ICAI – GITI/GITT

# ANOVA
## Estadística II

Eugenio Sánchez Úbeda

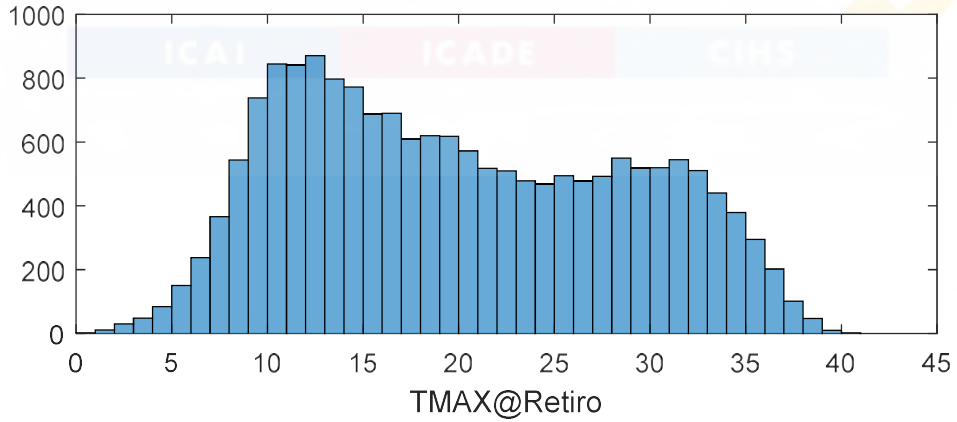
January 2024

1

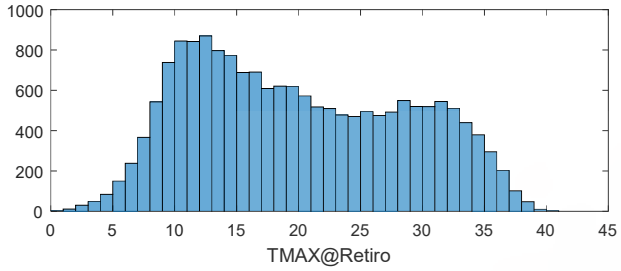
# Introduction

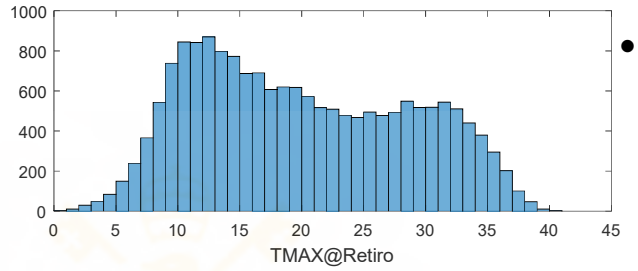
# Analysis of Variance (ANOVA) Introduction

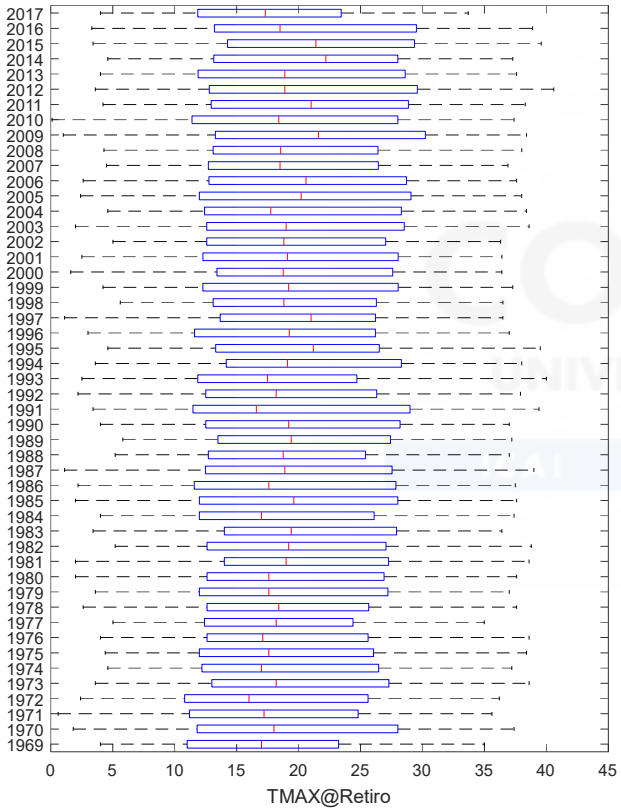
- ANOVA stands for ANalysis Of VAriance
- Sensible answer to questions of this type
  - Does the maximum temperature value in Madrid depend on the year?
  - Does the maximum temperature value in Madrid depend on the month?
  - Does the maximum temperature value in Madrid depend on the month and the year?

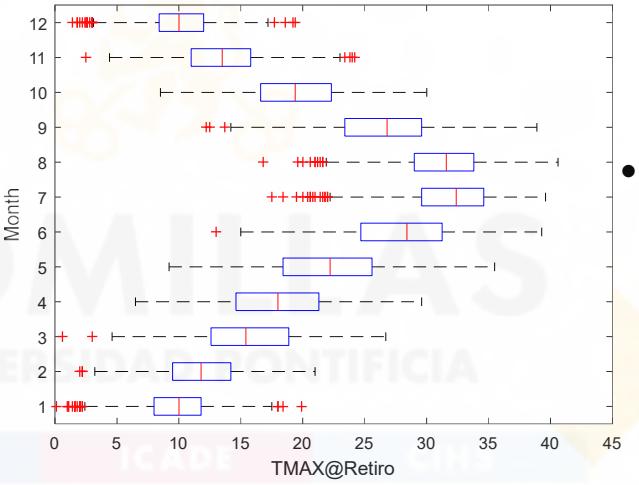


# Analysis of Variance (ANOVA) Introduction



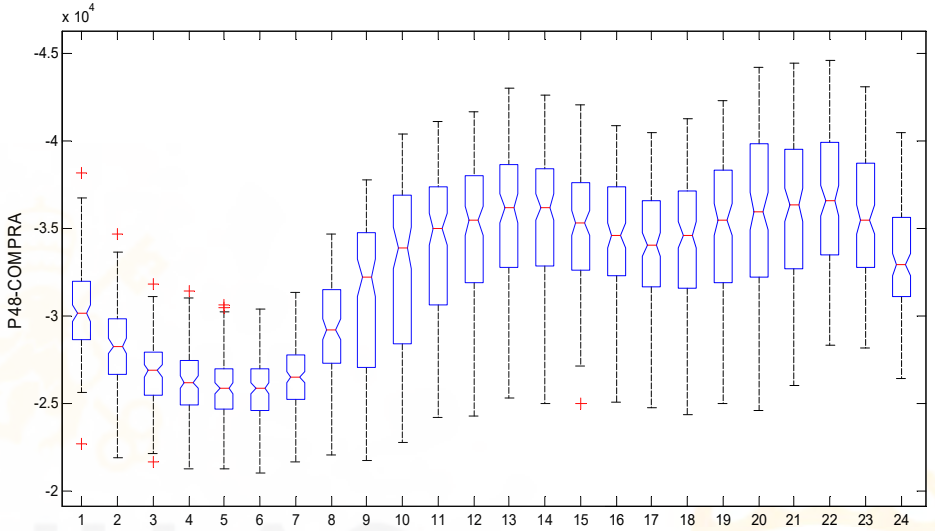




- The observed variance in the output is partitioned into components attributable to different sources of variation
- ANOVA provides a statistical test of whether the means of several groups are equal

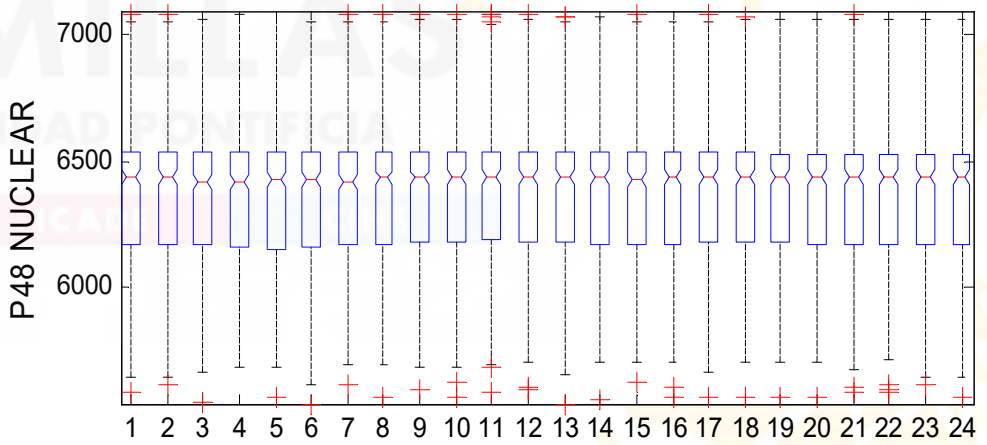
Strongly related to the box-and-whisker plots
Equality of means doesn't imply equality of medians

# Analysis of Variance (ANOVA)
## Example: Nuclear output & demand in electricity market

The expected value of the electric demand changes significantly with the hour



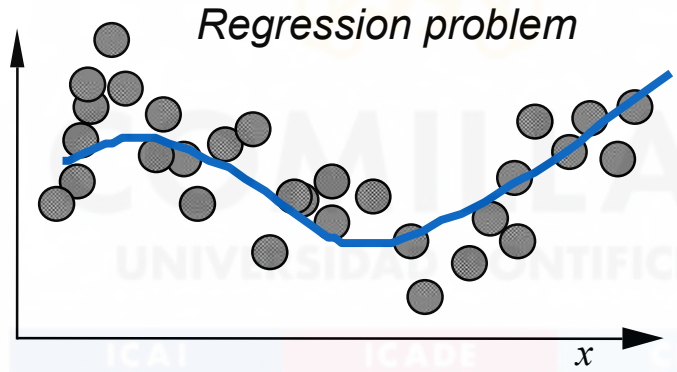

The expected value of the final program of total nuclear output doesn't change significantly with the hour (factor with 24 groups)




Two medians are significantly different at the 5% significance level if their intervals do not overlap. Interval endpoints are the extremes of the notches or the centers of the triangular markers. The extremes correspond to $q_2 - 1.57(q_3 - q_1)/\text{sqrt}(n)$ and $q_2 + 1.57(q_3 - q_1)/\text{sqrt}(n)$, where $q_2$ is the median (50th percentile), $q_1$ and $q_3$ are the 25th and 75th percentiles, respectively, and $n$ is the number of observations without any NaN values. When the sample size is small, notches may extend beyond the end of the box.

