The Detection of Disease by Statistic Test of Analyze of Variance

Kaouther. El kourd[1], Amer El kourd[2]
[1]Electronic institute, Biskra
University Biskra
Algeria
[2]Department of computer
Islamic University
Gaza Palestine
kaouther_youcef@yahoo.fr, el_kourd@yahoo.com

ABSTRACT: *From this article we used "application the analyses of variance", Anova technique, which equal to* mean squares dividing by the mean squares of error, *then applied the statistic study on the pathological image. After that, compeer the result obtained with student technique. The logicial applied is Matlab7.0 to extract the lesion by two way: distribution of Gaussian curve (hypothesis test of ho) and directly on the pathologique image.*

Keywords: Regression, Fitted, Anova, Student

Received: 22 November 2012, Revised 29 December 2012, Accepted 2 January 2013

1. Introduction

Analysis of variance became widely known after being included in Fisher's 1925 book Statistical Methods for Research Workers.

In statistics, analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance in a particular variable is partitioned into componentsattributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes *t*-test to more than two groups. Doing multiple two-sample t-tests would result in an increased chance of committing an error. For this reason, ANOVAs are useful in comparing two, three, or more means. [1]

As problematic what's the main of ANOVA technique? and what's it's performance if we compeer it for example with Student one?

2. Distribution & probability

2.1 Central aim of statistical tests

Determining the likelihood of a value in a sample, given that the Null Hypothesis is true: P (value| H_0)

- H_0: no statistically significant difference between sample & population (or between samples)

- H_1: statistically significant difference between sample & population (or between samples)

For example the distribution of events in a population for P (value | H_0) < 0.05.See figure (1)

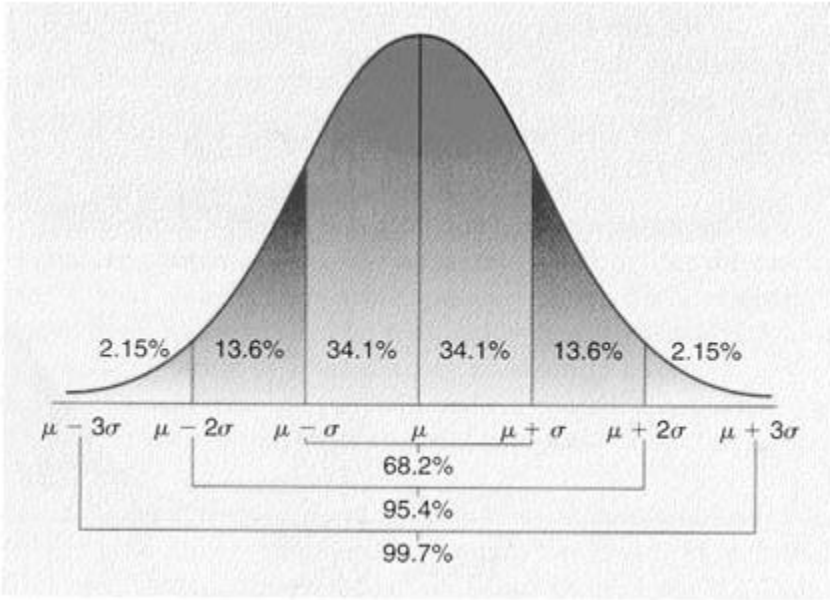

Figure 1. Distribution of events for p < 0.05 [2]

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \quad (1)$$

Wl

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

σ

μ : is Population mean

3. Variance

Instead of thinking about the group means, we can instead think about variances. Recall sample variance (equation (3)):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad (3)$$

F = Variance 1/Variance 2

$$F = \frac{s^2_{Model}}{s^2_{Error}} \quad (4)$$

F: is ANOVA (ANALYSIS OF VARIANCE)

Total variance = model variance(S^2 model) or between groups) + error variance(S^2E or within groups). [1] [2]

3. Correlation and Regression

Here we ask questions:

1) Is there a relationship between *x* and *y*?
Answer: yes if we applied the relation of covariance where it's relation is:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{n-1} \tag{5}$$

2) What is the strength of this relationship?

Answer: Pearson's role said: [1] [2]

- Covariance does not really tell us anything but we can measure the standardises the covariance value, and Divides the covariance by the multiplied standard deviations of X and Y: [1]

$$r_{xy} = \frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{(n-1)\ s_x s_y} \tag{6}$$

Where:

$s_x\ s_y$: Standard deviations

3) Can we describe this relationship and use it to predict *y* from *x*?

