# Sample size computing for factorial designs: An extension of the Athenian representative method 

Prof. CONSTANTINOS N. TSIANTIS<br>Systems Control Lab, Department of Energy Technology<br>Technological Educational Institution of Athens<br>Agiou Spyridonos Street, Aegaleo 12210 Athens<br>GREECE<br>cotsiant@live.com


#### Abstract

The paper considers sample size computing as a problem of representation. It clarifies the notion of representation and concentrates on the simultaneous representation of: i) a finite population divided into $r \times c$ cells provided by the crossing of levels of two factors, and ii) the respective set of values that a variable of interest takes upon the population subjects. Through a probabilistic approach, the sample size formula for the full factorial design, with non-zero mean differences among cells and interaction among factors, was derived. The formula verifies, that the statistical sample size $n_{s t}$ is the product of population representation $n_{p}$ and variance (SD) representation $n_{s}$ multiplied by the factor $R$ (sum of cell percentages). This law is an extension of the Athenian law of representation. It applies both for balanced and unbalanced factorial designs and operates as the representation standard in sample size computing for a multi-way (n-way) ANOVA.


Key-Words: - Population representation, Statistical representation, Total cell variance, Representation standards, ANOVAN.

## 1 Introduction

Sample size computing is one of the central issues in statistics and influences the cost of research and the validity of research outcomes. Most of the formulae in sample size computing are based today on the assumptions of normal curve, infinite population and the equality of variances for the population classes [8, 12, 13, 17]. These assumptions are rarely satisfied in practice and undermine the accuracy of statistical results in many cases [3,5].

During the last years many efforts have been developed to face this problem. Two basic principles were used in this respect: the probability principle (every element of the population has a known probability of being included in the sample) and the principle of representation [9, 16]. The latter principle, the roots of which go back to the ancient origins of political science, mathematics, philosophy and art, was the title of Kiaer's
monograph [7] published in Oslo, over a century ago, and which inspired the development of landmark works in the field $[1,4,6,10,11,18]$.

According to the representative principle, the statistical space as made of two distinct subspaces: one which refers to the population $N$, and the other to the set of values that a variable of interest Y takes upon the population subjects [15, 16]. It is exactly in respect to this two-dimensional space that sampling is defined as a problem of statistical representation, part of which constitutes the representation of population.

A closed mathematical formula for the representation of a finite stratified population was presented last year to the WSEAS American Conference on Applied Mathematics in Harvard [15]. The formula was called the "Athenian law of representation" (since it could justify the number of representatives in the parliament of ancient Athens) and signified that sample size is the product of
population representation and of representation of variance (standard deviation).

In the present article the previous formula of representation was extended for the case of twodimensional $r \times c$ full factorial design, with nonzero mean differences and interaction effects among cells. This case simulates that of multi-way (n-way) ANOVA design. By using the concept of total cell variance and a probabilistic expression of representation principle, the formula of sample size computing for the case of full factorial design was derived. The formula states that the sample size $n_{s t}$ is the product of population representation $n_{p}$ and variance (standard deviation) representation $n_{s}$ multiplied by R (the sum of cell percentages). It applies both for balanced and unbalanced factorial designs and operates as representation standard in sample size computing.

## 2 The problem of statistical representation

Table 1

|  | Column factor $\rightarrow$ Column levels (c) |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 | 2 | $\ldots$ | c-1 | c |  |
| Row <br> factor <br> $\downarrow$ Row <br> Levels <br> (r) | 1 |  |  |  |  |  | $\mu_{1 *}$ |
|  |  |  |  |  |  |  | $\mu_{2}{ }^{*}$ |
|  | $2$ |  |  | $\mathrm{N}_{\mathrm{ij}}$ |  |  | : |
|  |  |  |  | $\mathrm{w}_{\mathrm{ij}}$ |  |  | : |
|  |  |  |  | $\mathrm{v}_{\mathrm{ij}}$ |  |  |  |
|  | r-1r |  |  |  |  |  | $\mu_{\mathrm{r}-1}{ }^{*}$ |
|  |  |  |  |  |  |  | $\mu_{\mathrm{r}^{*}}$ |
|  |  | $\mu_{*}$ | $\mu_{* 1}$ | $\ldots$ | $\mu_{*}{ }_{c-1}$ | $\mu_{*}$ |  |

