MRDC Software Information Sheet

STATISTICAL OPTIONS IN MRDCL

Introduction

MRDCL contains a comprehensive range of options for producing table-based statistics. Statistical options are activated by using a number of format options that sometimes have different setting or work in conjunction with one or more other settings. Labelling is controlled in most cases by making use of special texts.

This document is split into two sections – a list of the format options that control the statistical outputs (with a brief description of how to use them) and the formulae used by MRDCL when calculating the results.

When calculating most statistics, MRDCL excludes records that have an *undefined* value – sometimes this is referred to as a null value. This allows statistics to be produced for those records where data is known. For example, the average of three records with the values 10, 20 and *undefined* is 15. MRDCL allows variables to be set to the value *u* – meaning undefined.

MRDC Software Information Sheet

Statistical Options

For each statistical option, the format option to be used is shown in brackets.

a) Most common table-based statistics

All of this group of options can be calculated from numeric variables (integers or real numbers) or from response variables (single or multi). For response variables, values must be assigned to each bit in the variable. Bits which do not have a value assigned are not counted in the calculation of the statistics. Label control <v> is used to control the values assigned to each bit of the variable.

- **Average/mean score (AVG)**

  The number of decimal places for the average/mean score can be controlled using format DPA. The decimal places can also be controlled by using label control <d>. The default text assigned to a mean score can be set by using the special label %avg.

- **Base for Statistics (BST)**

  This option allows you to show the base which have been used to calculate statistics (i.e. excluding records with an undefined value). The default text assigned can be set by using %bst.

- **Chi-squared test (CHI)**

  This option allows you to display the chi-squared test. There are two settings for this test. Format CHI1 shows the chi-squared value, the probability and degrees of freedom. Format CHI2 shows the calculated value even if more than 20% of the cells have been an expected value of less than 5. The decimal places can be controlled by using label control <d>. The default text assigned to a chi-squared test can be set by using the special label %chi. The default text for the degrees of freedom can be set by using the special label %dof. The default text for the probability can be set by the using the special label %pro.

- **Error variance (EVR)**

  This option allows you to display the error variance. The number of decimal places can be controlled by using format DPS. The decimal places can also be controlled by using label control <d>. The default text assigned to an error variance can be set by using the special label %evr.

MRDC Software Information Sheet

- **Mean over standard error (MSE)**

  This option allows you to display the mean divided by the standard error. The number of decimal places can be controlled by using format DPS. The decimal places can also be controlled by using label control <d>. The default text assigned to an error variance can be set by using the special label %mse.

- **Probabilites (PRO)**

  This option allows you to display the probabilities. It applies to chi-squared tests (CHI), Mann Whitney Wilcoxon test (MWW), Kolmogorov Smirnoff Tests (KST), Significance tests (SIG), T-Test (TTT) and F-Tests (TTF). The default text assigned to a probability can be set by using the special label %pro.

- **Standard deviation (SDV)**

  This option allows you to display the standard deviation. The number of decimal places can be controlled by using format DPS. The decimal places can also be controlled by using label control <d>. The default text assigned to an error variance can be set by using the special label %sdv.

- **Standard error (SER)**

  This option allows you to display the standard error. The number of decimal places can be controlled by using format DPS. The decimal places can also be controlled by using label control <d>. The default text assigned to an error variance can be set by using the special label %ser.

- **Sum of scores (SUM)**

  This option allows you to display the sum of scores. The decimal places can be controlled by using label control <d>. The default text assigned to an error variance can be set by using the special label %sum.

MRDC Software Information Sheet

b) Significance testing

A separate document is available from MRDC for fuller information about significance testing in MRDCL.

- **Significance tests (SIG)**

  There are three types of significance tests available under this format option. SIG1 applies the standard formula using combined variance. SIG2 uses a non-standard formula using separate variances. Column-based significance tests are carried out within header groups using format SHG (see separate section on Header Groups). Formats SLA/SLB/SLC/SLD allow you to test for up to four levels of significance. The default is two levels. The text to show the levels of significance can be set using special label %fsl. Significance tests are only carried out on bases of 30 or more by default. This can be changed by setting format option MCT.

- **T-Tests (TTT)**

  There are two T-Test settings. TTT1 causes a matrix to be computed and printed for a table as a matrix. A t-test is generated for each pair of columns (except the total column). TTT2 causes t-tests to be generated for each pair of columns with a header group (see separate section on header groups). The text to show the t-test can be set using special label %ttt.

- **F-Test (TTF)**

  This options allows you display a F-Test. The text to show the f-test can be set using special label %ttf.

- **T-Test Variance (TTV)**

  There are two T-Test variance settings. TTV1 causes the combined variance for any t-test produced. TTV2 causes separate variances to be calculated. The text to show the T-Test variance settings can be set using the special label %ttv.

MRDC Software Information Sheet

c) Header groups

Header group levels allow you to choose how significance data is calculated and compared with other data. The format SHG is used to control header groups.

