Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Considering the sample sizes as truncated Poisson random variables in mixed effects models

Célia Nunes, Elsa Moreira, Sandra S. Ferreira, Dário Ferreira & João T. Mexia

To cite this article: Célia Nunes, Elsa Moreira, Sandra S. Ferreira, Dário Ferreira & João T. Mexia (2019): Considering the sample sizes as truncated Poisson random variables in mixed effects models, Journal of Applied Statistics, DOI: 10.1080/02664763.2019.1641188

To link to this article: https://doi.org/10.1080/02664763.2019.1641188

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2019.1641188

Taylor & Francis
Taylor & Francis Group

Considering the sample sizes as truncated Poisson random variables in mixed effects models

Célia Nunes[a], Elsa Moreira[b], Sandra S. Ferreira[a], Dário Ferreira[a] and João T. Mexia[b]

[a]Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal; [b]CMA – Center of Mathematics and its Applications, Faculty of Science and Technology, New University of Lisbon, Lisbon, Portugal

ABSTRACT

When applying analysis of variance, the sample sizes may not be previously known, so it is more appropriate to consider them as realizations of random variables. A motivating example is the collection of observations during a fixed time span in a study comparing, for example, several pathologies of patients arriving at a hospital. This paper extends the theory of analysis of variance to those situations considering mixed effects models. We will assume that the occurrences of observations correspond to a counting process and the sample dimensions have Poisson distribution. The proposed approach is applied to a study of cancer patients.

ARTICLE HISTORY
Received 29 December 2018
Accepted 1 July 2019

KEYWORDS
Random sample sizes; mixed effects; *L* extensions models; *F*-tests; counting processes; cancer registries

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
62J12; 62J10; 62J99

1. Introduction

In some applications of analysis of variance in medicine, social sciences, economic or agriculture, etc., it is more appropriate to regard the sample sizes as random variables. These situations occur commonly when there is a fixed time span for collecting the observations, other examples arise when some other resource is limited. A motivating example is the collection of data from patients with several pathologies arriving at a hospital during a fixed time span. The number of patients for each pathology is not known in advance and a replication of the study during a different time period of the same length would result in a sample of different size. Therefore, if we plan to conduct just one study to compare the pathologies, it is more appropriate to consider the sample sizes as realizations, $n_1, \ldots, n_m$, of random variables, $N_1, \ldots, N_m$, [15,17,20]. Another important case arises when one of the pathologies is rare since, in that case, the desired number of patients in the sample set may not be achieved, [19]. In the cited studies, fixed effects ANOVA was applied. Now we extend the results to mixed effects models to deal with random sample sizes.

The current approach must be based on an adequate choice of the distribution of $N_1, \ldots, N_m$. In this paper, we will assume that the occurrence of observations corresponds to independent counting processes. An illustrative example of this is the aforementioned

CONTACT Célia Nunes celian@ubi.pt Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal

case, concerning the comparison of pathologies. This leads us to consider the assumption of $N_1, \ldots, N_m$ being independent and Poisson distributed with parameters $\lambda_1, \ldots, \lambda_m$, $N_i \sim P(\lambda_i)$, $i = 1, \ldots, m$ [12,15,17–20]. Since we need to have at least one observation per treatment, we will consider the random variables $\ddot{N}_i$, $i = 1, \ldots, m$, obtained truncating the random variables N_i for $N_i \geq 1$, $i = 1, \ldots, m$ (see Appendix 1). Through the independence of $\ddot{N}_i$, $i = 1, \ldots, m$, the variable $\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i$ has truncated Poisson distribution with parameter

$$\lambda = \sum_{i=1}^{m} \lambda_i.$$

For different situations, it will be more appropriate to consider other discrete distributions for random sample sizes, such as

- the Binomial distribution, when there exists an upper bound for the sample sizes, which however may not be attained (either owing to occurrences of failures or for some other reason). An illustrative example of this is when a planned number of patients are approached but only a proportion of them give consent to be included in the study [16,17];
- the Negative Binomial distribution, which can be used as an alternative to the Poisson distribution in cases in which the observations are overdispersed with respect to a Poisson distribution.

This paper is structured as follows. In Section 2, we present the formulation of the mixed models in the context of random sample sizes. The test statistics and their conditional and unconditional distributions are obtained in Section 3. Section 4 presents an application based on real medical data, namely on patients affected by cancer, in order to illustrate the usefulness of our approach. Finally, some concluding remarks are made in Section 5.

2. Model

When considering in mixed models that the sample size are random variables, very likely we will get different number of observations per treatment (combination of factor levels), that is, we have an unbalanced design. In order to cope with unbalanced situations a more broader class of models, designated as L extensions or L models, was developed some years ago in [3] and [14]. Using the L extensions in the formulation of the mixed models with random sample sizes, allow us to deal the lack of orthogonality originated by unbalanced situation.

Let us suppose that the m components of $\boldsymbol{Y}^o$ correspond to the treatments of a linear model and

$$\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{n}) = D(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m}) \tag{1}$$

be the block diagonal matrix with the principal blocks $\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m}$, where $\mathbf{1}_n$ denotes the vector with all n components equal to 1 and $\boldsymbol{n} = (n_1, \ldots, n_m)'$. Then

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{\varepsilon} \tag{2}$$

corresponds to a model with sample sizes $n_1, \ldots, n_m$, where $\boldsymbol{\varepsilon}$ is the error vector with null mean vector and variance–covariance matrix $\sigma^2 \boldsymbol{I}_n$, with $\boldsymbol{I}_n$ the $n \times n$ identity matrix and

$$n = \sum_{i=1}^{m} n_i.$$

Let's consider that

$$\boldsymbol{Y}^o = \sum_{i=0}^{w} \boldsymbol{X}_i \boldsymbol{\beta}_i, \tag{3}$$

where $\boldsymbol{\beta}_0$ is fixed with c_0 components and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w$ are random and independent, with null mean vectors and variance–covariance matrices $\sigma_1^2 \boldsymbol{I}_{c_1}, \ldots, \sigma_w^2 \boldsymbol{I}_{c_w}$, where c_i, $i = 1, \ldots, w$, denote the number of components of $\boldsymbol{\beta}_i$, $i = 1, \ldots, w$. Thus $\boldsymbol{Y}^o$ has mean vector and variance–covariance matrix given by

$$\boldsymbol{\mu}^o = \boldsymbol{X}_0 \boldsymbol{\beta}_0$$

$$\boldsymbol{V}^o = \sum_{i=1}^{w} \sigma_i^2 \boldsymbol{M}_i,$$

with $\boldsymbol{M}_i = \boldsymbol{X}_i \boldsymbol{X}_i'$, $i = 1, \ldots, w$, where matrices $\boldsymbol{X}_i$ have m rows and c_i, $i = 0, \ldots, w$, columns, see e.g. [5,8,23]. We point out that $\boldsymbol{Y}^o$ and $\boldsymbol{Y}$ are random vectors with m and n components, respectively, since $\boldsymbol{L}$ is an $n \times m$ matrix.

