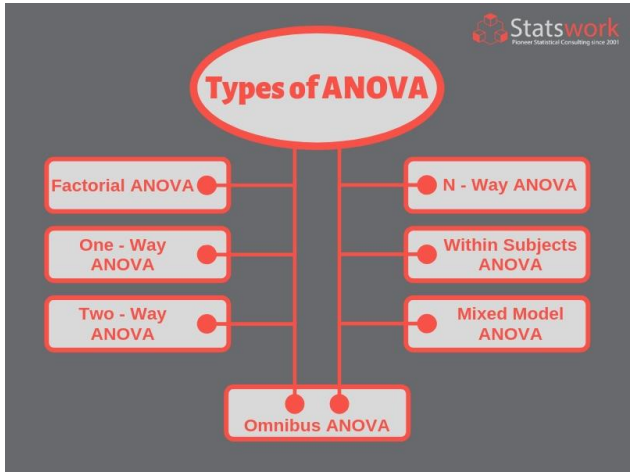
Analysis of Variance (ANOVA)

In the framework of the course:
"Applied Statistics"



Prof. Dr. Mohamed Samer
Engineering in Biosystems, Energy and Environment
Department of Agricultural Engineering
Faculty of Agriculture, Cairo University
E-Mails: msamer@agr.cu.edu.eg; samer@cu.edu.eg
Website: http://scholar.cu.edu.eg/samer/biocv

Introduction

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences between group means and their associated procedures e.g., "variation" among and between groups.

In ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are 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 a type I error.

For this reason, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

A type I error, also known as an error of the first kind, occurs when the null hypothesis (H_0) is true, but is rejected. It is asserting something that is absent, a false hit.

A type I error may be compared with a so-called *false positive* (a result that indicates that a given condition is present when it actually is not present) in tests where a single condition is tested for.

Factor analysis is the process by which a complicated system of many variables is simplified by completely defining it with a smaller number of "factors."

If these factors can be studied and determined, they can be used to predict the value of the variables in a system.

ANOVA is the method used to compare continuous measurements to determine if the measurements are sampled from the same or different distributions.

It is an analytical tool used to determine the significance of factors on measurements by looking at the relationship between a quantitative "response variable" and a proposed explanatory "factor."

This method is similar to the process of comparing the statistical difference between two samples, in that it invokes the concept of hypothesis testing.

Instead of comparing two samples, however, a variable is correlated with one or more explanatory factors, typically using the F-statistic.

From this F-statistic, the *P*-value can be calculated to see if the difference is significant.

ANOVA is a statistical tool used for comparing statistical groups using the dependent and the independent variables.

ANOVA is a technique that uses a sample of observations to compare the number of means.