A population of size $N$ is classified in respect to two factors, the raw factor R, with levels $1,2, \ldots$ r, and the column factor C , with levels $1,2, \ldots \mathrm{c}$. The number of cells if formed through the crossing of factorial levels $r \times c$. The number of subjects per level for the row factor is $\mathrm{N}_{\mathrm{r} 1}, \mathrm{~N}_{\mathrm{r} 2}, \ldots, \mathrm{~N}_{\mathrm{r}, \mathrm{r}-1}, \mathrm{~N}_{\mathrm{r}, \mathrm{r}}$ and the respective percentages $\mathrm{w}_{\mathrm{r} 1}, \mathrm{w}_{\mathrm{r} 2}, \ldots, \mathrm{w}_{\mathrm{r}, \mathrm{r}}$. The number of subjects per level for the column factor is $\mathrm{N}_{\mathrm{c} 1}, \mathrm{~N}_{\mathrm{c} 2}, \ldots, \mathrm{~N}_{\mathrm{c}, \mathrm{c}-1}, \mathrm{~N}_{\mathrm{cc}}$ and the respective percentages $\mathrm{w}_{\mathrm{c} 1}, \mathrm{w}_{\mathrm{c} 2}, \ldots, \mathrm{w}_{\mathrm{cc}} \quad\left(\Sigma \mathrm{N}_{\mathrm{ri}}=\mathrm{N}_{\mathrm{r}}, \Sigma \mathrm{N}_{\mathrm{cj}}=\mathrm{N}_{\mathrm{c}}\right.$, $\Sigma \mathrm{w}_{\mathrm{ri}}=1, \Sigma \mathrm{w}_{\mathrm{cj}}=1$, for $\mathrm{i}=1,2, \ldots, \mathrm{r}$, and $\left.\mathrm{j}=1,2, \ldots, \mathrm{c}\right)$. The population percentage for the cell (i,j) is $\mathrm{w}_{\mathrm{ij}},=\mathrm{w}_{\mathrm{i}}{ }^{*} \mathrm{w}_{\mathrm{j}}\left(\sum_{\mathrm{i}}^{\mathrm{ij}} \mathrm{=}\right)$ and the respective number of population subjects is $\mathrm{N}_{\mathrm{ij}}=\mathrm{W}_{\mathrm{ij}} \mathrm{N}$, where $\mathrm{N}=\mathrm{N}_{\mathrm{r}}{ }^{*} \mathrm{~N}_{\mathrm{c}}$. The population $N$ is examined in respect to some variable of interest Y . The respective statistical parameters are: grand mean $\mu$, standard deviation $\sigma$, column means $\mu_{*_{1},}, \mu_{*_{2}}, \ldots, \mu_{* c}$, raw means $\mu_{1^{*}}$,
$\mu_{2^{*}}, \ldots, \mu_{r^{*}}$, cell means $\mathrm{m}_{\mathrm{ij}}$, and cell standard deviations $\sigma_{\mathrm{ij}}$. A sample of size $n$, consisted of $n_{i j}$ subjects per cell, is drawn randomly from the population. What is the value of $n$ and its allocation ( $n_{i j}$ ) so that, on the basis of available information (data), the minimum valid representation of the population to be achieved?

## 3 Problem Solution

## 3. 1 The principle of representation

The essence of representation principle centres on the notions of similarity and probability. Similarity [2] refers to the similarity of structures between the space under representation and the representative space: the sample. The similarity aspect of representation is expressed through the equality of respective class proportions between the statistical space and the sample. The probability aspect is expressed by the equality: the probability that an element of cell-(i,j) from the statistical space be selected and included in the sample must be equal to its proportion $\lambda_{i j}$ in it $(\mathrm{i}=1,2, \ldots, r$ and $j=1,2, \ldots c)$.