# Analysis of Variance (ANOVA) Introduction

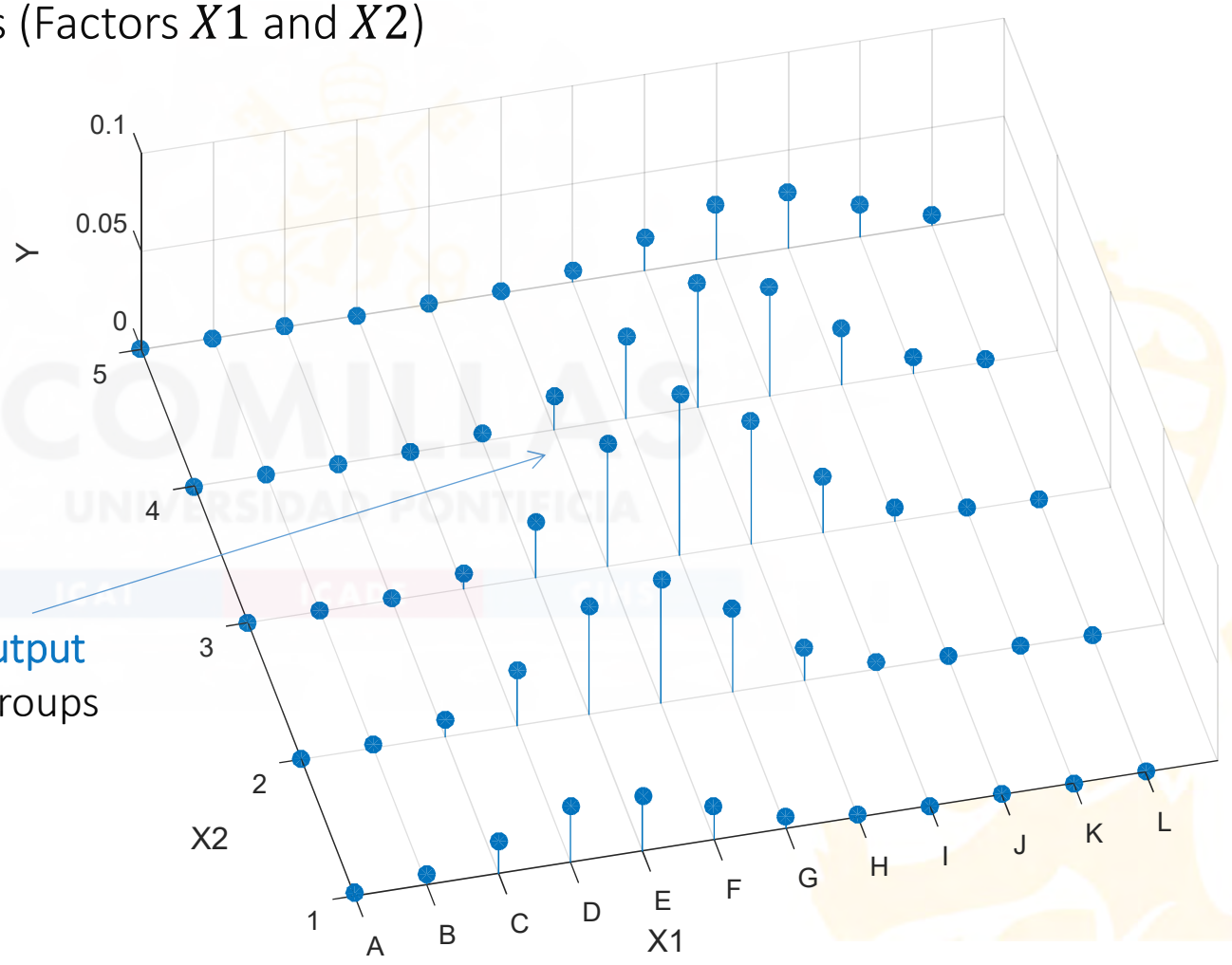
- The ANOVA method allows the study of the variation of a random (dependent) variable with the values taken by other variables (factors)
  - Factors (inputs) must take discrete values (although, by nature, they can be continuous)




- ANOVA studies the variation of the expectation of the output conditioned by the groups of factors

# Analysis of Variance (ANOVA) Introduction

- Synthetic example
  - Two inputs (Factors $X1$ and $X2$)




Expectation of the output conditioned by the groups of the factors

# Analysis of Variance (ANOVA)
# Main types

- One-way ANOVA
  - Only one factor (one input) with several groups (different values)
  - Used to test for differences among at least three groups since a *t*-test can cover the two-group case

- Two-way ANOVA
  - Two factors (i.e., two input variables)
  - More or less complex: without or with interaction

- N-way ANOVA
  - More than two factors
  - The number of possible interactions increases quickly

# 2

# One-way ANOVA

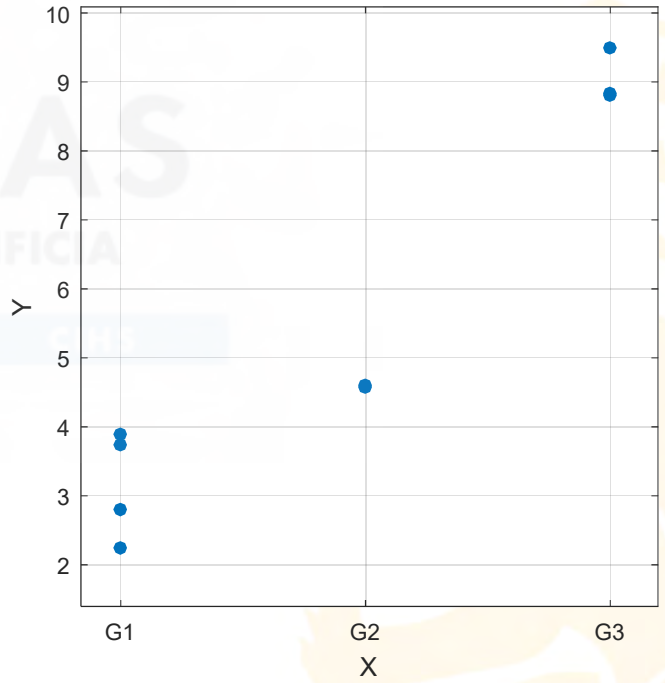
# One-way ANOVA
# Problem statement

- One continuous output variable $Y$ (regression problem)
- One discrete input variable $X$ (factor) with $m$ possible values (groups)
- For each group $i$ there are $n_i$ data points
- Example

| X | Y |
|---|---|
| G1 | 3.7481 |
| G1 | 2.8076 |
| G1 | 3.8886 |
| G1 | 2.2352 |
| G2 | 4.5977 |
| G2 | 4.5776 |
| G3 | 9.4882 |
| G3 | 8.8226 |
| G3 | 8.8039 |

Factor $X$ with groups G1, G2 & G3 ($m$=3)

$n$ = 9
$x_i$ = [G1, G2, G3]
$n_i$ = [4, 2, 3]



# One-way ANOVA
# Means model

- The theoretical regression model consists of two main terms:

Deterministic term

Random term (error)

Continuous output variable $Y$

$$Y = f(X) + \epsilon$$

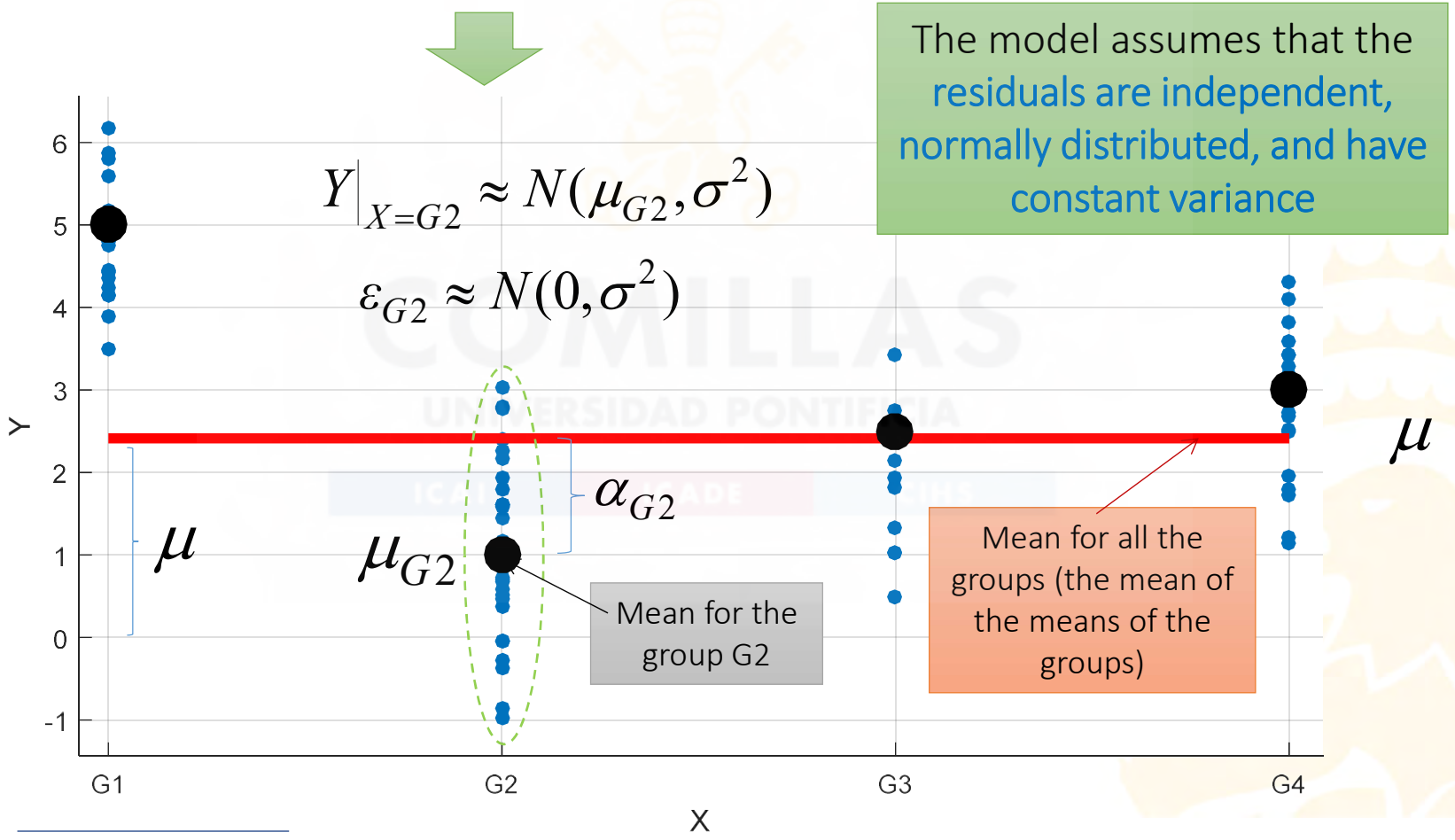
The model has the form of a lookup table. There is no continuous function that expresses how the output varies with the input, but a random variable for each group

$$Y|_{X=x_1} = \mu_1 + \varepsilon_1 = \mu + \alpha_1 + \varepsilon_1$$
$$\dots$$
$$Y|_{X=x_i} = \mu_i + \varepsilon_i = \mu + \alpha_i + \varepsilon_i$$
$$\dots$$
$$Y|_{X=x_m} = \mu_m + \varepsilon_m = \mu + \alpha_m + \varepsilon_m$$

$$i.i.d. \quad \varepsilon_i \approx N(0, \sigma^2)$$

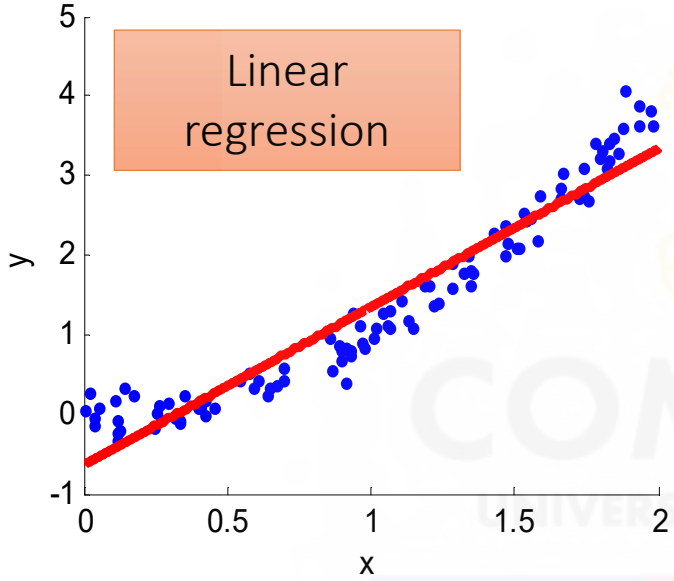
# One-way ANOVA
# Means model

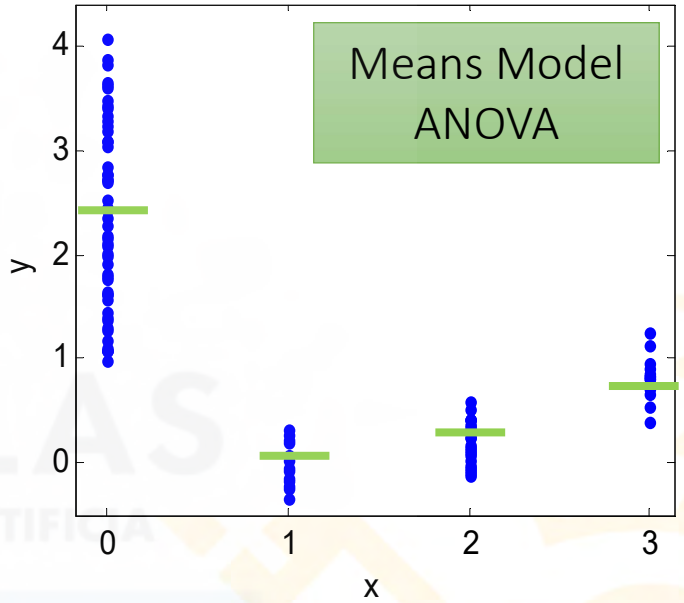
$$Y\big|_{X=x_i} = \mu_i + \varepsilon_i = \mu + \alpha_i + \varepsilon_i$$