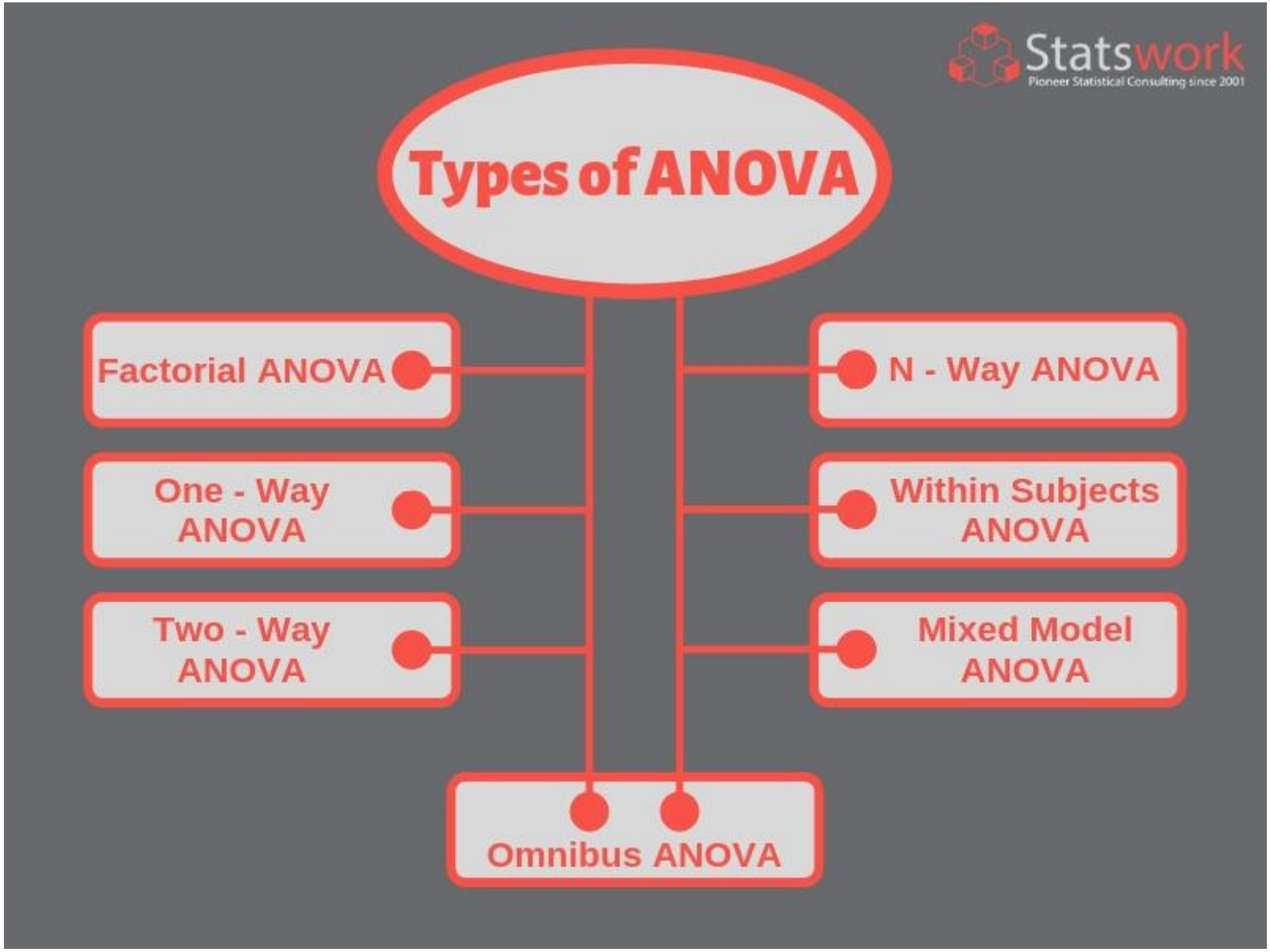
ANOVA calculates statistical differences between two or more means for either groups or variances.

The measured variables are called dependent variable e.g., Test score, while the variables which are controlled are termed as independent variable e.g., Test paper correction method.

ANOVA is a statistical technique used to compare datasets. It is referred as Fisher's ANOVA or Fisher's analysis of variance.

ANOVA is similar to that of *t*-test and *z*-test, which are used to compare mean along with relative variance.

However, in ANOVA, it is best suited when two or more populations/samples are compared.


Statswork
Pioneer Statistical Consulting since 2001
Types of ANOVA
Factorial ANOVA
One - Way ANOVA
Two - Way ANOVA
N - Way ANOVA
Within Subjects ANOVA
Mixed Model ANOVA
Omnibus ANOVA

Different Types Of ANOVA

Factorial ANOVA

This type of **ANOVA** show whether a combined independent variable can predict the value of the dependent variable. In one-way ANOVA, only a single factor is examined using the effects of different levels. Factorial ANOVA, on the other hand, allows you to understand the interactions between factors instead of requiring two different sets of experiments to determine the effects of the two factors. This ANOVA can use random numbers to design the factors.

