

# Description of QuaDS results files

A. Bouanich, A. El Ghaziri, P. Santagostini, A. Pernet, C. Landès and J. Bourbeillon

## Abstract

This document explains all the results files generated by QuaDS pipeline. For each results file, an explanation of the statistics used and a description of the output are provided. The explanations are presented both in general terms and through the specific example used for illustration in this package. We note that the choice of the appropriate test (parametric or non-parametric) will be automatically determined for the user through the pipeline.

## 1 Introduction

The QuaDS package offers a visualization technique that describes a factor of interest using both quantitative and qualitative variables. While the package provides an interactive visualization, it also generates a set of nine results files. To better understand these files, a comprehensive explanation is provided, detailing the results in each file along with the associated statistics. These files explore the relationship between the factor of interest and the variables in the study. By examining both quantitative and qualitative data, this study aims to identify the factors contributing to the explanation of the factor of interest. The interpretation of these results is also demonstrated using an illustrative example from the package. Understanding these results offers insight into the underlying concepts behind the QuaDS package's interactive visualization. We will begin by recalling the data used to illustrate the package, followed by a section for each file in the results directory, titled with the corresponding file name.

## 2 Illustration data

The factor of interest is the **chromosome**, with two categories: *Chr01* and *Chr02*. Four variables are used to understand the relationship with this factor: two quantitative variables, **nb TE** and **Gene expression**, and two qualitative variables, **type dupli** and **type TE**. In our case study, the significance threshold is set to 0.05. We note that the user can set the significance threshold.

## 3 File homoscedasticity.csv

We performed Bartlett's test to verify the equality of variances between the factor levels. The results table includes, for each variable, the Bartlett statistic and the associated  $p$ -value. If the  $p$ -value is strictly less than the significance threshold, it indicates that the assumption of homoscedasticity is not met. In our example, both quantitative variables, **nb TE** and **Gene expression**, satisfied homoscedasticity assumption.

Note that, if homoscedasticity is not verified, the v-test used later to describe the factor will not be calculated.

## 4 File normality.csv

We performed the Shapiro-Wilk test to assess whether each quantitative variable meets the normality assumption within the compared factor levels. If the  $p$ -value is strictly less than the significance threshold, it indicates that the normality assumption is not verified. In our example, **Gene expression** satisfied normality for both chromosome, while **nb TE** did not.

## 5 File anova.csv

A one-way analysis of variance (ANOVA) is performed for each quantitative variable in the study. The output is a table that lists the quantitative variables in rows and includes three columns: **eta-squared**, **p-value**, and **interpretation**.

Since our case study includes two quantitative variables, two one-way ANOVAs are performed. In each ANOVA, the following hypotheses are tested:

$H_0$ : The means of the two chromosomes are equal.

$H_1$ : The means of the two chromosomes are different.

In a more general scenario, the alternative hypothesis ( $H_1$ ) would be that at least one mean is different from the others.

In the **p-value** and **interpretation** columns, the  $p$ -value for the factor of interest is reported along with the corresponding interpretation of the hypothesis test:

- If the  $p$ -value is strictly lower than the significance threshold (*e.g.* 0.05 in the example), we reject  $H_0$ , indicating a significant difference between the means (in this case, between the two chromosomes). The interpretation is labeled as “Significant” in the table.
- If the  $p$ -value is greater than or equal to the significance threshold, we do not reject  $H_0$ , suggesting that the data do not provide sufficient evidence to reject  $H_0$ . The interpretation is labeled as “Not significant” in the table.

The **eta-squared** column represents the effect size [Coh88]. This metric indicates the proportion of variance in the quantitative variable that can be attributed to the factor. An effect size of zero indicates that the quantitative variable shows no variability regarding the factor, consistent with the null hypothesis.

In our example, **Gene expression** satisfied both normality and homoscedasticity assumptions. Therefore, a one-way ANOVA was performed, and the results indicate that this variable is significant, meaning that at least one factor level mean is significantly different from the others. The  $\eta^2$  was equal to 0.99, indicating a large effect size, suggesting that the chromosome factor strongly influences gene expression level.

## 6 File kruskal\_wallis.csv

When the conditions for a one-way ANOVA are not met, the non-parametric Kruskal-Wallis test is applied. The resulting output includes the observed test statistic, the associated  $p$ -value and the interpretation, which follows the same logic as ANOVA. If the  $p$ -value is strictly less than the significance threshold, it indicates a statistically significant difference for at least one level of the factor.

In our example, the variable **nb TE** was tested using the Kruskal-Wallis (since normality was not verified). The results indicate a significant difference across the distribution of the groups.