# One-way ANOVA
# Linear regression vs. Means Model

- Notice the difference




$$\mathrm{E}(Y|x=x_i)=\beta_0+\beta_1 x_i$$

- Continuous input
- Two coefficients
- Controlled flexibility (linear shape)

$$\mathrm{E}(Y|x=x_i)=\mu+\alpha_i$$

- Categorical input
- $1+m$ coefficients
- Very flexible (any shape)

# One-way ANOVA
# Coefficient estimation

- The structure of the additive model allows solution for the additive coefficients by simple algebra
- Estimation of the mean of the $i$-th group of the factor
  - It is the mean of the output variable of the $i$-th group

$$\hat{\mu}_i = \frac{\sum_{j=1,n_i} y_{ij}}{n_i}$$

Observation $j$ of the $i$ group

- Estimation of the global mean
  - It is the mean of the means of the groups, not the mean of all the observations (except if there are the same number of data in each group)

$$\hat{\mu} = \frac{\sum_{i=1,m} \hat{\mu}_i}{m}$$

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} \longrightarrow \sum_{i=1,m} \alpha_i = 0$$

# One-way ANOVA
# Variability decomposition

- ANOVA decomposes the variability of the output around its mean in two components:
  - Explained by the factor
  - Unexplained by the factor (i.e., the residual one)
- The variability is measured in terms of sums of squares

$$SS_{TOT} = SS_{EXP} + SS_{RES}$$

Total variability to be explained

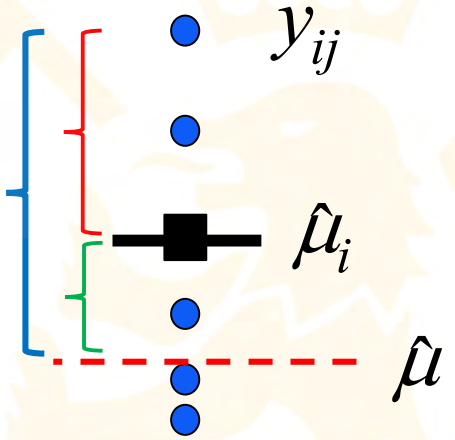
$$SS_{TOT} = \sum_{i=1,m} \sum_{j=1,n_i} (y_{ij} - \hat{\mu})^2$$

Variability between groups (explained)

$$SS_{EXP} = \sum_{i=1,m} \sum_{j=1,n_i} (\hat{\mu}_i - \hat{\mu})^2$$

Residual sum of squares or sum of squared residuals

$$SS_{RES} = \sum_{i=1,m} \sum_{j=1,n_i} (y_{ij} - \hat{\mu}_i)^2$$



# One-way ANOVA
# Variance comparison

- ANOVA compares two variances: the variance explained by the factor and the residual variance
- Explained variance or explained mean squares

$$\hat{S}^2_{EXP} = MS_{EXP} = \frac{SS_{EXP}}{m-1}$$

Variability between groups (explained)

- Residual variance or unexplained variance or unexplained mean squares

$$\hat{S}^2_{RES} = MS_{RES} = \frac{SS_{RES}}{n-m}$$

Residual sum of squares or sum of squared residuals

# One-way ANOVA
# Is the Means model significant?

- Is the model really a constant?
- It can be stated as a hypothesis test

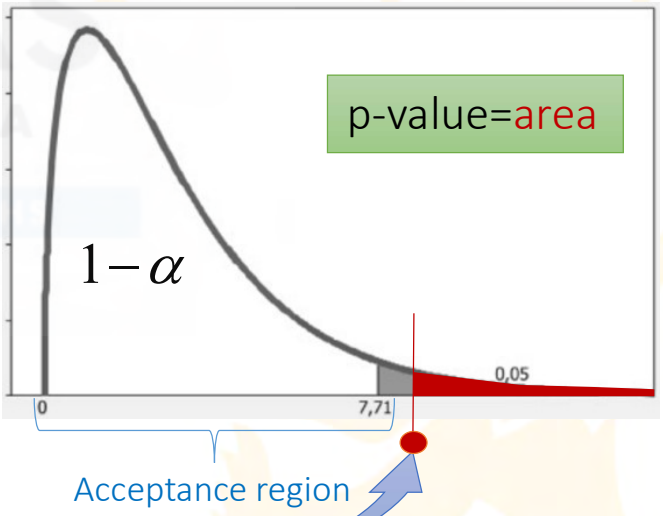
$$H_0: \mu_1 = \cdots = \mu_m$$
$$H_1: \exists i / \mu_i \neq \mu_j \text{ (at least one)}$$

If all these coefficients can be equal, then the means model is useless (the factor is not significant for explaining the output variance)

- If the null hypothesis is true, then

$$F = \frac{S^2_{EXP}}{S^2_{RES}} \approx F_{m-1,n-m}$$

Ratio between the explained variance and the unexplained variance (if $H_0$ is true, variances should be similar)




Sample value of the statistic

# One-way ANOVA
# ANOVA table

- The standard ANOVA table has this form

| Source | SS | df | MS | F | *p*-value |
|---|---|---|---|---|---|
| Group (Between) | SSR | $k-1$ | MSR = SSR/$(k-1)$ | MSR/MSE | $P(F_{k-1,N-k}) > F$ |
| Error (Within) | SSE | $N-k$ | MSE = SSE/$(N-k)$ | | |
| Total | SST | $N-1$ | | | |

| | |
|---|---|
| Groups (Between) | Variability due to the differences among the group means (variability between groups) |
| Error (Within) | Variability due to the differences between the data in each group and the group mean (variability within groups) |
| Total | Total variability |

COLUMNS

| | |
|---|---|
| **Source** | Source of the variability. |
| **SS** | Sum of squares due to each source |
| **df** | Degrees of freedom associated with each source. $N$ is the total # of observations, and $k$ is the # of groups. |
| **MS** | Mean squares for each source |
| **F** | $F$-statistic |
| **Prob>F** | $p$-value |

# One-way ANOVA
# ANOVA table

- Summarizes the information to compute the explained and residual variances and the result of the $F$-test of means equality
- Matlab example:

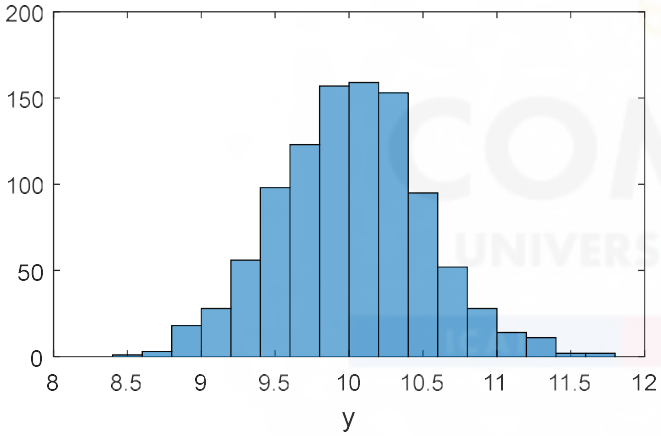
**ANOVA Table**

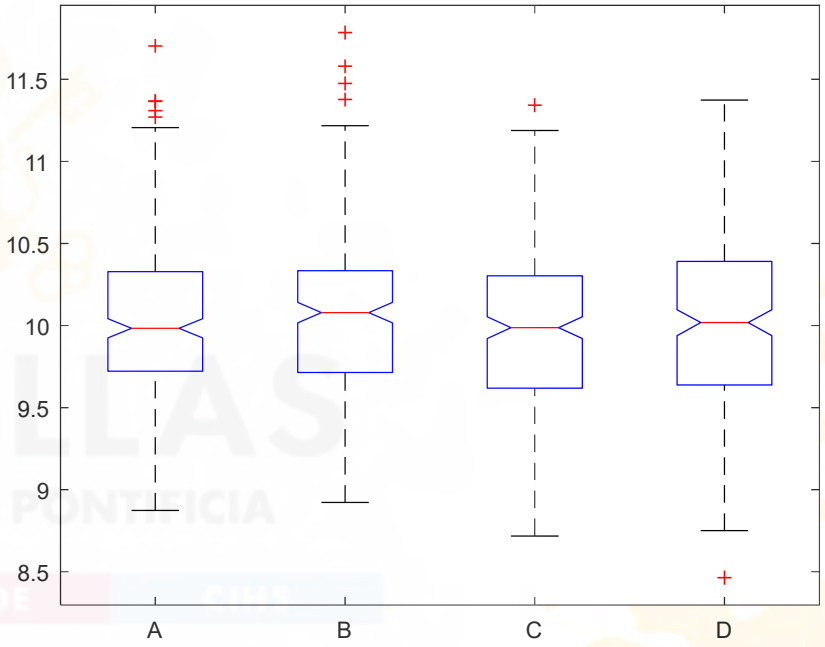
| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 50.537 | 2 | 25.2685 | 33.93 | 0.0005 |
| Error | 4.4686 | 6 | 0.7448 | | |
| Total | 55.0055 | 8 | | | |

# One-way ANOVA
# Illustrative synthetic cases

- C1: IDEAL with many data and small noise

```
n = 1000;
x = datasample('ABCD',n); % 4 groups
% output: independent of x
stdnoise = 0.5;
y = normrnd(10,stdnoise,n,1);
```




**ANOVA Table**

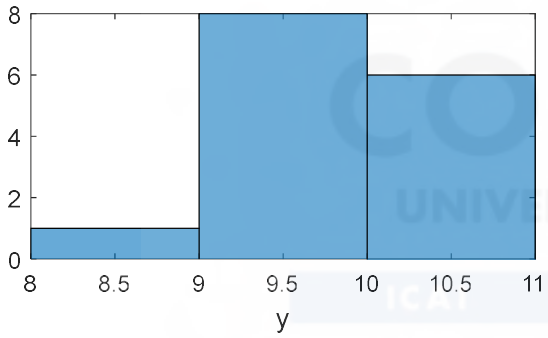
| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 0.76527 | 3 | 0.25509 | 1.0261 | 0.38021 |
| Error | 247.6184 | 996 | 0.24861 | | |
| Total | 248.3837 | 999 | | | |