The optimal usage of the representation principle demands the maximum possible exploitation of the information available for the population (its size and composition) and the variable of interest (data of measurement).

### 3.2 The law of statistical representation

### 3.2.1 General consideration

The population space $N$ is the sum of $r \times c$ subspaces or population cells $\left(N=\Sigma N_{i j}\right)$. If $\sigma$ could be expressed, also, as the sum of $r \times c$ cells of standard deviations (SDs), analogous to the SDs $\sigma_{\mathrm{Tij}}$ defined below, then a unified statistical space of $r \times c$ cells could be formed as product of N and $\sigma$. For that space the question of representative sample $n$ could be raised.

### 3.2.2 The population standard deviation $\sigma$ as sum of $r \times c$ cells

1. The total variability of a set of values can be expressed by the total sum of squares $\mathrm{SS}_{\text {Total }}$. This sum is partitioned into
(i) the column sum of squares SSc
(ii) the rows sum of squares SSr ,
(iii) the error sum of squares SSe , and
(iv) the interaction sum of squares SSint

If $\mathrm{w}_{i}$ is the population percentage for the $i$-level of raw factor $(i=1,2, \ldots, r)$, and $\mathrm{w}_{\mathrm{j}}$ the population
percentage for the $j$-level of column factor $(\mathrm{j}=1,2, \ldots, \mathrm{c})$, then the population percentage of the cell $(\mathrm{i}, \mathrm{j})$ is $\mathrm{w}_{\mathrm{ij}}=\mathrm{W}_{\mathrm{i}} \mathrm{W}_{\mathrm{j}}$. If $\delta_{i^{*}}$ is the difference of the row mean $\mu_{i^{*}}$ (no matter what the number of column) from the grand mean $\mu$, and $\delta_{*_{j}}$ the difference for the column mean $\mu_{*}$, then the sums of squares are written as follows:

$$
\begin{equation*}
S S R=\sum_{i j k}\left(\mu_{i \bullet}-\mu\right)^{2}=N \sum_{i=1}^{r} \sum_{j=1}^{c} w_{i j} \delta_{i \bullet}^{2} \tag{1}
\end{equation*}
$$

where $\Sigma \delta_{i^{*}}=0$.

$$
\begin{equation*}
S S C=\sum_{i j k}\left(\mu_{\bullet j}-\mu\right)^{2}=N \sum_{i=1}^{r} \sum_{j=1}^{c} w_{i j} \delta_{\bullet j}^{2} \tag{2}
\end{equation*}
$$

where $\Sigma \delta_{*_{\mathrm{i}}}=0$.

$$
\begin{align*}
S S \text { int } & =\sum_{i j k}\left(\mu_{i j}-\mu_{i \bullet}-\mu_{\bullet j}+\mu\right)^{2} \\
& =N \sum_{i=1}^{r} \sum_{j=1}^{c} w_{i j}\left(\mu_{i j}-\mu_{i \bullet}-\mu_{\bullet j}+\mu\right)^{2} \tag{3}
\end{align*}
$$

where $\Sigma\left(\mu_{\mathrm{ij}}-\mu_{\mathrm{i}^{*}}-\mu_{*_{\mathrm{j}}}+\mu\right)=0$.

$$
\begin{align*}
& S S E=\sum_{i j k}\left(Y_{i j k}-\mu_{i j}\right)^{2}=\sum_{i j k} \hat{\varepsilon}_{i j k}^{2}= \\
& =\sum_{i j}\left(N w_{i j}-1\right) \sigma_{i, j}^{2}=N \sum_{i j} w_{i j} \sigma_{i, j}^{2}-\sum_{i j} \sigma_{i, j}^{2} \tag{4}
\end{align*}
$$

where $\varepsilon_{\mathrm{ijk}}$ is assumed to be independent with mean zero and common variance $\sigma^{2}$. For unbalanced designs the above analysis has to be adjusted appropriately [14]. Also,

$$
\begin{equation*}
S S T=\sigma^{2}(N-1) \tag{5}
\end{equation*}
$$

The population variance then is

$$
\begin{align*}
& \sigma^{2}=\frac{N}{N-1}\left(\sum_{i=1}^{r} \sum_{j=1}^{c} w_{i j}\left(\delta_{i \bullet}^{2}+\delta_{\bullet j}^{2}+\operatorname{int}_{i, j}^{2}+\sigma_{i, j}^{2}\right)\right) \\
& -\frac{1}{N-1} \sum_{i j} \sigma_{i, j}^{2} \tag{6}
\end{align*}
$$

