

EUROPEAN ORGANISATION  
FOR THE SAFETY OF AIR NAVIGATION



**EUROCONTROL EXPERIMENTAL CENTRE**

**AN INTRODUCTION TO HUMAN IN THE LOOP EXPERIMENTS  
AND STATISTICAL ANALYSIS**

**EEC Note No. 07/06**

Project MMF

Issued: June 2006



**REPORT DOCUMENTATION PAGE**

<b>Reference:</b> EEC Note No. 07/06		<b>Security Classification:</b> Unclassified				
<b>Originator:</b>		<b>Originator (Corporate Author) Name/Location:</b> DeepBlue s.r.l Via Basento 52/D 00198 ROMA ITALY Telephone: +39 06 85 54 801				
<b>Sponsor:</b> Marc Bourgois Deputy Manager Innovative Research Area EUROCONTROL Experimental Centre		<b>Sponsor (Contract Authority) Name/Location:</b> EUROCONTROL Experimental Centre Centre de Bois des Bordes B.P.15 F – 91222 Brétigny-sur-Orge CEDEX FRANCE Telephone: +33 (0)1 69 88 75 00 WEB Site: <a href="http://www.eurocontrol.int">www.eurocontrol.int</a>				
<b>TITLE:</b> <p align="center"><b>AN INTRODUCTION TO HUMAN IN THE LOOP EXPERIMENTS AND STATISTICAL ANALYSIS</b></p>						
<b>Authors</b> Monica Tavanti (DEEPBLUE)	<b>Date</b> 06/2006	<b>Pages</b> viii + 22	<b>Figures</b> 1	<b>Tables</b> 12	<b>Annexes</b> -	<b>References</b> 10
	<b>Project</b> MMF	<b>Task No. Sponsor</b> C61PT/20004	<b>Period</b> 2006			
<b>Distribution Statement:</b> (a) Controlled by: Marc Bourgois (b) Special Limitations: None						
<b>Descriptors (keywords):</b> Experimental design, Statistical analysis						
<b>Abstract:</b> <p>This document constitutes an input to a larger project that aims to propose and evaluate Augmented Reality (AR) tools for towers. One of the objectives of the project is to empirically evaluate the acceptability and usability of the AR technologies, with controlled experiments. Thus, this document intends to provide a general introduction on hypothesis testing, based on lessons learned during earlier experiments, on how to plan, design and carry out simple experiments.</p> <p>This document was essentially written for students and for readers with very little knowledge of the topic. The content is simple and basic. All technical notions are fully explained and priority is given to practical suggestions rather than to theoretical explanations.</p>						



## FOREWORD

The increasing demand for higher levels of safety and capacity require the exploration of new technologies, especially at bottlenecks in the air traffic system, such as airports.

The Augmented Reality for Tower Control project has selected a specific innovative technology (Augmented Reality), a specific ATC setting (the Control Tower), and an approach that combines the technological push with the final users' needs and the operational constraints.

We believe that Augmented Reality (AR) could be a valuable technology to support and enhance the controller's view-out-of-the-window. However, how can AR be applied and tailored to this specific domain?

The potential of AR should be carefully studied and empirically evaluated. The present note provides a general introduction to experimental design and statistical analysis, which will be used as methodological support to carry out future evaluations.

Marc Bourgois  
Deputy Manager Innovative Research Area

**Page left intentionally blank**

## TABLE OF CONTENTS

<b>LIST OF FIGURES .....</b>	<b>VIII</b>
<b>LIST OF TABLES.....</b>	<b>VIII</b>
<b>1. PURPOSES OF THE DOCUMENT .....</b>	<b>1</b>
<b>2. OBLIGATIONS AND RECOMMENDATIONS .....</b>	<b>2</b>
<b>3. BEFORE THE STUDY .....</b>	<b>3</b>
<b>4. VARIABLES AND CONDITIONS.....</b>	<b>4</b>
4.1. EXPERIMENTAL AND NULL HYPOTHESIS.....	4
4.2. VALIDITY .....	4
4.3. MEASUREMENTS AND SCALES .....	5
<b>5. DESIGN AND CONTROL .....</b>	<b>6</b>
<b>6. A BRIEF INTRODUCTION TO STATISTICAL ANALYSIS .....</b>	<b>7</b>
6.1. STATISTICAL PROBABILITY .....	8
6.2. ERROR TYPES.....	9
<b>7. STATISTICAL TESTS.....</b>	<b>10</b>
7.1. DEGREES OF FREEDOM.....	11
7.2. ANALYSIS OF VARIANCE .....	11
7.2.1. One Way between-subjects ANOVA .....	12
7.2.2. One Way within-subjects ANOVA .....	13
7.3. NON PARAMETRIC TESTS: THE RANKING.....	15
7.3.1. The Kruskal-Wallis Test.....	16
7.3.2. The Friedman Test.....	17
<b>8. SUMMARY.....</b>	<b>18</b>
<b>9. RESUMING TABLES.....</b>	<b>19</b>
9.1. BETWEEN AND WITHIN SUBJECTS DESIGNS: PROS AND CONS .....	19
9.2. STATISTICAL TESTS .....	19
<b>10. INDEX OF THE TERMS.....</b>	<b>20</b>
<b>11. LITERATURE SOURCES AND REFERENCES .....</b>	<b>21</b>
<b>12. ACKNOWLEDGMENTS .....</b>	<b>22</b>

## LIST OF FIGURES

Figure 1: Normal distribution.....	10
------------------------------------	----

## LIST OF TABLES

Table 1: Counterbalancing .....	6
Table 2: Mean, median and mode.....	7
Table 3: Mean and SD for the two conditions.....	7
Table 4: Error types .....	9
Table 5: One Way ANOVA (between).....	12
Table 6: One Way ANOVA (within) .....	13
Table 7: Scores and ranks .....	15
Table 8: Scores and ranks .....	15
Table 9: Scores and ranks (Kruskall-Wallis).....	16
Table 10: Scores and ranks (Friedman) .....	17
Table 11: Pros and Cons.....	19
Table 12: Statistical tests .....	19

## 1. PURPOSES OF THE DOCUMENT

This document constitutes an input to a larger project that aims to propose and evaluate Augmented Reality (AR) tools for Control Towers. The AR project intends to provide support to the tasks carried out in the tower which, nowadays, can be limited or/and negatively influenced by poor visibility conditions (e.g. bad weather, occluded areas, etc.).