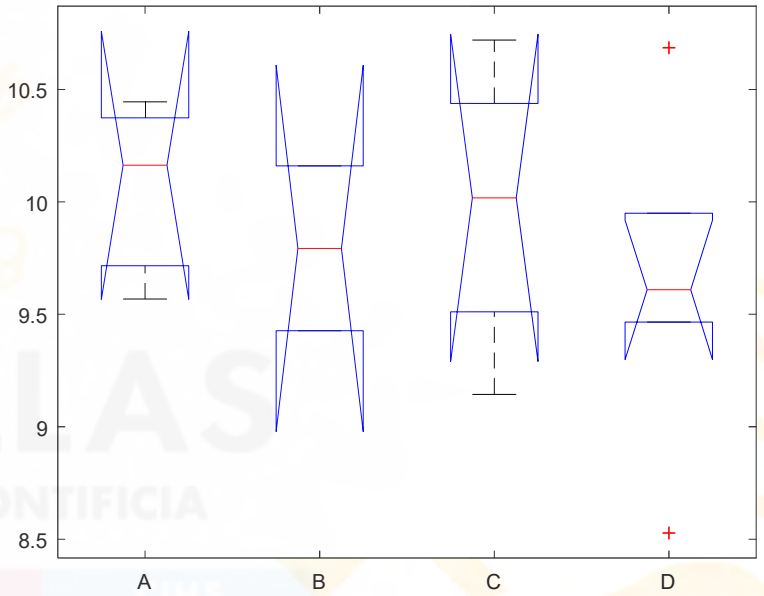
The factor is not significant for explaining the output variance

# One-way ANOVA
# Illustrative synthetic cases

- C2: IDEAL with few data and small noise

```
n = 15;
x = datasample('ABCD',n); % 4 groups
% output: independent of x
stdnoise = 0.5;
y = normrnd(10,stdnoise,n,1);
```




**ANOVA Table**

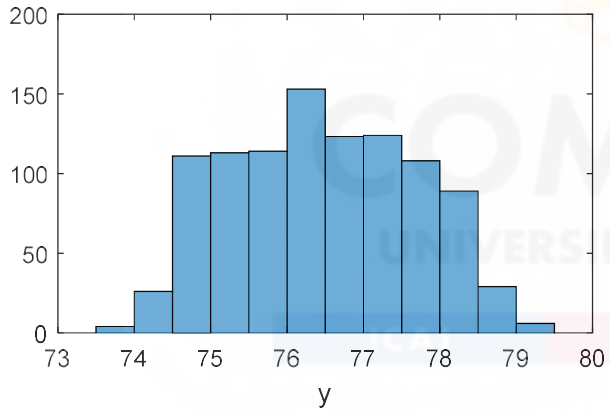
| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 0.45711 | 3 | 0.15237 | 0.37982 | 0.76952 |
| Error | 4.4128 | 11 | 0.40116 | | |
| Total | 4.8699 | 14 | | | |

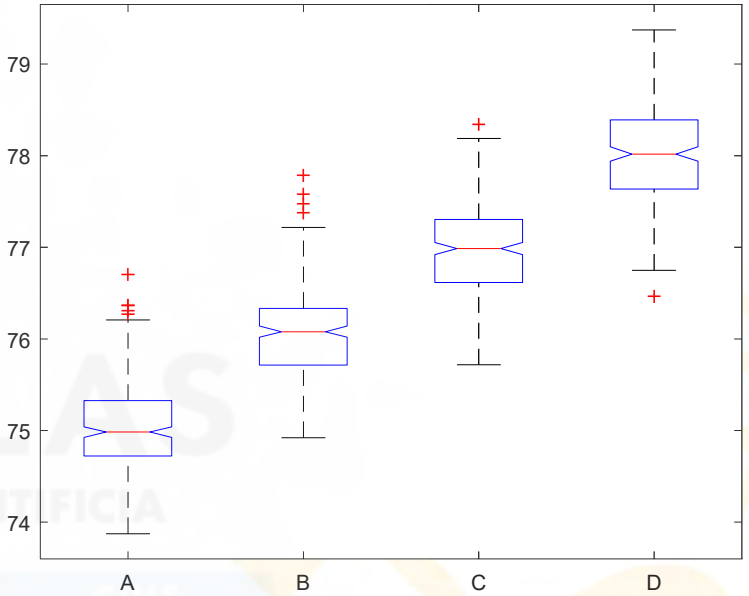
The factor is not significant for explaining the output variance

# One-way ANOVA
# Illustrative synthetic cases

- C3: IDEAL with many data and small noise

```
n = 1000;
x = datasample('ABCD',n); % 4 groups
% output: dependent of x
stdnoise = 0.5;
y = double(x') + normrnd(10,stdnoise,n,1);
```




**ANOVA Table**

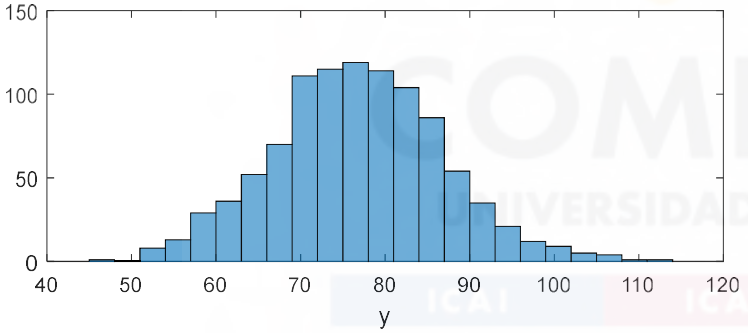
| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 1192.9455 | 3 | 397.6485 | 1599.4686 | 0 |
| Error | 247.6184 | 996 | 0.24861 | | |
| Total | 1440.5639 | 999 | | | |

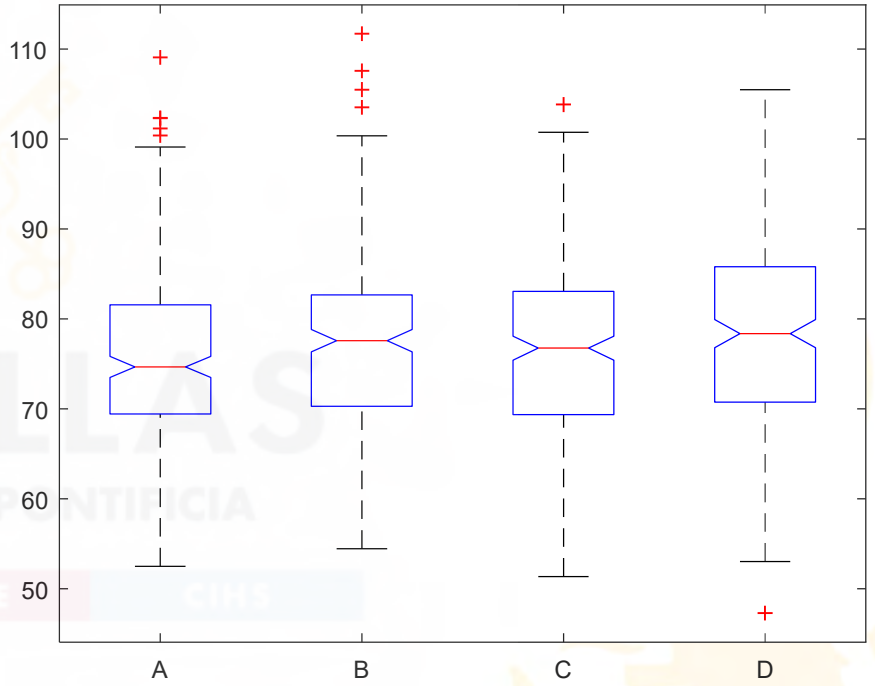
The factor is significant for explaining the output variance

# One-way ANOVA
# Illustrative synthetic cases

- C4: IDEAL with many data and large noise

```
n = 1000;
x = datasample('ABCD',n); % 4 groups
% output: dependent of x
stdnoise = 10;
y = double(x') + normrnd(10,stdnoise,n,1);
```




**ANOVA Table**

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 827.2125 | 3 | 275.7375 | 2.7728 | 0.040422 |
| Error | 99047.3715 | 996 | 99.4452 | | |
| Total | 99874.5841 | 999 | | | |

The factor is not significant for explaining the output variance

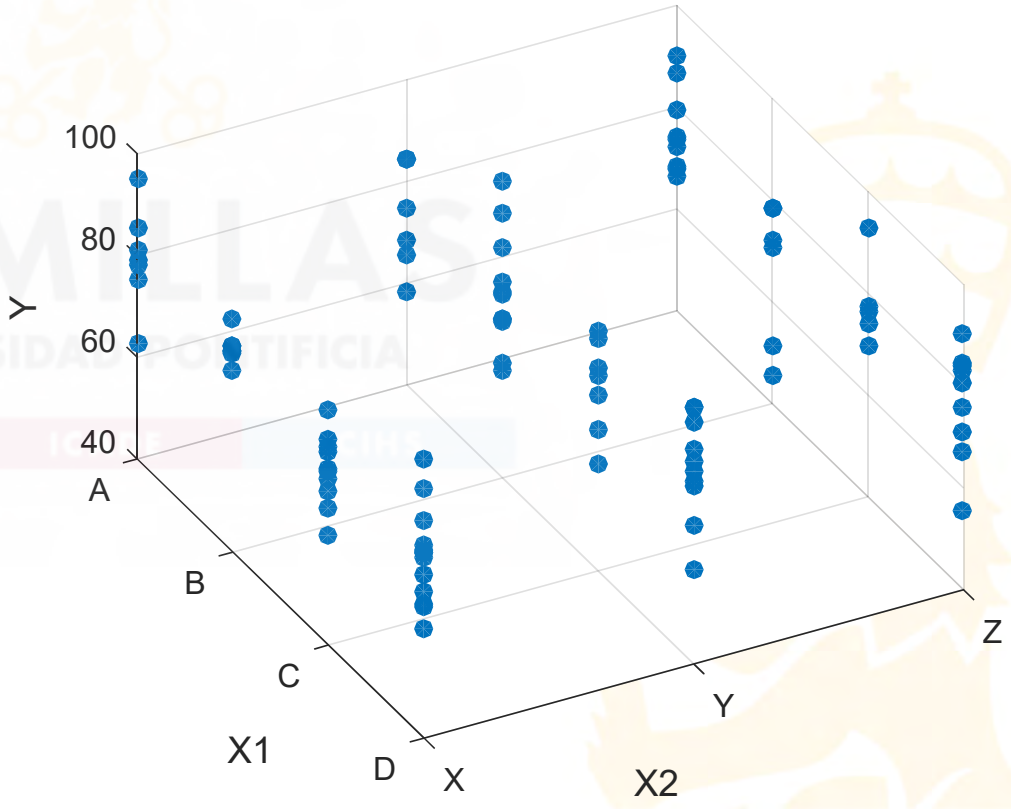
# 3

# Two-way ANOVA

# Two-way ANOVA
# Problem statement

- One continuous output variable $Y$ (regression problem)
- Two discrete input variables (factors), each one with a different number of possible values (groups)
- Example

| X1 | X2 | y |
|---|---|---|
| B | Y | 77.206 |
| C | Z | 62.565 |
| B | Y | 81.823 |
| A | X | 73.185 |
| D | X | 78.42 |
| B | Y | 79.618 |
| B | X | 84.537 |
| C | Z | 68.902 |
| B | Y | 61.329 |
| ... | ... | ... |



# Two-way ANOVA
# Means model without interaction

- It is a generalization of the same ideas of the one-way model
- It assumes that the group of a factor is independent of the group of the other factor

$$Y|_{X_1=x_i, X_2=x_j} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Global mean

