One-way random effects ANOVA with random sample sizes: An application to a Brazilian database on cancer registries

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

Citation: AIP Conference Proceedings **1648**, 110009 (2015); doi: 10.1063/1.4912416
View online: http://dx.doi.org/10.1063/1.4912416
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1648?ver=pdfcov
Published by the AIP Publishing

One-way Random Effects ANOVA with Random Sample Sizes: An Application to a Brazilian Database on Cancer Registries

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

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

Abstract. ANOVA is routinely used in many situations, namely in medical research, where the sample sizes may not be previously known. This leads us to consider the samples sizes as realizations of random variables. The aim of this paper is to extend one-way random effects ANOVA to those situations and apply our results to a Brazilian database on cancer registries.

Keywords: F-tests, Random effects model, Random sample sizes, Cancer registries.
AMS: 62J12, 62J10, 62J99.

INTRODUCTION

In many situations, such as in medical research, sample sizes may not be previously known. This often occurs when there is a specific time span for collecting the observations. The aim of this paper is to extend one-way random effects models to those situations and apply our results to a Brazilian database on cancer registries.

In these situations it is more correct to consider the samples sizes as realizations $n_1,...,n_r$ of independent random variables $N_1,...,N_r$, see [2, 3, 4, 5, 6, 7]. We assume that these random variables will be independent Poisson distributed with parameters $\lambda_1,...,\lambda_r$. Therefore $n=\sum_{i=1}^{r} n_i$ will be a realization of the random variable $N=\sum_{i=1}^{r} N_i$ which is Poisson distributed with parameter $\lambda=\sum_{i=1}^{r}\lambda_i$, see [2, 6, 7].

When $N_i=n_i$, $i=1,...,r$, the one-way random effects model can be written as, see e.g. [8],

$$Y_{i,j}=\mu+\alpha_i+e_{i,j},\ j=1,...,n_i,\ i=1,...,r, \tag{1}$$

where μ is a fixed unknown parameter and α_i and $e_{i,j}$, $j=1,...,n_i$, $i=1,...,r$, are independent normal variables with null mean values and variances, respectively, σ_α^2 and σ^2, $\alpha_i\sim N(0,\sigma_\alpha^2)$, $e_{i,j}\sim N(0,\sigma^2)$, $j=1,...,n_i$, $i=1,...,r$.

The model (1) can be written in matrix notation as, see e.g. [1],

$$\boldsymbol{Y}=\boldsymbol{\mu}+D(\mathbf{1}_{n_1},...,\mathbf{1}_{n_r})\boldsymbol{\alpha}+\boldsymbol{e}, \tag{2}$$

where $\boldsymbol{Y}=(Y_{1,1},...,Y_{1,n_1},Y_{2,1},...,Y_{2,n_2},...,Y_{r,1},...,Y_{r,n_r})'$, $\boldsymbol{\alpha}=(\alpha_1,...,\alpha_r)'$ and $D(\mathbf{1}_{n_1},...,\mathbf{1}_{n_r})$ denotes a block diagonal matrix with $\mathbf{1}_{n_1},...,\mathbf{1}_{n_r}$ along the blocks and $\mathbf{1}_m$ is the vector with all m components equal to 1. Therefore, $\boldsymbol{\alpha}$ and $\boldsymbol{e}$ are normal distributed, with null mean vectors and variance-covariance matrices, respectively, $\sigma_\alpha^2\boldsymbol{I}_r$ and $\sigma^2\boldsymbol{I}_n$, $\boldsymbol{\alpha}\sim N(\mathbf{0},\sigma_\alpha^2\boldsymbol{I}_r)$, $\boldsymbol{e}\sim N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$, with $\boldsymbol{I}_n$ the $n\times n$ identity matrix.