3. Test statistics and their distributions

In this section, we obtain the test statistics and their conditional distribution and unconditional distribution, under the assumption that we have random sample sizes. We will start by presenting some important results about L extensions.

Let us assume that $\boldsymbol{Y}^o$ has orthogonal block structure, so the matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w$ commute and they will be linear combinations of pairwise orthogonal projection matrices $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_\ell$, see [2]. Thus we have

$$\boldsymbol{M}_i = \sum_{j=1}^{\ell} b_{i,j} \boldsymbol{K}_j, \quad i = 1, \ldots, w,$$

and

$$\boldsymbol{V}^o = \sum_{j=1}^{\ell} \gamma_j \boldsymbol{K}_j,$$

where $\gamma_j = \sum_{i=1}^{w} b_{i,j} \sigma_i^2$, $j = 1, \ldots, \ell$. With $\boldsymbol{B} = [b_{i,j}]$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\ell)'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_w^2)'$, we also have

$$\boldsymbol{\gamma} = \boldsymbol{B}' \boldsymbol{\sigma}^2,$$

see e.g. [1,2,4,6]. Let's consider that the row vectors of $\boldsymbol{A}_j$, $j = 1, \ldots, \ell$, constitute an orthonormal basis for the range space of $\boldsymbol{K}_j$, $R(\boldsymbol{K}_j)$, $j = 1, \ldots, \ell$, then we have

$$\boldsymbol{K}_j = \boldsymbol{A}_j' \boldsymbol{A}_j, \quad j = 1, \ldots, \ell$$

$$\boldsymbol{I}_{g_j} = \boldsymbol{A}_j \boldsymbol{A}_j', \quad j = 1, \ldots, \ell,$$

with $g_j = \text{rank}(\boldsymbol{K}_j)$.

Let $\boldsymbol{L}^+$ the MOORE-PENROSE inverse of matrix $\boldsymbol{L}$, then the orthogonal projection matrices (OPM) on $\overline{\Omega} = R(\boldsymbol{L})$ and on its orthogonal complement $\overline{\Omega}^{\perp}$ are [22]

$$\boldsymbol{L}\boldsymbol{L}^+ = \boldsymbol{T}$$

$$\boldsymbol{I}_n - \boldsymbol{T}.$$

So, with $\boldsymbol{L} = D(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$, we have

$$\boldsymbol{L}^+ = D\left(\frac{1}{n_1}\mathbf{1}_{n_1}', \ldots, \frac{1}{n_m}\mathbf{1}_{n_m}'\right).$$

When $\boldsymbol{Y}^o$ is independent of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, i.e. $\boldsymbol{\varepsilon}$ is normal with null mean vector and variance–covariance matrix $\sigma^2 \boldsymbol{I}_n$, then $\boldsymbol{T}\boldsymbol{\varepsilon}$ and $(\boldsymbol{I}_n - \boldsymbol{T})\boldsymbol{\varepsilon}$ are also independent, since they have normal joint distribution and null cross-covariance matrices. Therefore

$$\boldsymbol{T}\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{T}\boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{T}\boldsymbol{\varepsilon}$$

and

$$\boldsymbol{Y}_{\overline{\Omega}^{\perp}} = (\boldsymbol{I}_n - \boldsymbol{T})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{T})\boldsymbol{\varepsilon}$$

are independent.

Since the column vectors of $\boldsymbol{L}$ are linearly independent we have [22]

$$\boldsymbol{L}^+\boldsymbol{L} = \boldsymbol{I}_m.$$

So we can consider [3]

$$\boldsymbol{Y}^{oo} = \boldsymbol{L}^+\boldsymbol{Y} = \boldsymbol{Y}^o + \boldsymbol{L}^+\boldsymbol{\varepsilon} = \boldsymbol{Y}^o + \boldsymbol{L}^+\boldsymbol{T}\boldsymbol{\varepsilon},$$

since $\boldsymbol{L}^+\boldsymbol{T} = \boldsymbol{L}^+\boldsymbol{L}\boldsymbol{L}^+ = \boldsymbol{L}^+$, independent of $\boldsymbol{Y}_{\overline{\Omega}^{\perp}}$, then independent of

$$S = \|\boldsymbol{Y}_{\overline{\Omega}^{\perp}}\|^2, \tag{4}$$

where $(1/\sigma^2)S$ has chi-square distribution with

$$g(n) = n - m$$

degrees of freedom, $S \sim \sigma^2 \chi^2_{g(n)}$.

Let us now observe that $\boldsymbol{Y}^{oo}$ has mean vector and variance–covariance matrix given by

$$\boldsymbol{\mu}^{oo} = \boldsymbol{\mu}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0$$

$$\boldsymbol{V}^{oo} = \boldsymbol{V}^o + \sigma^2\boldsymbol{L}^+(\boldsymbol{L}^+)' = \sum_{j=1}^{\ell} \gamma_j \boldsymbol{K}_j + \sigma^2\boldsymbol{L}^+(\boldsymbol{L}^+)'.$$

With $\boldsymbol{L} = D(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$, we will have

$$\boldsymbol{L}^+(\boldsymbol{L}^+)' = D(n_1^{-1}, \ldots, n_m^{-1})$$

and

$$\boldsymbol{Y}_j^{oo} = \boldsymbol{A}_j \boldsymbol{Y}^{oo}, \quad j = 1, \ldots, \ell,$$

has mean vector and variance–covariance matrix

$$\boldsymbol{\mu}_j^{oo} = \boldsymbol{A}_j\boldsymbol{\mu}^o = \boldsymbol{A}_j\boldsymbol{X}_0\boldsymbol{\beta}_0, \quad j = 1,\ldots,\ell$$
$$\boldsymbol{V}_j^{oo} = \gamma_j\boldsymbol{I}_{g_j} + \sigma^2\boldsymbol{A}_j(\boldsymbol{L}^+(\boldsymbol{L}^+)')\boldsymbol{A}_j', \quad j = 1,\ldots,\ell.$$