Effect of group $i$ of first factor

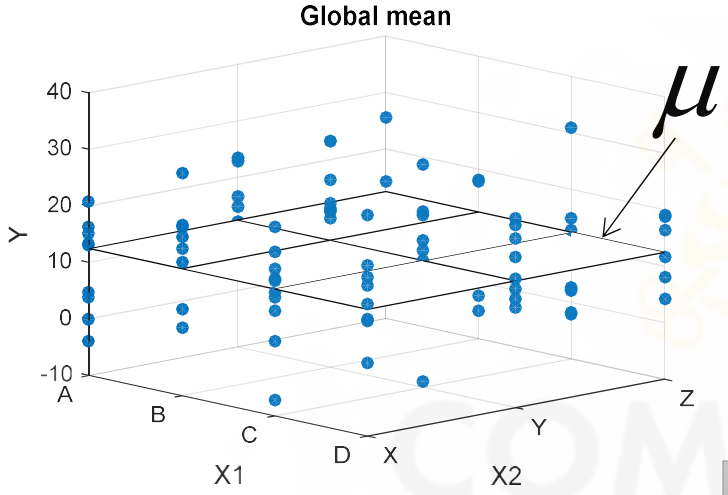
Effect of group $j$ of second factor

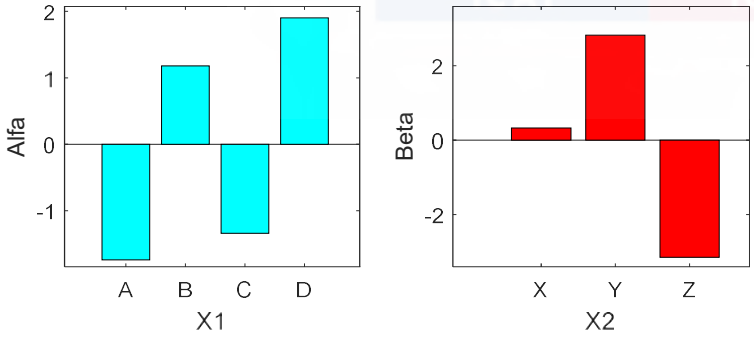
$$i.i.d. \quad \varepsilon_{ij} \approx N(0, \sigma^2)$$

$$\sum \alpha_i = 0$$
$$\sum \beta_j = 0$$

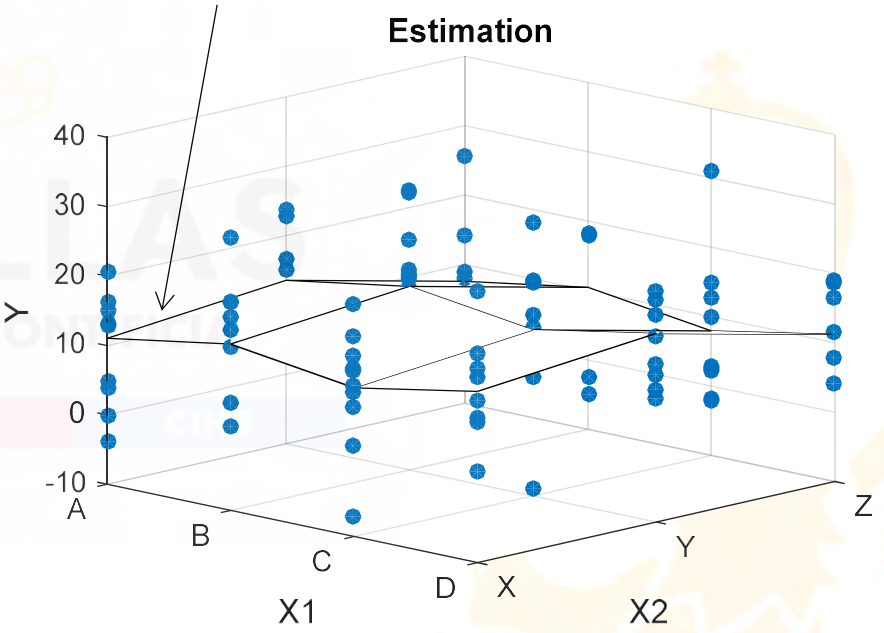
# Two-way ANOVA
# Means model without interaction

- Example




$$E\left(Y\middle|_{X_1=x_i,X_2=x_j}\right)=\mu+\alpha_i+\beta_j$$




6 coefficients have been estimated

# Two-way ANOVA
# Means model with interaction

- It is a generalization of the same ideas of the one-way model
- It assumes that the group of a factor is independent of the group of the other factor

$$Y\big|_{X_1=x_i, X_2=x_j} = \mu + \alpha_i + \beta_j + \theta_{ij} + \varepsilon_{ij}$$

Global mean

Effect of group ***i*** of first factor

Effect of group ***j*** of second factor

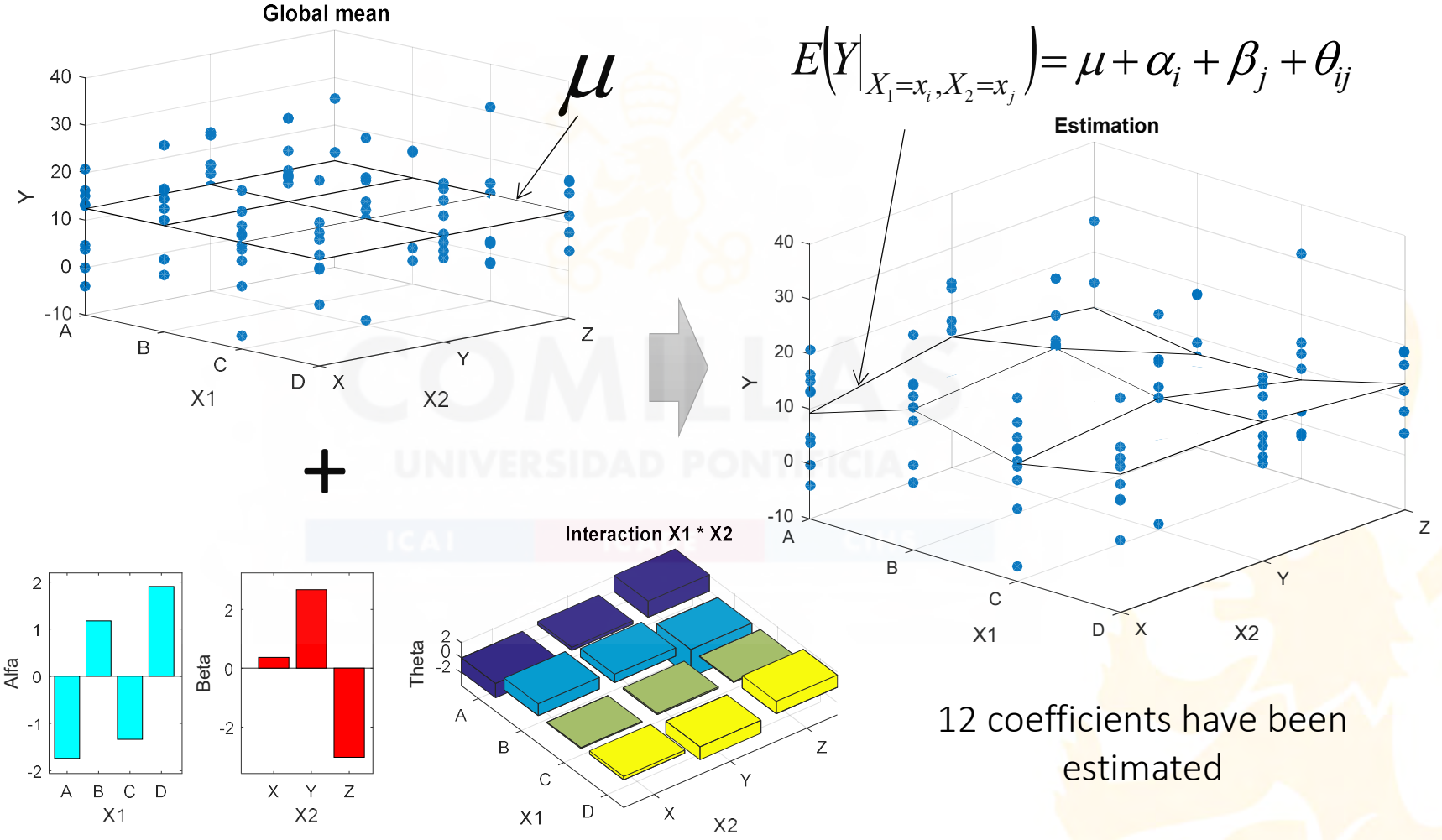
Effect of interaction of groups ***i*** and ***j*** of the two factors

$$\sum \alpha_i = 0$$
$$\sum \beta_j = 0$$
$$\sum \theta_{ij} = 0$$

$$i.i.d. \quad \varepsilon_{ij} \approx N(0, \sigma^2)$$

# Two-way ANOVA
# Means model with interaction

- Example




12 coefficients have been estimated

# Two-way ANOVA
# ANOVA table

- The variance explained by each factor as well as its interaction, can be computed
- There exist $F$-tests to determine if a factor is relevant or not, as well as the interaction
- Matlab examples:

Without interaction

**Analysis of Variance**

| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 51.98 | 3 | 17.3269 | 0.18 | 0.9109 |
| X2 | 109.59 | 2 | 54.793 | 0.56 | 0.5711 |
| Error | 9139.68 | 94 | 97.2306 | | |
| Total | 9297.32 | 99 | | | |

With interaction

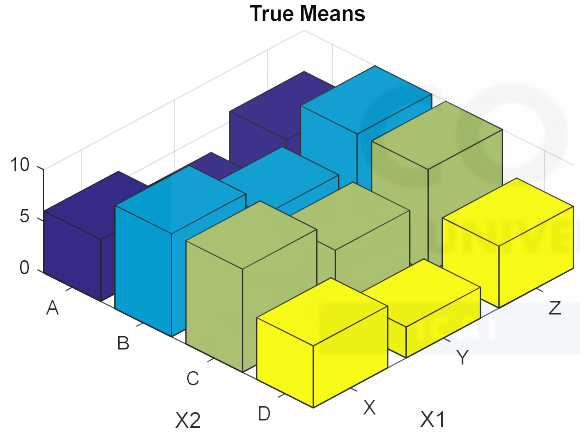
**Analysis of Variance**

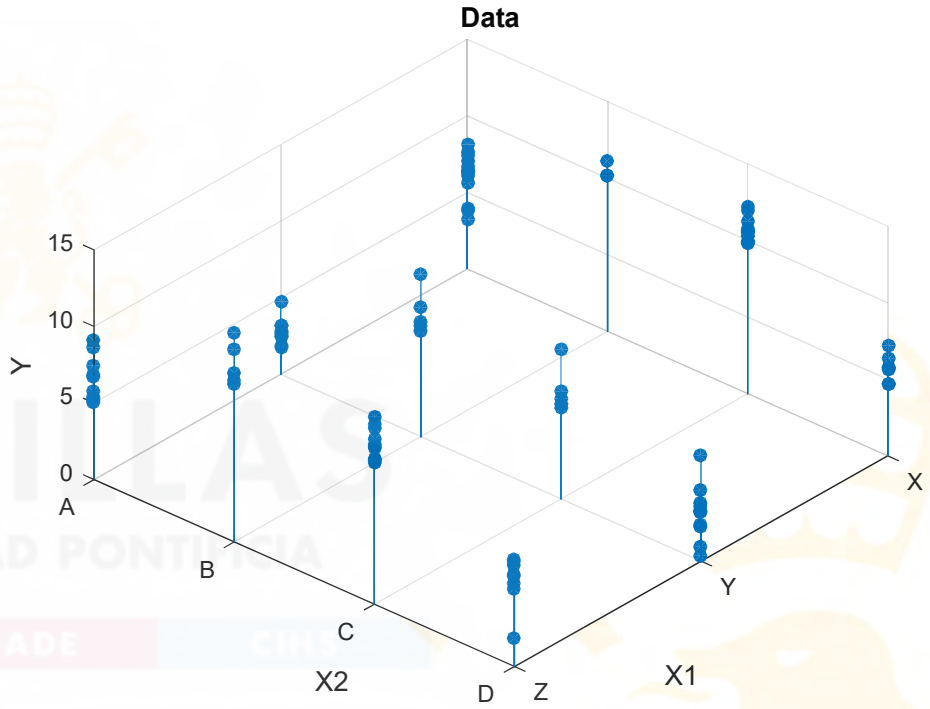
| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 438.09 | 3 | 146.029 | 1.46 | 0.2307 |
| X2 | 111.91 | 2 | 55.955 | 0.56 | 0.5733 |
| X1*X2 | 323.81 | 6 | 53.968 | 0.54 | 0.7765 |
| Error | 8796.12 | 88 | 99.956 | | |
| Total | 9741.94 | 99 | | | |