One – Way ANOVA

One-way ANOVA compares levels of a single factor i.e. one independent variable over the dependent variable.

Two – Way ANOVA

Two-way ANOVA is used to compare two or more factors i.e. effect of two independent variables on a single dependent variable. This can be used in understanding the interaction between the two independent variables. Both types of ANOVA have a single continuous response variable.

N – way ANOVA

Data classified in multiple independent variables are used in an N-way analysis of variance for example differences in age and gender can be checked simultaneously using two-way ANOVA. The N-way ANOVA can show whether there are effects of the independent variable and interactions between them. Interactions are usually seen when one independent variable depends on the second independent variable.

Within-Subjects ANOVA

Within-subject, ANOVA are factors where the same subjects are compared under different conditions or levels. These levels can be measurements for the same size. They can also be reiterations of the same outcome over time.

Mixed Model ANOVA

A combination of Within-unit ANOVA along with Between-unit ANOVA gives us a Mixed-unit ANOVA. It consists of at least two independent variables. One of these variables must vary with Between-unit ANOVA and one has to vary with Mixed-unit ANOVA.

Omnibus ANOVA Test

There is no significant difference in the groups which is the null hypothesis for an ANOVA. The other hypothesis states that there will be at least a single difference among the groups considered. The researcher must test the assumptions of ANOVA. After finding the data, F-ratio and the p-value must be calculated. If the p-value associated with F-ratio is smaller than 0.5 then the null hypothesis is rejected and the other hypothesis is given prominence. This means that the mean of all groups is not equal. After this, the researcher should consider doing the Post-hoc test to understand which groups are different from each other. Post- hoc test helps to identify errors and later places the items in a group.

F-Tests

The test is simply a ratio of two variances. Variances are a measure of how far the data is scattered. It is based on the population of the mean squares which is an estimate of the population variance.

T-Test

It is a test that determines whether there is a difference between the means of two groups which may have certain identical features. Mostly used in data set where flipping a coin or dice 100 times would be followed by distribution having unknown variances. It is a Hypothesis-testing tool and assumptions are tested using it.

Homogeneity Of Variance

It is an assumption where there are population variances in both T-tests as well as F-tests of two or more samples, which are equal.

Welch and the Brown-Forsythe Test

In certain cases, the variances cannot be assumed to be equal and at this juncture, the F test of ANOVA is not suitable so this is when Welch and the Brown-Forsythe test come into effect. The test adjusts the denominator of the F ratio and it has the same expectancy of the numerator when the null hypothesis is true.

Single-Factor ANOVA (One-Way):

One-way ANOVA is used to test for variance among two or more independent groups of data, in the instance that the variance depends on a single factor. It is most often employed when there are at least three groups of data, otherwise a t-test would be a sufficient statistical analysis.

Two-Factor ANOVA (Two-Way):

Two-way ANOVA is used in the instance that the variance depends on two factors. There are two cases in which two-way ANOVA can be employed:

- Data *without replicates*: used when collecting a single data point for a specified condition
- Data *with replicates*: used when collecting multiple data points for a specified condition (the number of replicates must be specified and must be the same among data groups)

When to Use Each ANOVA Type

- Example: There are three identical reactors (R1, R2, R3) that generate the same product.
- One-way ANOVA: You want to analyze the variance of the product yield as a function of the reactor number.
- Two-way ANOVA without replicates: You want to analyze the variance of the product yield as a function of the reactor number and the catalyst concentration.
- Two-way ANOVA with replicates: For each catalyst concentration, triplicate data were taken. You want to analyze the variance of the product yield as a function of the reactor number and the catalyst concentration.

ANOVA is a Linear Model. Though ANOVA will tell you if factors are significantly different, it will do so according to a linear model.

ANOVA will always assume a linear model; it is important to consider strong nonlinear interactions that ANOVA may not incorporate when determining significance.