We define now the total variance of a population cell by the equation

$$
\begin{equation*}
\sigma_{T i j}^{2} \equiv N w_{i j}\left(\delta_{i \bullet}^{2}+\delta_{\bullet j}^{2}+\operatorname{int}_{i, j}^{2}+\sigma_{i, j}^{2}\right)-\sigma_{i j}^{2} \tag{7}
\end{equation*}
$$

The whole population variance, therefore, is
$\sigma^{2}=\frac{1}{N-1}\left(\sum_{i=1}^{r} \sum_{j=1}^{c} \sigma_{T i j}^{2}\right)$
The respective population $\mathrm{SD} \sigma$ can be written then as the sum of SDs of cells $\sigma_{\mathrm{Tij}}$ :

$$
\begin{equation*}
\sigma \equiv k \sum_{i j} \sigma_{T i j} \tag{9}
\end{equation*}
$$

where k is a positive constant.
2. The proportion by which the total standard deviation $\sigma_{\mathrm{Tij}}$ of cell $(\mathrm{i}, \mathrm{j})$ contributes to the sum of cells total SDs is
$v_{i j}=\frac{\sigma_{T i j}}{\sum_{i, j} \sigma_{T i j}}$
where
$\sum_{i, j} v_{i j}=1$

### 3.2.3 Deriving the statistical law of representation

The problem has two dimensions. The first one refers to the finite population $N$ and its factorization into $N_{i j}$ cells $(i=1,2, \ldots, r, j=1,2, \ldots, c)$. The second one refers to the scores (the outcomes of measurement) of cell subjects and the respective means $\mu_{\mathrm{ij}}$ and standard deviations $\sigma_{\mathrm{Tij}}(i=1,2, \ldots, r$, $j=1,2, \ldots, c)$. The space of the problem is defined as the orthogonal product of above subspaces, that is:

$$
N_{i j} \sigma_{i^{\prime} j^{\prime}}=\left\{\begin{array}{lcc}
0 & \text { if } & i \neq i^{\prime} \text { and } j \neq j^{\prime}  \tag{12}\\
N_{i j} \sigma_{i j} & \text { if } & i=i^{\prime} \text { and } j=j^{\prime}
\end{array}\right.
$$

An element, therefore, of statistical space has the form $N_{i j}\left(k \sigma_{T i j}\right)$, where $N=\Sigma N_{i j}$ and $\sigma=\Sigma k \sigma_{T i j}$ ( $i=1,2, \ldots, r, j=1,2, \ldots, c$ ).
The probability that an element of the cell $(1,1)$ will be found in the statistical space is

$$
\begin{equation*}
p_{\text {prob } 11}=\frac{N_{11} \sigma_{T 11} k}{\sum_{i, j} N_{i j} \sigma_{T i j} k}=\frac{N w_{11} v_{11} \sum_{i j} \sigma_{T i j}}{N \sum_{i, j} w_{i j} v_{i j} \sum_{i j} \sigma_{T i j}}=\frac{w_{11} v_{11}}{\sum_{i=1}^{m} w_{i j} v_{i j}} \tag{13}
\end{equation*}
$$

where

$$
\begin{align*}
& N_{i j} \equiv w_{i j} N, \\
& \sigma_{T i j} \equiv v_{i j} \sum_{i j} \sigma_{T i j}  \tag{14}\\
& \sigma \equiv k \sum_{i j} \sigma_{T i j}
\end{align*}
$$

