Randomized Sample Size *F* Tests for the One-Way Layout

Cite as: AIP Conference Proceedings **1281**, 1248 (2010); https://doi.org/10.1063/1.3497917
Published Online: 17 September 2010

João T. Mexia, and Elsa E. Moreira

Randomized Sample Size F Tests for the One-Way Layout

João T. Mexia and Elsa E. Moreira

CMA - Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Abstract. Distributions and densities for F test statistics are obtained assuming random sample sizes, thus getting random degrees of freedom and non-centrality parameters. Classical optimum properties are extended to this new setup as well as Scheffé Theorem for simultaneous confidence intervals.

Keywords: Random sample sizes, F tests, simultaneous confidence intervals, optimum properties
AMS: 62J10

INTRODUCTION

When we cannot consider as known the dimensions $n_1,...,n_k$ of the samples for the k levels of the one-way layout it is more correct to consider these as realizations of some random variable. This situation arises mostly when we have a given time span for collecting the observation. This collecting is carried out separately for the different levels. A good example is the obtention of data for the comparison of pathologies. The data is obtained from the patients with each pathology as soon as they present themselves. Stochastically we will have independent counting processes which "generate" the samples. This leads us to assume that the sample sizes are values take by independent Poisson variables $N_1,...,N_k$ with parameters $\lambda_1,...,\lambda_k$, which will be the components of the random vector $\mathbf{N}$.

If we assume that the observations have equal mean values $\mu_1,...,\mu_k$ for each sample and that they are normal, independent, all with variance σ^2, we may think of using the F tests for testing $H_0 : \mu_1 = ... = \mu_k$, but now they will have only conditional F distributions. In the next section, we will derive the unconditional distributions. As we shall see, they are given by series whose terms corresponds to the vectors $\mathbf{n} = (n_1,...,n_k)$, thus we will study the truncation errors which occur when we restrict ourselves to samples with $\mathbf{n} \leq \mathbf{n}^0$. We also will consider simultaneous confidence intervals and extend some results on the optimal properties of the F tests.

DISTRIBUTIONS

Given the samples $x_{i,1},...,x_{i,n_i}$, $i = 1,...,k$, we get for H_0 the F test statistic

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

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 $n = \sum_{i=1}^{k} n_i$. This statistics will have, as conditional distribution, given $\mathbf{N} = \mathbf{n}$, the F distribution with $k-1$ and $n-k$ degrees of freedom and non-centrality parameter

$$\delta = \frac{1}{\sigma^2}[\sum_{i=1}^{k} n_i \mu_i^2 - n\mu.^2] \tag{2}$$

with

$$\mu. = \frac{1}{n}\sum_{i=1}^{k} n_i \mu_i \tag{3}$$

being the general mean value. This distribution will be given by

$$F(z|k-1, n-k, \delta) = \bar{F}(\frac{k-1}{n-k} z|k-1, n-k, \delta) \tag{4}$$

CP1281, *ICNAAM, Numerical Analysis and Applied Mathematics, International Conference 2010, Vol.II*
edited by T. E. Simos, G. Psihoyios, and Ch. Tsitouras

where $\bar{F}(.|k-1,n-k,\delta)$ is the distribution of $\mathscr{T}$ which is the quotient of a non-central chi-square with $k-1$ degrees of freedom and non-centrality parameter δ by a central chi-square with $n-k$ degree of freedom, both chi-squares being independent. In order to obtain the unconditional distribution of $\mathscr{F}$, we will use these last distributions since they are more tractable than the cental F distributions. Namely, see [2], we will have

$$\bar{F}(z|r,s,\delta) = Pr(\mathscr{T} \leq z) = e^{-\delta/2} \sum_{l=0}^{\infty} \frac{(\frac{\delta}{2})^l}{l!} \bar{F}(z|r+2l,s). \tag{5}$$

For more details see the Appendix.