ANOVA works by assuming each observation as overall mean + mean effect + noise.

If there are non-linear relationship between these (for example, if the difference between column 1 and column 2 on the same row is that column2 = column1^2), then there is the chance that ANOVA will not catch it.

Term or Variable	Description
Factor	Characteristics differentiating populations from one another
Group	Sample set that is influenced by the same factor
Levels	Populations (or treatments)
n_j	Number of elements in group j
x_{ji}	Element within group j
s_j^2	Variance within group j
N	Total number of samples
k	Total number of groups
$\bar{x}_j$	Mean of group j
$\bar{x}$	Mean of all samples within all groups
SSGroups	Sum of squares between groups
SSE	Sum of squared errors within groups
MSGroups	Mean square between groups
MSE	Mean square error within groups
SSTotal	Total sum of squares
df	Degrees of freedom

SETTING UP AN ANALYSIS OF VARIANCE TABLE

The fundamental concept in one-way analysis of variance is that the variation among data points in all samples can be divided into two categories: variation between group means and variation between data points in a group. The theory for analysis of variance stems from a simple equation, stating that the total variance is equal to the sum of the variance between groups and the variation within groups –

Total variation = variation between groups + variation within groups

An analysis of variance table is used to organize data points, indicating the value of a response variable, into groups according to the factor used in each case. For example, Table 1 is an ANOVA table for comparing the amount of weight lost over a three month period by dieters on various weight-loss programs.

Table 1 - Amount of weight lost by dieters on various programs over a 3 month period

Program 1	Program 2	Program 3
7	9	15
9	11	12
5	7	18
7		

A reasonable question is, can the type of program (a factor) be used to predict the amount of weight a dieter would lose on that program (a response variable)? Or, in other words, is any one program superior to the others?

MEASURING VARIATION BETWEEN GROUPS

The variation between group means is measured with a weighted sum of squared differences between the sample means and the overall mean of all the data. Each squared difference is multiplied by the appropriate group sample size, ni, in this sum. This quantity is called **sum of squares between groups** or **SS Groups**.

$$\text{SSGroups} = n_1(x_1 - x)^2 + n_2(x_2 - x)^2 + \ldots + n_k(x_k - x)^2 = \sum_{\text{yrcups}} n_j \left(\bar{x}_{jj} - \bar{x}'^2\right.$$

The numerator of the F-statistic for comparing means is called the **mean square between groups** or **MS Groups**, and it is calculated as -

$$\text{MSGroups} = \frac{\text{SSGroups}}{k-1}$$

MEASURING VARIATION WITHIN GROUPS

To measure the variation among data points within the groups, find the sum of squared deviations between data values and the sample mean in each group, and then add these quantities. This is called the **sum of squared errors**, **SSE**, or **sum of squares within groups**.

$$SSE = (n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_k - 1)\, s_k^2 = \sum_{\text{allgroups}} (n_j - 1)\, s_j^2$$

where

$$s_j^2 = \sum_{\text{groupj}} \frac{(x_{ij} - \bar{x}_j)^2}{n_j - 1} =$$

and is the variance within each group.

The denominator of the F-statistic is called the **mean square error**, **MSE**, or **mean squares within groups**. It is calculated as

$$MSE = \frac{SSE}{N - k} = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_k - 1)\, s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

MSE is simply a weighted average of the sample variances for the k groups. Therefore, if all n_i are equal, MSE is simply the average of the k sample variances. The square root of MSE (s_p), called the **pooled standard deviation,** estimates the population standard deviation of the response variable (keep in mind that all of the samples being compared are assumed to have the same standard deviation σ).

