Journal of Computational and Applied Mathematics 354 (2019) 112–122

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Exact critical values for one-way fixed effects models with random sample sizes☆

Célia Nunes [a,*], Gilberto Capistrano [b], Dário Ferreira [a], Sandra S. Ferreira [a], João Tiago Mexia [c]

[a] Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal
[b] School of Business and Development of Excellence-ENDEX, Pouso Alegre, Brazil
[c] Center of Mathematics and its Applications, Faculty of Science and Technology, New University of Lisbon, Portugal

ARTICLE INFO

Article history:
Received 18 September 2017
Received in revised form 19 March 2018

MSC:
62J12
62J10
62J99

Keywords:
ANOVA
Random sample sizes
Fixed effects models
Correct critical values
Cancer registries

ABSTRACT

Analysis of variance (ANOVA) is one of the most frequently used statistical analyses in several research areas, namely in medical research. Despite its wide use, it has been applied assuming that sample dimensions are known. In this work we aim to carry out ANOVA like analysis of one-way fixed effects models, to situations where the samples sizes may not be previously known. In these situations it is more appropriate to consider the sample sizes as realizations of independent random variables. This approach must be based on an adequate choice of the distributions of the samples sizes. We assume the Poisson distribution when the occurrence of observations corresponds to a counting process. The Binomial distribution is the proper choice if we have observations failures and there exist an upper bound for the sample sizes. We also show how to carry out our main goal by computing correct critical values. The applicability of the proposed approach is illustrated considering a real data example on cancer registries. The results obtained suggested that false rejections may be avoided by applying our approach.

1. Introduction

Analysis of variance (ANOVA) is one of the most frequently used statistical analyses in practical applications. It is routinely used in several research areas, namely in medical research. Usually, it has been applied assuming that sample dimensions are known. However in many relevant situations we not known beforehand these dimensions. This often occurs when there is a fixed time span for collecting the observations. A motivation example is the collection of data from patients with several pathologies arriving at a hospital during a fixed time period, see e.g. [1,2]. In this work we show how this may be overcome when we carry out ANOVA for one-way fixed effects models.

In these situations it is more appropriate, assuming there are m different levels, to consider the sample sizes as realizations, $n_1, \ldots, n_m$, of independent random variables, $N_1, \ldots, N_m$, [1–6]. By following this methodology we avoid the assumption of previously known the sample dimensions which renders our approach more realistic.

This new approach must be based on an adequate choice of the distribution of $N_1, \ldots, N_m$. These distributions are discrete with probability points as non negative integers. There are two families of such distributions, according to the non existence or existence of an upper bound for the sample sizes. Starting with no upper bound for the sample sizes we consider that we

☆ Selected papers of CMMSE-2017, Cadiz, Spain.

* Corresponding author.
E-mail address: celian@ubi.pt (C. Nunes).

https://doi.org/10.1016/j.cam.2018.05.057
0377-0427/

have counting processes. Namely we assume that the numbers collected in non overlapping intervals are independent and simultaneous arrivals are not to be expected. We are thus led to consider, possibly non homogeneous, Poisson counting processes. So for fixed collection periods our sample sizes will have Poisson distribution. Going over to the cases with upper bounds for sample sizes we use the Binomial distribution, which would correspond to samples collected in situations when there is a probability p of an observation failing. We assumed this probability to be the same for all treatments.

We are interested in obtaining the critical values for testing the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_m,$$

which may be rewritten as

$$H_0 : \boldsymbol{A}\boldsymbol{\mu} = \mathbf{0}, \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector of the treatment means with components $\mu_1, \ldots, \mu_m$, and $\boldsymbol{A} = [\boldsymbol{I}_{m-1} | -\mathbf{1}_{m-1}]$, with $\boldsymbol{I}_c$ the $c \times c$ identity matrix and $\mathbf{1}_c$ the vector with c components equal to 1.

This paper is structured as follows. Section 2 considers the two mentioned distributions for the samples sizes. Section 3 presents the test statistics and their conditional and unconditional distributions, under the assumption that we have random sample sizes. The presented approach is illustrated through an application on real medical data, using cancer registries, in Section 4. In Section 5 we show how to carry out our main goal by computing correct critical values. Section 6 presents the results of a simulation study, comparing and relating the performance of our approach with the common ANOVA. We conclude this work in Section 7, with some closing remarks.

2. Distributions of the sample sizes

In this section we consider two cases for the distributions of the sample sizes. These will be used to obtain the unconditional distributions of the statistics.

First we will assume that the occurrences of observations correspond to counting processes, leading us to consider the sample sizes, $N_1, \ldots, N_m$, as Poisson distributed. Then we will deal with situations when failures may occur on collections of observations and there exist the upper bounds for the sample sizes, inducing us to consider the Binomial distribution.

To avoid the existence of samples without observations and other highly unbalanced cases we assume minimums values for each samples dimension. In previous papers, see e.g. [2,3] and [6], it was only considered a minimum value for the global sample size.