Being $\boldsymbol{P}_j$ and $\boldsymbol{Q}_j$ the OPM on $R(\boldsymbol{A}_j\boldsymbol{X}_0)$ and $R(\boldsymbol{A}_j\boldsymbol{X}_0)^\perp$, with rank p_j and $f_j = g_j - p_j$, $j = 1,\ldots,\ell$, respectively and $\boldsymbol{S}_j$ and $\boldsymbol{W}_j$ the matrices which the row vectors constitute an orthonormal base to $R(\boldsymbol{A}_j\boldsymbol{X}_0)$ and $R(\boldsymbol{A}_j\boldsymbol{X}_0)^\perp$, $j = 1,\ldots,\ell$, we have

$$\boldsymbol{P}_j = \boldsymbol{S}_j'\boldsymbol{S}_j, \quad j = 1,\ldots,\ell$$
$$\boldsymbol{Q}_j = \boldsymbol{W}_j'\boldsymbol{W}_j, \quad j = 1,\ldots,\ell.$$

3.1. Fixed sample sizes

Let us now address the hypothesis tests for the canonical variance components [13], $\gamma_1,\ldots,\gamma_\ell$, assuming that, with $0 \le z < \ell$,

$$p_j < g_j, \quad j = z+1,\ldots.,\ell.$$

So, let's consider

$$\boldsymbol{Y}_j^\bullet = \boldsymbol{W}_j\boldsymbol{Y}_j^{oo} = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{Y}^o + \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+\boldsymbol{\varepsilon}, \quad j = z+1,\ldots.,\ell,$$

which has null mean vector and variance–covariance matrix $\gamma_j\boldsymbol{I}_{f_j} + \sigma^2\boldsymbol{B}_j, j > z$, with

$$\boldsymbol{B}_j = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}_j'\boldsymbol{W}_j', \quad j = z+1,\ldots,\ell.$$

We intend to test the hypothesis

$$H_{0,j} : \gamma_j = 0, \quad j = z+1,\ldots,\ell. \tag{5}$$

When $H_{0,j}$ holds, we have

$$pr(\boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{Y}^o = \boldsymbol{0}) = 1, \quad j = z+1,\ldots,\ell,$$

and consequently

$$pr(\boldsymbol{Y}_j^\bullet = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+\boldsymbol{\varepsilon}) = 1, \quad j = z+1,\ldots,\ell.$$

Therefore, when $H_{0,j}$ holds, $\boldsymbol{Y}_j^\bullet$ has null mean vector and variance–covariance matrix $\sigma^2\boldsymbol{B}_j$, $j = z+1,\ldots,\ell$, and $(1/\sigma^2)(\boldsymbol{Y}_j^\bullet)'(\boldsymbol{B}_j^{-1})\boldsymbol{Y}_j^\bullet$ has chi-square distribution with f_j degrees of freedom, $(\boldsymbol{Y}_j^\bullet)'(\boldsymbol{B}_j^{-1})\boldsymbol{Y}_j^\bullet \sim \sigma^2\chi^2_{f_j}, j = z+1,\ldots,\ell$ [10].

Since $\boldsymbol{Y}_j^{oo}$ is independent of S, $\boldsymbol{Y}_j^\bullet$ is also independent of $S, j = z+1,\ldots,\ell$. Due to this, when $H_{0,j}$ holds, the statistic

$$\mathcal{F}_j = \frac{g(n)}{f_j}\frac{(\boldsymbol{Y}_j^\bullet)'(\boldsymbol{B}_j^{-1})\boldsymbol{Y}_j^\bullet}{S}, \quad j = z+1,\ldots,\ell, \tag{6}$$

has central F distribution with f_j, $j = z+1,\ldots,\ell$, and $g(n)$ degrees of freedom, $F(\cdot\,|\,f_j, g(n))$, named as conditional distribution, and $\mathcal{F}_j$ might be used as the test statistic [21]. Moreover, the tests with the statistic $\mathcal{F}_j, j = z+1,\ldots,\ell$, are unbiased, e.g. [9,10].

3.2. Random sample sizes

Let us consider that $\boldsymbol{n}$ is the realization of a random vector $\ddot{\boldsymbol{N}} = (\ddot{N}_1, \ldots, \ddot{N}_m)'$, which means that the samples will have random dimensions. In this section, we will focus on the case where

$$\boldsymbol{L}(\ddot{\boldsymbol{N}}) = D(\mathbf{1}_{\ddot{N}_1}, \ldots, \mathbf{1}_{\ddot{N}_m}),$$

for this reason the previous results need to be unconditioned in order to $\ddot{\boldsymbol{N}}$.

Let us now suppose that we intend to test the hypothesis

$$H_0 : \boldsymbol{\theta} = \mathbf{0},$$

where $\boldsymbol{\theta}$ is a general parameter, and the test is unbiased whatever $\boldsymbol{n}$. So, denoting by $pr_{\boldsymbol{n},\boldsymbol{\theta}}(Rej_\alpha)$ $[pr_{\boldsymbol{n},\mathbf{0}}(Rej_\alpha)]$ the probability of rejecting H_0 for a significance level α, given $\boldsymbol{n}$ and the parameter $\boldsymbol{\theta}$ [the probability of rejecting H_0, given $\boldsymbol{n}$ and $\boldsymbol{\theta} = \mathbf{0}$], we have

$$pr_{\boldsymbol{n},\boldsymbol{\theta}}(Rej_\alpha) > pr_{\boldsymbol{n},\mathbf{0}}(Rej_\alpha). \tag{7}$$

Unconditioning (7) in order to $\ddot{\boldsymbol{N}}$, we still obtain

$$pr_{\boldsymbol{\theta}}(Rej_\alpha) > pr_{\mathbf{0}}(Rej_\alpha),$$

and the test still unbiased.

So, since the tests for the hypothesis $H_{0,j} : \gamma_j = 0$, $j = z+1, \ldots, \ell$, are unbiased whatever $\boldsymbol{n}$, we can conclude that they still remain unbiased after unconditioning.

Let us assume that the occurrence of observations corresponds to independent counting processes, which lead us to consider that $\ddot{N}_1, \ldots, \ddot{N}_m$ have truncated Poisson distribution with parameters λ_i, $i = 1, \ldots, m$. Furthermore, to perform inference we also consider that $\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i > m$.