The vector $\boldsymbol{Y}$, when $N=n$, will be conditionally normal with mean vector $\boldsymbol{\mu}=\mu\mathbf{1}_r$ and variance-covariance matrix given by, see [1] and [9], $\Sigma=\sigma_\alpha^2 D(\boldsymbol{J}_{n_1},...,\boldsymbol{J}_{n_r})+\sigma^2\boldsymbol{I}_n$, where $\boldsymbol{J}_m=\mathbf{1}_m\mathbf{1}_m'$. We put

$$\boldsymbol{Y}\underset{(N=n)}{\sim} N\left(\boldsymbol{\mu},\sigma_\alpha^2 D(\boldsymbol{J}_{n_1},...,\boldsymbol{J}_{n_r})+\sigma^2\boldsymbol{I}_n\right).$$

We are interested in test the hypotheses

$$H_0:\sigma_\alpha^2=0\ vs\ H_1:\sigma_\alpha^2>0. \tag{3}$$

In what follows we present the test statistic for testing these hypotheses and their conditional and unconditional distributions. We will consider the expression of the test statistic obtained in [5]. Then we present an application based on real medical data, particulary on cancer registries from São Paulo, Brazil. Finally we conclude with some final remarks.

Proceedings of the International Conference on Numerical Analysis and Applied Mathematics 2014 (ICNAAM-2014)
AIP Conf. Proc. 1648, 110009-1–110009-4; doi: 10.1063/1.4912416

STATISTIC AND THEIR DISTRIBUTIONS

In this section we will obtain the common conditional distribution of the test statistic and also it's unconditional distribution, under the assumption that we have a global minimum dimension for the samples, which avoids highly unbalanced cases.

When $N_i = n_i$, $i = 1, ..., r$, we have the averages $Y_{i,\bullet}$, $i = 1, ..., r$, for the samples $Y_{i,1}, ..., Y_{i,n_i}$, $i = 1, ..., r$. The sum of squares for the error will be given by, see e.g. [9],

$$S = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{i,\bullet})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{i,j}^2 - \sum_{i=1}^{r} \frac{T_i^2}{n_i}, \tag{4}$$

with $T_i = \sum_{j=1}^{n_i} Y_{i,j}$, $i = 1, ..., r$ and S will be the product by σ^2 of a central chi-square with $g(n) = n - r$ degrees of freedom, $S \sim \sigma^2 \chi^2_{g(n)}$. As we can see, for instance in [9], S can be written using the matrix formulation as

$$S = \boldsymbol{Y}' \left(\boldsymbol{I}_n - D\left(\frac{\boldsymbol{J}_{n_1}}{n_1}, ..., \frac{\boldsymbol{J}_{n_r}}{n_r} \right) \right) \boldsymbol{Y}. \tag{5}$$

When $N_i = n_i$, the sample means

$$Y_{i,\bullet} = \mu + \alpha_i + e_{i,\bullet},\ i = 1, ..., r,$$

have mean values μ and variances $\sigma_\alpha^2 + \frac{\sigma^2}{n_i}$, $i = 1, ..., r$. So, when the hypothesis H_0 holds,

$$Y_{i,\bullet} \quad \sim \quad N\left(\mu, \frac{\sigma^2}{n_i} \right), \ , \ i = 1, ..., r, \text{ and } \ Y_{\bullet,\bullet} = \tfrac{1}{n} \textstyle\sum_{i=1}^{r} n_i Y_{i,\bullet} \quad \sim \quad N\left(\mu, \frac{\sigma^2}{n} \right).$$

Let us consider

$$\boldsymbol{Z} = \boldsymbol{Y}_\bullet - \boldsymbol{Y}_{\bullet,\bullet} = \boldsymbol{B}\boldsymbol{Y}_\bullet,$$

where $\boldsymbol{Y}_\bullet = (Y_{1,\bullet}, ..., Y_{r,\bullet})'$, $\boldsymbol{Y}_{\bullet,\bullet} = (Y_{\bullet,\bullet}, ..., Y_{\bullet,\bullet})'$ and

$$\boldsymbol{B} = \boldsymbol{I}_r - \begin{bmatrix} \frac{n_1}{n} & \cdots & \frac{n_r}{n} \\ \vdots & \ddots & \vdots \\ \frac{n_1}{n} & \cdots & \frac{n_r}{n} \end{bmatrix} = \begin{bmatrix} \frac{n-n_1}{n} & \cdots & \frac{-n_r}{n} \\ \vdots & \ddots & \vdots \\ \frac{-n_1}{n} & \cdots & \frac{n-n_r}{n} \end{bmatrix} = \boldsymbol{I}_r - \frac{1}{n} \boldsymbol{1}_r \boldsymbol{n}'.$$