2.1. Counting processes

Let us assume that the occurrence of the observations corresponds to counting processes. An illustrative example of this is the collection of observations during a fixed time period in a study comparing, for example, several pathologies of patients arriving at a hospital. The number of patients for each pathology is not known in advance and the replication of the data collection during a different time period, of the same length, would result in a sample of different size. Another example is the approach presented in [1] where one of the pathologies is rare.

In these situations it is more appropriate to assume that the sample dimensions, $N_1, \ldots, N_m$, have Poisson distributions with parameters $\lambda_1, \ldots, \lambda_m$. We put $N_i \sim P(\lambda_i)$, $i = 1, \ldots, m$. Moreover $n = \sum_{i=1}^{m} n_i$ will be a realization of the random variable

$$N = \sum_{i=1}^{m} N_i$$

which, given the independence of N_i, $i = 1, \ldots, m$,

$$N \sim P(\lambda),$$

with $\lambda = \sum_{i=1}^{m} \lambda_i$. Furthermore the vector $\boldsymbol{n} = (n_1, \ldots, n_m)'$ will be a realization of $\boldsymbol{N} = (N_1, \ldots, N_m)'$.

For carrying out the inference we will assume that $N_i \geq n_i^{\bullet}$, $i = 1, \ldots, m$, which means that we have a minimum dimension for each sample. In this case the global minimum dimension will be $n^{\bullet} = \sum_{i=1}^{m} n_i^{\bullet}$ and $\boldsymbol{n}^{\bullet} = (n_1^{\bullet}, \ldots, n_m^{\bullet})'$. So, since we have m different treatments and considering all possible partitions of n into $n_1, \ldots, n_m$, we take

$$p_{\boldsymbol{n}^{\bullet}}(n) = pr(N = n | \boldsymbol{N} \geq \boldsymbol{n}^{\bullet}) = \sum_{n_1 = n_1^{\bullet}}^{n - \sum_{i=2}^{m} n_i^{\bullet}} \cdots \sum_{n_\ell = n_\ell^{\bullet}}^{n - (\sum_{i=1}^{\ell-1} n_i + \sum_{i=\ell+1}^{m} n_i^{\bullet})} \cdots$$

$$\sum_{n_m = n - \sum_{i=1}^{m-1} n_i}^{n - \sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^{\bullet}), \; n_i = n_i^{\bullet}, \ldots, \; i = 1, \ldots, m, \tag{2}$$

where, through the independence of N_i, $i = 1, \ldots, m$,

$$pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^{\bullet}) = \prod_{i=1}^{m} pr(N_i = n_i | N_i \geq n_i^{\bullet})$$
$$= \prod_{i=1}^{m} \frac{pr(N_i = n_i)}{pr(N_i \geq n_i^{\bullet})} = \prod_{i=1}^{m} \frac{e^{-\lambda_i}(\lambda_i^{n_i}/n_i!)}{1 - \sum_{u_i=0}^{n_i^{\bullet}-1} e^{-\lambda_i}(\lambda_i^{u_i}/u_i!)}$$
$$= \prod_{i=1}^{m} \frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i} - \sum_{u_i=0}^{n_i^{\bullet}-1} \frac{\lambda_i^{u_i}}{u_i!})}, \ n_i = n_i^{\bullet}, \ldots, \ i = 1, \ldots, m. \tag{3}$$

2.2. Observations failures

Let us now assume there exist upper bounds for the sample sizes, $r_1, \ldots, r_m$. These upper bounds are not always attained, since we may have observations failures. This situation may happen for instance when

- working with patients and, depending on the disease, there is a probability of having incomplete or absent reports;
- working with grapevines and there is a probability, that may depend on the treatment, some of them wither.

In these cases the Binomial distribution is the proper choice. So we assume that the sample dimensions, $N_1, \ldots, N_m$, have Binomial distributions with parameters $r_1, \ldots, r_m$ and $1 - p$, where p denotes the probability of an observation failure. This probability may be obtained from previous results. We put $N_i \sim B(r_i, 1 - p)$, $i = 1, \ldots, m$. Moreover, according to the reproducibility of Binomial distributions, we have

$$N \sim B\,(r, 1 - p)\,,$$

with $r = \sum_{i=1}^{m} r_i$.

Assuming that $N_i \geq n_i^{\bullet}$, $i = 1, \ldots, m$, and $n^{\bullet} = \sum_{i=1}^{m} n_i^{\bullet}$, we have $p_{\boldsymbol{n}^{\bullet}}(n)$ as defined in (2), with $n_i = n_i^{\bullet}, \ldots, r_i, \ i = 1, \ldots, m$, where $pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^{\bullet})$ is now given by

$$pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^{\bullet}) = \prod_{i=1}^{m} pr(N_i = n_i | N_i \geq n_i^{\bullet}) = \prod_{i=1}^{m} \frac{\binom{r_i}{n_i}(1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i=n_i^{\bullet}}^{r_i} pr(N_i = u_i)}$$
$$= \prod_{i=1}^{m} \frac{\binom{r_i}{n_i}(1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i=n_i^{\bullet}}^{r_i} \binom{r_i}{u_i}(1-p)^{u_i} p^{r_i - u_i}}, \ n_i = n_i^{\bullet}, \ldots, r_i, \ i = 1, \ldots, m. \tag{4}$$

3. Test statistic and their conditional and unconditional distributions

In this section, we start by presenting the test statistic and their conditional distribution (assuming fixed sample sizes). Then we will obtain the unconditional distribution, under the assumption that we have random sample sizes.