In order to avoid unbalanced cases we will assume that we have a global minimum dimension for the samples [12,20]. Therefore, considering $\ddot{N} > m^\bullet$, with $m^\bullet \geq m$, we may take the probability

$$\begin{aligned}
\ddot{p}_{n,m^\bullet} &= pr(\ddot{N} = n \,|\, \ddot{N} > m^\bullet) = \frac{pr(\ddot{N} = n)}{pr(\ddot{N} > m^\bullet)} \\
&= \frac{\ddot{p}_n}{pr(\ddot{N} > m)} \frac{pr(\ddot{N} > m)}{pr(\ddot{N} > m^\bullet)} = \frac{\ddot{p}_n}{1 - \ddot{p}_m} \frac{1 - \ddot{p}_m}{1 - \sum_{h=m}^{m^\bullet} \ddot{p}_h} \\
&= \ddot{p}_{n,m} \frac{1 - \ddot{p}_m}{1 - \sum_{h=m}^{m^\bullet} \ddot{p}_h}, \quad n = m^\bullet + 1, \ldots,
\end{aligned}$$

where

$$\ddot{p}_{n,m} = \frac{\ddot{p}_n}{1 - \ddot{p}_m}, \quad n = m^\bullet + 1, \ldots, \tag{8}$$

as defined in (A1), Appendix 1, which is dedicated to the truncated Poisson distribution.

Consequently, the unconditional distribution of $\mathcal{F}_j, j = z+1, \ldots, \ell$, when the hypothesis $H_{0,j}$ holds, will be given by, e.g. [12,20],

$$\begin{aligned}\overline{\overline{F}}_j(z) &= \sum_{n=m^\bullet+1}^{\infty} pr(\ddot{N} = n \,|\, \ddot{N} > m^\bullet) F(z \,|\, f_j, g(n)) \\ &= \sum_{n=m^\bullet+1}^{\infty} \ddot{p}_{n,m^\bullet} F(z \,|\, f_j, g(n)), \quad j = z+1, \ldots, \ell. \end{aligned} \tag{9}$$

4. An application to real data

In this section, we apply the proposed methodology to a dataset from patients affected by cancer. The data was collected from the U.S. Cancer Statistics Working Group [24] according to official guidelines and refer to the age of disease detection in 2009. We compare the results obtained using our approach and the common ANOVA.

We will consider a mixed model with one fixed and one random effects factors. The fixed effects factor will be the *Gender*, with two levels (*Male* and *Female*). Due to the large number of cancer types we resorted to the simple random sampling method to select three different types of cancer from the available list. Thus the random effects factor will be the *Type of Cancer* and the selected types constitute a random sample.

Table 1 illustrates the types of cancer which have been selected, the number of patients and the mean ages at the time of disease detection. This leads to $m = 2 \times 3 = 6$ different treatments. The global frequencies of these three types of cancer, for males and females, are provided in Appendix 2.

According to (3), in this particular example we have

$$\boldsymbol{Y}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \tag{10}$$

where $\boldsymbol{\beta}_0$ is fixed and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are random, independent, corresponding, respectively, to the random effects factor (*Type of cancer*) and interaction between the two factors. We have the design matrices

$$\begin{aligned}\boldsymbol{X}_0 &= \boldsymbol{I}_2 \otimes \mathbf{1}_3 \\ \boldsymbol{X}_1 &= \mathbf{1}_2 \otimes \boldsymbol{I}_3 \\ \boldsymbol{X}_2 &= \boldsymbol{I}_2 \otimes \boldsymbol{I}_3, \end{aligned}$$

where $\otimes$ denotes the Kronecker product, and

$$\boldsymbol{M}_1 = \boldsymbol{J}_2 \otimes \boldsymbol{I}_3$$

Table 1. Number of patients and sample mean ages.

	Number of patients		Sample means	
Type of cancer	Male	Female	Male	Female
Stomach (digestive system)	44	30	70.523	68.833
Melanomas of the skin	134	99	63.791	57.303
Non-Hodgkin lymphoma	123	105	63.382	66.286

$$\boldsymbol{M}_2 = \boldsymbol{I}_2 \otimes \boldsymbol{I}_3.$$

Let's assume that

$$\begin{aligned}\boldsymbol{M}_1 &= 2\boldsymbol{K}_1\\ \boldsymbol{M}_2 &= \boldsymbol{K}_1 + \boldsymbol{K}_2,\end{aligned}$$

which means that

$$\begin{aligned}\boldsymbol{K}_1 &= \tfrac{1}{2}\boldsymbol{M}_1 = \tfrac{1}{2}\boldsymbol{J}_2 \otimes \boldsymbol{I}_3\\ \boldsymbol{K}_2 &= \boldsymbol{M}_2 - \tfrac{1}{2}\boldsymbol{M}_1 = \left(\boldsymbol{I}_2 - \tfrac{1}{2}\boldsymbol{J}_2\right) \otimes \boldsymbol{I}_3\end{aligned}$$

and consequently the matrices $\boldsymbol{A}_j$, $j=1,2$, will be given by

$$\boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0\\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0\\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0\\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0\\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

and

$$\begin{aligned}\boldsymbol{A}_1\boldsymbol{X}_0 &= \tfrac{1}{\sqrt{2}}\mathbf{1}_2' \otimes \mathbf{1}_3\\ \boldsymbol{A}_2\boldsymbol{X}_0 &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \mathbf{1}_3.\end{aligned}$$

The matrices $\boldsymbol{Q}_j$, $j=1,2$, which are the OPM on $R(\boldsymbol{A}_j\boldsymbol{X}_0)^{\perp}$, $j=1,2$, will be given by

$$\begin{aligned}\boldsymbol{Q}_1 &= \boldsymbol{W}_1'\boldsymbol{W}_1 = \boldsymbol{I}_3 - \tfrac{1}{3}\boldsymbol{J}_3\\ \boldsymbol{Q}_2 &= \boldsymbol{W}_2'\boldsymbol{W}_2 = \boldsymbol{I}_3 - \tfrac{1}{3}\boldsymbol{J}_3,\end{aligned}$$

with $\boldsymbol{J}_r = \mathbf{1}_r\mathbf{1}_r'$ and

$$\boldsymbol{W}_1 = \boldsymbol{W}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Moreover, $f_1 = \operatorname{rank}(\boldsymbol{Q}_1) = 3$ and $f_2 = \operatorname{rank}(\boldsymbol{Q}_2) = 3$. Besides this, the OPM on $R(\boldsymbol{A}_j\boldsymbol{X}_0)$, $j=1,2$, are

$$\boldsymbol{P}_1 = \boldsymbol{A}_1\boldsymbol{X}_0(\boldsymbol{A}_1\boldsymbol{X}_0)^{+} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

$$\boldsymbol{P}_2 = \boldsymbol{A}_2\boldsymbol{X}_0(\boldsymbol{A}_2\boldsymbol{X}_0)^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

We will test the hypotheses

$$H_{0,j} : \gamma_j = 0, \quad j = 1, 2,$$

which are the hypotheses of absence of random effects and interaction between the two factors.

