Analysis of variance (ANOVA)

- Prof. Giuseppe Verlato
- Unit of Epidemiology & Medical Statistics, Department of Diagnostics & Public Health, University of Verona

ANALYSIS OF VARIANCE

To evaluate significance of differences between the means of independent samples:

t-test

2 sample means

ANOVA

3 or more sample means

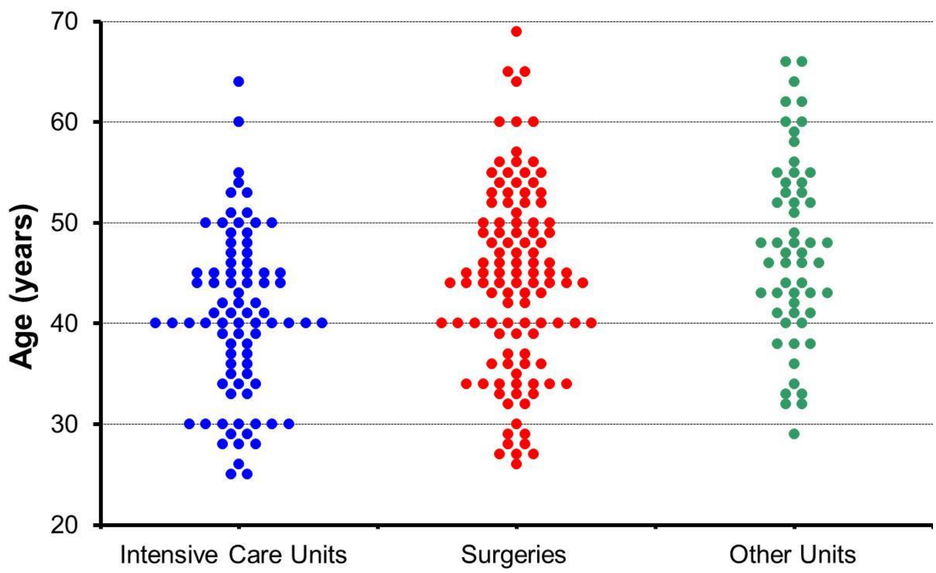
Analysis of variance - 1

There are k groups, each with a variable number of units.

Every statistical units is usually identified by two numbers in subscript: the 1° number identifies the group, the 2° number identifies the rank within the group.

group 1	group 2	group 3	………..	group k
x_{11}	x_{21}	x_{31}	………..	x_{k1}
x_{12}	x_{22}	x_{32}	………..	x_{k2}
x_{13}	x_{23}	x_{33}	………..	x_{k3}
x_{14}	x_{24}	x_{34}	………..	x_{k4}
x_{15}	x_{25}	x_{35}	………..	x_{k5}
x_{16}	x_{26}	x_{36}	………..	x_{k6}
x_{17}	x_{27}	x_{37}	………..	x_{k7}
x_{18}		x_{38}	………..	x_{k8}
		x_{39}	………..	x_{k9}
$\bar{x}_{1.}$	$\bar{x}_{2.}$	$\bar{x}_{3.}$	………..	$\bar{x}_{k.}$

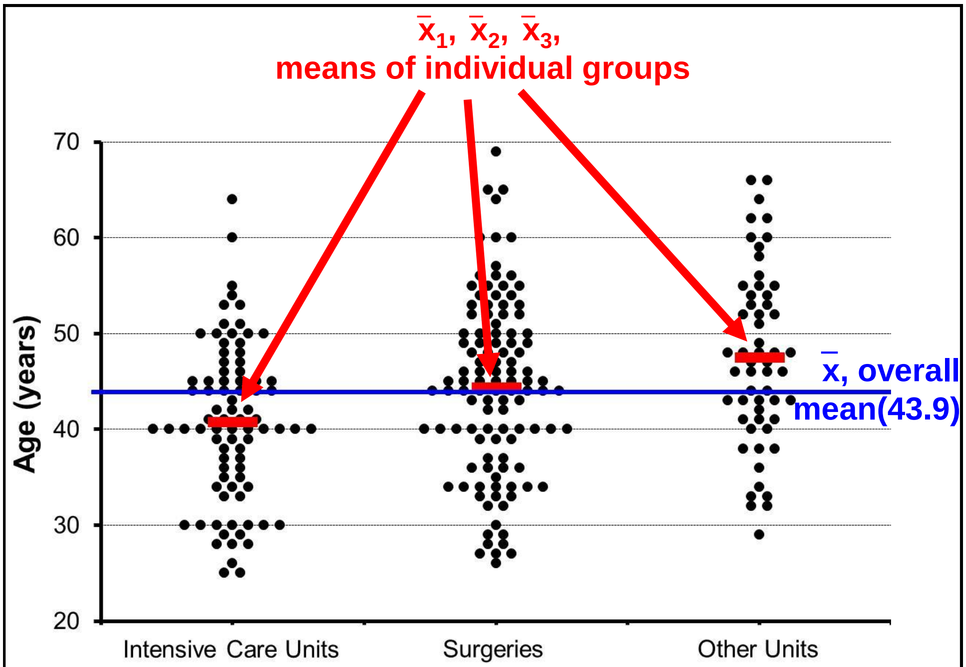
EXAMPLE: age of physicians working in the Intensive Care Units, Surgeries and other Units of a Veneto Hospital in 2002.