When $N_i = n_i$, $i = 1, ..., r$, $\boldsymbol{Z}$ will be normal distributed with null mean vector and variance-covariance matrix $\sigma^2 \boldsymbol{V}$, $\boldsymbol{Z} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$, with

$$\boldsymbol{V} = \boldsymbol{B} D\left(\frac{1}{n_1}, ..., \frac{1}{n_r} \right) \boldsymbol{B}' = D\left(\frac{1}{n_1}, ..., \frac{1}{n_r} \right) - \frac{1}{n} \boldsymbol{J}_r.$$

The sum of squares for the main effects of the factor will be given by $S_{num} = \boldsymbol{Z}' \boldsymbol{V}^- \boldsymbol{Z}$, where $\boldsymbol{V}^-$ denote a generalized inverse of matrix $\boldsymbol{V}$. When H_0 holds, we have

$$S_{num} = \boldsymbol{Z}' \boldsymbol{V}^- \boldsymbol{Z} \sim \sigma^2 \chi^2_g,$$

with $g = rank(\boldsymbol{V}) = r - 1$, see [5].

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

$$\Im = \frac{S_{num}}{S} = \frac{\boldsymbol{Z}' \boldsymbol{V}^- \boldsymbol{Z}}{S} \tag{6}$$

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

For carrying out the inference we assume that $N \geq n^\bullet$, which means that we assume that we have a global minimum dimension for the samples, see [2, 5, 6, 7]. So we take

$$p_{n^\bullet}(n) = pr(N = n | N \geq n^\bullet) = \frac{pr(N = n)}{pr(N \geq n^\bullet)} = \frac{e^{-\lambda} \lambda^n / n!}{1 - \sum_{u=0}^{n^\bullet - 1} e^{-\lambda} \lambda^u / u!} = \frac{\lambda^n}{n! \left(e^\lambda - \sum_{u=0}^{n^\bullet - 1} \lambda^u / u! \right)}, \ n = n^\bullet, ... \tag{7}$$

and we obtain the unconditional distribution given by

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

AN APPLICATION

The data used in this application were provided by the National Cancer Institute (INCA) and are from city of São Paulo, Brazil, 2010. We selected $r = 6$ different kinds of cancer, using simple random sampling. The Table 1 illustrates the kinds of cancer which have been selected and the number of patients.

TABLE 1. Kinds of cancers and number of patients

Kinds of Cancer	Number of patients
Body of Stomach	91
Spinal cord and other parties S.N.C.	42
Melanoma on the trunk	107
Encephalon	93
Ascending Colon	201
Upper lobe, bronchus or lung	155

As we saw, given $N = n$, the conditional distribution of $\Im$ is, when H_0 holds, a central $\overline{F}$ distribution with $g = 5$ and $g(n) = n-6$ degrees of freedom, $\overline{F}(.|5, n-6)$.

To carry out the calculations when we may assume that $\sum_{n=0}^{n^\bullet-1} p_{n^\bullet}(n) \simeq 0$, which means that, with high probability, we have $N \geq n^\bullet$, the unconditional distribution of the statistic will be given by

$$\overline{\overline{F}}(z) = \sum_{n=n^\bullet}^{\infty} p_{n^\bullet}(n)\overline{F}(z|5, n-6).$$

Moreover, to the monotony property of the $\overline{F}$ distribution, see [2], we have $\overline{F}(z|g, n-6) < \overline{F}(z|g, n^o-6)$, with $n < n^o$, so

$$\overline{F}(z|5, n^\bullet-6) \leq \overline{\overline{F}}(z) \leq 1,$$

which gives us a lower bound for $\overline{\overline{F}}(z)$. Thus, from $\overline{F}(z|5, n^\bullet-6)$, we can obtain upper bounds for the quantiles of the unconditional distribution $\overline{\overline{F}}(z)$. If we use these upper bounds as critical values we will have tests with sizes that do not exceed the theoretical values.

If the statistic's value exceeds the upper bounds, also exceeds the real critical value (obtaining considering random sample sizes) and in this case we reject the hypothesis. When the statistic's value is lower than the upper bound we must compute the real critical values or calculate the minimum value of $n^\bullet$ that leads to reject the null hypothesis.