Answer: yes by using regressing then Fitting a line using the Least Squares solution.

3.1 Regression

Description about the relationship between two variables where one is dependent and the other is independent.

3.2 Fitted Regression Line

The true regression line corresponding to equation (1.a) is usually never known. However, the regression line can be estimated by estimating the coefficients β_1 and β_0 for an observed data set. [2]

$$E(Y) = \beta_0 + \beta_1 x \tag{1.a}$$

The actual values of *y*, (which are observed as yield from the chemical process from time to time and are random in nature), are assumed to be the sum of the mean value, *E*(*Y*), and a random error term, *e*: The actual values of *y*, (which are observed as yield from the chemical process from time to time and are random in nature), are assumed to be the sum of the mean value, *E*(*Y*), and a random error term, *e*: equation (1.b). [1] [2].

$$\begin{aligned} Y &= E(Y) + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon \end{aligned} \tag{1.b}$$

The estimates, β_1 and β_0, are calculated using least squares. The estimated regression line, obtained using the values of, β_1 and β_0, is called the *fitted* line. The least square estimates, β_1 and β_0, are obtained using the following equations: [1]

$$\hat{\beta}_1 - = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$\hat{\beta}_0 - = \bar{y} - \hat{\beta}_1 x \tag{3}$$

Where $\bar{y}$ is the mean of all the observed values and $\bar{x}$ is the mean of all values of the predictor variable at which the observations were taken $\bar{y}$ is calculated using and is calculated using $\bar{y} = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} y_i$ and $\bar{x}$ is calculated using

$$\bar{x} = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} x_i$$

Once β_1 and β_0 are known, the fitted regression line can be written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4}$$

where $\tilde{y}$ is the *fitted* or *estimated* value based on the fitted regression model. It is an estimate of the mean value, $E(Y)$. The fitted value, $\tilde{y}_t$, for a given value of the predictor variable, x_i, may be different from the corresponding observed value, y_i. The difference between the two values is called the *residual*, e_i:

$$e_i = y_i - \hat{y}_i \tag{5}$$

To calculate the Statistic F_o, it must pass by the six titles: [2] [3] [4] [5] [6] [7]

3.2.1 Total Sum of Squares (SS_T)

On simple linear regression that the total sum of squares, SS_T, is obtained using the following equation:

$$SS_T - \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{6}$$

The total sum of squares in matrix notation is:

$$\begin{aligned} SS_T &= y'y - \left(\frac{1}{n}\right) y'Jy \\ &= y' \left[I - \left(\frac{1}{n}\right) J\right] y \end{aligned} \tag{7}$$

Where y is the vector of observed values, I is the identity matrix of order n; & J represents an $n \times n$ square matrix of ones,

3.2. 2 Model Sum of Squares (SS_R)

Similarly, the model sum of squares or the regression sum of squares, SS_R, can be obtained in

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= \hat{y}'\hat{y} - \left(\frac{1}{n}\right) y'Jy \\ &= y' \left[H - \left(\frac{1}{n}\right) J\right] y \end{aligned} \tag{8}$$

matrix notation as:

Where H is the hat matrix and is calculated using

$$H = X(X'X)^{-1}X' \tag{9}$$

3.2.3 Error Sum of Squares

The error sum of squares or the residual sum of squares, SS_E, is obtained in the matrix notation from the vector of residuals, e, as:

$$SS_E = SS_T - SS_R \tag{10}$$

$$\begin{aligned} &= (y - \hat{y})'(y - \hat{y}) \\ &= y'(I - H)y \end{aligned} \tag{11}$$

3.2.4 Mean Squares (MS_T)

Mean squares are obtained by dividing the sum of squares with their associated degrees of freedom. The number of degrees of freedom associated with the total sum of squares, SS_T, is $(n-1)$ since there are n observations in all, but one degree of freedom is lost in the calculation of the sample mean $\bar{y}$. The total mean square is:

$$MS_T = \frac{SS_T}{n-1} \tag{12}$$

3.2.5 Regression mean square (MS_R)

The number of degrees of freedom associated with the regression sum of squares, SS_R, is k. There are $(k + 1)$ degrees of freedom associated with a regression model with $(k + 1)$ coefficients, $\beta_0, \beta_1, \beta_2 ... \beta_k$