ANALYSIS OF VARIANCE - 2

In addition to the overall mean, $\bar{x}$, there are k means, one for each individual group, $\bar{x}_{1.}$, $\bar{x}_{2.}$, $\bar{x}_{3.}$, …., $\bar{x}_{k.}$

```
. tabstat age,by(Unit) stat(n mean sd skewness kurtosis)
Summary for variables: eta by categories of: reparto
         Unit |    N      mean         sd   skewness   kurtosis
-------------+-------------------------------------------------
IntensiveCare|   83   40.75904   8.187722   .1365529   2.816861
      Surgery |  112   44.35714   9.241365   .1166111   2.696551
        other |   56   47.44643   9.172924   .1151416   2.376357
-------------+-------------------------------------------------
        Total |  251   43.85657   9.198008   .1764211   2.703104
---------------------------------------------------------------
```



mean±SD   40.8 ± 8.2   44.4 ± 9.2   47.4 ± 9.2

Analysis of Variance - 3

Hypotheses
$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_0 \\ H_1: \text{at least one mean differs from the other ones} \end{cases}$$

To choose between the two hypotheses, should we perform several t-tests, comparing all possible pairs of means ?

NO, we shouldn't. Indeed this approach would cause an inflation (abnormal increase) of α (alpha), probability of type I error.

For example, if 20 different *t*-tests are performed, each with a nominal α of 0.05, one test should turn out to be significant just by chance:

$\alpha_{actual} = 1 - (1- \alpha_{nominal})^n$, where n = number of tests performed

For instance, if 6 t-tests are performed at an $\alpha_{nominal}$ of 0.05,

$$\alpha_{actual} = 1 - (1- 0.05)^6 = 1 - 0.95^6 = 1 - 0.735 = 0.265$$

ANALYSIS OF VARIANCE - 4

SIDAK's CORRECTION

$$\alpha_{corrected} = 1 - (1- \alpha_{nominal})^{1/n} = 1 - \sqrt[n]{(1- \alpha_{nominal})}$$

For instance if 6 t-tests are performed at an $\alpha_{nominal}$ of 0.05,

$$\alpha_{corrected} = 1 - (1- 0.05)^{1/6} = 1 - 0,95^{1/6} = 1 - 0.9915 = 0.0085$$