In this case we obtain

$$S = \sum_{j=1}^{91}(y_{1,j} - y_{1,\bullet})^2 + \sum_{j=1}^{42}(y_{2,j} - y_{2,\bullet})^2 + ... + \sum_{j=1}^{155}(y_{6,j} - y_{6,\bullet})^2 = 209006, \tag{9}$$

with the sample means $y_{1,\bullet} = 62.05; y_{2,\bullet} = 46.17; y_{3,\bullet} = 67.89; y_{4,\bullet} = 49.90; y_{5,\bullet} = 71.10; y_{6,\bullet} = 66.26$. The general mean obtained is equal to $y_{\bullet,\bullet} = 64.02$.

The numerator of the $\Im$ statistic, when the hypothesis holds, was defined by $S_{num} = \boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z} \sim \sigma^2\chi^2_5$. In this case we obtain

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}_\bullet = \begin{bmatrix} -1.965375 \\ -17.853653 \\ 3.867530 \\ -14.988060 \\ 7.084158 \\ 2.237745 \end{bmatrix}; \boldsymbol{V}^- = \begin{bmatrix} 88.6758568 & 0.9076503 & -4.3801736 & -2.5542037 & -26.409124 & -13.504328 \\ 0.9076503 & 43.3329148 & 0.3069035 & 0.8449923 & -7.814687 & -2.859674 \\ -4.3801736 & 0.3069035 & 99.9126454 & -4.6868906 & -34.691226 & -18.684709 \\ -2.5542037 & 0.8449923 & -4.6868906 & 90.2067458 & -27.384874 & -14.105984 \\ -26.4091237 & -7.8146868 & -34.6912261 & -27.3848744 & 95.674453 & -66.066861 \\ -13.5043283 & -2.8596742 & -18.6847092 & -14.1059835 & -66.066861 & 115.739094 \end{bmatrix}$$

and consequently $S_{num} = \boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z} = 47067.08$. Therefore, the statistic's value, $\Im_{Obs}$, is given by

$$\Im_{Obs} = \frac{47067.08}{209006} = 0.2252.$$

If we use the common conditional distribution of $\Im$, which corresponds to $\overline{F}(z|5, 683)$, we will obtain the quantiles, $z_{1-\alpha}$, given in Table 2. So, since $\Im_{Obs} > z_{1-\alpha}$, we can conclude that we reject H_0 for the usual levels of significance.

Let us now assume that $n^\bullet = 59$, which means that we have about 10 observations per treatment. The Table 2 shows the upper bounds for the quantiles, $z^u_{1-\alpha}$, for probability $1-\alpha$ of the unconditional distribution $\overline{\overline{F}}(z)$.

The results in this table may lead us to take a contrary decision that we had taken using the conditional distribution of the statistics for $\alpha = 0.05$ and $\alpha = 0.01$. Therefore the increase of the critical values points towards the possibility of non-rejection when random sample sizes are considered.

TABLE 2. The quantiles of the conditional distribution of the statistics and upper bounds for the quantiles

Values de α	0.1	0.05	0.01
$z_{1-\alpha}$	0.0136	0.0163	0.0223
$z^{u}_{1-\alpha}$	0.1848	0.2254	0.3193

TABLE 3. Minimum value of $n^{\bullet}$ that leads to reject the null hypothesis

Values of α	0.1	0.05	0.01
$n^{\bullet}$	51	60	79

Assuming the values of the test statistic remain unchanged, for ensuring rejection, we should have the total sample sizes presented in Table 3. Since for higher values of $n^{\bullet}$ we would get lower values for the quantiles, we have $\Im_{Obs} > z^{u}_{1-\alpha}$ for all $n^{\bullet} \geq 79$, which means that, in this case, we reject H_0 considering the usual levels of significance. Thus the kind of cancer has a significant random effect, so the ages of disease detection may differ significantly with it's kind.

FINAL REMARKS

When we cannot previously know the sample sizes it is much more correct to consider them as realizations of random variables. Through the application on cancer registries we can prove the relevance of the unconditional approach in order to possibly avoid false rejections. We also can conclude that when the samples dimensions increase, the conditional and unconditional approach converge to the same decision.

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, 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.
3. 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.
4. 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).
5. 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.
6. 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.
7. 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).
8. H. Scheffé, *The analysis of variance*. Wiley series in Probability and Statistics, John Wiley & Sons, New York, (1959).
9. S.R. Searle, G. Casella and C.E. McCulloch, *Variance Components.* Wiley series in Probability and statistics. John Wiley & Sons, New York, (1992).