However, one degree of freedom is lost because the deviations, $(\tilde{y}_t - \bar{y})$, are subjected to the constraints that they must sum to zero. $\sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2$ The regression mean square is:

$$MS_R = \frac{SS_R}{k} \tag{13}$$

The number of degrees of freedom associated with the error sum of squares is: $n - (k + 1)$, as there are n observations in all, but $(k + 1)$ degrees of freedom are lost in obtaining the estimates of $\beta_0, \beta_1, \beta_2 ... \beta_k$ to calculate the predicted values, $\tilde{y}_i$. The error mean square is:[2] [3] [4] [5] [6] [7]

$$MS_E = \frac{SS_E}{n - (k + 1)} \tag{14}$$

The error mean square, MS_E, is an estimate of the variance, σ^2, of the random error terms, e_i.

3.2.6 Mean square error (MS_E)

Is estimate variance $(\tilde{\sigma}^2)$ of random error e_i. see (equation 15)

$$\hat{\sigma}^2 = MS_E \tag{15}$$

3.2.7 Calculation of the Statistic F_o

Once the mean squares MS_R and MS_E are known, the statistic to test the significance of regression can be calculated as follows: [2]

$$F_o = \frac{MS_R}{MS_E} \tag{16}$$

4. Algorithm

1) It must converted MRI image IRM by (Dicom, Mrirco, SPM…) logigial to form (jpg.png, nii, img, hdr……..) for authorize the programmation with matlab.

2) choose a sample of image.

3) Analyse the sample -Analyze the data which is the protocol

- Estimate the linear regression with parameters estimates (image *x*).
- Extract the error between *y* & *y* estimate.
- Application the analyses of variance '*ANOVA*' .

4) Conception & results .

5) Comparison between Student & Anova.

5. Program execution

5.1 Analysis of data (the protocol radiologic)

Our protocol is for a patient aged 15 years; who is present in clinical context of late stature and puberty with pituitary dysfunction.

The patient is made by an MRI scan with injection of contrast medium. The machine is used radiologist type "*SIEMENS*", and with the field B = 1.5 Tesla. The sequences performed $T1$-weighted and $T2$.

MRI appearance of a tumor of the median line developed at the expense of the floor of the third ventricle and the left sidewall responsible for a moderate hydrocephalus by compression phenomenon of holes monro.

This aspect is mentioned first craniopharyngioma. Galiale tumor diagnosis remains in the range.

Gonalgiques disorders are due to the tumor development at the expense of the hypothalamic region.

Figure 1 present an image normal (left) & pathologique one (right).

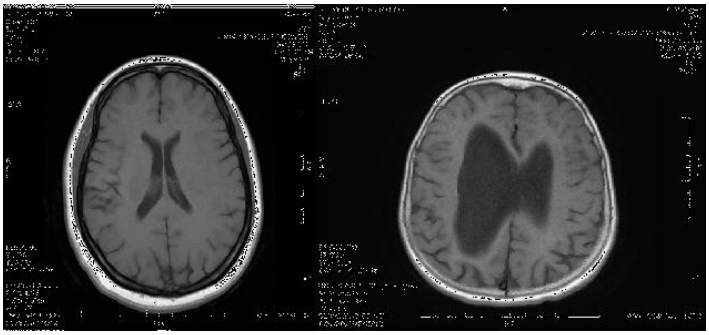

Figure 1. left. Normal image; right : Pathologique image

The figure (2) display the error (e) for one vector with length n = 200: see equation (5): $e = y - \hat{y}$

Where: y is the vector, & $\tilde{y}$ is data estimate.

Y is with Blue color; $\hat{y}$ is with Red color & error is with blue star.