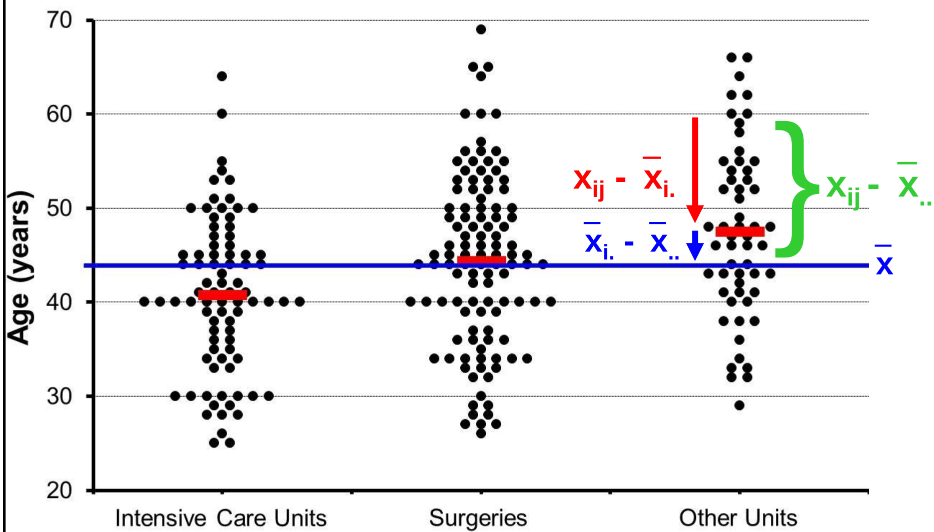
BONFERRONI's CORRECTION

An approximation consists in dividing α by the number of tests:

$$\alpha_{corrected} = 0.05/6 = 0.0083$$

Hence, it is appropriate to perform a global test, which simultaneously compare all groups: **analysis of variance**.

DECOMPOSING total SUM OF SQUARES in Analysis of Variance - 1



$x_{ij} - \bar{x}_{..}$ = deviation of a given observation (*j* value within group *i*) from the overall mean

$\bar{x}_{i.} - \bar{x}_{..}$ = deviation of the mean of group *i* from the overall mean

$x_{ij} - \bar{x}_{i.}$ = deviation of a given observation (*j* value within group *i*) from the mean of group *i*

DECOMPOSING total SUM OF SQUARES in Analysis of Variance - 2

For a given statistical unit:

$$\underbrace{(x_{ij} - \bar{x}_{..})}_{\text{Total variability}} = \underbrace{(\bar{x}_{i.} - \bar{x}_{..})}_{\text{Variability between groups}} + \underbrace{(x_{ij} - \bar{x}_{i.})}_{\text{Variability within groups}}$$

As regards the whole sample, it can be demonstrated that:

$$\underbrace{\sum_{i,j}(x_{ij} - \bar{x}_{..})^2}_{\text{Total Sum of Squares, SST}} = \underbrace{\sum_{i,j}(\bar{x}_{i.} - \bar{x}_{..})^2}_{\text{SSq between groups}} + \underbrace{\sum_{i,j}(x_{ij} - \bar{x}_{i.})^2}_{\text{SSq within groups}}$$

Computations involved in decomposing SSq

k groups should be compared, with different sizes $n_1, n_2, \ldots, n_i, \ldots n_k$. The following symbols adopted:

$$T_i = \sum_{j=1}^{n_i} x_{ij}\,, \quad G = \sum_{i=1}^{k} T_i\,, \quad N = \sum_{i=1}^{k} n_i$$

	group 1	group 2	group 3	…	group k	
	x_{11}	x_{21}	x_{31}	…	x_{k1}	
	x_{12}	x_{22}	x_{32}	…	x_{k2}	
	x_{13}	x_{23}	x_{33}	…	x_{k3}	
	x_{14}	x_{24}	x_{34}	…	x_{k4}	
	x_{15}	x_{25}	x_{35}	…	x_{k5}	
	x_{16}	x_{26}	x_{36}	…	x_{k6}	
	x_{17}	x_{27}	x_{37}	…	x_{k7}	
	x_{18}		x_{38}	…	x_{k8}	
			x_{39}	…	x_{k9}	
Total	T_1	T_2	T_3	…	T_k	G

DECOMPOSING TOTAL SUM OF SQUARES in Analysis of Variance - 3

Total SSq = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - (\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij})^2 / N$

SSq WITHIN groups = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} (\sum_{j=1}^{n_i} x_{ij})^2 / n_i$

SSq BETWEEN groups = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^{k} (\sum_{j=1}^{n_i} x_{ij})^2 / n_i - (\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij})^2 / N$

In Descriptive Statistics we learnt an alternative formula to compute SSq:

Sum of Squares = $\sum_i (x_i - \bar{x}_{..})^2 = \sum_i x_i^2 - (\sum_i x_i)^2 / N$

The same formula can be used to perform ANOVA. To compute TOTAL SSq, first data must be summed along columns (groups), and then column totals must be summed along the last row.

Total SSq = $\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N$

SSq WITHIN groups (residual SSq) is the algebraic sum of SSq of all individual groups.

