \mathbf{a}_t, \mathbf{b}_1, \ldots, \mathbf{b}_v$]) $\leq t+v-1$, while rank(\mathbf{P}) = t+v-1. This implies the desired condition and completes the proof of the lemma.

Proof of Theorem 1. In vector-matrix notation the sums of squares (1.3)–(1.6) may be expressed as

$$S_A(\mathbf{y}) = \|\mathbf{y} - \mathbf{P}_{1_n}\mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{P}_A\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{P}_A\mathbf{y} - \frac{1}{n} \mathbf{y}^T \mathbf{1}_n \mathbf{1}_n^T \mathbf{y} = \mathbf{y}^T \mathbf{P}_1 \mathbf{y},$$

$$S_B(\mathbf{y}) = \|\mathbf{y} - \mathbf{P}_{1_n}\mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{P}_B\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{P}_B\mathbf{y} - \frac{1}{n} \mathbf{y}^T \mathbf{1}_n \mathbf{1}_n^T \mathbf{y} = \mathbf{y}^T \mathbf{P}_2 \mathbf{y},$$

$$S_{A,B}(\mathbf{y}) = \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{P}_C\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{P}_C \mathbf{y} - \mathbf{y}^T \mathbf{P}_S,$$

$$S_E(\mathbf{y}) = \|\mathbf{y} - \mathbf{P}_C\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{P}_C \mathbf{y},$$

where \mathbf{P}_M denotes the orthogonal projector onto the column space of \mathbf{M} , \mathbf{C} is defined by (2.3), while \mathbf{P}, \mathbf{P}_1 and \mathbf{P}_2 are defined by (2.4)–(2.5) and (2.6). Now the theorem follows from Lemma 1.

3. Characterizing experimental designs by standard quadratic forms. Let $\mathbf{y} = (y_{111}, \ldots, y_{11n_{11}}; \ldots; y_{tv1}, \ldots, y_{tvn_{tv}})^T$ be subject to the normal linear model $\mathcal{N}(\mu \mathbf{1}_n + \mathbf{A}\alpha + \mathbf{B}\beta + \mathbf{C}\gamma, \sigma^2 \mathbf{I}_n)$, and let Q_A, Q_B, Q_E and $Q_{A,B}$ be defined by (1.7)–(1.9) and (1.11). We are interested in properties of the quadratic form $Q_{A,B}$ and its relations with the remaining quadratics.

THEOREM 2. The quadratic form $Q_{A,B}$ is nonnegative definite if and only if the allocation of the observations in cells is proportional, i.e. rank $(\mathbf{N}) = 1$.

Proof. Assume that $n_{ij} = n_{i.}n_{.j}/n$ for all i and j. Then

$$\sum_{i=1}^{t} \sum_{j=1}^{v} \frac{y_{ij}^2}{n_{ij}} - \sum_{i=1}^{t} \frac{y_{i.}^2}{n_{i.}} - \sum_{j=1}^{v} \frac{y_{.j}^2}{n_{.j}} + \frac{y^2}{n} = \sum_{i,j} n_{ij} \left(\frac{y_{ij}}{n_{ij}} - \frac{y_{.i}}{n_{i.}} - \frac{y_{.j}}{n_{.j}} + \frac{y}{n} \right)^2,$$

implying $Q_{A,B}(\mathbf{y}) \geq 0$ for all \mathbf{y} .

C. Stępniak

Conversely, assume that $Q_{A,B}(\mathbf{y}) \ge 0$ for all \mathbf{y} and set

$$y_{ij} = \begin{cases} n_{ij} & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases}$$

where k is fixed but arbitrary. Then $Q_{A,B}(\mathbf{y})$ reduces to $n_{.k}^2/n - \sum_i n_{.k}^2/n_{i.}$, and consequently, it may be presented in the form $Q_{A,B}(\mathbf{y}) = -n \operatorname{Var}(X)$, where X is a random variable taking values $x_i = n_{ik}/n_{i.}$ with probabilities $p_i = n_{i.}/n$ for $i = 1, \ldots, t$. Thus the condition $Q_{A,B}(\mathbf{y}) \ge 0$ for all \mathbf{y} implies (3.1) $n_{ik} = cn_{i.}, \quad i = 1, \ldots, t$,

for some c. On the other hand, by the definition $n_{k} = \sum_{i} n_{ik}$ we get

$$(3.2) c = \frac{n_k}{n}$$

Combining (3.1) and (3.2) we obtain $n_{ik} = n_{i.}n_{.k}/n$ for $i = 1, \ldots, t$. This completes the proof of necessity, and hence the proof of Theorem 2.

Now let us examine conditions for independence of the quadratic forms Q_A , Q_B , $Q_{A,B}$ and Q_E .