The probability that an element (of whatever cell) from the statistical space will be represented in the sample $n$ is

$$
\begin{align*}
p & =\frac{n}{\sum_{i, j} N_{i j} k \sigma_{T i j}}= \\
& =\frac{n}{\sum_{i, j} N w_{i j} k v_{i j} \sum_{i j} \sigma_{T i j}}=\frac{n}{N \sigma \sum_{i, j} w_{i j} v_{i j}} \tag{15}
\end{align*}
$$

Therefore, the probability $p_{11}$ that an element of the cell $(1,1)$ of statistical space will be represented in the sample $n$ equals to the product

$$
\begin{equation*}
p_{11}=p P_{p r o b ~} 11 \tag{16}
\end{equation*}
$$

Using Eqs 13 and 15 , the probability $\mathrm{p}_{11}$ becomes

$$
\begin{equation*}
p_{11}=\frac{w_{11} v_{11} n}{N \sigma\left(\sum_{i j} w_{i j} v_{i j}\right)^{2}}=\frac{w_{11} v_{11} n}{N \sigma R^{2}} \tag{17}
\end{equation*}
$$

where it was set
$R \equiv \sum_{i j} w_{i j} v_{i j}$
By selecting $\mathrm{n}_{11}$ elements at random from the statistical cell $(1,1)$, the probability $\mathrm{p}_{11 \mathrm{n}}$ of the event $\left(E_{11}\right)$ that an element of cell $(1,1)$ will be represented in the sample $n$ becomes:
$p_{11 n}=n_{11} \frac{w_{11} v_{11} n}{N \sigma R^{2}}$
But, according to the principle of representation, this probability must be equal to the probability implied by the proportion of respective cell in the sample $\lambda_{11}=n_{11} / \mathrm{n}$. That is,
$p_{11 n}=\lambda_{11}=n_{11} \frac{w_{11} v_{11} n}{N \sigma R^{2}}$
We repeat the aforementioned process for all the events $E_{i j}(i=1,2, \ldots, r, j=1,2, \ldots, c)$. Thus, the probabilities of individual events $E_{I I}, \ldots, E_{I c}, \ldots, E_{r l}$, $\ldots, E_{r c}$ are:
$p_{11 n}=\lambda_{11}=n_{11} \frac{w_{11} v_{11} n}{N \sigma R^{2}}, \ldots \ldots, p_{1 c n}=\lambda_{1 c}=n_{1 c} \frac{w_{1 c} \nu_{1 c} n}{N \sigma R^{2}}$
$p_{21 n}=\lambda_{21}=n_{21} \frac{w_{21} v_{21} n}{N \sigma R^{2}}, \ldots \ldots, p_{2 c n}=\lambda_{2 c}=n_{2 c} \frac{w_{2 c} v_{2 c} n}{N \sigma R^{2}}$
$p_{r 1 n}=\lambda_{r 1}=n_{r 1} \frac{w_{r 1} v_{r 1} n}{N \sigma R^{2}}, \ldots \ldots ., p_{r c n}=\lambda_{r c}=n_{r c} \frac{w_{r c} v_{r c} n}{N \sigma R^{2}}$
(21)

Since the events $E_{i j}$ are independent, the probability of their intersection is the product of their probabilities:
$P\left(E_{11} \cap E_{12} \cap \ldots \cap E_{1 c} \cdots E_{r 1} \cap E_{r 2} \cap \ldots \cap E_{r c}\right)=$
$=\left(p_{11 n}\right)\left(p_{12 n}\right) \ldots\left(p_{1 c n}\right) \ldots\left(p_{r 1 n}\right)\left(p_{r 2 n}\right) \ldots\left(p_{r c n}\right)=$
$=\lambda_{11} \lambda_{12} \ldots \lambda_{1 c} \ldots \lambda_{r 1} \lambda_{r 2} \ldots \lambda_{r c}=$
$=n_{11} w_{11} v_{11} \ldots n_{1 c} w_{1 c} v_{1 c} \ldots n_{r 1} w_{r 1} v_{r 1} \ldots n_{r c} w_{r c} v_{r c}\left(\frac{n}{N \sigma_{a} R^{2}}\right)^{r c}$