# Two-way ANOVA
# Illustrative synthetic cases

- C5: IDEAL with many data and small noise

```
n = 100;
stdnoise = 1.5;
constant = 2;
alfas = [3 0 3]';  % clear effect
betas = [1 5 5 1]; % clear effect
noise = normrnd(0,stdnoise,n,1);
```




**Analysis of Variance**

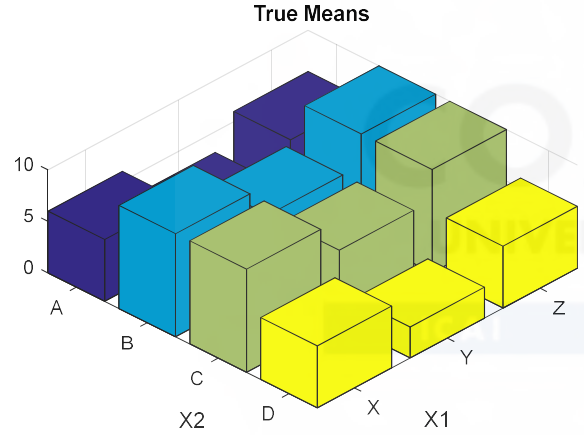
| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 200.1593 | 2 | 100.0796 | 58.884 | 2.6399e-17 |
| X2 | 522.1572 | 3 | 174.0524 | 102.4074 | 1.6341e-29 |
| Error | 159.7631 | 94 | 1.6996 | | |
| Total | 972.1857 | 99 | | | |

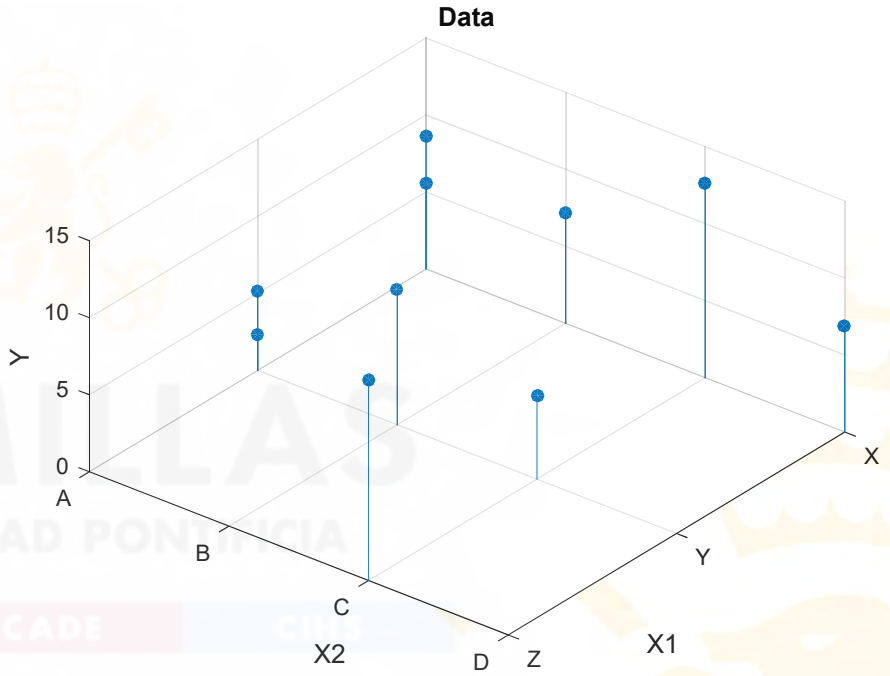
P-values indicate that both factors $X1$ and $X2$ affect the output $Y$

# Two-way ANOVA
# Illustrative synthetic cases

- C6: IDEAL with few data and small noise

```
n = 10;
stdnoise = 1.5;
constant = 2;
alfas = [3 0 3]';  % clear effect
betas = [1 5 5 1]; % clear effect
noise = normrnd(0,stdnoise,n,1);
```




**Analysis of Variance**

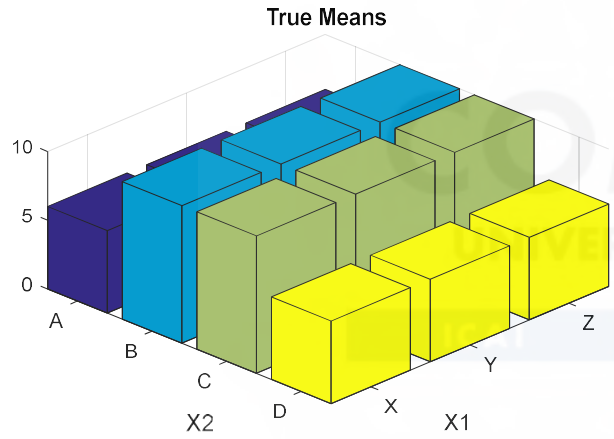
| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 30.0787 | 2 | 15.0394 | 2.1497 | 0.23229 |
| X2 | 22.1848 | 3 | 7.3949 | 1.057 | 0.45991 |
| Error | 27.9841 | 4 | 6.996 | | |
| Total | 100.7118 | 9 | | | |

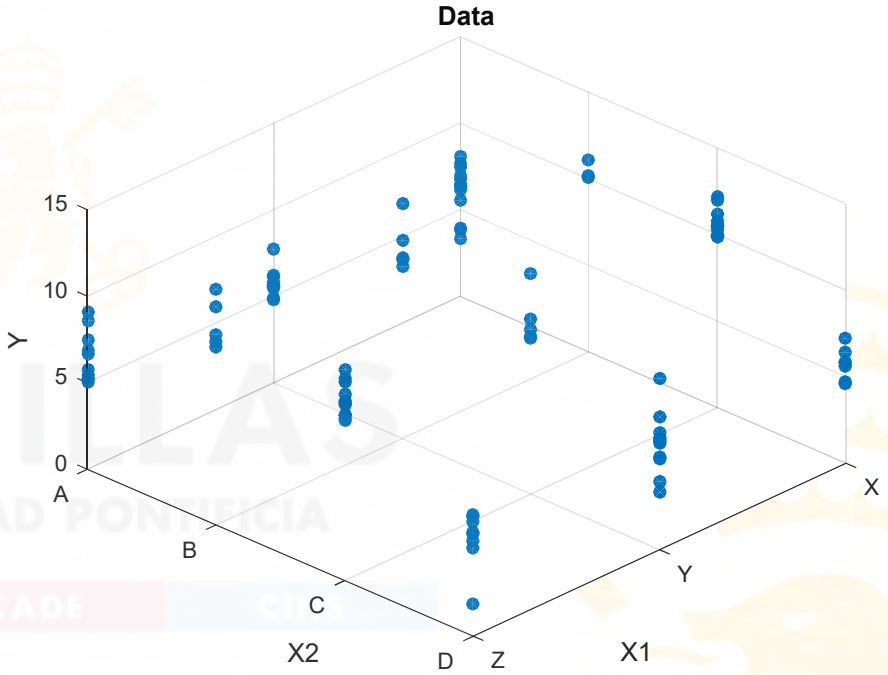
P-values indicate that there is no evidence of dependence of the output from both factors $X1$ and $X2$

# Two-way ANOVA
# Illustrative synthetic cases

- C7: IDEAL with many data and small noise

```
n = 100;
stdnoise = 1.5;
constant = 2;
alfas = [3 3 3]';  % No effect
betas = [1 5 5 1]; % clear effect
noise = normrnd(0,stdnoise,n,1);
```




**Analysis of Variance**

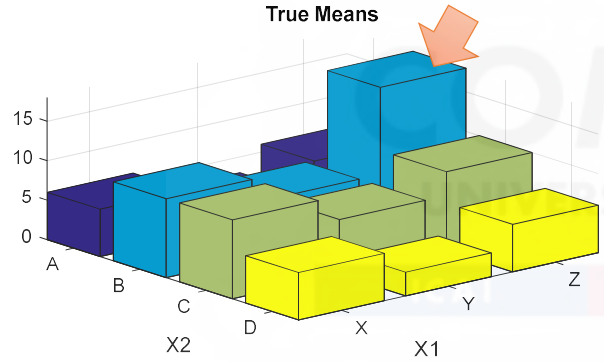
| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 0.73102 | 2 | 0.36551 | 0.21505 | 0.80689 |
| X2 | 522.1572 | 3 | 174.0524 | 102.4074 | 1.6341e-29 |
| Error | 159.7631 | 94 | 1.6996 | | |
| Total | 693.9 | 99 | | | |

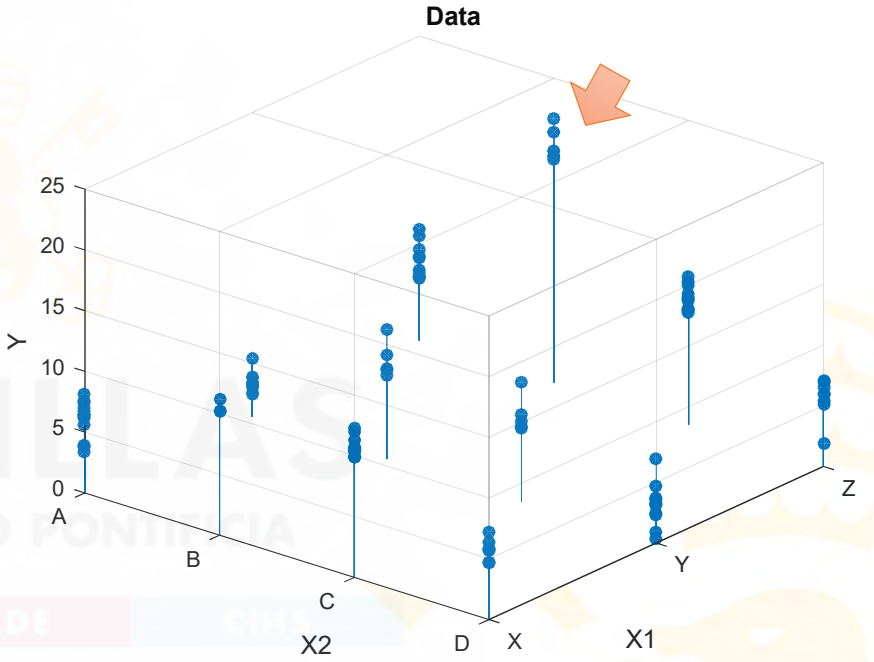
P-values indicate that factor $X2$ affects the output $Y$, but there is no evidence of $X1$ effect

# Two-way ANOVA
# Illustrative synthetic cases

- C8: IDEAL with many data and small noise

```
n = 100;
stdnoise = 1.5;
constant = 2;
alfas = [3 0 3]';  % clear effect
betas = [1 5 5 1]; % clear effect
thetas = zeros(3,4); thetas(3,2) = 8;
noise = normrnd(0,stdnoise,n,1);
```