We now obtain the unconditional distribution of $\mathscr{T}$. So, first we put

$$p(\mathbf{n}) = pr(\mathbf{N} = \mathbf{n}) = \prod_{i=1}^{k} pr(N_i = n_i) = \prod_{i=1}^{k} e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!}. \tag{6}$$

Since we only can carry out the test when $\mathbf{n} > \mathbf{0}$ (this is when $n_i > 0, i = 1,...,k$), we will be interested in the

$$q(\mathbf{n}) = pr(\mathbf{N} = \mathbf{n}|\mathbf{N} > \mathbf{0}) = \frac{pr(\mathbf{N} = \mathbf{n} \cap \mathbf{N} > \mathbf{0})}{pr(\mathbf{N} > \mathbf{0})} = \frac{p(\mathbf{n})}{pr(\mathbf{N} > \mathbf{0})} \tag{7}$$

where

$$pr(\mathbf{N} > \mathbf{0}) = \prod_{i=1}^{k} pr(N_i > 0) = \prod_{i=1}^{k} (1 - pr(N_i = 0)) = \prod_{i=1}^{k} (1 - e^{-\lambda_i}). \tag{8}$$

The unconditional distribution of $\mathscr{T}$ will then be

$$\bar{F}(z) = \sum_{\mathbf{n} > \mathbf{0}} q(\mathbf{n}) \bar{F}(z|k-1,n-k,\delta(\mathbf{n})) \tag{9}$$

with

$$\delta(\mathbf{n}) = \frac{1}{\sigma^2} \left[\sum_{i=1}^{k} n_i \mu_i^2 - \frac{1}{n} \left(\sum_{i=1}^{k} n_i \mu_i \right)^2 \right] \tag{10}$$

while for $\mathscr{F}$ the unconditional distribution will be

$$F(z) = \sum_{\mathbf{n} > \mathbf{0}} q(\mathbf{n}) \bar{F}(\frac{k-1}{n-k} z|k-1,n-k,\delta(\mathbf{n})). \tag{11}$$

TRUNCATION ERRORS

To compute the values of $\bar{F}(z)$ and $F(z)$ we must truncate the corresponding series. Since, whatever z, r and s, we have

$$0 < \bar{F}(z|r,s,\delta) < 1; 0 < z \tag{12}$$

the truncation errors, when we only consider the terms with $\mathbf{n} \leq \mathbf{n}^o$, will be bounded by

$$b(\mathbf{n}^o) = \sum_{\mathbf{n} \nleq \mathbf{n}^o} q(\mathbf{n}) = \frac{1 - \sum_{\mathbf{n} \nleq \mathbf{n}^o} p(\mathbf{n})}{pr(\mathbf{N} > \mathbf{0})} = \frac{1 - pr(\mathbf{N} \leq \mathbf{n}^o)}{pr(\mathbf{N} > \mathbf{0})} = \frac{1}{pr(\mathbf{N} > \mathbf{0})} - \frac{pr(\mathbf{N} \leq \mathbf{n}^o)}{pr(\mathbf{N} > \mathbf{0})} \tag{13}$$

and since $N_1,...,N_k$ are independent Poisson variables with parameters $\lambda_1,...,\lambda_k$, we have

$$b(\mathbf{n}^o) = \frac{1}{pr(\mathbf{N} > \mathbf{0})} - \frac{\prod_{i=1}^{k} (pr(N_i \leq n_i^o))}{pr(\mathbf{N} > \mathbf{0})} = \prod_{i=1}^{k} \frac{1}{1 - e^{-\lambda_i}} - \prod_{i=1}^{k} \frac{\sum_{n_i=0}^{n_i^o} e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!}}{1 - e^{-\lambda_i}}. \tag{14}$$

So, if we choose n_i^o such

$$\sum_{n_i=0}^{n_i^o} e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!} > 1 - \varepsilon, i = 1,...,k \tag{15}$$

we will have

$$b(\mathbf{n}^o) < \prod_{i=1}^{k} \frac{1}{1-e^{-\lambda_i}} - \prod_{i=1}^{k} \frac{1-\varepsilon}{1-e^{-\lambda_i}} = \frac{1-(1-\varepsilon)^k}{\prod_{i=1}^{k}(1-e^{-\lambda_i})} \approx \frac{k\varepsilon}{\prod_{i=1}^{k}(1-e^{-\lambda_i})}. \tag{16}$$