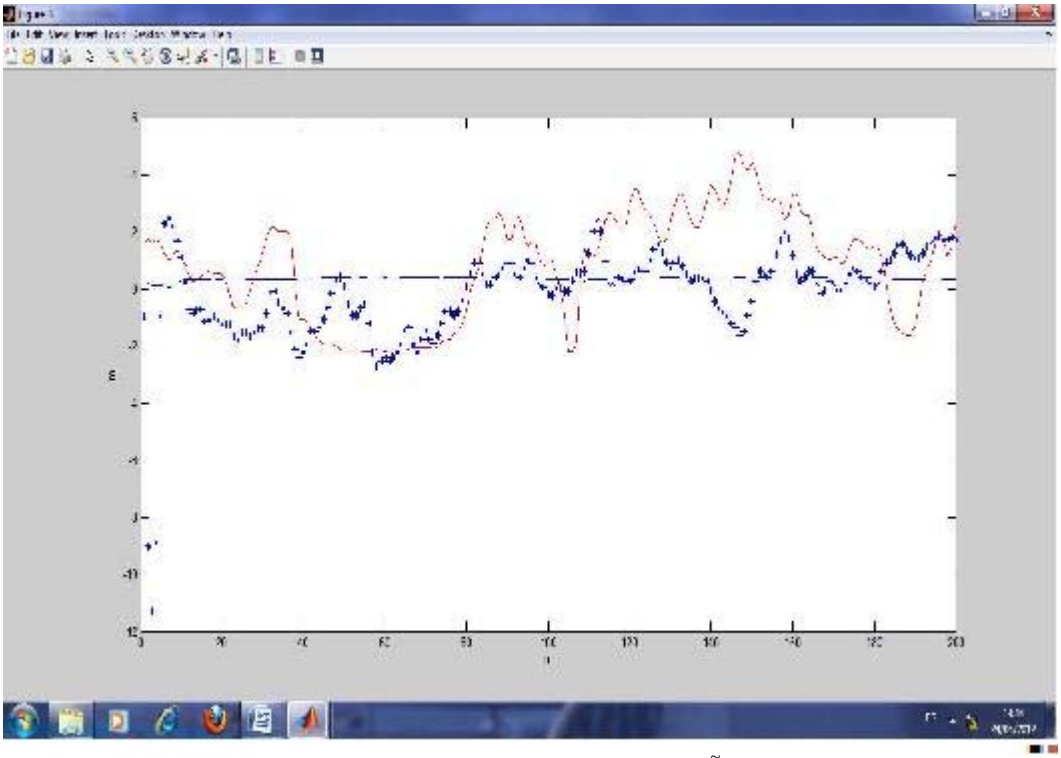

Figure 2. Presentation error e, y & $\tilde{Y}$

5.2 Calculate de Fcal: See equation (16) & figure (3)

$F\text{- }cal = MS_R / MS_E$

Where:

MS_R: The within-groups variation.

MS_E: the between-groups.