One of the objectives of the project is to empirically evaluate the acceptability and usability of the AR technologies, with controlled experiments. Thus, this document intends to provide a general introduction on hypothesis testing, on how to plan, design and carry out experiments.

The document was elaborated taking into account:

1. The previous knowledge of the AR project team members, concerning experimental design and statistics.
2. The need to give simple and accessible basic notions.
3. The need to explain technical notions in a very simple manner.
4. Give priority to practical suggestions rather than to theoretical explanations.

The readers will find here the information necessary to get acquainted with basic notions concerning the design of experiments. Moreover, commonly used statistical tests, namely, the Analysis of Variance (ANOVA) and non parametric alternatives will be briefly explained.

The types of designs described here are quite simple and all the examples concern experiments entailing only one independent variable, but more than two levels. This means that, despite their level of simplicity, these designs (and tests) allow to plan and perform quite sophisticated exploratory experiments.

This report aims to be a simple reference to gain some insights on the topic, and to some extent, it provides oversimplifications of many concepts. However, this basic knowledge can be used as a background to support the understanding of more complex designs and statistical tests. At the end of the document a reference list of books containing deeper and more extensive knowledge on the issue is provided.

## 2. OBLIGATIONS AND RECOMMENDATIONS

The researcher has responsibility towards the subjects involved in the study. The subjects have the right to privacy, confidentiality, they have the right of being informed about the nature of the research, and must be treated with respect. Any subject has the right to withdraw from a study at any point.

Any explanation that is required by the subjects should be provided (compatibly with the purposes of the experiment; i.e. do not reveal the hypotheses beforehand). Sometimes the participants might feel frustrated if they suspect that their performance was not good enough. Thus, the researcher should stress that the main goal of the research is assessing performance in general, or to evaluate the properties of some tools, etc. and make sure that every participant feels comfortable.

Obtain an informed consent, that is, a full agreement (written, when possible) from every participant who has to be acknowledged about the nature, the purposes (and possible risks) that the research could entail.

The researcher should be as neutral as possible; e.g. do not reveal her/his expectations about the experiment (either explicitly, or implicitly by researcher's behaviour).

Other obligations of the researchers are:

1. Perform a risk/benefit evaluation (e.g. possible effects on the participants, etc.).
2. Do not make any alteration on the data. If so, then explain what was done and why (e.g. how and why some of the participants' scores were removed).
3. Errors identified after the research have to be acknowledged to the readers.
4. Avoid plagiarism and do not write other people's work as if it were your own.
5. Anybody who contributed to a research should be acknowledged in all the written materials.
6. Share the data, making it available to the scientific community.

### 3. BEFORE THE STUDY

Here is some advice on how to structure and plan a study from the beginning.

- *Review past studies in the same area.* The knowledge of what was done before on the same topic is essential to every research.
- *Defining a specific topic.* The review of the state of the art, the specific problems and topics of interests can be identified.
- *Define the hypotheses.* In general terms, a research should be guided by some hypotheses. This is especially true when the objective is carrying out an experiment: the aim of an experiment is to test a hypothesis.
- *Pilot studies.* It is a good habit to run some preliminary trials, in order to check if the design and the procedure are appropriate and to verify if the equipment is working properly.

## 4. VARIABLES AND CONDITIONS

The experiments allow the manipulation of some factors in order to measure the possible effects of this manipulation.

In this context “factors” means variables, that is, some measurable properties of a certain event. A variable can be independent (causing a change of another variable), and dependent (affected by the independent variable). Suppose that a researcher wants to test whether a new teaching method will help young pupils in the study of mathematics. A “simple design” could be: two groups of pupils, one learning math with the new method and the other one using a traditional teaching scheme. The independent variable manipulated here is the teaching method and it has two levels (or conditions): *new* vs. *traditional* method; while the dependent variable could be the scores obtained by the two groups of pupils in a math examination.

Confounding variables are variables that can unpredictably affect the outcomes of an experiment: the research should identify possible and alternative explanations for the subjects’ behaviour and exclude every confusing factor from the experiment.

### 4.1. EXPERIMENTAL AND NULL HYPOTHESIS

The aim of an experiment is to test an experimental hypothesis, i.e., a prediction concerning the effects of an independent variable on the dependent variable). However, there is also a second hypothesis: the null hypothesis. Every experimental hypothesis is tested against a null hypothesis, i.e. the one that denies the experimental hypothesis (i.e. the manipulation of the independent variable will not produce the predicted effect).

### 4.2. VALIDITY

Validity refers to the conclusions that a researcher can establish concerning the causal relationship between the independent and the dependent variable. There are four types of validity that need to be taken into account: internal validity, construct validity, external validity, and statistical validity.

Internal validity, refers to the internal logic of the relationship identified between the independent and the dependent variable; the relationship has to be clearly demonstrated.

Construct validity, refers to the extent to which what is manipulated and measured is really what is needed to support a theory, that is a theoretical construct is properly addressed and that other possible theoretical explanations of the results should be excluded.

External validity refers to the generalisability of the results of a research to other subjects, settings, etc.

Statistical validity, this type of validity implies that the causal relationship (between an independent variable and dependent variable) discovered should not be due to chance, or to confounding variables or biases or noise (or, more in general, any systematic artefact of the design).