MEASURING THE TOTAL VARIATION

The total variation in all samples combined is measured by computing the sum of squared deviations between data values and the mean of all data points. This quantity is referred to as the **total sum of squares** or SS Total. The total sum of squares may also be referred to as SSTO. A formula for the sum of squared differences from the overall mean is

$$\text{SSTotal} = \sum_{\text{values}} (x_{ij} - \bar{x})^2$$

where x_{ij} represents the jth observation within the ith group, and $\bar{x}$ is the mean of all observed data values. Finally, the relationship between SS Total, SS Groups, and SS Error is

SS Total = SS Groups + SS Error

Overall, the relationship between the total variation, the variation between groups, and the variation within a group is illustrated by Figure 2.

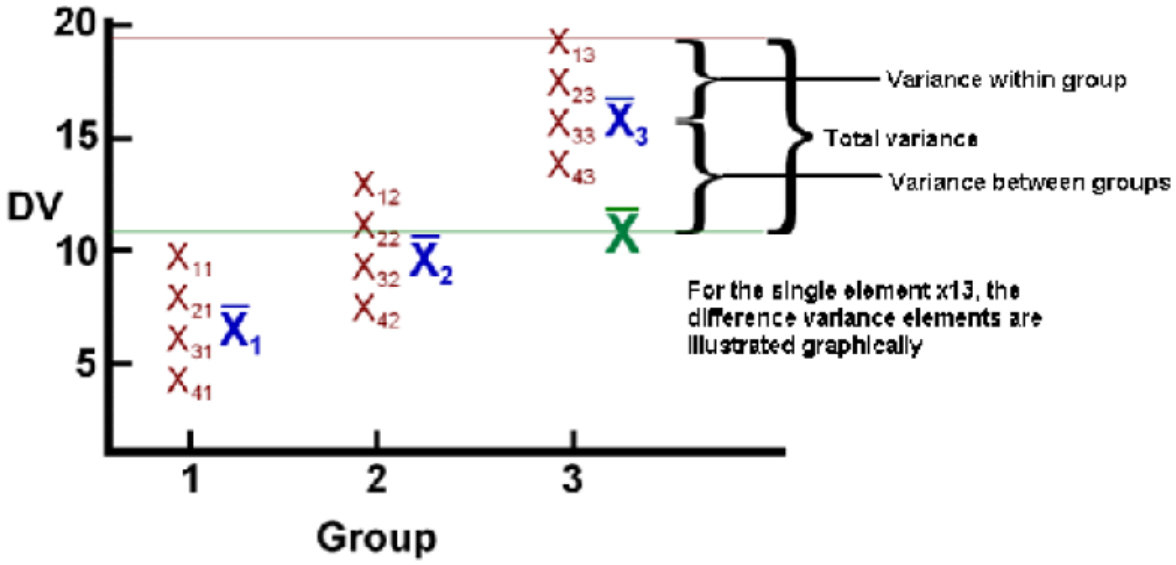


Figure 2: Visual illustration of variance in ANOVA, adapted from www.uwsp.edu/psych/stat/12/anova-1w.htm

A general table for performing the one-way ANOVA calculations required to compute the F-statistic is given below

Table 2 - One-Way ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-Statistic
Between groups	k-1	$SSGroups = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2$	$\frac{SSGroups}{k-1}$	$F = \frac{MSGroups}{MSE}$
Within groups (error)	N-k	$SSE = \sum_{groups} (n_i - 1)s_i^2$	$\frac{SSE}{N-k}$	
Total	N-1	$SSTO = \sum_{values} (x_{ij} - \bar{x})^2$		

TWO-WAY ANOVA CALCULATIONS

Like in one-way ANOVA analysis the main tool used is the square sums of each group.

Two-way ANOVA can be split between two different types: with repetition and without repetition.

With repetition means that every case is repeated a set number of times. Without repetition means there is one reading for every case.

These calculations are almost never done by hand. In the Exercises class you will usually use Excel or Mathematica to create these tables. Sections describing how to use these programs will be described later in the Exercises.

The calculations needed for two-way ANOVA with repetition

Using the SS values as a start the F-statistics for two-way ANOVA with repetition are calculated using the chart below where a is the number of levels of main effect A, b is the number of levels of main effect B, and n is the number of repetitions.