$$
\begin{equation*}
n^{2 r c} w_{11} \ldots w_{1 c} \ldots w_{r 1} \ldots w_{r c} v_{11} \ldots v_{1 c} \ldots v_{r 1} \ldots v_{r c}=\left(N \sigma R^{2}\right)^{m} \tag{23}
\end{equation*}
$$

whereupon the law of statistical representation for the factorial design is derived:

$$
\begin{equation*}
n \equiv n_{s t}=R \sqrt{\frac{N \sigma}{\frac{r c}{r_{11} \ldots w_{1 c} \ldots v_{r 1} \ldots w_{r c}}} \sqrt[r c]{\sqrt{v_{11} \ldots v_{1 c} \ldots v_{r 1} \ldots v_{r c}}}} \tag{24}
\end{equation*}
$$

By setting
$n_{p} \equiv \sqrt{\frac{N}{\sqrt{r c} \sqrt{w_{11} \ldots w_{1 c} \cdots w_{r 1} \ldots w_{r c}}}}$
we have the autonomous representation of population (independent of the set of variable's values) for the factorial design, analogous to the Athenian law of population representation [15].
By setting also
$n_{s} \equiv \sqrt{\frac{\sigma}{r c} \sqrt{v_{11} \ldots v_{1 c} \ldots v_{r 1} \ldots v_{r c}}}$
we have the representation of standard deviation (variance) for the factorial design. The, by using Eqs. 24, 25 and 26, we arrive at the general formula

$$
\begin{equation*}
n_{s t}=R n_{p} n_{s} \tag{27}
\end{equation*}
$$

In conclusion, the statistical sample size $\mathrm{n}_{\text {st }}$ is the product of population representation $n_{p}$ and standard deviation (variance) representation $n_{s}$ multiplied by the factor R: the sum of products percentages $\mathrm{w}_{\mathrm{ij}} \mathrm{v}_{\mathrm{ij}}$.
The allocation of sample size, according to the similarity principle (Neyman type) is then
$n_{i j}=\lambda_{i j} n_{s t}=\frac{N_{i j} \sigma_{T i j}}{\sum_{i, j} N_{i j} \sigma_{T i j}} n_{s t}=\frac{w_{i j} v_{i j}}{\sum_{i, j} w_{i j} v_{i j}} n_{s t}$

### 3.2.4 Application and discussion

A researcher deals with the issue of problem solving techniques and wants to test the impact of analogous methodology in the achievements of elementary and high school students [12]. For this purpose he/she selects 21 students from the elementary school and 21 students from the high school who are divided into subgroups of seven, and asked to solve some problems. In the first subgroup (A1), named the control group, he/she does not provide any kind of help or hint. In the second and third subgroup (A2 and A3), the experimental ones, the researcher presents a set of analogous problems, asking students to spend some time in dealing with them. To the third subgroup he/she provides the additional information that the administered problems are similar to the problems presented earlier. During the process of problem
solving the researcher recorded the number of mistakes that the students made. The test was repeated twice and the outcomes are illustrated in Table 2. The question is this: Is the given number of students and their allocation in treatment cells adequate so that the researcher can claim that his/her hypotheses about means are valid in respect to the representation principle for the given context? What should this number be on the basis of available test data?

| Table 2 |  | Factor A: METHOD |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | A1 |  | A2 |  | A3 |  |
|  |  | 38 | 37 | 20 | 18 | 10 | 11 |
|  |  |  | 45 |  |  | 8 | 8 |
|  |  |  |  |  | 20 | 12 | 9 |
|  |  |  |  |  |  | 5 | 7 |
|  |  | 33 |  |  | 29 | 6 | 4 |
|  |  | 31 | 35 |  |  | 8 | 12 |
|  |  | 33 |  | 36 |  | 15 | 9 |
|  |  | 34 | 34 | 29 |  | 37 | 40 |
|  |  |  |  |  | 48 | 45 | 46 |
|  |  | 37 | 35 |  | 29 | 48 | 46 |
|  |  |  | 27 |  |  | 29 | 27 |
|  |  |  | 43 |  |  | 34 | 33 |
|  |  |  |  |  |  | 32 | 34 |
|  |  |  |  |  |  | 51 | 52 |

## Solution

We have $\mathrm{N}=84, \mathrm{r}=2$ and $\mathrm{c}=3$. The cell population percentages are the same: $\mathrm{w}^{\mathrm{ij}}=\mathrm{w}=7 / 42=0.1667$.

