

## **Inferential Statistics**

### **t-test: paired and independent**

#### **Statistical Inference:**

allows the formation of **conclusions** about almost any parameter from a **sample** taken from a larger **population**

(i.e. are conclusions based on the sample valid for the whole population?)

*or*

allows the formation of conclusions about the **difference between populations** with regard to any given parameter.

There are two methods of reaching a statistical inference:

#### **a) Estimation**

In estimation, a **sample** from a population is studied and an inference is made about the **population** based on the sample.

The key to estimation is the probability with which particular values will occur during sampling - this allows the inference about the population to be made.

The values that occur are inevitably based on the sampling distribution of the population. The key to making an accurate inference about a population is **random sampling**, where:

**each possible sample of the same size has the same probability of being selected from the population.**

In real life, it is often difficult to take truly random samples from a population. Shortcuts are frequently taken, e.g. every third item on a list, or simply the first  $n$  results to be obtained. A better method is to use a table of random numbers or the Microsoft Excel **RAND** function.

Estimation is a relatively crude method of making population inferences. A much better method and the one which is normally used is **hypothesis testing**.

## b) Hypothesis Testing

To answer a statistical question, the question is translated into a **hypothesis - a statement that can be subjected to test**. Depending on the result of the test, the hypothesis is **accepted or rejected**.

The hypothesis tested is known as the **null hypothesis** ( $H_0$ ). This must be a true/false statement. For every null hypothesis, there is an **alternative hypothesis** ( $H_A$ ).

Constructing and testing hypotheses is an important skill, but the best way to construct a hypothesis is not necessarily obvious:

- The null hypothesis has priority and is not rejected unless there is strong evidence against it.
- If one of the two hypotheses is 'simpler' it is given priority so that a more 'complicated' theory is not adopted unless there is sufficient evidence against the simpler one (Occam's Razor: "If there are two possible explanations always accept the simplest")
- In general, it is 'simpler' to propose that there is no difference between two sets of results than to say that there is a difference.

The outcome of a hypothesis testing is "reject  $H_0$ " or "do not reject  $H_0$ ". If we conclude "do not reject  $H_0$ ", this does not necessarily mean that the null hypothesis is true, only that there is insufficient evidence against  $H_0$  in favour of  $H_A$ . Rejecting the null hypothesis suggests that the alternative hypothesis may be true.

In order to decide whether to accept or reject the null hypothesis, the level of significance ( $\alpha$ ) of the result is used ( $\alpha = 0.05$  or  $\alpha = 0.01$ ). This allows us to state whether or not there is a **"significant difference"** (technical term!) between populations, i.e. whether any difference between populations is a matter of chance, e.g. due to experimental error, or so small as to be unimportant.

Procedure for hypothesis testing:

1. Define  $H_0$  and  $H_A$ , based on the guidelines given above.
2. Choose a value for  $\alpha$ . This should be done before performing the test, not when looking at the result!
3. Calculate the value of the test statistic.
4. Compare the calculated value with a table of the critical values of the test statistic.

5. If the calculated value of the test statistic is LESS THAN the critical value from the table, accept the null hypothesis ( $H_0$ ). If the absolute (calculated) value of the test statistic is GREATER THAN or EQUAL to the critical value from the table, reject the null hypothesis ( $H_0$ ) and accept the alternative hypothesis ( $H_A$ ).

Note that:

- A significance test can never prove a null hypothesis, only fail to disprove it.
- A very small P value (e.g. 0.001) does not signify a large effect - it signifies that the observed data are highly improbable given the null hypothesis. A very small P value can arise when an effect is tiny but the sample sizes are large. Conversely a larger P value can arise when the effect is large but the sample size is small.

### Standard Scores (z Scores)

**z scores** define the position of a score in relation to the mean using the standard deviation as a unit of measurement.

**Standard scores are therefore useful for comparing datapoints in different distributions.**

$$z = (\text{score} - \text{mean}) / \text{standard deviation}$$

The z-score is the number of standard deviations that the sample mean departs from the population mean.