Source	SS	DF	Adj MS	F-Statistic
Main Effect A	From data given	a-1	SS/df	MS(A)/MS(W)
Main Effect B	From data given	b-1	SS/df	MS(B)/MS(W)
Interaction Effect	From data given	(a-1)(b-1)	SS/df	MS(A*B)/MS(W)
Within	From data given	ab(n-1)	SS/df	
Total	sum of others	abn-1		

The calculations needed for two-way ANOVA without repetition

Source	SS	DF	MS	F-Statistic
Main Effect A	From data given	a-1	SS/df	MS(A)/MS(E)
Main Effect B	From data given	b-1	SS/df	MS(B)/MS(E)
Error	From data given	(a-1)(b-1)	SS/df	
Total	sum of others	ab-1		

HYPOTHESES ABOUT MEDIANS

In general, it is best construct hypotheses about a population median, rather than the mean. Using the median accounts for the sample being skewed based on extreme outliers. Median hypotheses should also be used for dealing with ordinal variables (variables which are described only as being higher or lower than one other and do not have a precise value). When several populations are compared, the hypotheses are stated as -

H_0: Population medians are equal
H_a: Population medians are not all equal

KRUSKAL-WALLIS TEST FOR COMPARING MEDIANS

The **Kruskal-Wallis Test** provides a method of comparing medians by comparing the relative rankings of data in the observed samples. This test is therefore referred to as a rank test or non-parametric test because the test does not make any assumptions about the distribution of data.

To conduct this test, the values in the total data set are first ranked from lowest to highest, with 1 being lowest and N being highest. The ranks of the values within each group are averaged, and the test statistic measures the variation among the average ranks for each group. A p-value can be determined by finding the probability that the variation among the set of rank averages for the groups would be as large or larger as it is if the null hypothesis is true.

MOOD'S MEDIAN TEST FOR COMPARING MEDIANS

Another nonparametric test used to compare population medians is **Mood's Median Test**. Also called the Sign Scores Test, this test involves multiple steps.

1. Calculate the median (M) using all data points from every group in the study
2. Create a contingency table as follows

	A	B	C	Total
Number of values greater than M				
Number of values less than or equal to M				

3. Calculate the expected value for each data set using the following formula:

$$\text{expected} = \frac{(\text{rowtotal})(\text{columntotal})}{\text{grandtotal}}$$

4. Calculate the chi-square value using the following formula

$$\chi = \frac{(\text{actual} - \text{expected})^2}{\text{expected}}$$

A chi-square statistic for two-way tables is used to test the null hypothesis that the population medians are all the same. The test is equivalent to testing whether or not the two variables are related.

ANOVA AND FACTOR ANALYSIS IN PROCESS CONTROL

ANOVA and factor analysis are typically used in process control for troubleshooting purposes. When a problem arises in a process control system, these techniques can be used to help solve it. A factor can be defined as a single variable or simple process that has an effect on the system. For example, a factor can be temperature of an inlet stream, flow rate of coolant, or the position of a specific valve. Each factor can be analyzed individually to determine the effect that changing the input has on the process control system as a whole. The input variable can have a large, small, or no effect on what is being analyzed. The amount that the input variable affects the system is called the "factor loading", and is a numerical measure of how much a specific variable influences the system or the output variable. In general, the larger the factor loading is for a variable the more of an affect that it has on the output variable.