Therefore, $\mathrm{R}=\mathrm{w} \Sigma v^{\mathrm{ij}}=\mathrm{w}$. The denominator factor for the population is
$f_{w} \equiv \sqrt[r c]{w_{11} \ldots w_{1 c} \ldots w_{r 1} \ldots w_{r c}}=\left(w^{r c}\right)^{1 / r c}=w=0.1667$
The mean of the whole sample is $\mu=30.2738$. The raw means are $\mu_{1^{*}}=23.0714$ and $\mu_{2^{*}}=37.4762$. The column means are: $\mu_{*_{1}}=35.8571$, $\mu_{*_{2}}=30.7500, \mu_{*_{3}}=24.2143$. The cell means are:

$$
\mu(\mathrm{i}, \mathrm{j})=\left[\begin{array}{ccc}
36.5714 & 23.7857 & 8.8571 \\
35.1429 & 37.7143 & 39.5714
\end{array}\right]
$$

The standard deviation of the whole sample is $s=$ 12.5989. The cells standard deviations are:

$$
\mathrm{s}(\mathrm{i}, \mathrm{j})=\left[\begin{array}{ccc}
5.3308 & 6.0913 & 2.9835 \\
6.4074 & 8.2409 & 8.3548
\end{array}\right]
$$

The sums of squares per cell for the rows, columns, interactions and errors are respectively:

$$
\begin{aligned}
& \operatorname{SR}(\mathrm{i}, \mathrm{j})=\left[\begin{array}{lll}
726.2401 & 726.2401 & 726.2401 \\
726.2401 & 726.2401 & 726.2401
\end{array}\right] \\
& \mathrm{SC}(\mathrm{i}, \mathrm{j})=\left[\begin{array}{lll}
436.4306 & 3.1746 & 514.0496 \\
436.4306 & 3.1746 & 514.0496
\end{array}\right] \\
& \operatorname{Sint}(\mathrm{i}, \mathrm{j})=\left[\begin{array}{lll}
877.4306 & 0.7937 & 931.0020 \\
877.4306 & 0.7937 & 931.0020
\end{array}\right] \\
& \operatorname{Ser}(\mathrm{i}, \mathrm{j})=\left[\begin{array}{lll}
369.4286 & 482.3571 & 115.7143 \\
533.7143 & 882.8571 & 907.4286
\end{array}\right]
\end{aligned}
$$

The total variance per cell is then:

$$
\mathrm{s}_{\mathrm{T}(\mathrm{i}, \mathrm{j})}^{2}=\left[\begin{array}{lll}
2409.5 & 1212.6 & 2287.0 \\
2573.8 & 1613.1 & 3078.7
\end{array}\right]
$$

and the total standard deviation per cell is

$$
\mathrm{S}_{\mathrm{T}(\mathrm{i}, \mathrm{j})}=\left[\begin{array}{lll}
49.0870 & 34.8219 & 47.8227 \\
50.7328 & 40.1630 & 55.4862
\end{array}\right]
$$

Thus, the total SD percentages per cell are

$$
v_{(i, j)}=\left[\begin{array}{lll}
0.1765 & 0.1252 & 0.1720 \\
0.1824 & 0.1444 & 0.1995
\end{array}\right]
$$

The denominator factor for the total SDs per cells is then

$$
f v \equiv \sqrt[r c]{v_{11} \ldots v_{1 c} \ldots v_{r 1} \ldots v_{r c}}=0.1647
$$