Considering $\lambda_0 = Min\{\lambda_1, ..., \lambda_k\}$ we have

$$\prod_{i=1}^{k}(1-e^{-\lambda_i}) \geq (1-e^{-\lambda^o}) \tag{17}$$

so that

$$b(\mathbf{n}^o) < \frac{k\varepsilon}{(1-e^{-\lambda^o})^k}. \tag{18}$$

For different values of λ(the average dimension of the samples) and small ε, we present in Table 1 the minimal number n^o of terms to have

$$\sum_{n=0}^{n^0} e^{-\lambda} \frac{\lambda^n}{n!} > 1-\varepsilon. \tag{19}$$

So, given the number k of samples, we may chose the threshold ε such that the truncation error is controlled since it

TABLE 1. The minimal dimension of the samples n^o required to have the truncation error controlled

	λ				
minimum n^o	1	2	5	10	20
$\varepsilon = 10^{-4}$	6	9	15	24	39
$\varepsilon = 10^{-6}$	9	12	19	28	45
$\varepsilon = 10^{-8}$	11	14	22	32	50

will be bounded by $b(\mathbf{n}^o)$, when we require $n_i < n^o; i = 1, ..., k$.

For different values of k, λ and ε, we present in Table 2 the upper bound for the truncation error.

TABLE 2. Limit for the $b(\mathbf{n}^o)$

	k=2 λ					k=5 λ					k=10 λ				
ε	1	2	5	10	20	1	2	5	10	20	1	2	5	10	20
10^{-4}	5,0E-04	2,7E-04	2,0E-04	2,0E-04	2,0E-04	5,0E-03	1,0E-03	5,2E-04	5,0E-04	5,0E-04	9,8E-02	4,3E-03	1,1E-03	1,0E-03	1,0E-03
10^{-6}	5,0E-06	2,7E-06	2,0E-06	2,0E-06	2,0E-06	5,0E-05	1,0E-05	5,2E-06	5,0E-06	5,0E-06	9,8E-04	4,3E-05	1,1E-05	1,0E-05	1,0E-05
10^{-8}	5,0E-08	2,7E-08	2,0E-08	2,0E-08	2,0E-08	5,0E-07	1,0E-07	5,2E-08	5,0E-08	5,0E-08	9,8E-06	4,3E-07	1,1E-07	1,0E-07	1,0E-07

SIMULTANEOUS CONFIDENCE INTERVALS

Given $\mathbf{n}$ the vector of sample sizes, let $\mathbf{Y}.(\mathbf{n})$ be the vector of sample means and $S(\mathbf{n})$ the sum of sums of squares of residuals for the k samples. Then, $\mathbf{Y}.(\mathbf{n})$ and $S(\mathbf{n})$ are conditionally independent. Moreover, $\mathbf{Y}.(\mathbf{n})$ will be conditionally normal with mean vector μ, with components $\mu_1, ..., \mu_k$, and variance-covariance matrix $\sigma^2 D(\frac{1}{n_1}, ..., \frac{1}{n_k})$, while $S(\mathbf{n})$ will be the product of a chi-square with $n-k$ degrees of freedom. With $f_{1-q,k-1,n-k}$ the $(1-q)$-th quantile for a F distribution with $k-1$ and $n-k$ degrees of freedom, we have

$$pr\left[\bigcap_{\mathbf{a}\in\mathbb{R}^k}\left(|\mathbf{a}^T\mu - \mathbf{a}^T\mathbf{Y}.(\mathbf{n})| \leq \sqrt{(k-1)f_{1-q,k-1,n-k}\mathbf{a}^T D(\frac{1}{n_1}, ..., \frac{1}{n_k})\mathbf{a}\frac{S(\mathbf{n})}{n-k}}\right)\right] = 1-q \tag{20}$$

where $\bigcap$ indicates that all vectors $\mathbf{a} \in \mathbb{R}^k$ are considered, see [4].

Taking $N = \sum_{i=1}^{k} N_i$, we get

$$pr\left[\bigcap_{\mathbf{a}\in\mathbb{R}^k}\left(|\mathbf{a}^T\mu - \mathbf{a}^T\mathbf{Y}.(\mathbf{N})| \leq \sqrt{(k-1)f_{1-q,k-1,N-k}\mathbf{a}^T D(\frac{1}{N_1}, ..., \frac{1}{N_k})\mathbf{a}\frac{S(\mathbf{N})}{N-k}}\right)\right] = 1-q. \tag{21}$$

We thus obtained a version of the Scheffé theorem, for simultaneous confidence intervals, when sample sizes are random.