A simple equation for this would be:

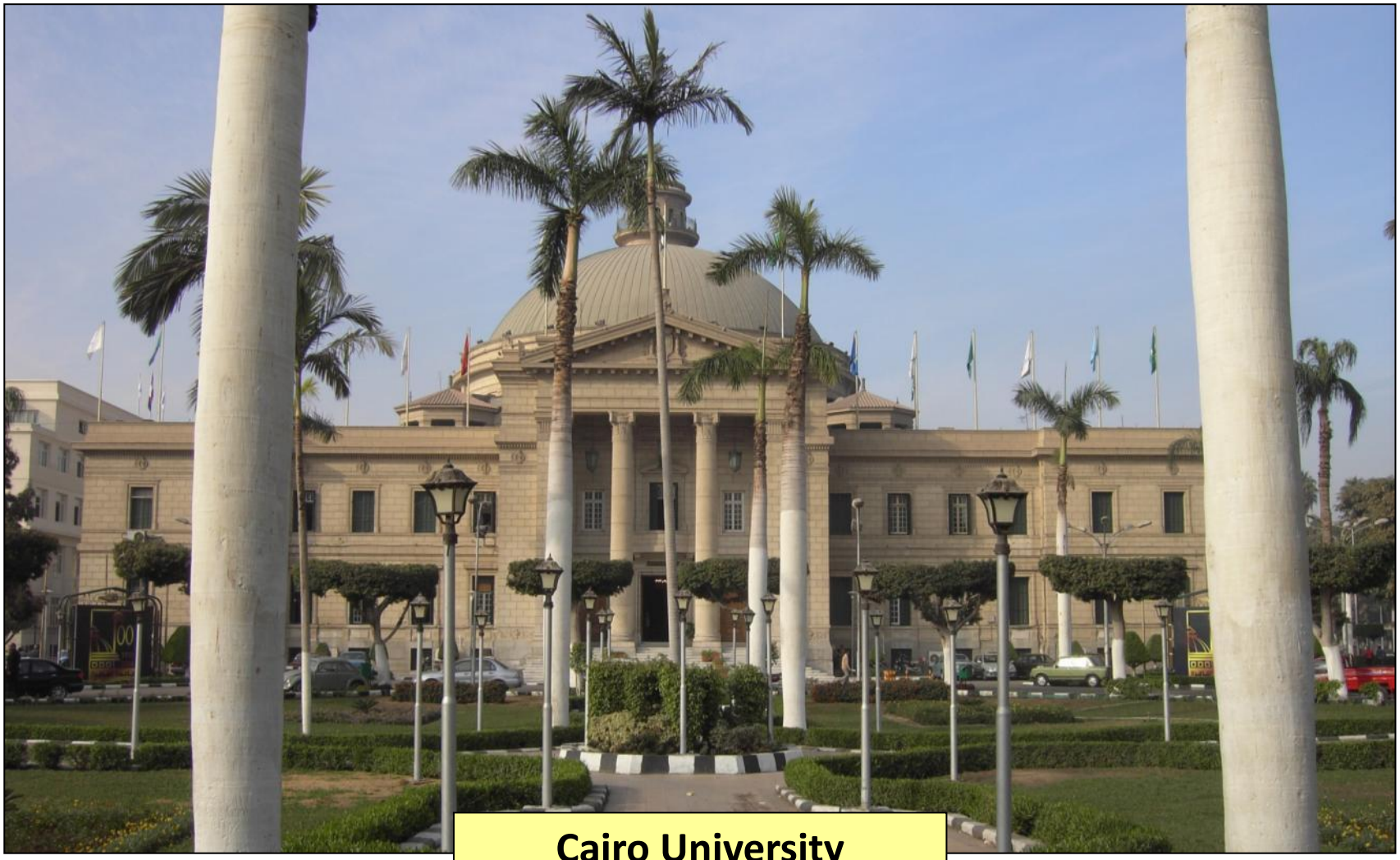
$$Output = f_1 * input_1 + f_2 * input_2 + ... + f_n * input_n$$

where f_n is the factor loading for the n^{th} input.

Factor analysis is used in this case study to determine the fouling in an alcohol plant reboiler. This article provides some additional insight as to how factor analysis is used in an industrial situation.

Exercises and Solutions

Thank You!



Cairo University