One-way Fixed Effects ANOVA with Missing Observations

Célia Nunes*, Gilberto Capistrano†, Dário Ferreira*, Sandra S. Ferreira* and João T. Mexia**

*Department of Mathematics and Center of Mathematics, University of Beira Interior, Covilhã, Portugal
†University Center of Itajubá - FEPI, Brazil and Center of Mathematics, University of Beira Interior, Portugal
**Department of Mathematics and CMA, Faculty of Science and Technology, New University of Lisbon, Portugal

Abstract. The aim of this paper is to extend the theory of F-tests with random sample sizes to situations when missing observations may occur. We consider the one-way ANOVA with fixed effects. This approach is illustrated through an application to patients affected by melanoma skin cancer, from three different states of Brazil.

Keywords: ANOVA, Random sample sizes, Missing observations, Cancer registries.
AMS: 62J12, 62J10, 62J99.

INTRODUCTION

ANOVA is routinely used in many situations within several areas, namely in medical research. However, in some of these situations, it is not possible to know previously the sample sizes. This often occurs when there is a given time span for collecting the observations. So in these situations, assuming there are m different treatments, it is more correct to consider the sample sizes as realizations $n_1,...,n_m$ of independent random variables $N_1,...,N_m$, see [3, 4, 5, 6, 9, 8].

In this paper we apply this to the case of samples with missing observations, which may happen, for instance, when working with patients we may have incomplete or absent reports. So now we assume that the sample dimensions $N_1,...,N_m$ have Binomial distributions with parameters $r_1,...,r_m$, the number of the designed observations, and $1-p$, the probability of a designed observation being taken. We put $N_i \sim B(r_i, 1-p)$, $i = 1,...,m$. Moreover $n = \sum_{i=1}^{m} n_i$ will be a realization of the random variable

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

which, through independence of N_i, $i = 1,...,m$,

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

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

We intend to test the hypothesis

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

which may be rewritten as

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

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

In what follows we obtain the test statistic and their conditional and unconditional distributions, under the assumption of missing observations may occur. To illustrate the usefulness of our approach we present an application to patients affected by melanoma skin cancer, from three different states of Brazil. The quantiles of the conditional and unconditional distributions was computed using *R* software.

STATISTIC AND THEIR DISTRIBUTIONS

Considering $N_i = n_i$, $i = 1,...,m$, we have the samples $Y_{i,1},...,Y_{i,n_i}$, $i = 1,...,m$, with averages $Y_{i,\bullet}$, $i = 1,...,m$. The sum of squares for the error will be given by, see e.g. [1] and [10],

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

If the observations are normal and independent with variance σ^2, when $N_i = n_i$, $i = 1,...,m$, S will be the product by σ^2 of a central chi-square with $g(n) = n - m$ degrees of freedom, $S \sim \sigma^2 \chi^2_{g(n)}$.

Proceedings of the International Conference on Numerical Analysis and Applied Mathematics 2014 (ICNAAM-2014)
AIP Conf. Proc. 1648, 110008-1–110008-5; doi: 10.1063/1.4912415

Moreover, S will be conditionally independent from the vector of treatment means, $\boldsymbol{Y}$, which has components $Y_{1,\bullet},...,Y_{m,\bullet}$. The vector $\boldsymbol{Y}$ will be normal with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\sigma^2 D(\frac{1}{n_1},...,\frac{1}{n_m})$, with $D(\frac{1}{n_1},...,\frac{1}{n_m})$ the diagonal matrix with principal elements $\frac{1}{n_1},...,\frac{1}{n_m}$.

So, when $N_i = n_i$, $i = 1,...,m$, see for instance [2],

$$S_{num} = (\boldsymbol{AY})' \left(\boldsymbol{A}D\left(\tfrac{1}{n_1},...,\tfrac{1}{n_m}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}) \tag{2}$$

will be the product by σ^2 of a noncentral chi-square with $g = m-1$ degrees of freedom and

$$\delta(n) = \frac{1}{\sigma^2}(\boldsymbol{A\mu})' \left(\boldsymbol{A}D\left(\frac{1}{n_1},...,\frac{1}{n_m}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{A\mu}) \tag{3}$$

non-centrality parameter, $S_{num} \sim \sigma^2 \chi^2_{g(n),\delta(n)}$. When H_0 holds, $\delta(n) = 0$ and $S_{num} \sim \sigma^2 \chi^2_{g(n)}$.

Therefore, when $N = n$ and H_0 holds, the conditional distribution of

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

will be a $\overline{F}(.|g, g(n))$ distribution, which is the distribution of the quotient of independent central chi-squares with g and $g(n)$ degrees of freedom, see e.g. [7].