Given $\ddot{N} = n$, when $H_{0,j}$, $j=1,2$ holds, the conditional distribution of

$$\mathcal{F}_j = \frac{g(n)}{3} \frac{(\boldsymbol{Y}_j^{\bullet})'(\boldsymbol{B}_j^{-1})\boldsymbol{Y}_j^{\bullet}}{S}, \quad j = 1, 2$$

is a central F distribution with $f_j = \text{rank}(\boldsymbol{Q}_j) = 3$, $j=1,2$, and $g(n) = n - 6$ degrees of freedom, $F(\cdot \mid 3, n-6)$.

In the calculations, we assume that

$$\sum_{n=0}^{m^{\bullet}} \ddot{p}_{n,m^{\bullet}} \simeq 0,$$

which means that, with high probability, we have $\ddot{N} > m^{\bullet}$, so $m^{\bullet} + 1$ is the global minimum dimension for the samples. Therefore the unconditional distribution of the statistics will be given by

$$\overline{\overline{F}}_j(z) = \sum_{n=m^{\bullet}+1}^{\infty} \ddot{p}_{n,m^{\bullet}} F(z \mid 3, n-6), \quad j = 1, 2. \tag{11}$$

Besides this, due to the monotony property of the F distribution [12], when $n < n^o$, we have

$$F(z \mid 3, n-6) < F(z \mid 3, n^o - 6), \tag{12}$$

so that

$$F(z \mid 3, m^{\bullet} + 1 - 6) \leq \overline{\overline{F}}_j(z) \leq 1$$

which gives us a lower bound for $\overline{\overline{F}}_j(z)$. Thus, from $F(z \mid 3, m^{\bullet} - 5)$, we can obtain upper bounds for the quantiles of the unconditional distributions $\overline{\overline{F}}_j(z)$, $j=1,2$. If we use these upper bounds as critical values, we will have tests with sizes that do not exceed the theoretical values.

Remarks:

- We can use these upper bounds for a preliminary test. If the test statistic exceeds the upper bound it also exceeds the real critical value (obtained when using the unconditional distribution). For the cases when the test statistic is lower than the upper bound

one must compute the critical value solving the equation $\overline{\overline{F}}_j(z) = 1-\alpha$, for $z, j = 1,2$. To solve it we may truncate the series in Equation (11) according to the rule established in [11,19]. This way, restricting the sum to the term $\overline{m} = \sum_{i=1}^{m} \overline{m}_i$, with $n_i \leq \overline{m}_i$, where n_i are the realizations of the $\ddot{N}_i$, $i = 1, \ldots m$, we will have

$$\overline{\overline{F}}_{j,\overline{m}}(z) = \sum_{n=m^{\bullet}+1}^{\overline{m}} \ddot{p}_{n,m^{\bullet}} F(z \mid 3, n-6), \quad i = 1, 2.$$

Considering ϵ small, we choose each $\overline{m}_i$ such that

$$\sum_{n_i=0}^{\overline{m}_i} \mathrm{e}^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!} > 1-\epsilon \Leftrightarrow \epsilon > 1 - \sum_{n_i=0}^{\overline{m}_i} \mathrm{e}^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!}, \quad i = 1, \ldots, m. \tag{13}$$

This inequality will be used to obtain the minimum value of $\overline{m}$ needed to $\overline{\overline{F}}_{j,\overline{m}}(z)$ be a good approximation for the distribution $\overline{\overline{F}}_j(z)$, $i = 1,2$, [11].

- Usually the analysis starts with a test of interaction and follows with the tests to the main effects whenever it is not significant. We do not follow this approach since we are interested in showing how these tests could be carried out through unconditioning [20].

4.1. Random effects factor

For the *second factor*, we have

$$\boldsymbol{Y}_1^{\bullet} = \boldsymbol{W}_1 \boldsymbol{A}_1 \boldsymbol{L}^{+} \boldsymbol{Y} = \begin{bmatrix} 1.1255 \\ -1.8846 \end{bmatrix},$$

where

$$\boldsymbol{L}^{+} = D\left(\tfrac{1}{44}\mathbf{1}'_{44}, \tfrac{1}{30}\mathbf{1}'_{30}, \tfrac{1}{134}\mathbf{1}'_{134}, \tfrac{1}{99}\mathbf{1}'_{99}, \tfrac{1}{123}\mathbf{1}'_{123}, \tfrac{1}{105}\mathbf{1}'_{105}\right),$$

with $\boldsymbol{L}^{+}\boldsymbol{Y}$ the vector of the sample means with components 70.523, 68.833, 63.791, 57.303, 63.382, 66.286 and

$$\boldsymbol{B}_1 = \boldsymbol{W}_1 \boldsymbol{A}_1 \boldsymbol{L}^{+} (\boldsymbol{L}^{+})' \boldsymbol{A}_1' \boldsymbol{W}_1' = \begin{bmatrix} 0.012453695 & -0.002286565 \\ -0.002286565 & 0.017972370 \end{bmatrix}$$

So, for the numerator of the statistic $\mathcal{F}_1$ we obtain

$$(\boldsymbol{Y}_1^{\bullet})' (\boldsymbol{B}_1^{-1}) \boldsymbol{Y}_1^{\bullet} = 262.120.$$

When $\ddot{N} = n$, $S = \| Y_{\overline{\Omega}^{\perp}} \|$ is the product by σ^2 of a central chi-square with $g(n) = n-6$ degrees of freedom, $\sigma^2 \chi^2_{n-6}$. In this case, we obtained $S = 131250.672$.