This result was included in our paper given the intimate connection there is between these simultaneous intervals and the F tests. Thus, see Scheffé(1959), the q level F test rejects the hypothesis $H_0 : \mu_1 = ... = \mu_k$ if and only if there exists $\mathbf{a}^o$ such that

$$\sqrt{(k-1)f_{1-q,k-1,n-k}\mathbf{a}^{oT}D(\frac{1}{n_1},...,\frac{1}{n_k})\mathbf{a}^o\frac{S(\mathbf{n})}{n-k}} < |\mathbf{a}^{oT}\mu - \mathbf{a}^{oT}\mathbf{Y}.(\mathbf{n})|. \tag{22}$$

OPTIMUM PROPERTIES OF THE F-TESTS

The F tests for fixed effects models, see for instance [1], are Uniformly Most Powerful (UMP) in the classes of tests such that,

1. the power is function of a non-centrality parameter;
2. they are invariant for transformation that leave the non- centrality parameter unchanged.

Moreover, see [3], if we have controlled heteroscedasticity, this is, a variance-covariance matrix $\sigma^2\mathbf{C}$, with $\mathbf{C}$ known, the F tests are UMP in the class of tests invariant for transformation

$$\mathbf{y}^+ = \mathbf{L}(\mathbf{y}+\mathbf{d}) \tag{23}$$

with $\mathbf{L}$ regular, and that, after reduction of the heteroscedasticity, they have power which depends only on non-centrality parameters. This last result is a generalization of (2.) since it replaces the requirement of homoscedasticity by that of controlled heteroscedasticity.

Since $\mathbf{C}$ is regular there exist, see [2], matrices $\mathbf{G}$ such that $\mathbf{GCG}^T = \mathbf{I}_n$ and so $\mathbf{GL}^{-1}(\mathbf{LCL}^T)(\mathbf{GL}^{-1})^T = \mathbf{I}_n$ so, working with

$$\mathbf{y}' = \mathbf{Gy} \tag{24}$$

or with

$$\mathbf{y}^{+'} = \mathbf{GL}^{-1}\mathbf{y}^+ \tag{25}$$

we are working with homoscedastic vectors. Now the previous results hold whatever the vector $\mathbf{n}$ of sample sizes. Representing by $pow(\delta|\mathbf{n})$ and $pow'(\delta|\mathbf{n})$ the powers of the F and of another test of the same size, belonging to one of the classes mentioned above, we have $pow'(\delta|\mathbf{n}) \leq pow(\delta|\mathbf{n})$, thus unconditioning we get

$$pow'(\delta) = \sum_{\mathbf{n}>\mathbf{0}} q(\mathbf{n})pow'(\delta|\mathbf{n}) \leq \sum_{\mathbf{n}>\mathbf{0}} q(\mathbf{n})pow(\delta|\mathbf{n}) = pow(\delta) \tag{26}$$

where $pow(\delta)$ and $pow'(\delta)$ are the unconditional powers of the F and the other test. We thus established

Proposition 0.1 *When we have a positive random vector of sample sizes, an F test is UMP in the classes of tests such that*

1. *the power is function of a non-centrality parameter;*
2. *are invariant for transformations that leave non-centrality parameters unchanged;*
3. *are invariant for transformations $\mathbf{y}^+ = \mathbf{L}(\mathbf{y}+\mathbf{d})$, with $\mathbf{L}$ regular, and that, after reduction of the heteroscedasticity, have power which depends only on non-centrality parameters.*

Clearly these optimum properties apply to the tests we have been studying.

REFERENCES

1. Lehmann E.L.(1959) Testing Statistical Hypotheses. 1ed. publ. by Wiley 1959. Springer. New York.
2. Mexia, J.T. (1995) Introdução à Inferência Estatística Linear. Centro de Estudos de Matemática Aplicada. Edições Universitárias Lusófonas. Lisboa.
3. Mexia, J.T. (1989) Controlled Heteroscedasticity, Quociente Vector Spaces and F Tests for Hypothesis on Mean Vectors. Trabalhos de Investigação Nž1. Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.
4. Scheffé, H. (1959) The Analysis of Variance. John Willey & Sons. New York.