When $N_i = n_i$, $i = 1, \ldots, m$, we have the samples $Y_{i,1}, \ldots, Y_{i,n_i}$, $i = 1, \ldots, m$, with averages $\overline{Y}_{i,\bullet}$, $i = 1, \ldots, m$. Assuming that the observations are normal and independent with variance σ^2, when $N_i = n_i$, $i = 1, \ldots, m$, the vector of treatment means, $\boldsymbol{Y}_{\bullet}$, which has components $\overline{Y}_{1,\bullet}, \ldots, \overline{Y}_{m,\bullet}$, will be normal with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\sigma^2 D(\frac{1}{n_1}, \ldots, \frac{1}{n_m})$, where $D(\frac{1}{n_1}, \ldots, \frac{1}{n_m})$ is the diagonal matrix with principal elements $\frac{1}{n_1}, \ldots, \frac{1}{n_m}$.

So, when $N_i = n_i$, $i = 1, \ldots, m$, see for instance [7,8], the sum of squares for testing the null hypothesis, $H_0 : \boldsymbol{A}\boldsymbol{\mu} = 0$, will be

$$S_{num} = (\boldsymbol{A}\boldsymbol{Y}_{\bullet})' \left(\boldsymbol{A} D\left(\frac{1}{n_1}, \ldots, \frac{1}{n_m} \right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{Y}_{\bullet}), \tag{5}$$

which corresponds to the product by σ^2 of a noncentral chi-square with $g = m - 1$ degrees of freedom and non-centrality parameter

$$\delta(n) = \frac{1}{\sigma^2} (\boldsymbol{A}\boldsymbol{\mu})' \left(\boldsymbol{A} D\left(\frac{1}{n_1}, \ldots, \frac{1}{n_m} \right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{\mu}), \tag{6}$$

$S_{num} \sim \sigma^2 \chi^2_{g,\delta(n)}$.

The sum of the sums for the error will be given by, see e.g. [9] and [10],

$$S = \sum_{i=1}^{m}\sum_{k=1}^{n_i}(Y_{i,k} - \overline{Y}_{i,\bullet})^2.$$

Moreover S will be the product by σ^2 of a central chi-square with

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

degrees of freedom, $S \sim \sigma^2\chi^2_{g(n)}$, and will be conditionally independent from S_{num}.

Therefore, when $N = n$, the conditional distribution of

$$\Im = \frac{S_{num}}{S} \tag{7}$$

will be a noncentral $\overline{F}$ distribution, which corresponds to the distribution of the quotient of independent chi-squares with g and $g(n)$ degrees of freedom and non-centrality parameters $\delta(n)$ and 0, denoted by $\overline{F}(.|g, g(n), \delta(n))$.

Moreover, using the mixtures method of Robbins [11] and Robbins and Pitman [12] (see also [13]),

$$\overline{F}(z|g, g(n), \delta(n)) = e^{-\delta(n)/2}\sum_{i=0}^{\infty}\frac{\delta(n)^i}{2^i i!}\overline{F}(z|g + 2i, g(n)),$$

which corresponds to the distribution of the test statistic when the sample sizes are $n_1, \ldots, n_m$.

Given $N = n$, when H_0 holds, $\delta(n) = 0$ and the conditional distribution of $\Im$ will be a central $\overline{F}$ distribution with g and $g(n)$ degrees of freedom, $\overline{F}(z|g, g(n))$.

3.1. Counting processes

Assuming that the occurrences of the observations corresponds to counting processes, when H_0 holds the unconditional distribution of $\Im$, defined by (7), will be given by, see e.g. [2] and [3],

$$\overline{\overline{F}}(z) = \sum_{n=n^\bullet}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) = \sum_{n=n^\bullet}^{\infty} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|g, g(n)), \tag{8}$$

considering $p_{\boldsymbol{n}^\bullet}(n)$ as defined in (2).

When we know that $N \leq \overline{n}$, we may not consider in (8) the terms for $n > \overline{n}$, and we have $\overline{\overline{F}}(z)$ bounded by

$$\overline{\overline{F}}_{\overline{n}}(z) \leq \overline{\overline{F}}(z) \leq \overline{\overline{F}}^*_{\overline{n}}(z), \tag{9}$$

where

$$\overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=n^\bullet}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) \tag{10}$$

and

$$\overline{\overline{F}}^*_{\overline{n}}(z) = \overline{\overline{F}}_{\overline{n}}(z) + \sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet). \tag{11}$$

So $\overline{n}$ denotes the upper bound needed to control the truncation error of the unconditional distribution $\overline{\overline{F}}(z)$. It is important to note that

$$\sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = 1 - \sum_{n=n^\bullet}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet). \tag{12}$$

Let us now consider the noncentral distributions. As we saw, when $N = n$, $S_{num} \sim \sigma^2\chi^2_{g,\delta(n)}$, thus we have

$$\Im \sim \overline{F}(z|g, g(n), \delta(n)),$$

and the corresponding unconditional distribution is given by

$$\overline{\overline{F}}^{\circ}(z) = \sum_{n=n^\bullet}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n), \delta(n)).$$

So we have

$$\overline{\overline{F}}^{\circ}_{\overline{n}}(z) \leq \overline{\overline{F}}^{\circ}(z) \leq \overline{\overline{F}}^{\circ *}_{\overline{n}}(z),$$

Table 1
Number of patients and sample means.