The size of the representative sample, therefore, is

$$
n_{s t}=R \sqrt{\frac{N s}{f_{w} f_{v}}}=\sqrt{\frac{N w s}{f_{v}}}=\sqrt{\frac{84 * 0.1667 * 12.5989}{0.1647}}
$$

that is, $\mathrm{n}_{\mathrm{st}}=32.7236<42$. This means that the sample selected by the researcher is of adequate size in respect to the representative principle. The allocation of sample size (Eq.28), since $\mathrm{w}_{\mathrm{ij}}=$ const (balanced design), is $n_{i j}=v_{i j} n_{s t}$. Thus we take
$\mathrm{n}_{(\mathrm{i}, \mathrm{j})}=\left[\begin{array}{lll}5.7757 & 4.0972 & 5.6269 \\ 5.9694 & 4.7257 & 6.5287\end{array}\right]$
This means that the maximum integer number of subjects per cell is 7 . For equal cells then we have
$\mathrm{n}_{(\mathrm{i}, \mathrm{j})}=\left[\begin{array}{lll}7 & 7 & 7 \\ 7 & 7 & 7\end{array}\right]$
The analysis of variance for this problem through the Matlab n-way ANOVA routine is illustrated in Table 3. Our calculations presented above provide the same results.

## 4 Conclusion

The principle of representation, including the aspects of similarity and probability, may reshape many types of problem solutions in statistics and other fields. Statistical representation centres on the problem of simultaneous representation of population and variance structure.

The law of statistical representation for the case of $r \times c$ full factorial design (with non-zero mean differences and interaction among classes) verifies, that the statistical sample size $n_{s t}$ is the product of population representation and SD representation multiplied by the sum of cell percentages. The method can be applied both for balanced and unbalanced factorial designs.
This law is an extension of the Athenian law of representation and operates as representation standard in sample size computing for the case of multi-way (n-way) ANOVA design.

## References:

[1] Deming, W. E.(1966), Some theory of sampling, New York: Dover Publications, 226-230.
[2] Edelman, S.(1998), Representation is representation of similarities, Behavioral and Brain Sciences, Vol.21, 449-498.
[3] Glass, G.V., Reckham, P.D. \& Sanders, J.R. (1972), Consequences of failure to meet the assumptions underlying the use of analysis of variance and covariance. Review of Educational Research, 42, 237-288. .
[4] Horvitz, D.G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, Vol. 47, 663-685.
[5] Ioannidis, JP (2005). Why most published research findings are false. PLoS Med.; e124. doi: 10.1371/journal.pmed. 0020124.
[6] Jensen, A. (1926). The representative method in practice, Bulletin of the International Statistical Institute, 22, pt. 1, 359-439.
[7] Kiaer, A.W. (1895). The Representative Method of Statistical Surveys, English translation, 1976, Oslo: Statistik Sentralbyro.
[8] Kirk, R.E.(1978), Introductory Statistics, Wadsworth Publishing, 308-312.
[9] Kish, L.(2002). New Paradigms (Models) for Probability Sampling, Survey Methodology, Vol. 28, No. 1, pp. 31-34.
[10] Kruskal, W.H., and Mosteller, F. (1979). Representative sampling, I: Nonscientific literature. International Statistical Review, 47, 13-24.
[11] Neyman, J.(1934), On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection, Journal of Royal Statistical Society, Vol.97, No.4, 558-625.
[12] Roussos, P., J. Tsaousis, Applied Statistics for the Social Sciences (in Greek), Athens: Hellinika Grammata, 2002, p. 310.
[13] Stigler, S.M. (1986). History of Statistics, Cambridge: Harvard University Press.
[14] Searle, S.R. (1987), Linear Models for Unbalanced Data, John Wiley \& Sons, New York.
[15] Tsiantis, C.N. (2008), "Computing the sample size of multivariate populations: the Athenian law of representation and sampling", American Conference on Applied Mathematics, Harvard Graduate School of Education, Cambridge, MA, USA, March 24-26, WSEAS Press. pp.301-306.
[16] Traat I., Bondesson L. and Meister, K. (2004), Sampling design and sample selection through distribution theory, Journal of Statistical Planning and Inference 123, 395-413.
[17] Tryfos, P (1996). Sampling methods for applied research, New York: John Wiley, 104105.
[18] Yates, F. (1949), Sampling Methods for censuses and Surveys, New York: Hafner Publishing Co.

