Statistics and Probability Letters 83 (2013) 1520–1526

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

F tests with random sample sizes. Theory and applications

Elsa E. Moreira [a,*], João T. Mexia [a], Christoph E. Minder [b]

[a] CMA - Center of Mathematics and Applications, Faculty of Sciences and Technology, Nova University of Lisbon, Campus Caparica, 2829-516 Caparica, Portugal

[b] Horten Zentrum, Medical Faculty, University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 24 January 2012
Received in revised form 27 December 2012
Accepted 26 February 2013
Available online 5 March 2013

Keywords:
ANOVA
F distribution
Pathologies comparison
Poisson distribution
Power analysis

ABSTRACT

Given a fixed time span for collecting observations in a study comparing, for example, the pathologies of patients entering a hospital sequentially, it is advisable to consider the sample sizes of the ANOVA levels as random variables. Using this approach, more powerful tests are developed, leading to lower critical values. The approach is used to obtain the minimum duration of data collection to ensure a pre-fixed power for the F test.

1. Introduction

The theoretical developments presented in this paper were motivated by a real case situation in the field of medicine. Consider the numbers of patients with a fixed spectrum of pathologies arriving at an hospital during a given time span. We assume that the data on pathology are collected from each patient as soon as he/she arrives. We hypothesize that the pathologies are distinguishable using some (continuous) measurement. The number of patients with a given pathology is not known in advance; i.e., if we were to repeat the counting during a later time period of the same length, the numbers of patients obtained with those pathologies would differ from those in the first counting. So, if we plan to conduct just one study to compare the pathologies, it is more correct to consider the sample sizes as realizations of random variables. The minimum data collection duration to ensure a desired power with a given probability will be obtained. Such a situation arises frequently, because studies are often planned to run over a fixed prespecified time span. At the end of the study interval, F test statistics can be obtained to test hypotheses about the different mean values.

In what follows, it is assumed that the sample sizes $n_1, \ldots, n_k$ for the k pathologies are realizations of independent Poisson variables with parameters $\lambda_1, \ldots, \lambda_k$, and that the observations $x_{i,j},\ i = 1, \ldots, k, j = 1, \ldots, n_i$, in these samples are normal and independent, with mean values $\mu_1, \ldots, \mu_k$ and variance σ^2. As a result, the F statistic to test the null hypotheses $H_0 : \mu_1 = \cdots = \mu_k$ will have conditional F distribution on the number of observations, as will be seen in the next section.

Thus, when each of the k pathologies is observed at least once, the conditional distribution of the F test statistic has $k-1$ and $n-k$ degrees of freedom, with n the sum of all $n_i,\ i = 1, \ldots, k$, and non-centrality parameter δ, which is null when the mean values $\mu_1, \ldots, \mu_k$ for different pathologies are equal, δ being a measure of distance of the alternatives from H_0. Finally, we may look for the minimum duration that ensures, with a given probability, that all pathologies have at least a minimum number of observations that allow the conditional power of the α level test for a given δ to be sufficiently high. In

* Corresponding author. Tel.: +351 212948300x10867.
E-mail address: efnm@fct.unl.pt (E.E. Moreira).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.02.020

Section 2, we present the theory of F tests with random sample sizes. Section 3 is devoted to power analysis. In Section 3.1, we present an algorithm to determine the minimum sample sizes $\dot{n}_i,\ i = 1, \ldots, k$, to have a sufficiently powerful F test. In Section 3.2, the minimum time span for attaining these $\dot{n}_1, \ldots, \dot{n}_k$ with probability p is obtained. In Section 4, an application with simulated data is presented to illustrate the methodology. Section 5 presents a verbal summary of the results obtained.

2. *F* tests with random sample sizes

Let the vector of sample sizes $\mathbf{n} = (n_1, \ldots, n_k)$ be a realization of the vector $\mathbf{N} = (N_1, \ldots, N_k)$ with the components of $\mathbf{N}$ independent Poisson variables with means $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$. Assuming that, for $\mathbf{n} > \mathbf{0}$, the samples $x_{i,1}, \ldots, x_{i,n_i}$, $i = 1, \ldots, k$ are normal with mean vector $\mu_1, \ldots, \mu_k$ and common variance σ^2, $H_0 : \mu_1 = \cdots = \mu_k$ may be tested using the statistic

$$\mathcal{F} = \frac{n-k}{k-1} \frac{\sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n}}{\sum_{i=1}^{k} S_i},$$

where $T_i = \sum_{j=1}^{n_i} x_{i,j}$, $S_i = \sum_{j=1}^{n_i} x_{i,j}^2 - T_i^2/n_i$, $T = \sum_{i=1}^{k} T_i$ and $n = \sum_{i=1}^{k} n_i$.

Conditionally on $\mathbf{N} = \mathbf{n}$, the $\mathcal{F}$ statistic will have an F distribution with $k-1$ and $n-k$ degrees of freedom and non-centrality parameter

$$\delta = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i (\mu_i - \mu_\cdot)^2,$$

where

$$\mu_\cdot = \frac{1}{n} \sum_{i=1}^{k} n_i \mu_i$$

is the general mean value (Hocking, 2003; Mexia, 1990).

In a previous paper (Mexia and Moreira, 2010), the unconditional distribution for the $\mathcal{F}$ statistic was derived. This distribution is given by the following series:

$$\dot{F}(z) = \sum_{\mathbf{n}>\mathbf{0}} q(\mathbf{n}) F(z|k-1, n-k, \delta(\mathbf{n})), \tag{1}$$

whose terms correspond to all the vectors $\mathbf{n} = (n_1, \ldots, n_k)$ with $n_i > 0,\ i = 1, \ldots, k$, with

$$q(\mathbf{n}) = \prod_{i=1}^{k} \frac{e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!}}{1 - e^{-\lambda_i}}.$$

When the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ holds, $\delta(\mathbf{n}) = 0$, $\forall \mathbf{n} > 0$ and

$$\mathcal{F} \sim \dot{F}_0(z) = \sum_{\mathbf{n}>\mathbf{0}} q(\mathbf{n}) F(z|k-1, n-k). \tag{2}$$

Notice that, in Eq. (1), the number of degrees of freedom $n-k$ for the $\mathcal{F}$ denominator does not depend on $\mathbf{n}$.

To actually compute the values of $\dot{F}_0(z)$, the corresponding series in Eq. (2) must be truncated. In a previous paper we showed that, restricting the sum to samples with $\mathbf{n} \le \mathbf{n}^o$, the truncation error is bounded above by

$$\frac{k\varepsilon}{(1 - e^{-\lambda^o})^k},$$

if the components n_i^o of $\mathbf{n}^o$ are such that

$$\sum_{n_i=0}^{n_i^o} e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!} > 1 - \varepsilon, \quad i = 1, \ldots, k, \tag{3}$$

and ε is small (Mexia and Moreira, 2010). Using this inequality, we may obtain the minimal sample sizes needed to control the truncation error for the distribution $\dot{F}_0(z)$ in Eq. (2).

Table 1 exhibits the minimal sample sizes n^o needed to satisfy inequality (3) for $k = 3$, various $\lambda^o = Min\{\lambda_1, \ldots, \lambda_k\}$, and various small ε. The table shows that the truncation errors are controlled even for small sample sizes. For instance, if the minimum of the $\lambda_i's$ is $\lambda_0 = 1$, for $\varepsilon = 10^{-6}$, the sample sizes should not be less than 9. In this example, the truncation error has an upper bound of 0.000012. The tables of critical values in common use show three decimal places of precision; thus truncating the distribution with an error of 0.000012 suffices to get a critical value of good accuracy.