SSq of group 1= $\sum_j (x_{1j} - \bar{x}_{1.})^2 = \sum_j x_{1j}^2 - (\sum_j x_{1j})^2 / n_1$

SSq of group 2= $\sum_j (x_{2j} - \bar{x}_{2.})^2 = \sum_j x_{2j}^2 - (\sum_j x_{2j})^2 / n_2$

SSq of group 3= $\sum_j (x_{3j} - \bar{x}_{3.})^2 = \sum_j x_{3j}^2 - (\sum_j x_{3j})^2 / n_3$

SUM of SSq = $\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = \sum_i \sum_j x_{ij}^2 - \sum_i (\sum_j x_{ij})^2 / n_i$

SSq BETWEEN groups= $\sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2$

SSq BETWEEN groups is computed as the difference between TOTAL SSq and SSq WITHIN groups.

total SSq - SSq within =

$$\sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N \quad - \quad [\sum_i \sum_j x_{ij}^2 - \sum_i (\sum_j x_{ij})^2 / n_i] =$$

$$\sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N \quad - \quad \sum_i \sum_j x_{ij}^2 + \sum_i (\sum_j x_{ij})^2 / n_i =$$

$$\sum_i (\sum_j x_{ij})^2 / n_i \quad - \quad (\sum_i \sum_j x_{ij})^2 / N$$

Inference on variance

Assumption n.1: Homoscedasticity
Variance is constant in all k groups.

If the assumption holds, it is possible to improve the estimate of common variance by pooling variance estimates, obtained from individual groups. **Within-group variance**, also called **residual variance**, can be estimated by dividing the sum of numerators (SSq of individual groups) by the sum of denominators (degrees of freedom of individual groups).

$$\text{Residual variance} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 / (N-k)$$

It can be demonstrated that the expected value for residual variance is the common variance σ^2.

Between-groups variance is estimated by dividing SSq between groups by the appropriate number of degrees of freedom, i.e. by the number of groups minus one (k-1).

$$\text{Between-groups variance} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2 / (k-1)$$

Assumption n. 2: Observations (i.e. errors) are independent from each other.

If the assumption is true, the expected value for between-groups variance is equal to:

$$E(\text{between var}) = \sigma^2 + \sum_{i=1}^{k} n_i (\mu_{i.} - \bar{\mu})^2 / (k-1)$$

If H_0 is true, the expected value for between-groups variance is exactly σ^2 (residual variance); otherwise its value is greater than σ^2.

Assumption n.3: observations (errors) are normally distributed within each group.

If the assumption is true, a significance test can be properly performed. Indeed, both **within-groups variance (residual variance)** and **between-groups variance** are independent estimates of σ^2 under H_0.

The most suited test to compare these two variances is F-test:

F = **var BETWEEN** / **var WITHIN**

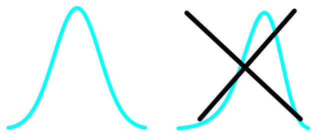
Under H_0 the test statistic follows the F distribution by Fisher-Snedecor with degrees of freedom (k-1) and (N-k).

The observed value is compared with a critical threshold $F_{\alpha,(k-1),(N-k)}$, reported by specific tables.

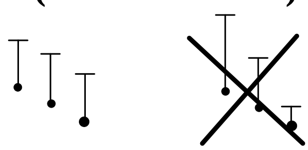
If $F > F_{\alpha,(k-1),(N-k)}$, H_0 is rejected ($P<\alpha$); otherwise H_0 is accepted.

ASSUMPTIONS OF PARAMETRIC TESTS

1. Normality



2. Homoscedasticity (stable variance)



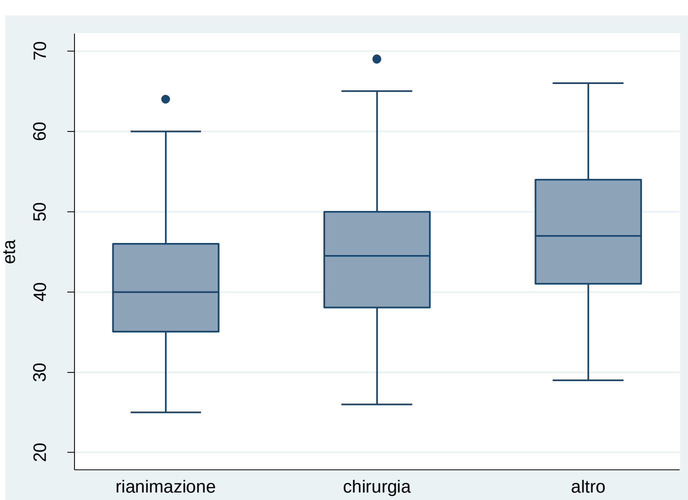
3. Independence of observations (errors)