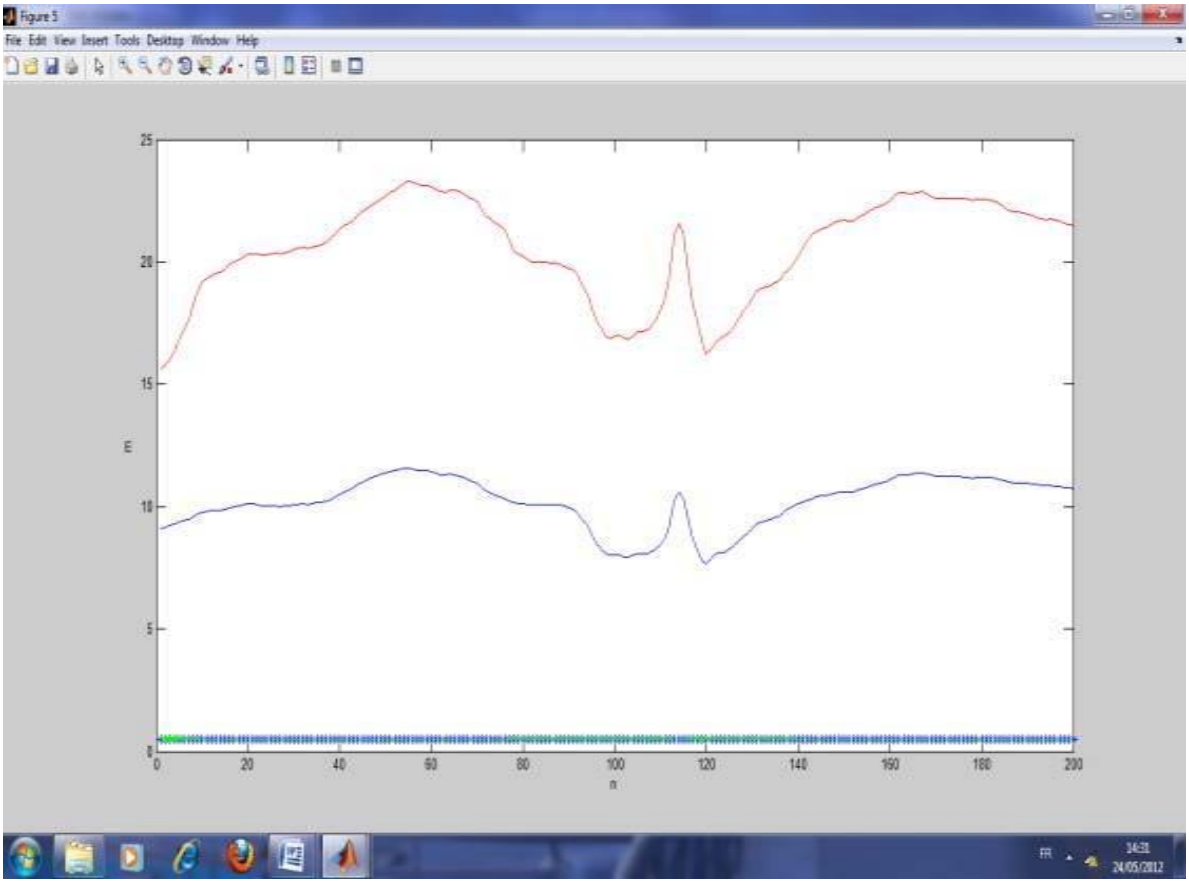

Figure 3. Presentation de F-cal (point blue), MS_R (red) color & MS_E (blue color)

5.3 Distribution Gaussian (figure 4)

From table of test (fisher & student for $\alpha = 0.01$) we have .

* Degree of liberty: ddl: $v = n - 1 = 200 - 1 = 199$.
* F – tab = 1.
* Hypothesis ho will be :

$$\begin{cases} \text{If } F\text{– cal} \le F\text{– tab} \Rightarrow f - \text{test with (anova)} \\ \rightarrow \text{accept } H_j. \\ \text{If } F\text{–cal} > \text{F– tab} \Rightarrow \textit{test } F\text{– test reject } H_o. \end{cases}$$

* Replace F-tab (± 1) on the gauss curve.

5.4 Surface 50×50

For surface 50×50 with probability alpha = 0.01); figure 1 present pathologique image, to see clearly the lesion; we have on it translate the figure1 to binary to extract with precise the place of lesion with anova technique (color white). See figure (2). In figure 3 the curve display the error of number of samples: (10 to 35), the length more then 35 or less then 10 becomes the error near the zeros. In this matrix; Fisher table (ftab) = 1.85, so all results from fisher calculate (fcal = fo) superior then this value will be reject (c.a.d ho : reject), & all results inferior then 1.85 will be accept (ho : accept). Figure 5 present the result but with student technique; where here there isn't any detection of lesion in t-test (student)).

5.5 Surface 100×100

For second example with surface 100×100; figure 2 give excellent result of the place of lesion in front of the figure 5 which present the lesion worse .ftab here equal 1.65 (see figure 4) where ho accept all results inferior then this value. For figure 3; the error present the same problem in the same place as surface (50×50) (see protocol) but with more amplitude.