There are also other important issues related to validity. For instance, any statistically significant effect should also be relevant, i.e. is also significant in practical terms.

### 4.3. MEASUREMENTS AND SCALES

The notion of variable is closely related to the notion of measurement, a rather obvious statement. “Measuring” means assigning numerical values to an event or to an object according to a certain rule.

It is common to distinguish four scales of measurement that differ in the manner (the rule) the values are assigned:

Nominal scales. In this scale, the numbers are labels allowing the classification of objects and events according to some categories (e.g. female=1, male=2, or vice versa).

Ordinal scales. This scale allows the ranking of the events or objects according to an order (e.g. from “very sweet”=10, to “very bitter” =1, etc.). This type of scale (e.g. the ordinal position) does not provide any information about the differences between its composing elements; in other words, I cannot say that “very sweet” is ten times sweeter than “very bitter”.

Interval scales. With the interval scale, instead, the differences among the values have meanings, since it is assumed that here there are equal intervals between the values. A classical example of interval scale is the temperature. The difference between 30° and 20° Celsius is the same as between 40° and 50°.

Ratio scales. Similarly to the interval scale, this scale gives information concerning the magnitude of the differences between the measured events, but the ratio scale gives additional information: it is provided with a true zero point. Taking again the example of the temperature: would it be meaningful to state that: 20° is twice as warm as 10°? If the same example is provided with centimetres, then the statement makes more sense.

## 5. DESIGN AND CONTROL

In section 4, we referred to the terms «a simple design», mentioning the fact that the researcher had selected two groups of pupils, one receiving the new teaching method and the other one receiving a traditional method. This means, for example, that ten students are allocated to one experimental condition, and ten students to the other condition. This is a between-subjects design.

When using different groups of subjects it is necessary to bear in mind that there could be a variability linked to the subjects (e.g. some students remarkably good at math).

- Suggestion I: In order to deal with this type of problems, all the subjects have to be assigned randomly to each condition. Random assignment means that every subject must have equal chances to end up in one of the conditions. Random assignment is essential. Following non-random strategies may lead to biased results.

Another way to tackle the variability of the subjects is assigning the same subjects to different conditions. This is a within-subjects design. Also this design implies some problems, mostly related to order effects and carry-over effects<sup>1</sup>.

- Suggestion II: Counterbalance the order of the treatment. In our example<sup>2</sup>, the treatment corresponds to the type of teaching method (e.g. treatment A and B = method *new* and *traditional*). So, having twenty pupils as subjects, half of the subjects should receive first the treatment A and then the treatment B; and the other half, should receive first the treatment B and then the treatment A. For more than two conditions, the conditions should be arranged so that the order of presentation is different for each subject, so as to avoid any systematic learning effect). For example, having three conditions:

Table 1: Counterbalancing

Subj 1	A	B	C
Subj 2	B	C	A
Subj 3	C	A	B
Subj 4	A	C	B
Subj 5	B	A	C
Subj 6	C	B	A

Control is very important as it helps to exclude that other variables (than the selected independent variable) will affect the results.

A last remark concerns the number of the subjects. How many subjects should participate in an experiment?

- Suggestion III: There is one rule of thumb: as many subjects as possible. Usually, in many experiments, a number between 10 and 15 for each condition is used; however, when feasible, increasing this amount is beneficial.

<sup>1</sup> That is, when *something* is transferred from one condition to another.

<sup>2</sup> In this context, we kept the example of the teaching method for practical reasons, even if it is not a good example. In fact, when *learning* is involved, carry-over and learning effects could bias the results of the experiment. Thus, when *learning* is the chosen task, a between-subjects design would be more appropriate.

## 6. A BRIEF INTRODUCTION TO STATISTICAL ANALYSIS

Once the experiment is carried out and the scores are available, the data has to be summarised in a descriptive manner. Given a list of scores (e.g. math scores of our example) we can put them together and try to get some “impressions” about them. There are three “summary” statistics giving some ideas about the central measure of our data. The mean (or average, the sum of the scores divided by their number); the median (the middle point of the ordered scores); the mode, which is the most frequently occurring value among the scores (cf. Table 2).

Table 2: mean, median and mode

Mean	Median	Mode
10	1	10
8	3	8
5	4	5
3	5	3
5	5	5
7	7	7
1	8	1
4	9	4
9	10	9
<b>Mean:</b>	<b>Median:</b>	<b>Mode:</b>
<b>5.8</b>	<b>5</b>	<b>5</b>

Table 3: mean and SD for the two conditions

Group_new		Group_traditional	
10		5	
8		2	
7		8	
8		5	
7		8	
9		2	
6		6	
8		10	
10		2	
<b>Mean</b>	<b>8.1</b>	<b>Mean</b>	<b>5.3</b>
<b>SD</b>	<b>1.36</b>	<b>SD</b>	<b>2.96</b>

The mean, mode and median give information about the central tendency of the scores, which is, where the scores tend to aggregate and cluster together.

But, there are also some “markers” of variability, that is, the opposite tendency of the scores to spread and differentiate away from each other, namely the variance and the standard deviation.

The variance ( $s^2$ ) tells us the degree to which an individual score deviates from the mean of the scores. It is calculated starting from the Sum of the Squared deviates (SS), obtained subtracting each score from the mean and squaring each value. Then all the values are summed up; this sum is then divided by the number of scores. In other words:

$$s^2 = \frac{\sum (x_i - M_x)^2}{N} = \frac{SS}{N}$$

Where  $x_i$  represents the score,  $M_x$  is the mean, and  $N$  is the number of scores.

The standard deviation (SD) is the square root of the variance.

A simple example on the meaning of variability and spread: if you have a group of nine scores and the value of any of the scores is five, both variance and SD will be zero: that is, there is no variability in the scores.