Therefore, the statistic's value, $\mathcal{F}_{1,Obs}$, is given by

$$\mathcal{F}_{1,Obs} = \tfrac{529}{3} \tfrac{262.120}{131250.672} = 0.352.$$

If we use the common conditional distribution of $\mathcal{F}_1$, which corresponds to $F(z \mid 3, 529)$, since $n = 535$, we will obtain the quantiles given in Table 2.

Table 2. The quantiles of the conditional distribution.

Values of α	0.1	0.05	0.01
$z_{1-\alpha}$	2.094	2.622	3.819

Table 3. Upper bounds for the quantiles.

	Values of α	0.1	0.05	0.01
$z^u_{1-\alpha}$	$m^\bullet = 11$	3.289	4.757	9.779
	$m^\bullet = 15$	2.728	3.708	6.552
	$m^\bullet = 18$	2.560	3.410	5.739

Table 4. The quantiles of the truncated unconditional distribution.

	Values of α	0.1	0.05	0.01
$z^t_{1-\alpha}$	$m^\bullet = 11$	3.255	4.693	9.583
	$m^\bullet = 15$	2.720	3.695	6.518
	$m^\bullet = 18$	2.555	3.402	5.722

So, since $\mathcal{F}_{1,Obs} < z_{1-\alpha}$, we do not reject $H_{0,1}$ for the usual levels of significance.

Let's assume that we have 12 [16 and 19] observations as global minimum dimensions for the samples, which means that we consider $m^\bullet + 1 = 12 \Leftrightarrow m^\bullet = 11[m^\bullet = 15$ and $m^\bullet = 18]$. Table 3 shows the upper bounds for the quantiles with probability $1 - \alpha$, $z^u_{1-\alpha}$, of the unconditional distribution $\overline{\overline{F}}_1(z)$.

It is to be expected that the quantiles for random sample sizes (obtained when using the unconditional distribution) to exceed the classical ones (obtained when using common conditional distribution), since the first ones take into account a new source of variation. Then, since in this case we do not reject the hypothesis using the classical quantiles the same result is expected when using the quantiles for random sample sizes and consequently the upper bound approach. This interpretation leads us to not reject $H_{0,1}$.

The quantiles for the unconditional distribution are approximated by truncation of the infinite series indicated in Equation (11). We obtained the minimum value $\overline{m} = 38$ for a truncation error not greater than 10^{-8} ($\epsilon \leq 10^{-8}$). To carry out the computation, we assumed that λ_i, $i = 1, \ldots, 6$, are the daily average of occurrences per year. So we have $\lambda_1 = 0.13$, $\lambda_2 = 0.09$, $\lambda_3 = 0.37$, $\lambda_4 = 0.28$, $\lambda_5 = 0.34$, $\lambda_6 = 0.29$.

The obtained quantiles with probability $1 - \alpha$, $z^t_{1-\alpha}$, of the truncated unconditional distribution

$$\overline{\overline{F}}_{1,\overline{m}}(z) = \sum_{n=m^\bullet+1}^{38} \ddot{p}_{n,m^\bullet} F(z \mid 3, n-6) \tag{14}$$

are presented in Table 4.

Results in Table 4 agree with those in Table 3, i.e. $H_{0,1}$ is not rejected therefore the random factor is not significant.

Table 5. Minimum value $m^\bullet$ that leads to reject the hypothesis $H_{0,2}$.

Values of $1-\alpha$	0.1	0.05	0.01
$m^\bullet$	8	9	15

4.2. Interaction

For the *interaction* between the fixed factor and the random one, we have

$$\boldsymbol{Y}_2^\bullet = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^+\boldsymbol{Y} = \begin{bmatrix} 7.8572 \\ 0.0512 \end{bmatrix}$$

and

$$\boldsymbol{B}_2 = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}_2'\boldsymbol{W}_2' = \begin{bmatrix} 0.012453695 & -0.002286565 \\ -0.002286565 & 0.017972370 \end{bmatrix}.$$

For the numerator of the statistic $\mathcal{F}_2$, we obtain

$$(\boldsymbol{Y}_2^\bullet)'(\boldsymbol{B}_2^{-1})\boldsymbol{Y}_2^\bullet = 5084.346.$$

Therefore, the statistic's value, $\mathcal{F}_{2,Obs}$, is given by

$$\mathcal{F}_{2,Obs} = \frac{529}{3}\frac{5084.346}{131250.672} = 6.831.$$

If we use the common conditional distribution of $\mathcal{F}_2$, which corresponds to $F(z \mid 3, 529)$, we obtain the quantiles given in Table 2. Since $\mathcal{F}_{2,Obs} > z_{1-\alpha}$, we reject $H_{0,2}$ for the usual levels of significance.

Considering the truncated unconditional distribution, $\overline{\overline{F}}_{2,\overline{m}}$, which correspond to $\overline{\overline{F}}_{1,\overline{m}}$ defined in (14), we obtained the quantiles, $z^t_{1-\alpha}$, given in Table 4. The results in this table lead us to:

- reject $H_{0,2}$ for $\alpha = 0.1$ and 0.05 and do not reject for $\alpha = 0.01$, considering $m^\bullet + 1 = 12$;
- reject $H_{0,2}$ for the usual level of significance, considering $m^\bullet + 1 = 16$ or 19.

Table 3 shows the upper bounds for the quantiles with probability $1-\alpha$, $z^u_{1-\alpha}$, of the unconditional distribution. These results agree with those based on the quantiles of the truncated unconditional distribution. Assuming the values of the test statistic remain unchanged, then we should have the total sample sizes presented in Table 5 for ensuring rejection.

Since for higher values of $m^\bullet$ we would get lower values for the quantiles, we have $\mathcal{F}_{Obs,2} > z^u_{1-\alpha}$ for all $m^\bullet \geq 15$. In this case, we reject $H_{0,2}$ considering the usual levels of significance, which means that the interaction between factors is significant.

4.3. Conclusion

Our discussion shows the relevance of the unconditional approach in avoiding false rejections. As we saw, the inference results for some situations depends on the approach. Since

the unconditional approach is more secure, when testing the interaction the null hypothesis is not rejected when $m^{\bullet} = 11$ and $\alpha = 0.01$, whereas the common conditional approach would lead to a false rejection.

The results in Tables 3 and 4 show that for higher minimum sample sizes, we get smaller upper bounds and quantiles of the unconditional distribution. Due to this, we may conclude that with the increase of the minimum sample sizes, the decision based on both approaches is similar.

To finish we would like to note that all the computations were performed using the R `software`.