For carrying out the inference we will assume that we have a minimum dimension for each sample, which avoids highly unbalanced cases, see e.g. [9]. Therefore we will consider that $N_i \geq n_i^\bullet$, $i = 1,...,m$, and the global minimum dimension will be $n^\bullet = \sum_{i=1}^m n_i^\bullet$. So we will take

$$\begin{aligned} p_{\boldsymbol{n}^\bullet}(n) = pr(N = n|\boldsymbol{N} \geq n^\bullet) = \sum_{n_1=n_1^\bullet}^{n-\sum_{i=2}^m n_i^\bullet} ... \\ \sum_{n_\ell=n_\ell^\bullet}^{n-(\sum_{i=1}^{\ell} n_i + \sum_{i=\ell+1}^m n_i^\bullet)} ... \\ \sum_{n_m=n_m^\bullet}^{n-\sum_{i=1}^{m-1} n_i} pr(N = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet),\ n_i = n_i^\bullet, ..., \\ r_i,\ i = 1,...,m, \end{aligned} \tag{4}$$

where, through the independence of N_i, $i = 1,...,m$, and $\boldsymbol{n}^\bullet = (n_1^\bullet,...,n_m^\bullet)'$,

$$pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^m pr(N_i = n_i|N_i \geq n_i^\bullet) = \prod_{i=1}^m \frac{\binom{r_i}{n_i}(1-p)^{n_i} p^{r_i-n_i}}{\sum_{u_i=n_i^\bullet}^{r_i} pr(N_i = u_i)} = \prod_{i=1}^m \frac{\binom{r_i}{n_i}(1-p)^{n_i} p^{r_i-n_i}}{\sum_{u_i=n_i^\bullet}^{r_i} \binom{r_i}{u_i}(1-p)^{u_i} p^{r_i-u_i}}. \tag{5}$$

The unconditional distribution of $\Im$, when the hypothesis H_0 holds, will be given by, see e.g. [3] and [9],

$$\overline{\overline{F}}(z) = \sum_{n=n^\bullet}^{r} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) = \sum_{n=n^\bullet}^{r} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|g, g(n)). \tag{6}$$

AN APPLICATION TO REAL DATA

The data used in this application were provided by the National Cancer Institute (INCA) and are from patients affected by melanoma skin cancer, from three different states of Brazil, 2008.

The factor considered is the *State*, with three levels *Espírito Santo*, *Mato Grosso do Sul* and *Sergipe*, belonging to the regions southeast, central region and northwest of Brazil, respectively. The following table illustrate the number of patients for each state.

We will test the hypothesis

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

TABLE 1. Number of patients

State	Number of patients
Espírito Santo	14
Mato Grosso do Sul	16
Sergipe	24

with

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

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

$$S_{num} = (\boldsymbol{AY})' \left(\boldsymbol{AD}\left(\tfrac{1}{14}, \tfrac{1}{16}, \tfrac{1}{24}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}) \ ,$$

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 will obtain

$$\left(\boldsymbol{AD}\left(\frac{1}{14}, \frac{1}{16}, \frac{1}{24}\right)\boldsymbol{A}'\right)^{-1} = \begin{bmatrix} 10.3704 & -4.1481 \\ -4.1481 & 11.2593 \end{bmatrix} \text{ and } \boldsymbol{Ay} = \begin{bmatrix} -12.3810 \\ -3.8542 \end{bmatrix},$$

with the sample means $\overline{y}_{1,\bullet} = 51.2857$; $\overline{y}_{2,\bullet} = 59.8125$; $\overline{y}_{3,\bullet} = 63.6667$. For the numerator of the statistic we have

$$S_{num} = 1361.022.$$

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

$$S = \sum_{j=1}^{14} (y_{1,j} - \overline{y}_{1,\bullet})^2 + \sum_{j=1}^{16} (y_{2,j} - \overline{y}_{2,\bullet})^2 + \sum_{j=1}^{24} (y_{3,j} - \overline{y}_{3,\bullet})^2 = 13074.628.$$

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

$$\Im_{Obs} = \frac{1361.022}{13074.628} = 0.1041.$$

Given $N = n$, when H_0 holds, the common conditional distribution of $\Im$ is a central $\overline{F}$ distribution with $g = 2$ and $g(n) = 54 - 3 = 51$ degrees of freedom, $\overline{F}(z|2,51)$. The quantiles, $z_{1-\alpha}$, of this conditional distribution are given in Table 2. So we can conclude that 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.

TABLE 2. The quantiles of the conditional distribution.

Values of α	0.1	0.05	0.01
$z_{1-\alpha}$	0.0945	0.1247	0.1979

To carry out the computations we are led to use our previous information assuming that we have $r_1 = 18$, $r_2 = 20$ and $r_3 = 30$ designed observations for each level and the probability of a designed observation being taken is $1 - p = 0.8$. This means that $N_1 \sim B(18, 0.8)$, $N_2 \sim B(20, 0.8)$ and $N_3 \sim B(30, 0.8)$. Through the independence of N_i, $i = 1, 2, 3$, $N \sim B(68, 0.8)$.