## 6.1. STATISTICAL PROBABILITY

Let's take again the example of the teaching methods. There are two groups, one for each condition; the researcher carried out the experiment, collected the results, summarised the data. Looking at the Table 3, it seems that the averages of the scores are different (8.1 vs. 5.3); but are they significantly different?

Statistical tests tell us the probability whether an outcome of a research is due by chance. More exactly, they can tell the percentage of probability that a certain effect is due to chance. This critical probability is called *alpha* (written also as:  $\alpha$  or  $p$ ). An important decision concerns to what degree the researcher is ready to take the risk that a certain result is random. For example, if we take a decision criterion so that  $p=.05$ , then we also accept the fact that there is a 5% probability that the results are just random; otherwise, we may want to accept only a 1 % probability that the results are just random ( $p=.01$ ), etc.

- Suggestion IV: The *alpha* level should be set before the experiment (e.g.  $p=.05$ ) and kept, no matter how significant the results are.

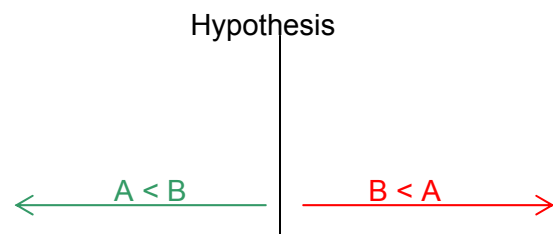
These probabilities levels are provided in statistical tables. These tables report the critical values for a certain statistic and the associated levels of significance; in other words, they support the identification of a certain level of significance for any given difference.

Another important issue is whether the experimental hypothesis is "open" or "closed". To be more precise, we must decide whether the experimental hypothesis is two-tailed (or bi-directional) or one-tailed (or directional). This means that the researcher investigating the teaching method could predict if the new teaching method will affect the math scores (in either ways, thus the prediction does not go into a specific direction) or if the new teaching method will increase the math scores of the pupils, or vice versa (thus making a prediction going into a specific direction). Here it is a graphical explanation of the concept.

In the picture, there are two directional hypotheses:

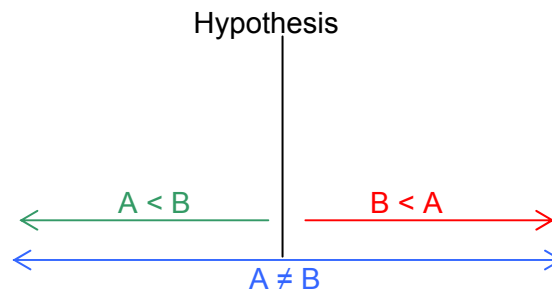
A is smaller than B ( $A < B$ )

B is smaller than A ( $B < A$ )



These two hypotheses are mutually exclusive (logically, the one excludes other).

However, we could also have a "more general" or bi-directional hypothesis, that is: A is different than B ( $A \neq B$ ). In this case we do not have any precise hypothesis regarding the *direction* of this difference: A is different than B, we leave "open" the following questions: is B smaller than A? Or, is B bigger than A? Intuitively, we know that this "open" hypothesis comprises two precise, directional hypotheses. If  $A \neq B$ , then it could be that  $A < B$  or it could be that  $B < A$ . If we look at the next picture we see that the bi-directional hypothesis "contains" the two directional hypotheses.



Probability values are additive. Deciding that a hypothesis goes in either directions, means that the probabilities of a random difference is given by the added probabilities of random differences arising in both directions. Predicting an outcome that goes in either direction means doubling the probabilities that a certain difference occurs by chance.

## 6.2. ERROR TYPES

There are two ways in which a researcher could be wrong about a hypothesis: considering again the null hypothesis (usually written as:  $H_0$ ):

Table 4: Error types

		$H_0$ False	$H_0$ True
<i>Decision</i>	Reject $H_0$	Correct	Type I error
	Fail to reject $H_0$	Type II error	Correct

A Type I error occurs when the null hypothesis is rejected even though it is true; vice versa accepting the null hypothesis when it is false means committing a Type II error.

The researcher should decide whether it is more or less risky to make a Type II error (or a Type I error). Knowing this, the researcher should be aware of the “limits” that a research involves, and decide whether being more or less conservative.

- Suggestion V: Usually, for an exploratory study, a .05 level is acceptable as decision criterion.
- Suggestion VI: Setting a smaller  $\alpha$  will help reducing Type I errors.

The goal of an experimental study is testing hypotheses and speculating whether a certain effect (or result) observed in the sample can be “generalisable” to a broader number of subjects. This simple fact implies that it must be evaluated whether a p level of .05 is acceptable for the purposes of the research.

## 7. STATISTICAL TESTS

There is distinction to make, between “parametric and non-parametric” statistical tests. Traditionally, the choice between these two “types” is dictated by some rules. Usually, parametric tests require the following assumptions to be satisfied:

1. The scores are measured in an interval or ratio scale.
2. The scores for each condition come from a population having normal distribution.
3. There is homogeneity of variance.

It has been explained in section 4.3 the meaning of interval and ratio scale. Let’s explain the remaining assumptions with some examples.

Usually the scores are distributed so that there are more “middle values” than “extreme values”. In other words, if we could account, for example, the height of a population, we would discover that very short and very tall people are less frequent than “middle height people”. A normal distribution curve is bell-shaped, that is, it is “elevated” in the middle where the highest frequency occurs, and it looks symmetrical.

This concept is illustrated in Figure 1. The bars represent the scores (with middle range scores in the middle and extreme scores at the tails).

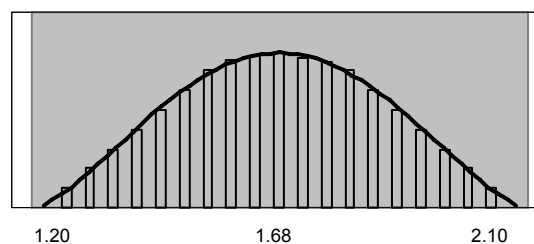


Figure 1: Normal distribution

Plotting the frequency distribution of the data (i.e. the scores) with a histogram helps us to discover whether the data approximates (or “look like”) the normal distribution. Another way to check the normality of the distribution is to use a normality test. An example of normality test is the Shapiro-Wilk W test.