**Analysis of Variance**

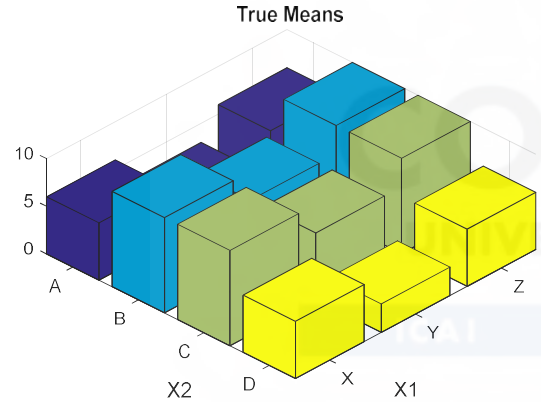
| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 400.2043 | 2 | 200.1022 | 115.3607 | 2.5535e-25 |
| X2 | 772.3107 | 3 | 257.4369 | 148.4148 | 2.583e-34 |
| X1*X2 | 202.2468 | 6 | 33.7078 | 19.4329 | 2.6226e-14 |
| Error | 152.6428 | 88 | 1.7346 | | |
| Total | 1645.5884 | 99 | | | |

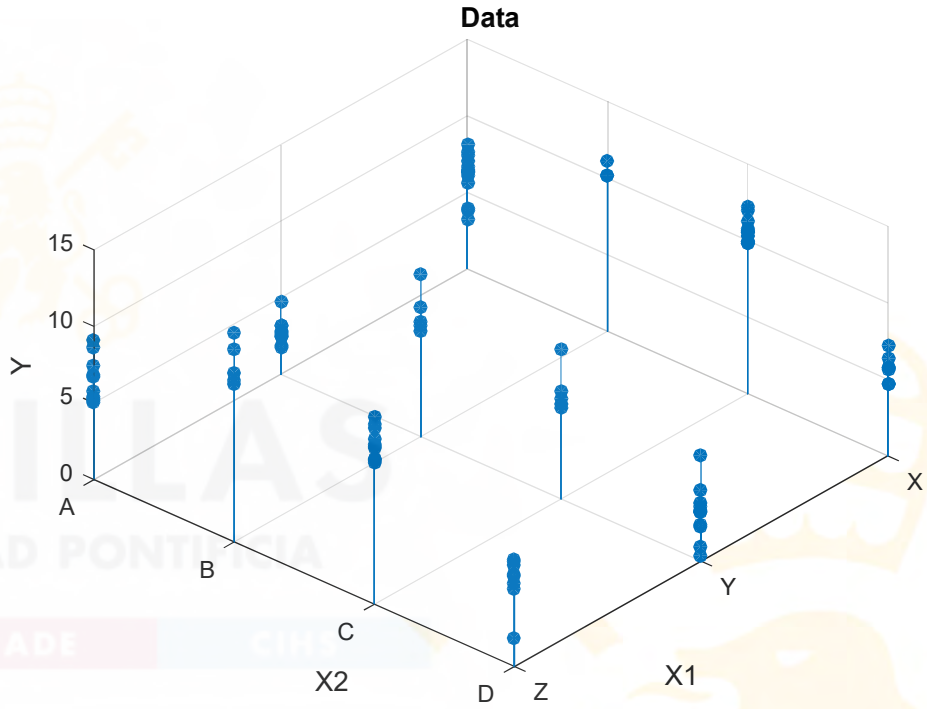
P-values indicate that both factors $X1$ and $X2$ affect the output $Y$, as well as the interaction of both

# Two-way ANOVA
# Illustrative synthetic cases

- C5bis: IDEAL with many data and small noise

```
n = 100;
stdnoise = 1.5;
constant = 2;
alfas = [3 0 3]';  % clear effect
betas = [1 5 5 1]; % clear effect
noise = normrnd(0,stdnoise,n,1);
```




**Analysis of Variance**

| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| X1 | 179.1113 | 2 | 89.5556 | 51.6296 | 1.465e-15 |
| X2 | 480.9503 | 3 | 160.3168 | 92.4241 | 4.1843e-27 |
| **X1*X2** | 7.1203 | 6 | 1.1867 | 0.68416 | **0.66284** |
| Error | 152.6428 | 88 | 1.7346 | | |
| Total | 972.1857 | 99 | | | |

P-values indicate that both factors $X1$ and $X2$ affect the output $Y$, but there is no evidence of an interaction effect of both

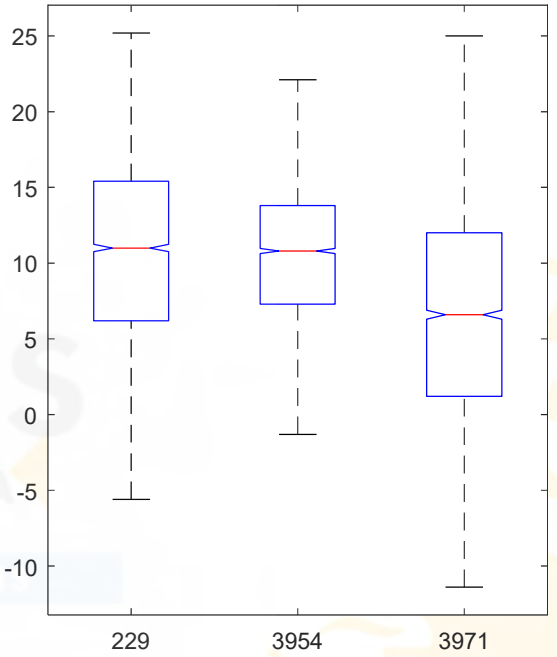
# 4

# Quiz

# Quiz
# Question 1

- Para analizar la dependencia de la temperatura con la variable TOWN, con valores posibles 229, 3954 y 3971, se ha utilizado ANOVA. En función de los resultados obtenidos se puede afirmar que:

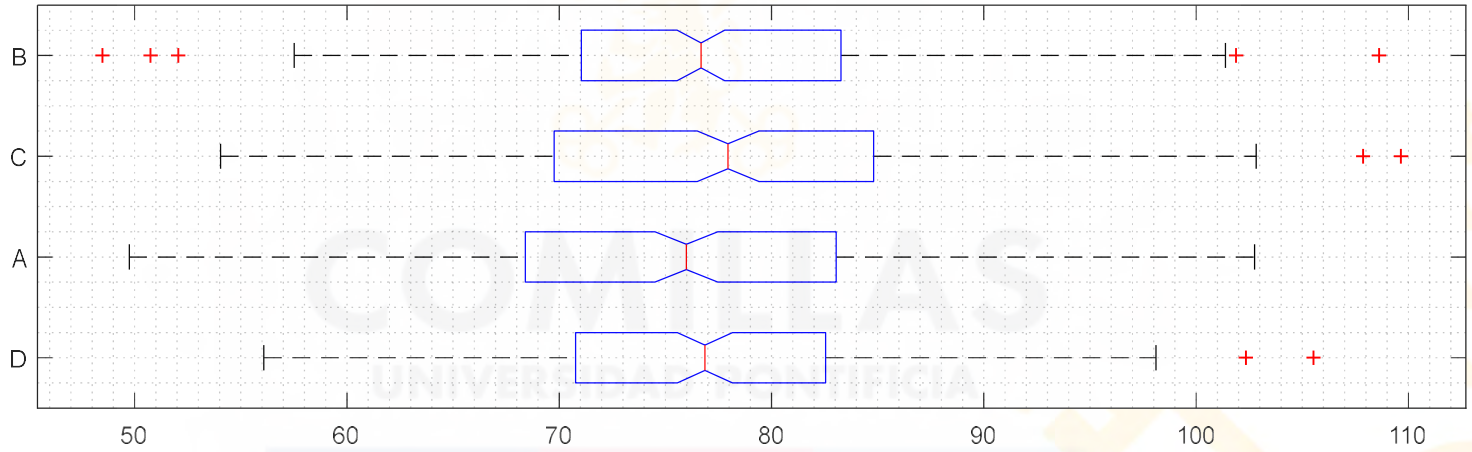
| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 33424.2 | 2 | 16712.1 | 500 | 1.48659e-207 |
| Error | 329492.1 | 9858 | 33.4 | | |
| Total | 362916.3 | 9860 | | | |




A. Hay diferencias significativas en la temperatura esperada según el valor que tome la variable TOWN.
B. No se puede rechazar la hipótesis nula del contraste asociado.
C. Se pueden considerar iguales todas las medias de la temperatura para los diferentes grupos del factor TOWN.

# Quiz
# Question 2

- Se dispone de un conjunto de datos con dos variables X e Y, en donde Y es una variable continua y X puede tomar los valores A, B, C o D. Observando los diagramas de cajas de la variable Y para cada valor de X se puede afirmar que los intervalos de confianza de la mediana al 95% permiten concluir que:




A. La mediana del grupo C es significativamente diferente del resto de medianas.
B. La mediana del grupo A es significativamente diferente de la mediana del grupo D.
C. Las medianas de todos los grupos no son significativamente diferentes.

# Quiz
# Question 3

- Para analizar la dependencia de la temperatura con las variables TOWN y MONTH se ha utilizado ANOVA , incluyendo interacción. En función de los resultados obtenidos se puede afirmar que:

| Source | Sum Sq. | d.f. | Mean Sq. | F | Prob>F |
|---|---|---|---|---|---|
| TOWN | 33498.6 | 2 | 16749.3 | 1577.69 | 0 |
| MONTH | 213005.1 | 11 | 19364.1 | 1823.99 | 0 |
| TOWN*MONTH | 12181.2 | 22 | 553.7 | 52.15 | 0 |
| Error | 104305.8 | 9825 | 10.6 | | |
| Total | 362916.3 | 9860 | | | |

A. No se puede rechazar la hipótesis nula del contraste asociado al factor TOWN.
B. Hay diferencias significativas en la temperatura esperada según el valor que tome la variable MONTH, con independencia del otro factor y la interacción.
C. La temperatura esperada no depende de la interacción entre ambos factores.

# Quiz
# Question 4

- Para analizar la dependencia de la temperatura con las variables TOWN y MONTH se ha utilizado ANOVA, incluyendo interacción. Según el modelo de medias ajustado, el valor de la temperatura que deberíamos esperar en el mes 7 para la estación '229' es

A. 0.24 grados centígrados.
B. 17.40 grados centígrados.
C. 7.92 grados centígrados.

| GROUP | MEANS |
|---|---|
| Constant | 9.245262 |
| TOWN=229 | 1.374123 |
| TOWN=3954 | 1.232115 |
| TOWN=3971 | -2.606238 |
| MONTH=1 | -5.788631 |
| MONTH=2 | -5.621378 |
| MONTH=3 | -3.877042 |
| MONTH=4 | -1.163534 |
| MONTH=5 | 1.260711 |
| MONTH=6 | 4.611528 |
| MONTH=7 | 6.546016 |
| MONTH=8 | 6.885323 |
| MONTH=9 | 4.653009 |
| MONTH=10 | 1.243746 |
| MONTH=11 | -3.151559 |
| MONTH=12 | -5.598189 |

| GROUP | MEANS |
|---|---|
| TOWN=229 * MONTH=1 | -0.650825 |
| TOWN=229 * MONTH=2 | -0.577535 |
| TOWN=229 * MONTH=3 | -0.190730 |
| TOWN=229 * MONTH=4 | 0.106741 |
| TOWN=229 * MONTH=5 | 0.398756 |
| TOWN=229 * MONTH=6 | 0.557605 |
| TOWN=229 * MONTH=7 | 0.236391 |
| TOWN=229 * MONTH=8 | 0.280238 |
| TOWN=229 * MONTH=9 | 0.697605 |
| TOWN=229 * MONTH=10 | 0.284897 |
| TOWN=229 * MONTH=11 | -0.590049 |
| TOWN=229 * MONTH=12 | -0.553095 |
| TOWN=3954 * MONTH=1 | 1.932760 |
| TOWN=3954 * MONTH=2 | 1.321953 |
| TOWN=3954 * MONTH=3 | 1.066332 |
| TOWN=3954 * MONTH=4 | -0.019028 |
| TOWN=3954 * MONTH=5 | -0.869630 |
| TOWN=3954 * MONTH=6 | -1.724461 |
| TOWN=3954 * MONTH=7 | -2.194361 |
| TOWN=3954 * MONTH=8 | -2.264492 |
| TOWN=3954 * MONTH=9 | -0.821868 |
| TOWN=3954 * MONTH=10 | 0.432640 |
| TOWN=3954 * MONTH=11 | 1.196404 |
| TOWN=3954 * MONTH=12 | 1.943751 |
| TOWN=3971 * MONTH=1 | -1.281934 |
| TOWN=3971 * MONTH=2 | -0.744418 |
| TOWN=3971 * MONTH=3 | -0.875602 |
| TOWN=3971 * MONTH=4 | -0.087713 |
| TOWN=3971 * MONTH=5 | 0.470873 |
| TOWN=3971 * MONTH=6 | 1.166855 |
| TOWN=3971 * MONTH=7 | 1.957970 |
| TOWN=3971 * MONTH=8 | 1.984255 |
| TOWN=3971 * MONTH=9 | 0.124262 |
| TOWN=3971 * MONTH=10 | -0.717538 |
| TOWN=3971 * MONTH=11 | -0.606355 |
| TOWN=3971 * MONTH=12 | -1.390656 |