Eyes, ears and teeth of the same patients are not independent

t-test
ANOVA
ANCOVA
correlation & regression

Bartlett's test for equal variances: chi2(2)=1.4959 Prob>chi2=0.473

Age variability does not significantly differ among Intensive Care Units, Surgeries and other Units.



Verifying the assumption of NORMALITY

```
. bysort UNIT:sfrancia AGE
--------------------------------------------------------------
-> UNIT = Intensive Care
                  Shapiro-Francia W' test for normal data
    Variable |    Obs       W'          V'         z       Prob>z
-------------+------------------------------------------------
         AGE |     83     0.98700       1.009      0.019   0.49260
--------------------------------------------------------------
-> UNIT = Surgery
                  Shapiro-Francia W' test for normal data
    Variable |    Obs       W'          V'         z       Prob>z
-------------+------------------------------------------------
         AGE |    112     0.99192       0.801     -0.461   0.67752
--------------------------------------------------------------
-> UNIT = other
                  Shapiro-Francia W' test for normal data
    Variable |    Obs       W'          V'         z       Prob>z
-------------+------------------------------------------------
         AGE |     56     0.99273       0.412     -1.784   0.96275
```

Table of one-way Analysis of Variance

Source of variability	Degrees of freedom	SSq	Variance	Test statistic (F value)
Between groups	k-1	$\Sigma\, n_i(\bar{x}_{i.} - \bar{x}_{..})^2$	SSq between/(k-1)	σ^2 between/ σ^2 within
Within groups	N-k	$\Sigma_i\, \Sigma_j(x_{ij} - \bar{x}_{i.})^2$	SSq within/(N-k)	
TOTAL	N-1	$\Sigma_i\, \Sigma_j(x_{ij} - \bar{x}_{..})^2$		

If H_0 is false, i.e. at least one mean differs from the others, then:

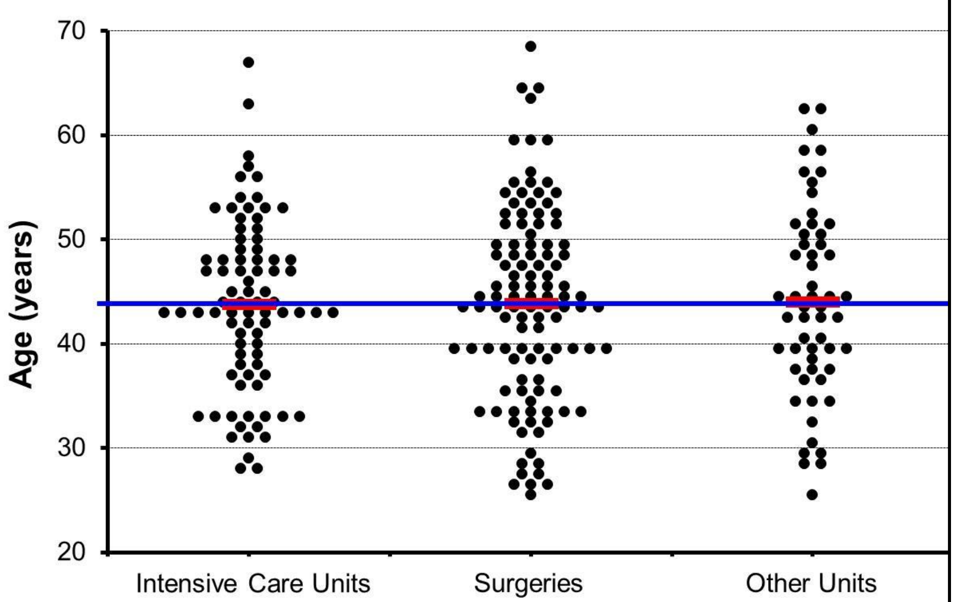
σ^2 between > σ^2 within ≈ σ^2

Does the age of physicians (n=251) significantly differ among Intensive Care units, Surgeries and other Units ?