The third assumption allowing the choice of a parametric test is the homogeneity of variance<sup>3</sup>. This means that the variability in the conditions should be (more or less) the same. In order to exactly test whether the requirement of homogeneity of variance is satisfied, a dedicated test could be used, for example, the Levene’s test.

- Suggestion VII: Non-parametric tests require less rigid requirements to be satisfied if compared to parametric tests. If the assumptions are weakly met and/or more than one requirement is violated, then, a non-parametric test would be more suitable<sup>4</sup>.

<sup>3</sup> The “impact” of this assumption can be lessened by always having an equal number of subjects in each condition.

<sup>4</sup> This is and should be taken as a simple suggestion. There are other options to deal with data not compliant with the assumptions for parametric tests (e.g. non normal data can be transformed).

## 7.1. DEGREES OF FREEDOM

I will follow the examples and the definitions given in all the books, by saying that the "degrees of freedom" (df) means: the freedom to vary.

Assume having two dishes A and B and that we need to place some beans into the dishes, given the condition that the total sum of the beans to be placed into the two dishes is 10. At the beginning of this process we are free to put any number of beans into the first dish (e.g. 4); but then we won't be free to place any number of beans into the dish B; this choice is constrained by the fact that the total sum has to be 10 (necessarily we have 6 beans left). This means that we have only 1 df.

Now, let's assume that we have four dishes (A, B, C and D), and that we need to fill them with 20 beans. We are free to place any amount of beans into the dishes A, B and C (e.g. 5, 3 and 6), but the number of beans for the dish D is constrained: there are 6 beans available for that dish. In this case we have 3 df.

Now, let's take a sample of 4 subjects, and consider a situation in which only the number of subjects (4), the scores of three subjects (e.g. 2, 3 and 6) and the sum of the scores (20) are known. The score of the fourth subject is also predictable. This means that this fourth score has no freedom to vary, and that there are 3 df (4-1).

The df is an important concept to many statistical tests. In the following sections the appropriate way to calculate the necessary df will be given.

## 7.2. ANALYSIS OF VARIANCE

A commonly used parametric test is the analysis of variance (ANOVA). In this report, two types of ANOVA will be described: one that can be used for between-subjects design and the other one for within-subjects design. The two tests described here can be used to seek differences between two or more levels of one independent variable (e.g. comparing the math scores after three different teaching methods have been used).

As its name suggests, this test is based on the concept of variance, which (cf. section 6) is the degree to which each value deviates from the mean of a certain group.

The ANOVA allows making estimations concerning the variance of a population, "comparing":

1. the variance coming within the levels of the independent variable (thus, variation due error, e.g. individual differences);
2. the variance coming from between the levels of the independent variable (or variation between the means).

This "comparison" is a ratio (called F), and given by:

$$F = \frac{\text{Variance between the means}}{\text{Error (unexplained variance)}}$$

Thus, the F ratio compares the variance between the levels of an independent variable (i.e. difference between groups due the independent variable) and the variance within the levels of that variable (or error, since it cannot be explained). If the  $H_0$  is true, the estimated variances between the means are more or less the same (F approximating 1) and thus the independent variable did not produce any effect on the dependent variable.

In order to calculate the F ratio, for every source of variance (“between and within”), the Mean Squares, or MS have to be calculated. The MS are obtained dividing the SS (cf. section 6) by the appropriate degree of freedom.

### 7.2.1. One Way between-subjects ANOVA

The One Way between-subjects ANOVA split the variance into two sources:

$$F = \frac{\text{between groups variance}}{\text{within groups variance (error)}}$$

Table 5 reports all the “ingredients” of the One Way between-subjects ANOVA.

**Table 5: One Way ANOVA (between)**

Sources of variance	SS	df	MS	F	p (alpha)
Between groups	SS <sub>bg</sub>	df <sub>bg</sub>	SS <sub>bg</sub> / df <sub>bg</sub>	MS <sub>bg</sub> / MS <sub>wg</sub>	...level...
Within groups (or Error)	SS <sub>wg</sub>	df <sub>wg</sub>	SS <sub>wg</sub> / df <sub>wg</sub>		
Total	SS <sub>tot</sub>	df <sub>tot</sub>			

Where, the df are given by:

1.  $df_{bg} = k - 1$  (where k is number of conditions)
2.  $df_{tot} = N - 1$  (where N is the total number of subjects -scores-)
3.  $df_{wg} = N - k$

As an example, having three conditions (thus, three independent groups of subjects), and 10 subjects in each group, the degrees of freedom will be:

1.  $df_{bg} = 3 - 1 = 2$
2.  $df_{tot} = 30 - 1 = 29$
3.  $df_{wg} = 30 - 3 = 27$

The Sum of the Squares between groups (SS<sub>bg</sub>), is obtained by finding the deviation of each condition mean from the overall mean; every deviation should be squared and multiplied by the number of scores; then all the data should be summed up to have the total. The total Sum of the Squares (SS<sub>tot</sub>) is obtained squaring the deviation of every score from the overall mean, and then summing up them up to have the total. The SS within groups (SS<sub>wg</sub>) is given by:  $SS_{tot} - SS_{bg}$ .

Then, we can calculate the MS, for “between groups” (or MS<sub>bg</sub>) and “within groups” (MS<sub>wg</sub>.) dividing the SS by the adequate df (as shown in the table 5). The F ratio will be given by:

$$F = \frac{MS_{bg}}{MS_{wg}}$$

When the F value is obtained, we have to bear in mind two things:

What *alpha* level was chosen (e.g. .05);

The degrees of freedom of F (according to the previous example we have  $F_{2,27}$ , where 2 is  $df_{bg}$  and 27 is  $df_{wg}$ ).