THEOREM 3. Let $\mathbf{y} = (y_{111}, \ldots, y_{11n_{11}}; \ldots; y_{tv1}, \ldots, y_{tvn_{tv}})^T$ be subject to the normal linear model $\mathcal{N}(\mu \mathbf{1}_n + \mathbf{A} \boldsymbol{\alpha} + \mathbf{B} \boldsymbol{\beta} + \mathbf{C} \boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n)$, and let Q_A, Q_B , Q_E and $Q_{A,B}$ be defined by (1.7)–(1.9) and (1.11), respectively. Then Q_E is always independent of the other quadratic forms. Moreover the following conditions are equivalent:

- (a) $\operatorname{rank}(\mathbf{N}) = 1$,
- (b) Q_A is independent of Q_B ,
- (c) Q_A is independent of $Q_{A,B}$,
- (d) Q_B is independent of $Q_{A,B}$.

Proof. The first assertion follows immediately from the fact that in the simple normal sample the sample mean is independent of the sample variance.

Next, by Craig's theorem (e.g. Mathai and Provost, 1992, p. 209 or Harville and Kempthorne, 1997) two quadratic forms $\mathbf{y}^T \mathbf{M}_1 \mathbf{y}$ and $\mathbf{y}^T \mathbf{M}_2 \mathbf{y}$ are independent if and only if $\mathbf{M}_1 \mathbf{M}_2 = \mathbf{0}$. Recall that the matrices of the quadratics Q_A , Q_B , and $Q_{A,B}$, may be represented in the form $\mathbf{P}_1, \mathbf{P}_2$ and $\mathbf{P}_0 - \mathbf{P}_1 - \mathbf{P}_2$, where \mathbf{P}_1 and \mathbf{P}_2 are defined by (2.5) and (2.6), while

$$\mathbf{P}_0 = \sum_{i,j} rac{1}{n_{ij}} \, \mathbf{c}_{ij} \mathbf{c}_{ij}^T - rac{1}{n} \, \mathbf{1}_n \mathbf{1}_n^T.$$

(a) \Leftrightarrow (b). We observe that

$$\mathbf{P}_1 \mathbf{P}_2 = \sum_{j=1}^{v} \left(\sum_{i=1}^{t} \frac{n_{ij}}{n_{i.} n_{.j}} \mathbf{a}_i - \frac{1}{n} \mathbf{1}_n \right) \mathbf{b}_j^T,$$

and, by (2.2),

$$\mathbf{P}_1 \mathbf{P}_2 = \sum_{j=1}^v \sum_{i=1}^t \left(\frac{n_{ij}}{n_{i.} n_{.j}} - \frac{1}{n} \right) \mathbf{a}_i \mathbf{b}_j^T.$$

Thus we get the equivalence of (a) and (b).

(a) \Leftrightarrow (c) and (a) \Leftrightarrow (d). Since \mathbf{P}_i is idempotent for i = 0, 1, 2, and $\mathbf{P}_0\mathbf{P}_i = \mathbf{P}_i$ for i = 1, 2, we get $(\mathbf{P}_0 - \mathbf{P}_1 - \mathbf{P}_2)\mathbf{P}_1 = \mathbf{P}_2\mathbf{P}_1$ and $(\mathbf{P}_0 - \mathbf{P}_1 - \mathbf{P}_2)\mathbf{P}_2 = \mathbf{P}_1\mathbf{P}_2$. Now the desired results follow from the equivalence of (a) and (b).

This completes the proof. \blacksquare

Acknowledgements. Thanks are due to the referee for his careful reading and valuable comments which led to improving the presentation of this work.

References

- D. Harville and O. Kempthorne (1997), An alternative way to establish the necessity part of the classical result on the statistical independence of quadratic forms, Linear Algebra Appl. 127, 157–171.
- E. L. Lehmann (1986), Testing Statistical Hypotheses, 2nd ed., Wiley, New York.
- W. Oktaba and H. Mikos (1970), Mathematical models for proportional data, Zastos. Mat. 11, 151–172 (in Polish).
- A. M. Mathai and S. B. Provost (1992), Quadratic Forms in Random Variables, Dekker, New York.
- H. Scheffé (1959), Analysis of Variance, Wiley, New York.
- S. R. Searle (1971), Linear Models, Wiley, New York.
- C. Stępniak (1983), Optimal allocation of units in experimental designs with hierarchical and cross classification, Ann. Inst. Statist. Math. A 39, 461–473.

Institute of Mathematics University of Rzeszów Al. Rejtana 16A 35-959 Rzeszów, Poland E-mail: cees@univ.rzeszow.pl Statistics and Econometrics Unit Maria Curie-Skłodowska University Pl. Marii Curie-Skłodowskiej 5 20-031 Lublin, Poland

Received on 3.11.2006; revised version on 6.3.2007

(1841)