Source of variability	Degrees of freedom	SSq	Variance	*F* value (significance)
Between groups	2	1546.10	773.05	9.779
Within groups	248	19604.73	79.05	($p<0.001$)
TOTAL	250	21150.84		

H_0 is rejected: age significantly differ among different Units.

Modified example: the groups have about the same mean



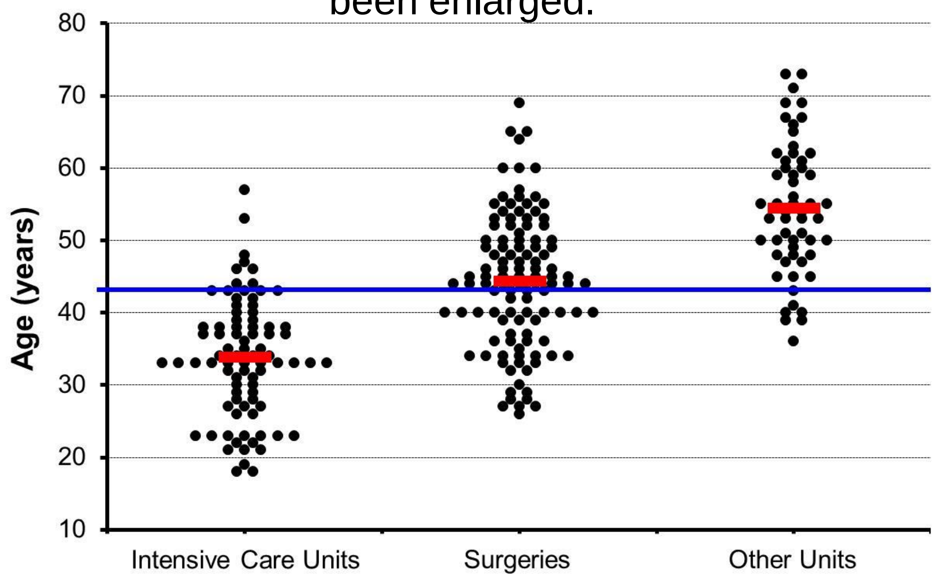
mean±SD 43.8 ± 8.2 43.9 ± 9.2 43.95 ± 9.2

Does the age of physicians (n=251) significantly differ among Intensive Care units, Surgeries and other Units ?

Source of variability	Degrees of freedom	SSq	Variance	*F* value (significance)
Between groups	2	1.21	0.60	0.008
Within groups	248	19604.73	79.05	($p=0.992$)
TOTAL	250	19605.94		

H_0 is accepted: age does not significantly differ among different Units.

Modified example: differences among groups have been enlarged.



mean±SD	33.8 ± 8.2	44.4 ± 9.2	54.45 ± 9.2

Does the age of physicians (n=251) significantly differ among Intensive Care units, Surgeries and other Units ?

Source of variability	Degrees of freedom	SSq	Variance	*F* value (significance)
Between groups	2	14628.6	7314.3	92.53
Within groups	248	19604.7	79.05	($p<0.001$)
TOTAL	250	34233.3		

H_0 is rejected: age significantly differ among different Units.

If the *F*-test (global test) is significant, it is feasible to compare individual means. For this purpose several tests are available, named as multiple comparisons or "post hoc" analysis.

1) Multiple comparisons are designed in order to avoid inflation of α, probability of type I error.

2) Multiple comparisons use residual variance, computed in the frame of analysis of variance, as an estimate of random variability.

MAIN TYPES OF MULTIPLE COMPARISONS

1. Scheffè's test: this is the most conservative test, i.e. the test with the highest protection against multiple testing bias and the lowest probability of detecting significant results. It allows to compare both individual means and pooled means.
2. Tukey's test: it allows to compare all possible pairs of means (pairwise comparisons).
3. Dunnett's test: it allows to compare individual means with a control mean.
4. Bonferroni's and Sidak's corrections are type of multiple comparisons.

H_0 is rejected: age significantly differ among different Units.

Comparison of AGE by UNIT (Bonferroni)

	IntensiveCare	Surgery
Surgery	3.59811	
	0.017	
other	6.68739	3.08929
	0.000	0.104