Type of cancer	Number of patients	Sample means
Soft tissues of the thorax	18	49.5000
Intestinal tract	22	61.7727
Nasal cavity	25	62.4000

with

$$\overline{\overline{F}}^{\circ}_{\overline{n}}(z) = \sum_{n=n^{\bullet}}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet})\overline{F}(z|g, g(n), \delta(n))$$

and

$$\overline{\overline{F}}^{\circ *}_{\overline{n}}(z) = \overline{\overline{F}}^{\circ}_{\overline{n}}(z) + \sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}).$$

Let us note that

$$\overline{\overline{F}}^{\circ *}_{\overline{n}}(z) - \overline{\overline{F}}^{\circ}_{\overline{n}}(z) = \overline{\overline{F}}^{*}_{\overline{n}}(z) - \overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}),$$

so we may use the same value $\overline{n}$ that was used for the central case.

3.2. Observations failures

Assuming that we have the upper bounds for the sample sizes, $r_1, \ldots, r_m$, which are not always attained since failures may occur, the unconditional distribution of $\Im$, when H_0 holds, will be given by

$$\overline{\overline{F}}(z) = \sum_{n=n^{\bullet}}^{r} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet})\overline{F}(z|g, g(n)) = \sum_{n=n^{\bullet}}^{r} p_{\boldsymbol{n}^{\bullet}}(n)\overline{F}(z|g, g(n)), \tag{13}$$

with $r = \sum_{i=1}^{m} r_i$ and $p_{\boldsymbol{n}^{\bullet}}(n)$ as defined in (2), where in this case $n_i = n_i^{\bullet}, \ldots, r_i,\ i = 1, \ldots, m$.

4. Application

In this section we evaluate our approach under a real data example. To construct this experiment we resort to a dataset which was provided by the Brazilian National Cancer Institute (INCA) [14]. The dataset gathers information regarding the age of patients with cancer disease. The data considered is from 2010 and refers to the City of São Paulo, Brazil.

Two situations will be considered, first assuming that the entries in the samples correspond to independent counting processes and then assuming that we may have observations failures.

All computational procedures, namely the quantiles of the conditional and unconditional distribution as well as all the computations in Sections 5 and 6, were performed using *R* software.

In our model the factor considered is the *Type of Cancer*, with three levels: *Soft tissues of the thorax*, *Intestinal tract* and *Nasal cavity*. Tables A.1–A.3 in Appendix show the frequencies of these three types of cancers, grouped by age. Table 1 illustrates the number of patients and the sample mean age for each type of cancer.

We will test the hypothesis

$$H_0 : \boldsymbol{A}\boldsymbol{\mu} = 0,$$

with

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

The numerator of the $\Im$ statistic is now given by

$$S_{num} = (\boldsymbol{A}\boldsymbol{Y}_{\bullet})'\left(\boldsymbol{A}D\left(\frac{1}{18}, \frac{1}{22}, \frac{1}{25}\right)\boldsymbol{A}'\right)^{-1}(\boldsymbol{A}\boldsymbol{Y}_{\bullet}),$$

which is, when H_0 holds, the product by σ^2 of a central chi-square with $g = m - 1 = 2$ degrees of freedom, $S_{num} \sim \sigma^2\chi^2_2$. So, we obtain

$$\left(\boldsymbol{A}D\left(\frac{1}{18}, \frac{1}{22}, \frac{1}{25}\right)\boldsymbol{A}'\right)^{-1} = \begin{bmatrix} 13.0154 & -6.0923 \\ -6.0923 & 14.5538 \end{bmatrix} \text{ and } \boldsymbol{A}\boldsymbol{y}_{\bullet} = \begin{bmatrix} -12.9000 \\ -0.6273 \end{bmatrix},$$

Table 2
The quantiles of the conditional distribution.

Values of α	0.1	0.05	0.01
$z_{1-\alpha}$	0.07711	0.10146	0.16016

where $\boldsymbol{y}_\bullet$ has components $\bar{y}_{1,\bullet} = 49.5000$; $\bar{y}_{2,\bullet} = 61.7727$; $\bar{y}_{3,\bullet} = 62.4000$. Therefore for the numerator of the statistic we obtain

$$S_{num} = 2073.021.$$

The denominator of the statistic is, when $N = n$, the product by σ^2 of a central chi-square with $g(n) = n - 3$ degrees of freedom, $S \sim \sigma^2 \chi^2_{n-3}$. In this case we obtain

$$S = \sum_{j=1}^{18}(y_{1,j} - \bar{y}_{1,\bullet})^2 + \sum_{j=1}^{22}(y_{2,j} - \bar{y}_{2,\bullet})^2 + \sum_{j=1}^{25}(y_{3,j} - \bar{y}_{3,\bullet})^2 = 26632.364.$$

So, the statistic's value, $\Im_{Obs}$, is given by

$$\Im_{Obs} = \frac{2073.021}{26632.364} = 0.07784.$$

Given $N = n$, when H_0 holds, the common conditional distribution of $\Im$ is a central $\overline{F}$ distribution with $g = 2$ and $g(n) = 65 - 3 = 62$ degrees of freedom, since $n = 65$, $\overline{F}(z|2, 62)$.