Since this technique normalizes distributions, z-scores can be used to compare data from different sets, e.g. a student's performance on two different exams (e.g. did Joe Blogg's performance on module 1 and module 2 improve or decline?):

- JoeB scored 71.2% on module 1 (mean = 65.4%, SD = 3.55)  
 $z = (71.2 - 65.4) / 3.55 = 1.63$
- JoeB scored 66.8% on module 2 (mean = 61.1%, SD = 2.54)  
 $z = (66.8 - 61.1) / 2.54 = 2.24$

**Conclusion:** JoeB did better, relatively speaking, on module 2 than on module 1, even though his mark was lower on this course.

Note that the z-score is only informative when it refers to a normal distribution - calculating a z-score from a skewed dataset may not produce a meaningful number.

Comparing z-scores for different distributions is also meaningless unless:

- The datasets being compared are as similar as possible (e.g. response to different doses of a drug under the same physiological conditions).
- The shapes of the distributions being compared are as similar as possible.

**There are two ways to obtain z-scores (standard scores) using MSEXcel:**

1. Perform the calculations using the spreadsheet:  **$z = (\text{score} - \text{mean}) / \text{standard deviation}$**
2. Use the **STANDARDIZE** function which performs this calculation for you, but assumes the mean and standard deviation are known

**Warning:** The MSEXcel **ZTEST** function and [Analysis Toolpak Z-test option](#) do not calculate standard scores! **ZTEST** gives the 2-tailed P-value (probability) for a normal distribution. This function can be used to test if a particular observation is drawn from a certain population. The ToolPak Z-test option is a modified version of the t-test

## Comparing Two Populations

A common experimental design in HCI involves **comparing** experimental results with those obtained under **control conditions**.

To interpret this type of experiment, we must be able to make objective decisions about the nature of any differences between the experimental and control results - is there a **statistically significant difference** or are the results due to experimental error or random chance (sampling error)?

A frequently used test of statistical significance is the:

### Student's t-test (t-test)

The **Student's t-test** (or simply **t-test**) was developed by [William Gosset - "Student"](#) in 1908. Gossett was a chemist at the Guinness brewery in Dublin and developed the t-test to ensure that each batch of Guinness was as similar as possible to every other batch!

The t-test is used to compare two groups and comes in at least 3 flavours:

- **Paired t-test:** used when each data point in one group corresponds to a matching data point in the other group.
- **Unpaired t-test:** used whether or not the groups contain matching datapoints:
  - Two-sample assuming equal variances
  - Two-sample assuming unequal variances

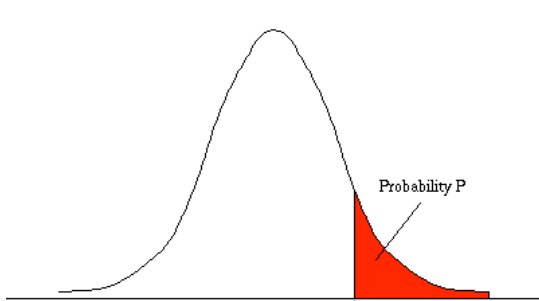
### Assumptions:

The t-test is a **parametric test** that assumes that the data analyzed:

- are [continuous, interval data](#) comprising a whole population or sampled randomly from a population.
- Have a [normal distribution](#)
- If  $n < 30$ , the [variances](#) in the two groups should be similar (t-tests can be used to compare groups with different variance if  $n > 30$ ).
- Sample size should not differ hugely between the groups.

**If you use the t-test under other circumstances, the results will be meaningless!** In other situations, **non-parametric tests** should be used to compare the groups, e.g. the [Wilcoxon signed rank test for paired data](#) and the [Wilcoxon rank sum test for unpaired data](#). To comparing **three or more groups**, use other tests such as ANOVA.