5. Final remarks

The approach followed in this paper is more realistic than the usual F tests for the situations where it is not possible to known in advance the sample sizes. To do that, we have to make assumptions regarding the distribution of the sample sizes based on previous knowledge of the sample collection and incorporate this source of variation into the mixed model. We choose the Poisson distribution since it would correspond to Poisson processes for observation collection and the underlying assumption for these (independent and stable increments and not clustering) seems realist. Moreover, the L extensions fit easily in the assumption of random sample sizes. These model formulation have been used to solve the unbalance originated by different number of observations per treatment, which cause non-orthogonality in fixed and mixed effects models. We included an application with cancer data to illustrate how straightforward it is to apply our approach in a medical context. The comparative results show that when random sample sizes are considered the critical values may exceed those of classical ANOVA (obtained when using the common F conditional distribution). So, we can conclude that this approach avoids working with incorrect critical values and thus carrying out tests without the proper level. We would like also to highlight that our methodology is not restricted to the medical domain and yet may be applied to several other research areas.

Acknowledgments

The authors would like to thank the anonymous referees for useful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially supported by the FCT- Fundação para a Ciência e Tecnologia, under the projects UID/MAT/00212/2019 and UID/MAT/00297/2019.

References

[1] R.A. Bailey, S.S. Ferreira, D. Ferreira, and C. Nunes, *Estimability of variance components when all model matrices commute*, Linear Algebra Appl. 492 (2016), pp. 144–160.

[2] F. Carvalho, J.T. Mexia, C. Santos, and C. Nunes, *Inference for types and structured families of commutative orthogonal block structures*, Metrika 78 (2015), pp. 337–372.

[3] S. Ferreira, D. Ferreira, E. Moreira, and J.T. Mexia, *Inference for L orthogonal models*, J. Interdiscip. Math. 12 (2009), pp. 815–824.
[4] S.S. Ferreira, D. Ferreira, C. Nunes, and J.T. Mexia, *Estimation of variance components in linear mixed models with commutative orthogonal block structure*, Rev. Colomb. Estadist. 36 (2013), pp. 261–271.
[5] F. Heinzl and G. Tutz, *Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm*, Stat. Modell. 13 (2013), pp. 41–67.
[6] A.M. Houtman and T.P. Speed, *Balance in designed experiments with orthogonal block structure*, Ann. Statist. 11 (1983), pp. 1069–1085.
[7] N.L. Johnson and S. Kotz, *Discrete Distributions*, John Wiley & Sons, New York, 1969.
[8] A.I. Khuri, T. Mathew, and B.K. Sinha, *Statistical Tests for Mixed Linear Models*, John Wiley & Sons, New York, 1998.
[9] E.L. Lehmann, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959.
[10] J.T. Mexia, *Best linear unbiased estimates, duality of F tests and the Scheffé multiple comparison method in presence of controlled heterocedasticity*, Comput. Stat. Data Anal. 10 (1990), pp. 271–281.
[11] J.T. Mexia and E. Moreira, *Randomized sample size F tests for the one-way layout.* 8th International Conference on Numerical Analysis and Applied Mathematics 2010. AIP Conf. Proc. 1281(II), 2010, pp. 1248–1251.
[12] J.T. Mexia, C. Nunes, D. Ferreira, S.S. Ferreira, and E. Moreira, *Orthogonal fixed effects ANOVA with random sample sizes*, Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling (ASM'11), 2011, pp. 84–90
[13] A. Michalski and R. Zmyślony, *Testing hypothesis for variance components in mixed linear models*, Statistics 27 (1996), pp. 297–310.
[14] E. Moreira, J.T. Mexia, M. Fonseca, and R. Zmyślony, *L models and multiple regressions designs*, Statist. Papers 50 (2009), pp. 869–885.
[15] E.E. Moreira, J.T. Mexia, and C.E. Minder, *F tests with random sample size. Theory and applications*, Stat. Probab. Lett. 83 (2013), pp. 1520–1526.
[16] C. Nunes, G. Capistrano, D. Ferreira, S.S. Ferreira, and J.T. Mexia, *One-way fixed effects ANOVA with missing observations*, Proceedings of the 12th International Conference on Numerical Analysis and Applied Mathematics, AIP Conf. Proc. 1648, 2015, p. 110008.
[17] C. Nunes, G. Capristano, D. Ferreira, S.S. Ferreira, and J.T. Mexia, *Exact critical values for one-way fixed effects models with random sample sizes*, J. Comput. Appl. Math. 354 (2019), pp. 112–122. doi:10.1016/j.cam.2018.05.057.
[18] C. Nunes, D. Ferreira, S.S. Ferreira, and J.T. Mexia, *F Tests with Random Sample Sizes*. 8th International Conference on Numerical Analysis and Applied Mathematics. AIP Conf. Proc. 1281(II), 2010, pp. 1241–1244
[19] C. Nunes, D. Ferreira, S.S. Ferreira, and J.T. Mexia, *F-tests with a rare pathology*, J. Appl. Stat. 39 (2012), pp. 551–561.
[20] C. Nunes, D. Ferreira, S.S. Ferreira, and J.T. Mexia, *Fixed effects ANOVA: An extension to samples with random size*, J. Stat. Comput. Simul. 84 (2014), pp. 2316–2328.
[21] H. Scheffé, *The Analysis of Variance*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 1959.
[22] J.R. Schott, *Matrix Analysis for Statistics*, John Wiley & Sons, New York, 1997.
[23] S.R. Searle, G. Casella, and C.E. McCulloch, *Variance Components*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 1992.
[24] U.S. Cancer Statistics Working Group, United States Cancer Statistics: 1999–2010 Incidence and Mortality Web-based Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, 2013. Available at https://nccd.cdc.gov/uscs/.

Appendices

Appendix 1. Truncated Poisson distributions

This appendix presents some results about the truncated Poisson distribution, which are useful in obtaining the unconditional distribution of the test statistics.

Since we need to have at least one observation per treatment, we will consider the common form of truncated Poisson distribution, which corresponds to the omission of the zero class, e.g. [7]. So we have $N_i \geq 1,\ i = 1, \ldots, m$. To perform inference, we also consider that $N > m$, where $N = \sum_{i=1}^{m} N_i$.