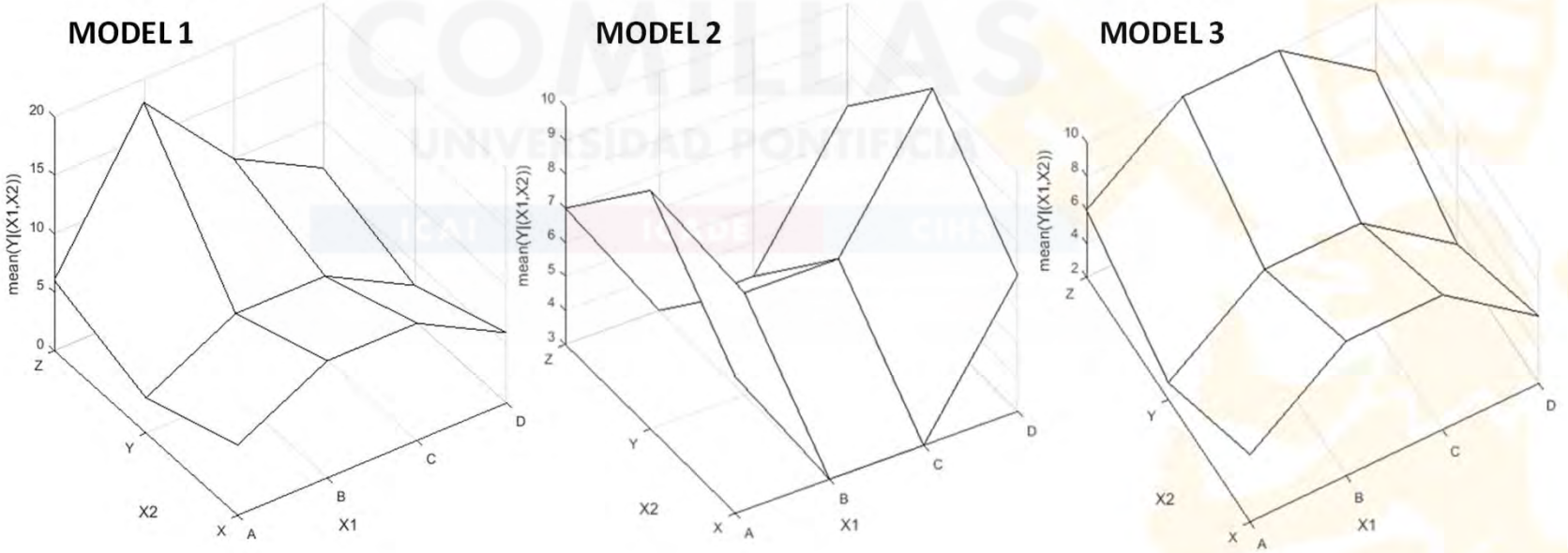
# Quiz
# Question 5

- Se ha realizado un análisis ANOVA con dos factores X1 y X2. A partir de los coeficientes del modelo de medias ajustado se puede deducir que la gráfica que se corresponde con el mismo es:

  A. La gráfica que tiene como título 'MODEL 2'.
  B. La gráfica que tiene como título 'MODEL 1'.
  C. La gráfica que tiene como título 'MODEL 3'.

| TERM | COEF |
|---|---|
| 'Constant' | 2 |
| 'X1=A' | 1 |
| 'X1=B' | 5 |
| 'X1=C' | 5 |
| 'X1=D' | 1 |

| TERM | COEF |
|---|---|
| 'X2=X' | 3 |
| 'X2=Y' | 0 |
| 'X2=Z' | 3 |



# Quiz
# Question 6

- Se quiere determinar si un sistema de reconocimiento de personas, recientemente instalado para controlar el aforo de centros comerciales, estima por igual el nº de personas según la hora del día y el día de la semana. La opción más razonable para hacerlo es utilizar:
  A. Análisis Discriminante.
  B. Regresión lineal.
  C. ANOVA.

# Quiz Answers

- Q1-A
- Q2-C
- Q3-B
- Q4-B
- Q5-C
- Q6-C

# 5

# Real examples

# ANOVA real examples
# Underwater Wireless Sensor Networks

IEEE SENSORS JOURNAL, VOL. 15, NO. 10, OCTOBER 2015



## An Enhanced K-Means and ANOVA-Based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks

Hassan Harb, Abdallah Makhoul, and Raphaël Couturier

***Abstract*— Underwater wireless sensor networks (UWSNs) have recently been proposed as a way to observe and explore aquatic environments. Sensors in such networks are used to perform pollution monitoring, disaster prevention, or assisted navigation and to send monitored data to the sink. Compared with the traditional sensor networks, sensors in UWSNs consume more energy due to the acoustic technology used in under water communications. Node clustering is a common method to organize data traffic and reduce in-network communications while improving scalability and energy consumption. In this paper, we present a new clustering method to handle the spatial similarity between node readings. We suppose that readings are sent periodically from sensor nodes to their appropriate cluster heads (CHs). Then, a two-tier data aggregation technique is proposed. At the first level, each node periodically cleans its readings in order to eliminate redundancies before sending its data set to its CH. Once the CH receives all data sets, it applies an enhanced K-means algorithm based on a one-way ANOVA model to identify nodes generating identical data sets and to aggregate these sets before sending them to the sink. Our proposed approach is validated via experiments on real**

submerged wrecks, oceanographic data collection, disaster prevention, etc [1], [2]. Underwater sensor nodes are small devices with constrained energy and little memory [3]. Moreover, these nodes use acoustic signals that can travel to longer distance than radio waves due to lower frequency [4]–[6]. Hence, unlike traditional sensor networks, sensors in UWSNs consume more energy due to the acoustic technology used in under water communications. In addition, they are costly and difficult to replace. Therefore, there are increasing demands for innovative methods to improve energy efficiency and to prolong the network lifetime. The node clustering and the data aggregation at the level of cluster heads (CHs) are two common methods to organize data traffic and reduce in-network redundancies while improving scalability and energy consumption. Indeed, nodes clustering makes a network look smaller and extends its lifetime by reducing data transmissions between the nodes and the sink [7]; while data aggregation is considered to be the best way to minimize the

# ANOVA real examples
# Underwater Wireless Sensor Networks

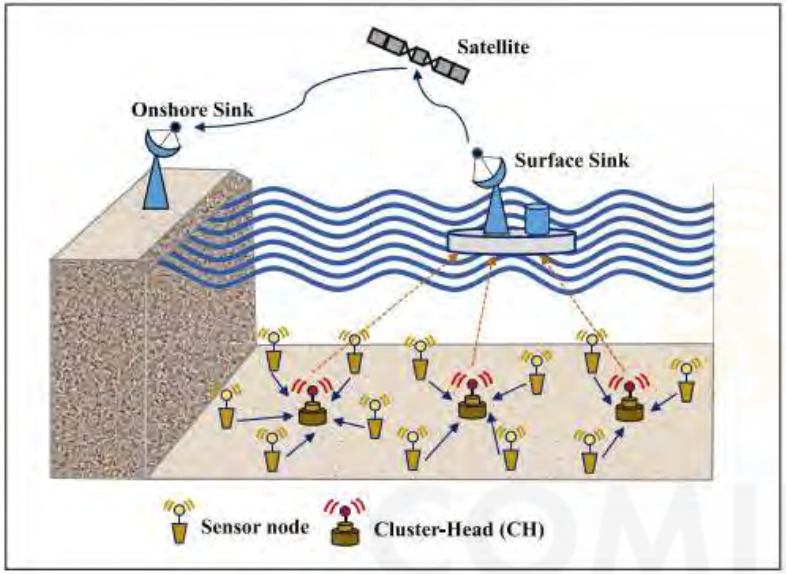



Fig. 1. Cluster-based network architecture for 2-D UASN.

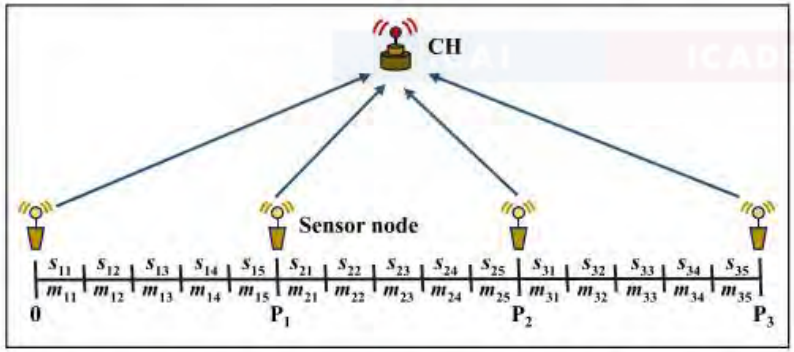



Fig. 2. Illustrative example of periodic UASN (PUASN).

## IV. Variance Study

Studying the variance between measurements in the data sets is an effective way to find nodes that generate redundant data. The ANOVA model provides a statistical tests of whether or not the means of several sets are equal. In the typical application of ANOVA, the null hypothesis ($H_0$) supposes that the variance between sets is not significant. Consequently, the test result ($R$) of the ANOVA is the ratio of the computed variance based on the measurements in the sets. $R$ can be calculated in different manners depending on the statistic tests (presented in the next section) proposed in the ANOVA model. The sets are considered duplicated if the result $R$ is less than a threshold $T$ (significance level) for some desired false-rejection probability (risk $\alpha$).

At each period, we suppose that the CH receives $n$ sets from its sensor nodes, each set contains $\tau$ measures. Also, we assume that measures in each set $M_j$ are independent, with mean $\overline{Y_j}$ and that the variances of sets are equal $\sigma_n^2 = \sigma^2$. Then the measure's variables can be written as follows:

$$m_{ji} = \overline{Y_j} + \epsilon_{ji}; \quad j = 1, \ldots, n; \quad i = 1, \ldots, |M_j|$$

Where $\epsilon_{ji}$ are the residuals which are independent and are normally distributed following $N(0, \sigma^2)$.

# A

# Dictionary

# ANOVA
# Dictionary

- **Box plot –** diagrama de cajas
- **Homocedasticity** – homocedasticidad (varianza constante)
- **Lookup table –** Tabla de consulta
- **Notch –** muesca (en el diagrama de cajas)
- **Null hypothesis –** hipótesis nula
- **Significance level** – nivel de significatividad
- **Sum of squares** – suma de cuadrados
- **Whisker** – bigote (en el diagrama de cajas)

*Thank you for your attention*

Eugenio Sánchez Úbeda