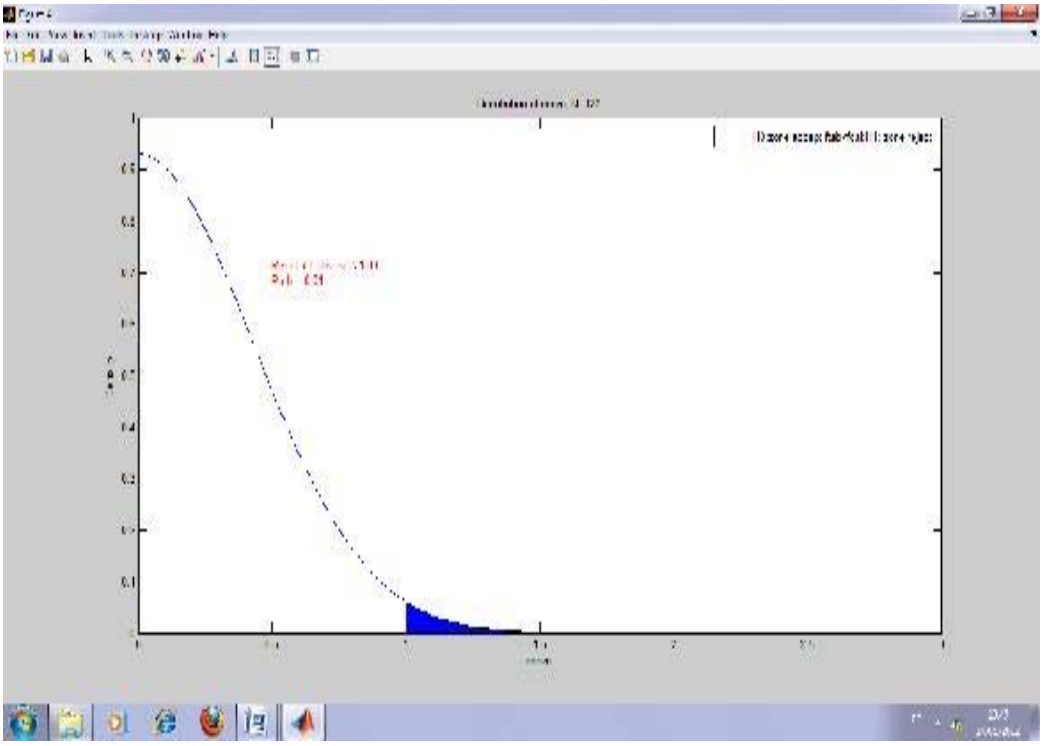

Figure 4. Gaussian Distribution of anova-test

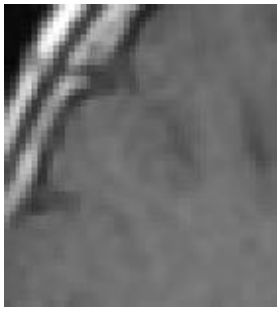

Figure 5.1. Pathologic image

Figure 5.2. Detection of lesion with anova

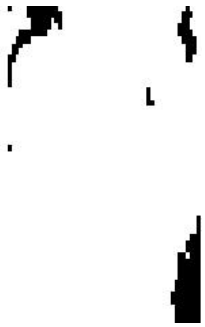

Figure 5.3. Pathologic image

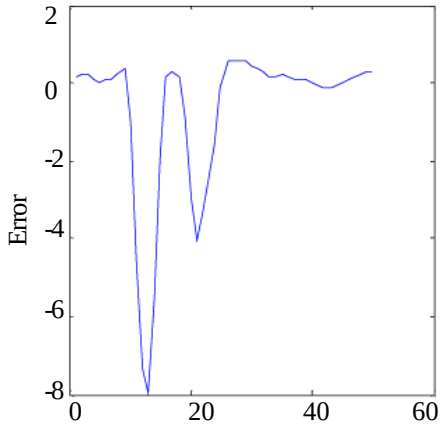

Figure 6. Nomber of sample

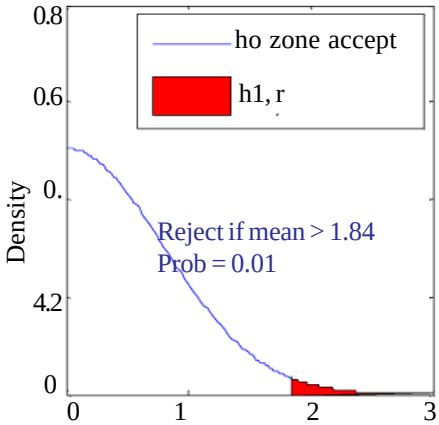

Figure 7. Anova ftab

Figure 8.1. Pathologic image

Figure 8.2. Detection of lesion with anova

Figure 8.3. Pathologic image

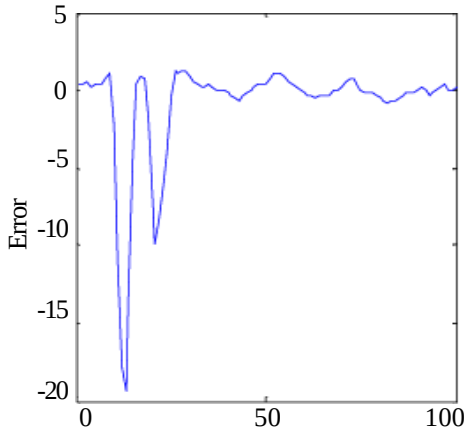

Figure 9. Nomber of sample

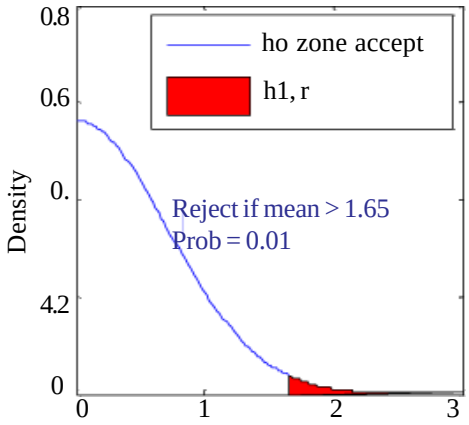

Figure 10. Anova ftab

5.6 Surface 200 × 200

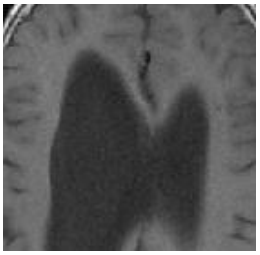

Figure 11.1. Pathologic image

Figure 11.2. Detection of lesion with anova

Figure 11.3. Pathologic image with student

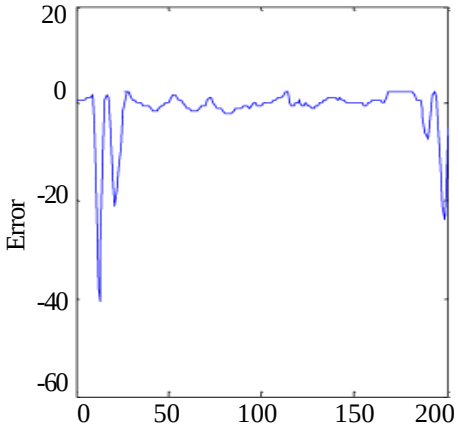

Figure 12. Nomber of sample

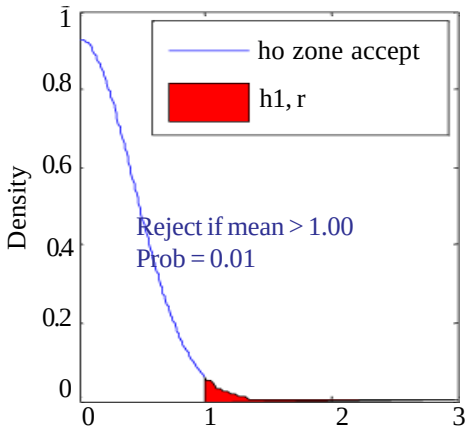

Figure 13. Anova ftab

6. Conclusion

From these article we can distinct:

- Our study is applied for comparison between two models (normal image & pathological one) for medical image & (new model and old one) in the other specialty as: biologic, architect, medicament….ect.

- Anova techniques is used for small surface but student is used for small & middle surface

- The result with Anova is more precise in front of ttest.

We propose using as perspective the technique of anova but for regression non linear.

References

[1] http://en.wikipedia.org/w/index.php?title = Factorial_ANOVA&redirect = no, 12/6/2012 at 15:00.

[2] http://www.weibull.com/DOEWeb/introduction.ht;1 2/6/2012, 15-45.

[3] Research Methods I, ANOVA and Multiple Regression,

[4] UFR SPSE-Master, Université Paris X – Nanterre, PMP EF 205, Méthodes Statistiques pour l'analyse de données en psychologie. Chapitre 5. Comparaison deplusieursmoyennespourdes échantillonsindépendants (ANOVA), 2008-2009.

[5] Viviane Kostrubiec. Les comparaisons multiples: entre mythe et réalité, Laboratoire Adaptations Perceptivo-Motrices et Apprentissage (EA 3191), Université Paul Sabatier–Toulouse III.

[6] Jon Roiser, Predrag Petrovic. (2006). t-tests, ANOVA and regression, February 1st.

[7]Henson, R., Penny, W. (2005). ANOVAs and SPM, Institute of Cognitive Neuroscience,Wellcome Department of Imaging Neuroscience, University College London. July 12.