# Experimental Design and Analysis

## Statistical Inference

All the techniques described so far are intended either to describe or summarise data.

For many purposes this is sufficient, but often we need to go further:

* We may wish to *prove* that one set of scores differs significantly from another set
* We may wish to prove that there is a difference in a certain direction, e.g., that the set of scores *A* is better in some way than the set of scores *B*

This is known as drawing *statistical inference*.

## Significance

When designing experiments we try to keep all possible factors stable with the exception of one, the *independent variable*, which we deliberately manipulate in some way.

We then measure another variable, the *dependent variable*, to see how it has been affected by the change(s) in the independent variable.

However, we cannot assume that *all* changes in the dependent variable are due to our manipulation of the independent variable.

Some changes will almost certainly occur by chance.

The purpose of statistical testing is to determine the likelihood that the results occurred by chance.

We can never prove beyond doubt that any differences observed are the result of changes in the independent variable rather than mere chance occurrences.

However, we can determine just how likely it is that a given result could have occurred by chance, and then use this figure to indicate the reliability of our findings.

Before testing, we formulate two hypotheses:

* *Any differences arise purely as a result of chance variations*

This is known as the *null hypothesis*

* *Any differences arise - at least in part - as a result of the change(s) in the independent variable*

This is known as the *alternate* or *experimental hypothesis*

Statistical tests allow us to determine the likelihood of our results having occurred purely by chance.

Thus they allow us to decide whether we should accept the *null hypothesis* or the *alternate hypothesis*.

We usually express probability on a scale from 0 to 1. For example:

*p* ≤ 0.05

When accompanying a statistical finding, this indicates that the likelihood of the observed difference having occurred as a result of chance factors is less than one in 20.

This is known as the *significance level*.

When testing, we choose a level of significance in advance, then test for this level of significance.

* If the outcome is greater than our chosen significance level (e.g., *p* > 0.05) we accept the null hypotheses.
* If the outcome is equal to or less than our chosen significance level (e.g., *p* ≤ 0.05) we accept the alternate hypotheses.

What is an appropriate level of significance to test for?

* If we choose a relatively high value of significance, we are more likely to obtain significant results, but the results will be wrong more often.

This is known as a *Type 1* error.

* If we choose a very low value for significance, we can place more confidence in our results. However, we may fail to find a correlation when it does in fact exist.

This is known as a *Type 2* error.

There is no 'correct' value for significance. However, the value *p* ≤ 0.05 is commonly used and is regarded as a good compromise.

Where a more stringent test is required (e.g., because the result is controversial and/or challenges accepted theories), a value of *p* ≤ 0.01 is sometimes used.

## One-Tailed and Two-Tailed Predictions

In formulating our prediction, we must also decide whether to predict the *direction* of any observed difference or not.

* If we predict only that there *will* be a difference, we are using a two-tailed test.
* If we predict the *direction* of the difference, we are using a one-tailed test.

For example, suppose we choose *p* ≤ 0.05.

* If we use a two-tailed test, we are predicting that the difference between the test statistics will be such that one falls in the top or bottom 2.5% bands of the distribution relative to the other.
* If we use a one-tailed test, we are predicting that the difference between the test statistics will be such that one falls in (e.g.) the top 5% band of the distribution relative to the other.



For a given value of *p*, the requirements for a one-tailed test are less extreme than for a two-tailed test.

Most statistical tests can be used to test either one- or two-tailed predictions.

Having considered all these factors, we can now decide which test should be used to analyse our data.

## Choice of Test

When choosing a test, the following factors should be taken into account:

* **Two-sample** or ***k*-sample**

Most tests compare two groups of samples, e.g., the results obtained from comparative tests on two different systems.

Some tests can be used to compare more than two groups of samples, e.g., the results obtained from comparative tests on three or four different systems.

* **Related measures** or **independent measures**

Different tests are used depending upon whether the two (or more) groups from which the data is drawn are related or not.

* **Nominal**, **ordinal**, **interval** or **parametric** data.

These three factors - number of groups, relationship between groups and quality of data - are the principal factors to be taken into account when designing a study and choosing a statistical test.

There are tests available to suit each combination of these factors, for example:

|  |  |  |
| --- | --- | --- |
|   | **Related Samples Tests** | **Independent Samples Tests** |
| **2-Sample Tests** | **t-test (related-samples)***parametric data* | **t-test (independent-samples)** *parametric data* |
| **Wilcoxon***interval-scaled data* |   |
| **Sign test***ordinal-scaled data* | **Mann-Whitney***ordinal-scaled data* |
|   | ***X*2 test***nominal-scaled data* |
| ***k*-Sample Tests** | **Page's *L* test***ordinal-scaled data* | **Jonckheere trend test***ordinal-scaled data* |

Various software packages are available to carry out these and similar tests.

Therefore, the main task facing the designer of a usability test is to choose the right test, in accordance with the data being gathered.