As previously mentioned, we assumed that $N_i \sim P(\lambda_i)$, $i = 1, \ldots, m$ and $N \sim P(\lambda)$. So we have

$$p_{r,i} = pr(N_i = r \,|\, N_i \geq 1) = \frac{pr(N_i = r)}{pr(N_i \geq 1)} = \frac{\mathrm{e}^{-\lambda_i}\lambda_i^r/r!}{1 - \mathrm{e}^{-\lambda_i}} = \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}\frac{\lambda_i^r}{r!}, \quad r \geq 1,\ i = 1, \ldots, m.$$

Therefore, the moment generating function of N_i, when $N_i \geq 1$, $i = 1, \ldots, m$, will be

$$\varphi_i(u) = \sum_{r=1}^{\infty} \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}\frac{\lambda_i^r\,\mathrm{e}^{ru}}{r!} = \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}(\mathrm{e}^{\lambda_i\,\mathrm{e}^u} - 1), \quad i = 1, \ldots, m,$$

and the probability generating functions

$$\chi_i(z) = \varphi_i(\ln z) = \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}(\mathrm{e}^{\lambda_i z} - 1), \quad i = 1, \ldots, m.$$

With $\ddot{N}_i$, $i = 1, \ldots, m$, the truncated variables N_i, $i = 1, \ldots, m$, when $N_i \geq 1$, and considering

$$\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i,$$

we will obtain the probability generating function

$$\begin{aligned}\ddot{\chi}(z) &= \prod_{i=1}^{m} \chi_i(z) = \left(\prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}\right)\prod_{i=1}^{m}(\mathrm{e}^{\lambda_i z} - 1) \\ &= \left(\prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}\right)\sum_{\mathcal{C}\subseteq\overline{\overline{m}}}(-1)^{m-\sharp(\mathcal{C})}\,\mathrm{e}^{(\sum_{i\in\mathcal{C}}\lambda_i)z}, \quad i = 1, \ldots, m,\end{aligned}$$

where $\overline{\overline{m}} = \{1, \ldots, m\}$ and $\sharp(\mathcal{C})$ denotes the cardinal of $\mathcal{C}$, any subset of $\overline{\overline{m}}$.

Therefore we will have

$$\ddot{p}_r = pr(\ddot{N} = r) = \frac{1}{r!}\left(\prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda_i}}{1 - \mathrm{e}^{-\lambda_i}}\right)\sum_{\mathcal{C}\subseteq\overline{\overline{m}}}(-1)^{m-\sharp(\mathcal{C})}\left(\sum_{i\in\mathcal{C}}\lambda_i\right)^r, \quad r = m, \ldots.$$

It is interesting to observe that we have

$$\ddot{\chi}^{\langle s\rangle}(0) = 0, \quad s = 1, \ldots, m-1,$$

where $\langle s\rangle$ denotes the derivative of order s, which results from

$$j_1 + \cdots + j_m = s; \quad s = 1, \ldots, m-1,$$

$\boldsymbol{j} = (j_1, \ldots, j_m)'$ have one or more null components and $\chi_i(0) = 0$, $i = 1, \ldots, m$.

Indeed, with $\mathcal{P}_s^{(m)}$ the family of partitions of s with cardinal m, we have

$$\ddot{\chi}^{\langle s\rangle}(0) = \sum_{\boldsymbol{j}\in\mathcal{P}_s^{(m)}} \frac{(\sum_{i=1}^{m} j_i)!}{\prod_{i=1}^{m} j_i!}\prod_{i=1}^{m} \chi_i^{<j_i>}(0), \quad s = 1, \ldots$$

and, if $s < m$, whatever $\boldsymbol{j} \in \mathcal{P}_s^{(m)}$,

$$\prod_{i=1}^{m} \chi_i^{\langle j_i \rangle}(0) = 0$$

since $\boldsymbol{j}$ has at least one null component. So, since $\ddot{\chi}^{<s>}(0) = s!\ddot{p}_s$, we obtain

$$\ddot{p}_s = \frac{1}{s!}\ddot{\chi}^{<s>}(0) = 0, \quad s \leq m-1.$$

Furthermore, the only non-null term of $\ddot{\chi}^{<m>}(0)$ corresponds to $\boldsymbol{j} = \mathbf{1}_m$, so

$$\ddot{p}_m = pr(\ddot{N} = m) = \frac{1}{m!}\ddot{\chi}^{<m>}(0) = \prod_{i=1}^{m} \chi_i^{<1>}(0) = \prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda_i}\lambda_i}{1-\mathrm{e}^{-\lambda_i}}$$

and

$$pr(\ddot{N} > m) = 1 - \ddot{p}_m = 1 - \prod_{i=1}^{m} \frac{\mathrm{e}^{-\lambda_i}\lambda_i}{1-\mathrm{e}^{-\lambda_i}}.$$

Considering $\ddot{\boldsymbol{N}}$ the random vector with components $\ddot{N}_1, \ldots, \ddot{N}_m$, we have $\ddot{N} > m$, which means there exists at least one $\ddot{N}_i > 1,\ i = 1, \ldots, m$, if and only if $\ddot{\boldsymbol{N}} > \mathbf{1}_m$ so

$$pr(\ddot{\boldsymbol{N}} > \mathbf{1}_m) = 1 - \ddot{p}_m.$$

We also have

$$\ddot{p}_{r,m} = pr(\ddot{N} = r \mid \ddot{N} > m) = \frac{\ddot{p}_r}{1-\ddot{p}_m}, \quad r = m+1, \ldots. \tag{A1}$$

Appendix 2. Frequency tables of types of cancer

Table A1. Males with stomach (digestive system) cancer.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	0	0	0	0	0	0	1
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	1	2	4	5	6	7	7	6	5

Table A2. Females with stomach (digestive system) cancer.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	0	0	0	0	0	1	1
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	2	2	2	3	3	3	4	4	5

Table A3. Males with melanomas of the skin.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	0	0	1	2	2	4	6
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	8	12	14	17	16	16	14	12	10

Table A4. Females with melanomas of the skin.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	0	1	2	4	4	6	7
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	10	10	10	10	8	7	7	6	7

Table A5. Males with non-Hodgkin lymphoma.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	1	1	1	2	2	3	5
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	8	10	12	14	15	14	14	12	9

Table A6. Males with non-Hodgkin lymphoma.

Age	1–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Mean age	2	7	12	17	22	27	32	37	42
Patients	0	0	0	1	1	1	2	2	3
Age	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85+
Mean age	47	52	57	62	67	72	77	82	87
Patients	5	7	9	11	13	13	13	12	12