## Paired t-test:



The paired t-test is used to investigate the relationship between two groups where there is a **meaningful one-to-one correspondence between the data points in one group and those in the other**, e.g. a variable measured at the same time points under experimental and control conditions. It is **NOT** sufficient that the two groups simply have the same number of datapoints!

The advantage of the paired t-test is that the formula procedure involved is fairly simple:

1. Start with the hypothesis ( $H_0$ ) that the mean of each group is equal ( $H_A$ : the means are not equal). How do we test this? By considering the variance (standard deviation) of each group.
2. Set a value for  $\alpha$  (significance level, e.g. 0.05).
3. Calculate the difference for each pair (i.e. the variable measured at the same time point under experimental and controlled conditions).
4. Plot a histogram of the differences between data pairs to confirm that they are normally distributed - if not, **STOP!**
5. Calculate the mean of all the differences between pairs ( $d_{av}$ ) and the standard deviation of the differences (SD)
6. The value of  $t$  can then be calculated from the following formula:

$$t = \frac{d_{av}}{SD/\sqrt{N}}$$

where:

- $d_{av}$  is the mean difference, i.e. the sum of the differences of all the datapoints (set 1 point 1 - set 2 point 2, ...) divided by the number of pairs

- **SD** is the standard deviation of the differences between all the pairs
- **N** is the number of pairs.

**N.B. The sign of t (+/-) does not matter, assume that t is positive.**

7. The significance value attached to the resulting value of t can be looked up in a [table of the t distribution](#) (or obtained from appropriate software). To do this, you need to know the "**degrees of freedom**" (**df**) for the test. The degrees of freedom take account of the number of independent observations used in the calculation of the test statistic and are needed to find the true value in a probability table. For a paired t-test

$$df = n-1 \quad (\text{number of pairs} - 1)$$

8. To look up t, you also need to determine whether you are performing a **one-tailed or two-tailed test**. In any statistical test we can never be 100% sure that we have to reject (or accept) the null hypothesis. There is, therefore, the possibility of making an error:

		<b>NULL HYPOTHESIS:</b>	
		True:	False:
<b>Decision:</b>	Reject	Type I error	Correct
	Accept	Correct	Type II error

**Falsely rejecting a true  $H_0$**  is called a **type I error** (finding an innocent person guilty). The probability of committing a type I error is always equal to  $\alpha$ . **Failure to reject a false  $H_0$**  is called a **type II error** (finding a guilty person innocent). The probability of committing a type II error depends on the probability of retaining a false  $H_0$ . The "**power**" of a statistical test refers to the probability of claiming that there is a significant difference when this is true. As scientists are cautious, it is considered "worse" to make a type I error than a type II error - we thus reduce the possibility of making a type I error by having a stringent rejection limit, i.e. 5%. However, as we reduce the possibility of making one type of error, we increase the possibility of making the other type. Whether you use a one- or two-tailed test depends on your testing hypothesis:

**One-tailed test:** Used where there is some basis (e.g. previous experimental observation) to predict the direction of the difference, e.g. expectation of a significant difference between the groups.

**Two-tailed test:** Used where there is no basis to assume that there may be a significant difference between the groups - this is the test most frequently used.

Note that  $H_A$  states 'there is a difference ....', it does not state why there is a difference or whether the difference between the two groups is greater or less than. If  $H_A$  had specified the nature of the difference, this would have been a one-tailed hypothesis. However, since  $H_A$  does not specify the nature of the difference, hence we can either accept a reduction or an increase. This is therefore a two-tailed hypothesis. For a variety of reasons two-tailed hypotheses are safer than one-tailed. Statistical tables are sometimes tabulated only for one-tailed hypotheses. To convert them to two-tailed, double  $\alpha$ .

**See chart of t-values ...**

**If the calculated value of  $t$  is greater than the critical value,  $H_0$  is rejected**, i.e. there is evidence of a statistically significant difference between the groups. **If the calculated value of  $t$  is less than the critical value,  $H_0$  is accepted**, i.e. there is no evidence of a statistically significant difference between the two groups.