## 7 File quantitative\_results.csv

This file explains the levels of the factor of interest as a function of the quantitative variables. We denote the quantitative variable by  $X$  and a level of the factor of interest by  $q$ . We test whether the variable  $X$  characterizes the level  $q$  through a hypothesis test:

$H_0$ : The variable  $X$  does not characterize the level  $q$ . (or  $H_0$ : The mean of  $X$  for the  $q$  level is equal to the overall mean).

$H_1$ : The variable  $X$  characterizes the level  $q$ .

The first column represents the factor levels, and the second column represents the quantitative variable used. The seven following columns are:

1. **Mean in category**: the mean of the quantitative variable  $X$  for the  $q$  level values:  $\bar{X}_q$ .
2. **Overall mean**: the global mean of the quantitative variable  $X$ :  $\bar{X}$ .
3. **Standard deviation in category**: the standard deviation of the  $q$  level values:  $s_q$

4. **Overall standard deviation:** the standard deviation of the quantitative variable  $X$ :  $s$
5. **v-test:** the statistic used to test the hypothesis, calculated using the following equation:

$$\text{v-test} = \frac{\bar{X}_q - \bar{X}}{\sqrt{\frac{s^2}{I_q} \cdot \frac{I - I_q}{I - 1}}} \quad (1)$$

where:

- $I_q$ : number of individuals with the level  $q$ .
- $I$ : total number of individuals.

The v-test represents a standardized difference between the mean of the category and the overall mean ([HLP17], page 155).

6. **p-value:** The  $p$ -value associated with the hypothesis test. If the  $p$ -value is strictly lower than the significance threshold (*e.g.*, 0.05),  $H_0$  is rejected, indicating that the variable explains the category. If the  $p$ -value is greater than or equal to the significance threshold, it is not reported and is replaced with “not significant”.
7. **Interpretation:** if the  $p$ -value is greater than or equal to the significance threshold, “not significant” is reported also as the interpretation. When the  $p$ -value is strictly lower than the threshold:
  - if the v-test  $> 0$ , the mean of the category is higher than the overall mean (reported as “above average” in the table),
  - if the v-test  $< 0$ , the mean of the category is lower than the overall mean (reported as “below average” in the table).

In this context, this means that the number of transposable elements (**nb TE**) is significantly higher on *chromosome 2* than the overall average, while the variable **Gene expression** is significantly higher on *chromosome 1*.

## 8 File Chi2.csv

For each qualitative variable, a Chi-square test of independence is conducted with the factor of interest if the expected frequencies in contingency table cells are sufficiently large ( $\geq 5$ ) [Coc52]. The resulting output is a table listing the qualitative variables in rows with three columns: **Chi2 Statistic**, **p-value** and **interpretation**. For each qualitative variable, the following hypothesis is tested:

- $H_0$ : The variable is independent of the factor of interest.
- $H_1$ : The variable is not independent of the factor of interest.

The value of the statistic is reported in the column **Chi2 Statistic**, and the associated  $p$ -value is reported in the column **p-value**, with the interpretation in the column **interpretation**. If the  $p$ -value is strictly less than the threshold (*e.g.*, 0.05 in this example), we reject  $H_0$ . This indicates that the variable is dependent on the factor (*e.g.*, Chromosome in this case), and the interpretation value is “Significant”.

If the  $p$ -value is greater than or equal to the significance threshold (*e.g.*, 0.05), suggesting that the data do not provide sufficient evidence to reject the hypothesis that variable is independent of the factor. The interpretation here is “Not significant”.

In our example, the variable **type TE** is significant, which means that this variable is dependent on the **Chromosome**.

## 9 File `fisher_exact.csv`

If the expected frequencies in contingency table cells between the qualitative variable and the factor of interest are very low ( $< 5$ ), Fisher's exact test is used. The interpretation is similar to that of the Chi-square test of independence. The output of this test includes the  $p$ -value and the interpretation of each qualitative variable as either ("Significant" or "Not significant").

In our example, the variable **type dupli** is significant, which means that this variable is dependent on the **Chromosome**.

## 10 File `qualitative_results.csv`

This file provides a description of the relation between each level of the factor of interest and the qualitative variables in the study. We start by introducing some notations:

- $n_{kj}$ : number of individuals with level  $j$  for a qualitative variable and level  $k$  for the factor of interest.
- $n_k$ : number of individuals with the level  $k$  of the factor of interest.
- $n_j$ : number of individuals with the level  $j$  of the qualitative variable.
- $n$ : total number of individuals.

In the `catdes` function, "Cla" represents the class to describe (the factor of interest), while "Mod" refers to the qualitative variables modalities. To ensure consistency with the function's output, we maintain this notation: "Cla" for the factor and "Mod" for qualitative variable levels. Each row represents a combination of a level of the factor of interest and a level of a qualitative variable. Six columns describe the relationship between them.