The quantiles, $z_{1-\alpha}$, of the conditional distribution are given in Table 2. These quantiles are obtained considering $z_\alpha = \frac{2}{62} f_{1-\alpha,2,62}$, where $f_{1-\alpha,2,62}$ corresponds to the $1 - \alpha$ quantile of a central F distribution with 2 and 62 degrees of freedom. So we can conclude that using the common approach we reject H_0 for $\alpha = 0.1$, since $\Im_{Obs} > z_{1-\alpha}$, and we do not reject for $\alpha = 0.05$ and 0.01.

4.1. Counting processes

To carry out the computation we are led to use our previous information assuming that λ_i, $i = 1, 2, 3$ correspond to the average numbers of occurrences per year. So we take $\lambda_1 = 18$; $\lambda_2 = 22$ and $\lambda_3 = 25$, which means that $N_1 \sim P(18)$, $N_2 \sim P(22)$ and $N_3 \sim P(25)$. Let us also assume that we have at least 5 observations per level, which means that $n_i^\bullet = 5$, $i = 1, 2, 3$, $n^\bullet = \sum_{i=1}^3 n_i = 15$ and consequently $\boldsymbol{n}^\bullet = (5, 5, 5)'$.

To compute the quantiles for the unconditional distribution we obtain the minimum value $\overline{n} = 97$ (considering in expression (10)) such that

$$\sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = 1 - \sum_{n=n^\bullet}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) < 10^{-4}. \tag{14}$$

Therefore, the infinite series in (8) is truncated not considering the terms for $n > 97$. So, when H_0 holds, we have the distribution

$$\overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=15}^{97} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|2, n-3)$$
$$= \sum_{n=15}^{97} \sum_{n_1=5}^{n-10} \sum_{n_2=5}^{n-(n_1+5)} \sum_{n_3=n-(n_1+n_2)}^{n-(n_1+n_2)} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|2, n-3),$$

with

$$pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^{3} \frac{\lambda_i^{n_i}/n_i!}{e^{\lambda_i} - \sum_{u_i=0}^{4} \frac{\lambda_i^{u_i}}{u_i!}}.$$

The quantiles, $z^t_{1-\alpha}$, for the probability $1 - \alpha$, of this distribution are presented in Table 3. Since $\Im_{Obs} < z^t_{1-\alpha}$, we can conclude that we do not reject H_0 for the usual level of significance. So these results lead us to take a contrary decision that we had taken using the conditional approach, for $\alpha = 0.1$.

Table 3
The quantiles of the unconditional distribution (counting processes).

Values of α	0.1	0.05	0.01
$z^t_{1-\alpha}$	0.07856	0.10341	0.16341

4.2. Observations failures

To carry out the computations we use our previous information assuming that the probability of a failure is equal to 0.2 ($p = 0.2$). Therefore the probability of collecting a designed observation may be taken as $1 - p = 0.8$, $i = 1, 2, 3$, and consequently $r_1 = 22$, $r_2 = 27$ and $r_3 = 31$ (since $\frac{n_i}{r_i} \simeq 1 - p, i = 1, 2, 3$). This means that $N_1 \sim B(22, 0.8)$, $N_2 \sim B(27, 0.8)$ and $N_3 \sim B(31, 0.8)$. According to the reproducibility of Binomial distributions we will have $N \sim B(80, 0.8)$.

Let us assume, as before, that we have at least 5 observations per level, so $n_i^\bullet = 5, i = 1, 2, 3, \boldsymbol{n}^\bullet = (5, 5, 5)'$ and $n^\bullet = 15$. Therefore we have

$$\begin{aligned}\overline{\overline{F}}(z) &= \sum_{n=15}^{80} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|2, n-3)\\ &= \sum_{n=15}^{80}\sum_{n_1=5}^{n-10}\sum_{n_2=5}^{n-(n_1+5)}\sum_{n_3=n-(n_1+n_2)}^{n-(n_1+n_2)} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|2, n-3),\end{aligned}$$

with

$$pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^{3} \frac{\binom{r_i}{n_i} 0.8^{n_i} 0.2^{r_i - n_i}}{\sum_{u_i=5}^{r_i} \binom{r_i}{u_i} 0.8^{u_i} 0.2^{r_i - u_i}}.$$

The obtained quantiles, $z^b_{1-\alpha}$, for probability $1 - \alpha$ of this distribution, are presented in Table 4. We conclude that we do not reject H_0 for the usual levels of significance, which agree with the counting processes's results.

Table 4
The quantiles of the unconditional distribution (observations failures).