When SHG0 is set, significance is indicated by markers using the markers specified in formats SMA and SMB (default is * and ** for lower and upper limit). When SHG1 is set, calculations are carried out for each column within each of the first levels of heading (a \ or *** heading in MRDCL). When SHG2 is set, calculations are carried out for each column within each of the second levels of heading (above the first level of heading). When SHG3 is set, calculations are carried out within each of the third levels of heading. When SHG-1 is set, each column is tested against every other column.

MRDC Software Information Sheet

d) Statistics using distributions

There are a small group of statistical options which work with distributions. These are as follows:

- **Show distribution (DIS)**

  This format allows you to show or hide the distribution of values when producing a value distribution table. A value distribution takes the following form:

  T#1(f=dis/rna)=$value(100) * $banner,

  This will show the first 100 values found for the variable $value. If more than 100 different values are found, the user will see a warning message and should increase the value. Format RNA is used to rank the values in ascending order. Format RNA must be used with other formats detailed in this section. NDIS would hide the distribution – NDIS is usually used with other statistics options.

- **Show highest value (ILH)**

  This format will display the highest value for the tabulated variable. It must be used with format RNA. The text shown can be controlled by the special label %ilh.

- **Show lowest value (ILL)**

  This format will display the lowest value for the tabulated variable. It must be used with format RNA. The text shown can be controlled by the special label %ill.

- **Median (MED)**

  This format will display the median value for the tabulated variable. It must be used with format RNA. The text shown can be controlled by the special label %med. Where the median value falls between two values, the average of the two values is displayed.

- **Mode (MOD)**

  This format will display the mode for the tabulated variable. It must be used with format RNA. The text shown can be controlled by the special label %mod. Where more than one value shares the same mode, no value is displayed.

MRDC Software Information Sheet

e) Effects of weighting

MRDCL allows two types of weighting – respondent weighting and quantity weighting. Respondent weighting refers to the factor applied to an entire record. Quantity weighting refers to the value applied to a weighting a table – often referred to as a volumetric table – for example, a table shown in terms of units used rather than respondents. Tables can be weighted by both types of weighting.

Statistical calculations generally use the weighed figures. Where appropriate the un-weighted or effective number of respondents is used. This is shown as N_e in the formulae. If the recommended format ESS is used then the effective sample size used in calculations is:

$$N_e = \frac{(\sum W)2}{\sum W^2}$$

Where **W** is the weight applied to each respondent. If ESS has not been set, then N_e is the un-weighted number of respondents.

Format UNE allows the user to show the Effective Sample Size instead of the Unweighted Base on a weighted table.

Users are advised to seek the advice of an expert statistician when working with weighted data.

MRDC Software Information Sheet

f) Other statistical tests

- **Kolmogorov Smirnov Tests (KST)**

  This format produces a Kolmogorov Smirnov test on a table.

- **Mann-Whitney-Wilcoxon Test (MWW)**

  This format produces a Mann-Whitney-Wilcoxon test on a table.

- **Sum of Squares (SSQ)**

  This format produces the sum of squares for a value.

MRDC Software Information Sheet

Formulae

- **Mean score or average**

  Where:

  x is each value or score value.

  n is the (weighted) number of respondents with that value.

  N is the (weighted) base (sum of n).

$$\overline{x} = \frac{\sum nx}{N}$$

- **Standard error**

$$se = \frac{s}{\sqrt{N_e}}$$

- **Standard deviation**

  This is calculated as:

$$s = \sqrt{\frac{\sum nx^2 - \frac{(\sum nx)^2}{N}}{N-1}}$$

  This is normally written as:

$$s = \sqrt{\frac{(x - \bar{x})^2}{N-1}}$$

MRDC Software Information Sheet

- **Variance**

  This is calculated as:

$$s^2 = \frac{\sum nx^2 - \frac{(\sum nx)^2}{N}}{N-1}$$

  This is normally written as:

$$s^2 = \frac{(x - \bar{x})^2}{N-1}$$

- **Error Variance:**

$$sv = \frac{s^2}{N_e}$$

- **Mean over standard error**

$$mse = \frac{\bar{x}}{se}$$

- **Means comparison t-test**

  When comparing two column means with TTV1:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{(N_1 - 1){s^2}_1 + (N_2 - 1){s^2}_2}{N_1 + N_2 - 2}\left(\frac{1}{N_{1e}} + \frac{1}{N_{2e}}\right)}}$$

MRDC Software Information Sheet

When comparing two column means with TTV2:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s^2{}_1}{N_{1e}} + \frac{s^2{}_2}{N_{2e}}}}$$

- **Proportions comparison Z test**

When comparing two column percentages in the same row. Where: p is each proportion p_t is the combined proportion from both columns added together N_t is the combined base from both columns added together When comparing proportions (percentages) with SIG1:

$$z = \frac{p_1 - p_2}{\sqrt{P_t(1 - P_t)\left(\frac{1}{N_{1e}} + \frac{1}{N_{2e}}\right)}}$$

With SIG2:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1 - p_1)}{N_{1e}} + \frac{p_2(1 - p_2)}{N_{2e}}}}$$