Let us assume that we have at least 10 observation per level, thus $n_i^\bullet = 10$, $i = 1, 2, 3$, $\boldsymbol{n}^\bullet = (10, 10, 10)'$ and $n^\bullet = 30$. Therefore we have

$$p_{\boldsymbol{n}^\bullet}(n) = \sum_{n_1=10}^{n-20} \sum_{n_2=10}^{n-(n_1+10)} \sum_{n_3=10}^{n-(n_1+n_2)} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet),\ n_i = 10, ..., r_i,\ i = 1, 2, 3, \tag{7}$$

with

$$pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})=\prod_{i=1}^{3}\frac{\binom{r_i}{n_i}(1-p)^{n_i}p^{r_i-n_i}}{\sum_{u_i=10}^{r_i}\binom{r_i}{u_i}(1-p)^{u_i}p^{r_i-u_i}}. \tag{8}$$

The unconditional distribution of $\Im$, when the hypothesis H_0 holds, will be given by

$$\overline{\overline{F}}(z)=\sum_{n=30}^{68}\sum_{n_1=10}^{n-20}\sum_{n_2=10}^{n-(n_1+10)}\sum_{n_3=10}^{n-(n_1+n_2)}pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|2,n-3). \tag{9}$$

The obtained quantiles, $z^{u}_{1-\alpha}$, for probability $1-\alpha$ of this distribution are presented in Table 3. Since $\Im_{Obs}<z^{u}_{1-\alpha}$, we do not reject H_0 for the usual levels of significance.

TABLE 3. The quantiles of the unconditional distribution.

Values of α	0.1	0.05	0.01
$z^{u}_{1-\alpha}$	1.2169	1.6030	2.5363

These results lead us to take a contrary decision that we had taken using the common conditional approach for $\alpha=0.1$.

We saw that the inference depends on the approach and since the unconditional approach is more secure, we conclude that the factor is not significant. This means that the age of disease detection is not significantly different in these three states.

FINAL REMARKS

In this paper we tried to open a new field based on the use of the binomial distribution to one-way fixed effects model for missing observations. Through the application we showed the relevance of the unconditional approach in avoiding false rejections.

During our treatment we worked with $\overline{F}$ distributions since they are more treatable and statistically equivalent. This equivalence enabled us to consider our tests as F-tests.

ACKNOWLEDGMENTS

This work was partially supported by Center of Mathematics, University of Beira Interior, through the project PEst-OE/MAT/UI0212/2014 and by CMA, Faculty of Science and Technology, New University of Lisbon, through the project PEst- OE/MAT/UI0297/2014.

REFERENCES

1. A.I. Khuri, T. Mathew, B.K. and B.K. Sinha, *Statistical Tests for Mixed Linear Models*. John Wiley & Sons, Inc., New York, (1998).
2. 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), 271-281 (1990).
3. J.T. Mexia, C. Nunes, D. Ferreira, S.S. Ferreira and E. Moreira, "Orthogonal fixed effects ANOVA with random sample sizes", in *Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling (ASM'11)*, Corfu, Greece, 2011, pp. 84-90.
4. C. Nunes, D. Ferreira, S.S. Ferreira and J.T. Mexia, "F Tests with Random Sample Sizes", in *8th International Conference on Numerical Analysis and Applied Mathematics,* AIP Conf. Proc., 2010, 1281(II), pp. 1241-1244.
5. C. Nunes, D. Ferreira, S.S. Ferreira and J.T. Mexia, F-tests with a rare pathology. *Journal of Applied Statistics* **39**(3), 551–561 (2012).
6. C. Nunes, D. Ferreira, S.S. Ferreira, M.M. Oliveira and J.T. Mexia, "One-way random effects ANOVA: An extension to samples with random size", in *10th International Conference on Numerical Analysis and Applied Mathematics,* AIP Conf. Proc., 2012, 1479, pp. 1678-1681.
7. C. Nunes, D. Ferreira, S.S. Ferreira and J.T. Mexia, Control of the truncation errors for generalized F distributions. *J. Stat. Comput. Simul.* **82**(2), 165-171 (2012).
8. C. Nunes, G. Capistrano, D. Ferreira and S. S. Ferreira, "ANOVA with Random Sample Sizes: An Application to a Brazilian Database on Cancer Registries", in *11th International Conference on Numerical Analysis and Applied Mathematics,* AIP Conf. Proc., 2013, 1558, 825-828.

9. C. Nunes, D. Ferreira, S. S. Ferreira and J.T. Mexia. Fixed effects ANOVA: an extension to samples with random size. *Journal of Statistical Computation and Simulation*, **84**(11), 2316-2328, (2014).
10. S.R. Searle, G. Casella and C.E. McCulloch, *Variance Components*. Wiley series in Probability and statistics. John Wiley & Sons, New York, (1992).