Values of α	0.1	0.05	0.01
$z^b_{1-\alpha}$	0.07871	0.10361	0.16365

In summary, we draw the following conclusions: The classical F-tests provide quantiles that are slightly smaller than the ones given by the unconditional approach, leading us to take a contrary decision for $\alpha = 0.1$. Therefore the proposed methodology, beyond being more realistic when the sample sizes are unknown, gives more precise critical values leading to a decrease in the probability of false rejections.

5. Computing critical values

This section presents a new way to compute correct critical values, which may be important to avoid working with incorrect test levels. We assume that the occurrence of observations corresponds to counting processes.

As previously shown in Section 3.1,

$$\overline{\overline{F}}(z) = \sum_{n=n^\bullet}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) = \sum_{n=n^\bullet}^{\infty} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|g, g(n)),$$

with $p_{\boldsymbol{n}^\bullet}(n)$ as defined in (2).

So, throught (9) we have

$$\overline{\overline{F}}_{\overline{n}}(z) \leq \overline{\overline{F}}(z) \leq \overline{\overline{F}}^*_{\overline{n}}(z),$$

and consequently

$$f^*_{\overline{n},1-\alpha} < f_{1-\alpha} < f_{\overline{n},1-\alpha},$$

with $f_{\overline{n},1-\alpha}, f_{1-\alpha}$ and $f^*_{\overline{n},1-\alpha}$ the $(1-\alpha)$th quantiles for these distributions, see [3].

Therefore, the approximate quantile value can be taken by

$$\widetilde{f}_{1-\alpha} = \frac{f_{\overline{n},1-\alpha} + f^*_{\overline{n},1-\alpha}}{2}, \tag{15}$$

which can be used as a critical value for the usual values of α, see [2].

5.1. Computing lower bounds for Poisson parameters

Since the parameters $\lambda_i, i = 1, \ldots, m$, are unknown, we now show how to deal with them in order to compute the critical values.

Nunes et al. [2] showed that the unconditional distribution increases with $\lambda_i, i = 1, \ldots, m$. Therefore the corresponding quantiles decrease and we will use lower bounds for these parameters.

The lower bounds will be the minimum values of $\lambda_i, i = 1, \ldots, m$, such that

$$e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!} = \alpha, \ \ i = 1, \ldots, m, \tag{16}$$

with α the usual level of significance.

So, considering

$$g_{n_i}(\lambda_i) = \frac{1}{n_i!}e^{-\lambda_i}\lambda_i^{n_i},$$

we obtain

$$\frac{dg_{n_i}(\lambda_i)}{d\lambda_i} = -\frac{1}{n_i!}e^{-\lambda_i}\lambda_i^{n_i} + \frac{1}{n_i!}e^{-\lambda_i}n_i\lambda_i^{n_i-1} = \frac{1}{n_i!}e^{-\lambda_i}\lambda_i^{n_i-1}(-\lambda_i + n_i),$$

which means that $g_{n_i}(\lambda_i)$ increases with λ_i, when $\lambda_i < n_i, i = 1, \ldots, m$. Therefore we will work with the left solution, $\lambda_{\alpha,i}$, of Eq. (16) to obtain the $(1-\alpha)$% confidence interval for $\lambda_i, i = 1, \ldots, m$,

$$[\lambda_{\alpha,i}; +\infty[.$$

Table 5
The lower bounds for $\lambda_i, i = 1, 2, 3$.

Values of α	0.05	0.01
$\lambda_{\alpha,1}$	13.5	10.5
$\lambda_{\alpha,2}$	17.5	13.75
$\lambda_{\alpha,3}$	20.5	16.25

5.2. Computation results

In order to compute the critical values we consider that $n_i, i = 1, 2, 3$, corresponds to the number of patients presented in Table 1. We also assume that we have the same minimum dimension for each sample ($n_i^\bullet = 5, \ i = 1, 2, 3$).

The obtained lower bounds, $\lambda_{\alpha,i}$, for the $\lambda_i, i = 1, 2, 3$, considering $\alpha = 0.05$ and $\alpha = 0.01$, are presented in Table 5.

Resorting to (10), we have

$$\overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=15}^{\overline{n}} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|2, n-3)$$

and

$$\begin{aligned}\overline{\overline{F}}^*_{\overline{n}}(z) &= \overline{\overline{F}}_{\overline{n}}(z) + \left(1 - \sum_{n=15}^{\overline{n}} p_{\boldsymbol{n}^\bullet}(n)\right)\\ &= \overline{\overline{F}}_{\overline{n}}(z) + q,\end{aligned}$$

where

$$q = 1 - \sum_{n=15}^{\overline{n}} p_{\boldsymbol{n}^\bullet}(n),$$

which do not depend on z. Assuming the lower bounds for $\lambda_i, i = 1, 2, 3$ we obtained the minimum value $\overline{n} = 80$ [$\overline{n} = 66$] such that $q < 10^{-4}$ for $\alpha = 0.05$ [$\alpha = 0.01$]. We computed the quantiles of $\overline{\overline{F}}^*_{\overline{n}}(z)$ replacing $1-\alpha$ by $(1-\alpha) - q$, assuming that $q = 10^{-4}$.

The obtained critical values, $\widetilde{f}_{1-\alpha}$, defined in (15), are presented in Table 6.