At this point, given the degrees of freedom 2 and 27, we can compare the calculated F, with the one reported in the appropriate statistical table where the critical values (for any given *alpha* level) are provided<sup>5</sup>.

If the calculated F value is equal to or larger than the F value of the table, then the null hypothesis can be rejected. But, if the calculated F is smaller than the one indicated in the table, we must accept the null hypothesis<sup>6</sup>.

The ANOVA tested a “two-tailed” hypothesis, it checks for overall differences among conditions and gives only a “preliminary” result, which has to be further analysed (i.e. with three conditions, a directional hypothesis is only one among of the possible divergent hypotheses).

If the ANOVA identified a significant difference among the means of the three groups, then it is possible to proceed to pairwise comparisons, comparing one by one all the three conditions (e.g. 1 vs. 2; 1 vs. 3; 2 vs. 3), by using other statistical tests (an example of post-hoc test that can be used in this case, is the HSD Tukey).

- Suggestion VIII: If the test allowed discovering a significant difference among conditions, pairwise comparisons testing directional hypotheses can be carried out.

### 7.2.2. One Way within-subjects ANOVA

This type of ANOVA can be used to seek differences between more than two levels of one independent variable administered to the same subjects. The same general logic explained in the previous section applies, since the total variation in the data is attributable to two sources: the differences between the groups and the differences within the groups. However, this last source of variance is further split. More specifically, this type of ANOVA allows distinguishing the part of the  $SS_{wg}$  that can be due to individual differences. This part (called  $SS_{subj}$ ) is removed from the analysis and the remaining  $SS_{error}$  is used as the measure of unexplained variation.

Table 6: One Way ANOVA (within)

Sources of variance	SS	Df	MS	F	p (alpha)
Between groups	$SS_{bg}$	$df_{bg}$	$SS_{bg}/df_{bg}$	$MS_{bg} / MS_{error}$	...level...
Within groups	$SS_{wg}$	$df_{wg}$			
Subjects	$SS_{subj}$	$df_{subj}$			
Error	$SS_{error}$	$df_{error}$	$SS_{error}/df_{error}$		
Total	$SS_{tot}$	$df_{tot}$			

<sup>5</sup> Usually, books of statistical analysis include the necessary statistical tables in the appendixes.

<sup>6</sup> Checking the tables is somehow “old-fashioned” nowadays. Statistical software report the exact alpha value for every test performed.

The  $SS_{bg}$ ,  $SS_{wg}$ , and  $SS_{tot}$  can be calculated in the same manner as for the between-subjects ANOVA. The  $SS_{subj}$  can be obtained by taking the average of each subject's scores for all the conditions deduct it from the overall mean, then square the result and multiply it by the number of conditions. In other words:

$$SS_{subj} = K(M_{subji} - M)^2$$

Where K is the number of conditions.

The  $SS_{error}$  is given by:  $SS_{wg} - SS_{subj}$

The df are calculated as it follows. Assuming that we have a sample of 10 subjects, each of them participating in an experiment in which three levels of one independent variable are evaluated, we will have:

1.  $df_{tot} = N_{tot} - 1$  (where  $N_{tot}$  is total number of scores)  $\rightarrow df_{tot} = 30 - 1 = 29$
2.  $df_{bg} = k - 1$   $\rightarrow df_{bg} = 3 - 1 = 2$
3.  $df_{subj} = N - 1$  (N is the number of subjects)  $\rightarrow df_{subj} = 10 - 1 = 9$
4.  $df_{wg} = N_{tot} - K$   $\rightarrow df_{wg} = 30 - 3 = 27$
5.  $df_{error} = df_{wg} - df_{subj}$   $\rightarrow df_{error} = 27 - 9 = 18$

When the  $SS_{bg}$  and the  $SS_{error}$  are divided by the corresponding degrees of freedom (cf. Table 6), the required MS values can be obtained. The F statistic is given by:

$$F = \frac{MS_{bg}}{MS_{error}}$$

When the F value is obtained, the same procedure and the same logic used for the between-subjects ANOVA apply. Having in mind the alpha level and the appropriate degrees of freedom ( $F_{2,18}$ ) the calculated F value has to be compared with the critical F reported in the statistical table.

Now, let's imagine that we performed an experiment aiming to explore whether there is a difference (for example, in terms of reaction times) between three different interaction metaphors to be used with a three-dimensional environment. For the study, a within-subjects design was used, that is, the same subjects performed the tasks with all the three metaphors. The study was exploratory, so we decided to choose an alpha level of .05.

Then, we carried out the ANOVA to see if there was any significant difference in the performance with the three metaphors, and we obtained the F value.

The calculated F appeared smaller than the critical F reported in the table, at the level of significance chosen. Thus, we have to reject the experimental hypothesis: no significant difference was found among the three metaphors. Since we cannot accept the experimental hypothesis, more detailed analyses with pairwise comparisons will not be performed.

### 7.3. NON PARAMETRIC TESTS: THE RANKING

As stated before, the non parametric tests allow to perform statistical analyses when the conditions described in section 7 are poorly respected, or if more than one assumption is not met. One of the differences between parametric and non parametric tests is that the most common non parametric tests compare the subjects' performance not using the original scores but the ranks of the scores.

For instance, if we have a group of scores, they have to be ranked in order to determine which score is higher or lower. Usually the rank is assigned so that the smallest score goes first (cf. Table 7, where the smallest score is 3, which is ranked 1; the biggest score is 12, which is ranked 7).

Table 7: Scores and ranks

Scores	Ranks
6	4
3	1
12	7
4	2
7	5
5	3
8	6

In this case the ranking is very easy, being all the scores different. When we have the same value for more than one score (i.e. tied scores), then the average of the ranks have to be assigned. In other words, having three scores of 1 (1, 1, 1), the rank value should be:  $1+2+3/3=2$ , then the next assigned rank should be 4.

