

## Concept “Significant difference”

In order to explain significance, we start with a population  $P$ , a socio-demographic variable (abbreviated “socio-demo”), consisting of a number of categories  $K_i | 1 \leq i \leq t$ , and a pre-selected KPI. Using significance analysis we examine, for each socio-demo category, significant differences in average KPI scores between this category and all other categories viewed as one.

The concept “significant difference” should be understood within a statistical framework. As a generic example we consider the population of employees in a particular company, and “profession” with the categories of blue collars, white collars and management as socio demo. In a significance analysis for a specified KPI we examine by how much the average KPI score for blue collars differs from that for white collars and managers, by how much the average score for white collars differs from that of blue collars and managers, and finally by how much the average score for managers differs from that of workers and employees.

At each step in the analysis we therefore determine by how much the average KPI scores for two separate population subgroups (the first formed by the pre-selected socio demographic category and the second by the other categories taken together) differ. These groups are referred to as  $G_1$  and  $G_2$ . Given that in statistics we always work with exact numbers, we use population averages  $\mu_{G_1}$  and  $\mu_{G_2}$  or population medians  $\eta_{G_1}$  and  $\eta_{G_2}$  in determining the average score per group. A population average  $\mu_{G_1}$  or  $\mu_{G_2}$  is in this way the average KPI score for all elements of the population belonging to group  $G_1$  or  $G_2$ . The population median  $\eta_{G_1}$  or  $\eta_{G_2}$  is the value such that for each element randomly drawn out of  $G_1$  or  $G_2$  there is a chance of 0.5 that its KPI score is smaller than  $\eta_{G_1}$  or  $\eta_{G_2}$ . At each step of the analysis we determine for each KPI category by how much  $\mu_{G_1}$  and  $\mu_{G_2}$  or  $\eta_{G_1}$  and  $\eta_{G_2}$  differ from each other. (Whether we use the difference between  $\mu_{G_1}$  and  $\mu_{G_2}$  or between  $\eta_{G_1}$  and  $\eta_{G_2}$  will depend on the size of groups  $G_1$  and  $G_2$ .)

In statistics, values calculated from a population are also referred to as population parameters. In the real world such population parameters are generally not precisely known. In many cases a population is described using conceptual terminology, for example “the Dutch working population” or else contains too many elements, for example “all inhabitants of the Netherlands” for them to be taken into account. To enable them to make estimates for population parameters, statisticians therefore start with a sample. This is a set of objects selected at random from the population such that it is representative for this population. In most cases such a sample that is representative of a population  $P$  is designated with the symbol  $S_p$ .

In the generic example a sample would thus consist of a number of *randomly* selected employees from the company. This sample can then serve to estimate the population parameters. Each such estimate is also referred to as a statistic. Of crucial importance, however, is that every such statistic has an associated distribution function. This is a mathematical function which attaches to each possible value of a statistic the probability that this value is observed.

For the practical calculations in the *ZebraZone Toolset* we therefore start with a sample  $S_p$ , which is representative for P. The population P consists in this case the population for which the significance analysis is being carried out. An estimate of  $\mu_{G_1}$  is then calculated by summing all KPI scores of the sample elements in  $G_1$  and dividing the result by the total number of elements in  $G_1$ . We call this the sample average for the KPI for subgroup  $G_1$  of  $S_p$  and refer to it as  $m_{G_1}$ . Similarly  $m_{G_2}$  represents the sample average for group  $G_2$ .

To obtain an estimate of  $\eta_{G_1}$ , the elements of  $S_p$  belonging to  $G_1$  are first ranked from small to large. If  $n_1$  is an uneven number,  $\eta_{G_1}$  is taken to be the  $(n_1 + 1)/2$ nd element of this series. If  $n_1$  is an even number, we use the average of the  $n_1/2$ nd and  $(n_1 + 2)/2$ nd element in this series as an estimate. This figure is also referred to as the sample median for the elements of  $S_p$  belonging to  $G_1$ , and is given the symbol  $med_{G_1}$ . Similarly  $med_{G_2}$  represents the estimate of  $\eta_{G_2}$ . In our generic example,  $m_{\text{blue collars}}$  represents the sample average of KPI scores for all blue collars and  $med_{\text{blue collars}}$  the sample median for the KPI scores of all blue collars.

Besides functioning as estimates, statistics are in general also used to compare population parameters. The significance calculation in the *ZebraZone Toolset* follows this philosophy completely. The fact that  $\mu_{G_1}$  and  $\mu_{G_2}$  are identical does not necessarily imply that  $m_{G_1}$  and  $m_{G_2}$  will have exactly the same values. Sample averages are no more than approximations for population averages. Of course one would expect that if the population averages are the same, the difference between  $m_{G_1}$  and  $m_{G_2}$  will be relatively small. The same logic applies entirely to the population medians  $\eta_{G_1}$  and  $\eta_{G_2}$  with associated estimates  $med_{G_1}$  and  $med_{G_2}$ . To determine the threshold for which the measured difference between  $m_{G_1}$  and  $m_{G_2}$  or  $med_{G_1}$  and  $med_{G_2}$  indicates a significant difference between  $\mu_{G_1}$  and  $\mu_{G_2}$  or  $\eta_{G_1}$  and  $\eta_{G_2}$ , we rely on a number of statistical procedures. Such procedures are also referred to as statistical tests.