Comparing these critical values with those obtained in Section 4.1 we find that these ones are slightly higher than the previous ones. This appears to be a reasonable "price" for the increase of robustness due to use of a more complete model (obtained with the estimation of the parameters $\lambda_i, i = 1, \ldots, m$ instead of assuming their values). Moreover the use of lower bound for the $\lambda_i, i = 1, 2, 3$, parameters decreases the required values for $\overline{n}$.

Table 6
Correct critical values.

Values of α	0.05	0.01
$\widetilde{f}_{1-\alpha}$	0.13476	0.28704

6. A simulation study

In this section we carry out a simulation study to compare and relate the performances of our approach with those of common ANOVA.

In these simulations we considered three sets of values for λ_1, λ_2 and λ_3. These sets are presented in Table 7. For each λ_i, $i = 1, 2, 3$, in each set we generated a Poisson distributed sample with size 30. Thus for each set of λ_i, $i = 1, 2, 3$, we had 30 triplets of sample sizes from which we obtained the corresponding conditional critical values as well as lower bounds for the Poisson parameters (which correspond to the left solution of Eq. (16)), both for a 5% level. Assuming to have at least 5 observations per level and $\overline{n}$ such that the truncation errors do not exceed 10^{-4} we also obtained the 5% unconditional critical values for each sample triplets. The minimum value of $\overline{n}$ for each set of λ_i, $i = 1, 2, 3$, presented in Table 7, allows us to conclude that we have a good control of the truncation error.

Thus for each initial sets of parameters values we had, also at 5% level, 30 pairs for conditional and unconditional critical values. These pairs are shown in Figs. 1–3. We see that there is a close linear relation between conditional and unconditional critical values, which is confirmed by the Pearson correlation coefficients presented in Table 7.

The close relation points the possibility of estimating unconditional critical values from the conditional ones using the empirical regression, whose equations are presented also in Table 7, where y denotes the unconditional and x the conditional critical values. We also checked that the assumptions underlying the analysis of residues for the adjusted linear regressions did hold.

Table 7
Simulation results.

	λ_1	λ_2	λ_3	Minimum $\overline{n}$	Correlation coefficient	Empirical regressions
Set 1	9	11	13	55	0.995	$y = 0.079 + 1.13x$
Set 2	18	22	25	100	0.999	$y = -0.038 + 1.704x$
Set 3	36	44	50	184	0.999	$y = -0.012 + 1.363x$

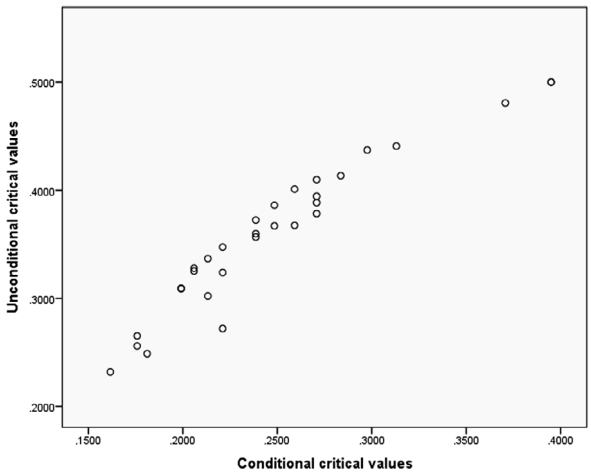

Fig. 1. Scatterplot for Set 1.

7. Closing remarks

In this paper we propose to assume the sample sizes as realizations of random variables when they are not known in advance. In light of this we were able to obtain precise critical values, thus overcoming the fact that using the usual approach only approximate critical values may be obtained when the sample dimensions are unknown. To conduct our approach we resorted to the Poisson and Binomial distributions as the adequate choices for the distributions of the sample sizes. We open room to a new field based on the assumption that we have a minimum dimension for each sample considering one-way ANOVA. Through the application presented we can confirm that the quantiles may exceed those of the common ANOVA

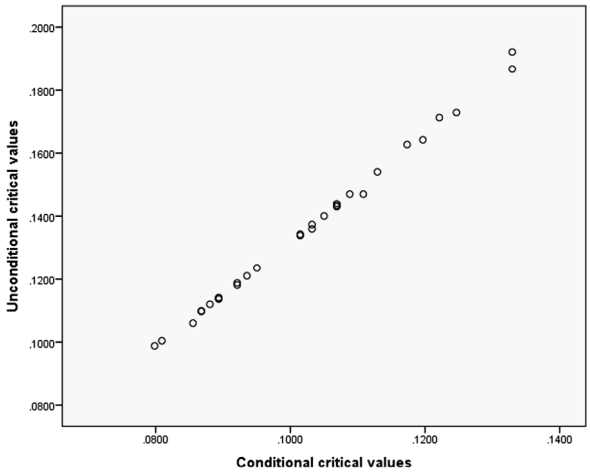

Fig. 2. Scatterplot for Set 2.