There are different ways to assign ranks depending on the type of design.

For between-subjects designs the ranks have to be assigned as if they were a single set of ranks. For example, if we compare the scores of two groups, the ranks will be:

Table 8: Scores and ranks

Scores Group A	Ranks	Scores Group B	Ranks
1	1	2	2,5
4	5	6	7
5	6	3	4
2	2,5 <sup>7</sup>	7	8

For within-subjects design the way to assign ranks to scores is a bit more complicated and varies depending on the test.

<sup>7</sup> The score 2 of the Group A, and the score 2 of the Group B are tied.

### 7.3.1. The Kruskal-Wallis Test

This test corresponds to the between-subjects ANOVA and should be used when the requirements for the ANOVA are not met (cf. Section 7).

So, as an example, we have three groups (A, B, and C), each with 10 controllers who are engaged in a simple task, like rating the easy of use of three types of radar displays (according to a scale so that 1 means “very bad” and 10 means “very good”).

Our hypothesis is that there is a difference among the means of the ratings of the displays. The null hypothesis is that there is no difference. For this study we decided to set .05 as the decision criterion.

Table 9: Scores and ranks (Kruskall-Wallis)

SCORES			RANKS		
A	B	C	A	B	C
6	3	1	5	2.5	1
7	5	3	6	4	2.5
...	...	...	...	...	...
30	...	...	...	...	...
Rank totals (=sum of ranks)			70	40	15

As shown in Table 9, the smallest value (1) corresponds to the “first” rank (1); while the largest value (7) corresponds to the “last” rank (6). The aim of this test is to assess whether there is a significant difference among the rank totals of three groups.

Once the ranks are assigned, the following equation has to be used to obtain a value called H:

$$H = \frac{12}{N(N+1)} \left( \sum \frac{T^2}{n} \right) - 3(N+1)$$

where N is the number of all subjects; n is the number of subjects in each group;  $T^2$  is the squares of the rank totals for each condition.

Once the H is known and the degrees of freedom are calculated (in this case, the df are given by the number of conditions minus one, that is 3 - 1 = 2 df), the H we found has to be compared to the H reported in the corresponding statistical table. For the level of significance decided (.05), if the calculated H is equal to or larger than the critical H in the table, then the difference is significant. Let’s assume that the calculated H is larger than the critical H reported in the table.

Thus we can say that: df=2, H=10<sup>8</sup>, p<.05; we can reject the null hypothesis. This test only checks for overall differences among conditions (i.e. a two-tailed hypothesis). Thus, if the calculated value of H is significant, we can proceed to pairwise comparisons<sup>9</sup>.

<sup>8</sup> This value was taken (almost randomly), as an example. However, 9.9 is indeed larger than the critical H value, for df=2.

<sup>9</sup> In this case the most appropriate test is the Mann-Whitney U test, a non parametric test for one independent variable with two levels.

### 7.3.2. The Friedman Test

This test should be used for within-subjects designs entailing two or more conditions. If the within-subjects ANOVA cannot be used because the requirements described in section 7 for a parametric test are not adequately satisfied, the Friedman test is an adequate choice.

Suppose that we want to compare three interaction metaphors and that we ask to a group of 8 controllers to rate each metaphor using a scale from 1 to 5. Since it is a within-subjects design, every controller takes part to three different conditions, one for each metaphor. We decided to set an alpha level of .05.

Also for the Friedman test the original scores should be ranked. But, differently from the Kruskal-Wallis, this time, the scores ranked entail each subject across each condition (to use a “graphical explanation”, the ranking is done “horizontally” for each subject). For example, in the Table 10, the scores of the subject 1 are ranked (in the order) 1, 3 and 2. The scores of the subject 2 are ranked: 1, 3 and 2, etc.

Table 10: Scores and ranks (Friedman)

Subject	Scores Metaphor A	Scores Metaphor B	Scores Metaphor C	Rank Metaphor A	Rank Metaphor B	Rank Metaphor C
1	2	5	4	1	3	2
2	1	5	3	1	3	2
3	3	5	5	1	2.5	2.5
4	...	...	...	...	...	...
...	...	...	...	...	...	...
8	...	...	...	...	...	...
Rank totals				12	15	17

Once the ranks are assigned the  $\chi_r^2$  value has to be found. The following equation should be used:

$$\chi_r^2 = \left( \left( \frac{12}{Nk(k+1)} \right) (\sum T^2) \right) - (3N(k+1))$$

Where N is the number of subjects (8); k is the number of conditions (3);  $T^2$  is the squares of the rank totals for each condition. Once the  $\chi_r^2$  value is found, it can be compared to the critical  $\chi_r^2$  reported in the corresponding statistical table. In order to find the right  $\chi_r^2$  in the table, the number of subjects (N) and the number of conditions (k) are required. The calculated value of  $\chi_r^2$  is significant (at any given level of p) if it is equal to or larger than the critical  $\chi_r^2$ . Let's assume now, that our  $\chi_r^2$  is smaller than the critical  $\chi_r^2$  at the level of significance decided (which was .05, cf. above). This means that we must accept the null hypothesis. This also means that our analysis cannot further proceed.

- Suggestion IX: For an exploratory study, if a statistical test that checks for overall differences among conditions does not allow rejecting the null hypothesis, the pairwise comparisons (comparing each condition against each other) cannot be performed.

## 8. SUMMARY

As a conclusion, the “main steps” to follow to start and perform an experiment are summarized here.