Given that we are using estimates of population parameters, the results from such statistical tests are not 100 percent conclusive. Although  $\mu_{G_1}$  and  $\mu_{G_2}$  are equal there is always a chance that this is not reflected in the measured sample, leading one to conclude that  $\mu_{G_1}$  and  $\mu_{G_2}$  are different. Such an error is generally referred to as a type I error. Apart from such a type I error, it is possible for a difference between  $\mu_{G_1}$  and  $\mu_{G_2}$  not to be reflected in the test. Concluding that the population averages are identical leads in this case to what is generally indicated as a type II error. The probability that a type I error in a test occurs is called the significance level of the test and is denoted by  $\alpha$ . The probability of making a type II error is denoted as  $\beta$ . The number  $(1-\beta)$  thus represents the probability of the difference between  $\mu_{G_1}$  and  $\mu_{G_2}$  or  $\eta_{G_1}$  and  $\eta_{G_2}$  also appear in the test. In statistical theory this is referred to as the power of the test.

In most tests, significance level as opposed to power can be pre-determined. Both values power and significance level are, however, mutually dependant. In general, the higher the significance level (lower  $\alpha$ ) is, the lower the power of the test. This means that a test with a significance level of 0.05 will in general be less powerful than a test with an  $\alpha$  of, for example, 0.03.

In order to guarantee a reliable significance analysis we have to choose an  $\alpha$  which gives us enough power. In practice a significance level of 0.05 is generally sufficient to guarantee a proper power value. The analyses in the *ZebraZone Toolset* follow this rule.

As well as the choice of significance level, the choice of the type of test plays a crucial role in optimizing the test's reliability. Depending on the number of observations we opt for either the parametric *Student t-Test* or its non-parametric counterpart, the Wilcoxon's Rank-Sum Test. Whilst the Student t-Test relies upon some a priori assumptions such as normality, the Wilcoxon's Rank Sum test does depend on additional requirements. Nevertheless this last test generally has less power than the Student t-Test. Only where we have a small number of observations does the Wilcoxon's test's power outperform that of the Student t-Test. We therefore have the following two cases:

❑ **Case one: both groups  $G_1$  and  $G_2$  contain more than 30 elements**

In this case we use the "Central Theorem Limit"<sup>1</sup>. This implies that the probability distributions of both  $m_{G_1}$  and  $m_{G_2}$  follow the bell-shaped Gauss curve<sup>2</sup>. In this way the crucial *normality assumption* for the Student t-Test is met. General statistical theory implies furthermore that the Student t-Test will have more power than the Wilcoxon's Rank sum test. The significance analysis will thus rely on the Student t-Test with a significance level of 0.05 or 5%. In essence this test examines the difference between population averages  $m_{G_1}$  and  $m_{G_2}$  for groups  $G_1$  and  $G_2$ . Using  $m_{G_1}$  and  $m_{G_2}$  and sample data we calculate the value of a t statistic and an interval I. If t is contained in this interval, we conclude that the population averages  $\mu_{G_1}$  and  $\mu_{G_2}$  do not differ significantly. If not, we conclude the contrary.

❑ **Case two: one of the two groups  $G_1$  and/or  $G_2$  contains less than 30 elements**

The "Central Limit Theorem" no longer applies and the Student t-Test is no longer reliable. In this case we opt for an alternative; the non-parametric Wilcoxon's Rank-Sum Test, which does not depend on the normality condition. Basically the Wilcoxon's Rank-Sum test examines the difference between population medians  $\eta_{G_1}$  and  $\eta_{G_2}$ . Based on the sample data a number W and an interval I are calculated. If W is contained in the interval I, we conclude that the medians  $\eta_{G_1}$  and  $\eta_{G_2}$  are identical and if not that  $\eta_{G_1}$  and  $\eta_{G_2}$  differ. As in the case where both groups contain more than 30 elements, the significance level is set at 0.05 or 5%.

---

<sup>1</sup> One of the fundamental theorems of statistics. Globally this theorem states that if a sample S is sufficiently large the probability distribution of the sample average of any measurable quantity on S follows a bell-shaped Gauss curve.

<sup>2</sup> The bell-shaped Gauss curve is the most important probability distribution in statistics. In fact it is the mathematical translation of the phenomenon of average occurrence. For example there are a large number of people of average weight, a smaller number are heavier or lighter, and a few exceptions are much heavier or lighter.