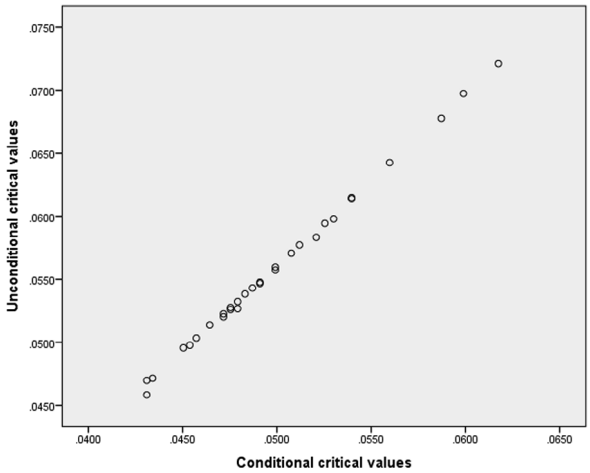

Fig. 3. Scatterplot for Set 3.

when random sample sizes are considered, giving relevance to the unconditional approach in avoiding false rejections. We showed how to obtain correct critical values, giving us the possibility to carry out tests with proper level. In Section 6 we conclude there exists a close linear relation between conditional and unconditional critical values. We intend to deepen the study in a future work and check the possibility of estimating the unconditional critical values from the conditional ones.

In our approach we worked with $\overline{F}$ instead of F distribution, since $\overline{F}$ leads to useful monotony properties that lighten the treatment.

Acknowledgments

The authors would like to thank the anonymous referees for useful comments and suggestions. This work was partially supported by national funds of FCT —Foundation for Science and Technology, Portugal under UID/MAT/00212/2013 and UID/MAT/00297/2013.

Appendix. Frequency tables of types of cancers

See Tables A.1–A.3.

Table A.1
Soft tissues of the thorax cancer.

Ages	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	2	1	2	1	1	2
Ages	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	0	0	0	2	1	1	2	1

Table A.2
Intestinal tract cancer.

Ages	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	1	1	1	0
Ages	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	3	3	2	1	3	0	1	2	4

Table A.3
Nasal cavity cancer.

Ages	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	0	0	2	0
Ages	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	4	0	4	1	3	3	1	4

References

[1] C. Nunes, D. Ferreira, S.S. Ferreira, J.T. Mexia, *F*-tests with a rare pathology, J. Appl. Stat. 39 (3) (2012) 551–561. http://dx.doi.org/10.1080/02664763.2011.603293.

[2] C. Nunes, D. Ferreira, S.S. Ferreira, J.T. Mexia, Fixed effects ANOVA: an extension to samples with random size, J. Stat. Comput. Simul. 84 (11) (2014) 2316–2328. http://dx.doi.org/10.1080/00949655.2013.791293.

[3] J.T. Mexia, C. Nunes, D. Ferreira, S.S. Ferreira, E. Moreira, Orthogonal fixed effects ANOVA with random sample sizes, in: Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling (ASM'11), 2011, pp. 84–90..

[4] E.E. Moreira, J.T. Mexia, C.E. Minder, F tests with random sample size. Theory and applications, Stat. Probab. Lett. 83 (6) (2013) 1520–1526. http://dx.doi.org/10.1016/j.spl.2013.02.020.

[5] C. Nunes, G. Capistrano, D. Ferreira, S.S. Ferreira, J.T. Mexia, One-way fixed effects ANOVA with missing observations, in: Proceedings of the 12th International Conference on Numerical Analysis and Applied Mathematics, in: AIP Conf. Proc., vol. 1648, 2015, 110008. http://dx.doi.org/10.1063/1.4912415.

[6] G. Capistrano, C. Nunes, D. Ferreira, S.S. Ferreira, J.T. Mexia, One-way random effects ANOVA with random sample sizes: An application to a brazilian database on cancer registries, in: Proceedings of the 12th International Conference on Numerical Analysis and Applied Mathematics, in: AIP Conf. Proc., vol. 1648, 2015, 110009. http://dx.doi.org/10.1063/1.4912416.

[7] E.L. Lehmann, Testing Statistical Hypotheses, John Wiley & Sons, Inc., New York, 1959.

[8] J.T. Mexia, Best linear unbiased estimates, duality of *F* tests and the Scheffé multiple comparison method in presence of controled heterocedasticity, Comput. Statist. Data Anal. 10 (3) (1990) 271–281.

[9] A.I. Khuri, T. Mathew, B.K. Sinha, Statistical Tests for Mixed Linear Models, in: Wiley series in Probability and Statistics, John Wiley & Sons, New York, 1998.

[10] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, in: Wiley series in Probability and statistics, John Wiley & Sons, New York, 1992.

[11] H. Robbins, Mixture of distribution, Ann. Math. Stat. 19 (1948) 360–369.

[12] H. Robbins, E.J.G. Pitman, Application of the method of mixtures to quadratic forms in normal variates, Ann. Math. Stat. 20 (1949) 552–560.

[13] C. Nunes, D. Ferreira, S.S. Ferreira, J.T. Mexia, Control of the truncation errors for generalized F distributions, J. Stat. Comput. Simul. 82 (2) (2012) 165–171. http://dx.doi.org/10.1080/00949655.2011.631924.

[14] Brazilian National Cancer Institutr (INCA), 2010. http://www2.inca.gov.br/wps/wcm/connect/inca/portal/home (accessed 15.09.05).