1. Formulate the hypothesis.
2. Identify the independent and the dependent variables.
3. Check for confounding variables (or sources of noise).
4. Design the experiment (type of design, counterbalancing, etc.).
5. Check the measurement scale and think about the possible options of statistical tests.
6. Check the validity.
7. Perform a pilot test.
8. Get a “general impression” and perform some checks (e.g. flaws in the design; validity; appropriateness of the variables; confounding variables; suitability and functioning of the equipment; correctness of the recordings; etc.)
9. If required, refine the experiment and carry out additional pilot tests.
10. Define the  $\alpha$  level.
11. Collect the participants (at least 10 for each condition).
12. Use the random assignment to allocate the subjects to each condition.
13. Counterbalance (when required).
14. Run the experiment.
15. Collect the data.
16. Analyse the data with descriptive statistics.
17. Check the distribution and the homogeneity of variance.
18. Perform the statistical analysis as it was planned (using the most appropriate statistical test).
19. Accept / reject the null hypothesis, respecting the chosen  $\alpha$  level.

The last important step is the analysis of the results. When reporting the results of an experiment, the researcher should explain what happened and why, that is, finding an explanation of the results; eventually, also comparing the findings with the results found in the literature.

Moreover, any other possible explanation for the results should be explored and reported.

As a matter of fact, sometimes, an experiment is not sufficient to address completely the topic that it aimed to tackle and definitive conclusions cannot be drawn. The analysis of the results can provide some hints and ideas to perform more experiments, to test new hypotheses, or to address the problems according to a different point of view.

## 9. RESUMING TABLES

### 9.1. BETWEEN AND WITHIN SUBJECTS DESIGNS: PROS AND CONS

In section 5 two types of experimental design were explained. The following table reports some additional information expressed in the form of pros and cons that can help the researcher in the choice between the two design types.

Table 11: Pros and Cons

	Pros	Cons
Within-subjects design	<ul style="list-style-type: none"> <li>Reduce variability linked to subjects' differences</li> <li>Need less subjects</li> </ul>	<ul style="list-style-type: none"> <li>Carry-over effects</li> <li>Counterbalancing complex to manage when many conditions</li> </ul>
Between-subjects design	<ul style="list-style-type: none"> <li>No carry-over effects &amp; less boredom/tiring</li> <li>Convenient when some attributes have to be considered: male vs. females; controllers vs. non controllers, etc.</li> </ul>	<ul style="list-style-type: none"> <li>More variability</li> <li>Need more subjects</li> </ul>

### 9.2. STATISTICAL TESTS

As mentioned in section 1, this document aims to provide a quick and easy reference to support the beginners in the understanding of experimental design and of (some) statistical tests. However, it should be clear that there are also other tests that can be used, depending on the requirements of the experiment. Some of those tests (e.g. the t-test) are included in the following resuming table. The table provides a short summary of the most common parametric and non parametric tests according to the experimental design chosen for the study.

Table 12: Statistical tests

	TESTS	
	<u>Parametric</u>	<u>Non parametric</u>
Two conditions (Within)	t-test (related)	Wilcoxon
Two Conditions (Between)	t-test (unrelated)	Mann-Whitney U test
Three or more Conditions (Within)	One way Within ANOVA	Friedman
Three or more Conditions (Between)	One way Between ANOVA	Kruskal-Wallis

## 10. INDEX OF THE TERMS

This section includes a small index of the main terms used in the document. For every term the page number is given.

Analysis of variance	11, 19
Between-subjects design	6, 19
Carry-over effects	6, 19
Conditions	4, 6, 7, 10, 12, 13, 14, 15, 16, 17
Confounding variables	4, 18
Counterbalancing	6, 18, 19
Critical probability	8
Degrees of freedom	11, 12, 13, 14, 16
Dependent variable	4
Error types	9
Experimental hypothesis	4, 8, 14
Friedman	17, 19
Independent variable	4
Kruskal-Wallis	16, 17, 19
Levene's test	10
Mean	7
Measures of central tendency	7
Median	7
Mode	7
Non parametric tests	15
Null hypothesis	4, 9, 13, 16, 17, 18
Random assignment	6, 18
Ranking	15, 17
Scales	5
Shapiro-Wilk W test	10
Standard deviation	7
Tied scores	5
Validity	4, 18
Variance	7, 10, 11
Within-subjects design	6, 19

## 11. LITERATURE SOURCES AND REFERENCES

The main literature sources used for this document are:

1. Baroni, M.R., (1984). Il soggetto umano nelle ricerche di psicologia, Borla, Roma.
2. Ercolani A.P., Areni A., (1995). Statistica per la ricerca in psicologia, Il Mulino, Bologna.
3. Luccio, R., (1996). Tecniche di ricerca e analisi dei dati in psicologia, Il Mulino, Bologna.
4. Wickens, C. D., Gordon, S. E., Liu, Y., (1997). An introduction to human factors engineering, Addison Wesley, NY.
5. A good source of information (for this document, but also in more general terms) is the Richard Lowry's site at Vassar. He provides detailed accounts of many concepts. His explanation of ANOVA is simple and essential.
6. The following books, available at the EEC library, provide more extensive and detailed knowledge concerning the design of experiments and statistical analysis.
7. Brodsky, B. E., Darkhovsky, B. S., (2000). Non parametric statistical diagnosis, Kluwer Academic Publishers, The Netherlands.
8. Dean, A., Voss, D., (2000). Design and analysis of experiments, Springer, NY.
9. Frigon, L., Mathews, D., (1997). A practical guide to experimental design, John Wiley and Sons Inc., NY.
10. Hicks, C. R., Turner, K. V., (1999). Fundamental concepts in the design of experiments Fifth Edition, Oxford University Press, NY.

## 12. ACKNOWLEDGMENTS

I would like to thank Marc Bourgois and Patrizia Marti, and also Alistair Jackson, Guglielmo Guastalla, Elsa Freville and Dirk Schaefer for providing useful comments and feedback. Also thanks to my former advisor, Mats Lind, who firstly introduced me to this topic, providing useful comments and suggestions.