1. The first statistic in the table corresponds to **Cla/Mod**:

$$Cla/Mod = \frac{n_{kj}}{n_j} \times 100. \quad (2)$$

This is the percentage of individuals with level  $j$  of the qualitative variable that also have level  $k$  of the factor of interest.

2. The second statistic in the table corresponds to **Mod/Cla**:

$$Mod/Cla = \frac{n_{kj}}{n_k} \times 100. \quad (3)$$

This is the percentage of individuals with level  $k$  of the factor of interest that also have level  $j$  of the qualitative variable.

3. The third statistic in the table corresponds to **Global**:

$$Global = \frac{n_j}{n} \times 100. \quad (4)$$

This is the percentage of individuals that have level  $j$  in total.

4. The fourth and fifth columns (**p-value** and **v-test**) indicate the  $p$ -value and the associated statistical test based on a hypergeometric distribution ([HLP17, LPM06]). The formula of the hypergeometric law is:

$$p_k(j) = \sum_{x=n_{kj}}^{x=n_j} P(N=x) \text{ with } P(N=x) = \frac{C_{n_k}^x \times C_{n-n_k}^{n_j-x}}{C_n^{n_j}} \quad (5)$$

The  $p$ -value is the calculated as follows:

$$\text{if } \frac{n_{kj}}{n_k} > \frac{n_j}{n} : p\text{-value} = 2 \times P(N > n_{kj}) \quad (6)$$

$$\text{if } \frac{n_{kj}}{n_k} \leq \frac{n_j}{n} : p\text{-value} = 2 \times P(N \leq n_{kj}) \quad (7)$$

and the **v-test** is then calculated as follows:

$$\text{if } \frac{n_{kj}}{n_k} > \frac{n_j}{n} : v\text{-test} = 1 \times \frac{P(X < p\text{-value})}{2} \quad (8)$$

$$\text{if } \frac{n_{kj}}{n_k} \leq \frac{n_j}{n} : v\text{-test} = -1 \times \frac{P(X \leq p\text{-value})}{2} \quad (9)$$

5. The last column corresponds to the results **interpretation**. A positive value of the v-test indicates that level  $j$  of the qualitative variable is considered over-represented for level  $k$  of the factor of interest (and under-represented for a negative value of the v-test).

In our example, for the variable **type dupli** in *Ch01* for example, the levels *unique homolog* and *singleton* are over-represented, while the *multiple homolog* level is under-represented. The reverse is true for *Ch02*.

## 11 File weight.csv

This file provides a count of over- and under-represented levels for each qualitative variable identified as dependent on the factor of interest (using the Chi-square test in section 8). It allows a fair comparison of the contribution of the variables in explaining the factor, independently of the number of levels.

Each row is associated with a significant variable (column **variable**), and ten characterisations are provided:

1. **number mod over**: the number of over-represented levels of the variable in the entire analysis.
2. **number mod under**: the number of under-represented levels of the variable in the entire analysis.
3. **number mod over & under**: the number of over- and under-represented levels of the variable in the entire analysis.
4. **sum of all mod of all groups**: the number of level combinations between the variable and the factor of interest.
5. **ratio over/mod**: the percentage of over-represented levels of the variable in the analysis,  $\frac{\text{number mod over}}{\text{sum of all mod of all groups}} \times 100$ .
6. **contribution over/mod**: the rank of the variable compared to the other qualitative variables for the ratio over/mod values.
7. **ratio under/mod**: the percentage of under-represented levels of the variable in the analysis,  $\frac{\text{number mod under}}{\text{sum of all mod of all groups}} \times 100$ .
8. **contribution under/mod**: the rank of the variable compared to the other qualitative variables for the ratio under/mod values.
9. **ratio over & under/mod**: the percentage of over- and under-represented levels of the variable in the analysis,  $\frac{\text{number mod over and under}}{\text{sum of all mod of all groups}} \times 100$ .
10. **contribution over & under/mod**: the rank of the qualitative variable compared to the other qualitative variables for the ratio over&under/mod values.

In our example, the **type dupli** variable has more levels contributing to the characterisation of the factor in both over-represented and under-represented categories (with contributions in over, under and over&under/mod equal to 1). This indicates that the **type dupli** variable provides greater differentiation between the two chromosomes than the **type TE** variable.

## References

- [Coc52] William G. Cochran. The  $\chi^2$  test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345, 1952.
- [Coh88] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*, chapter The Analysis of Variance, pages 273–406. Academic Press, New York, 2nd edition, 1988.
- [HLP17] François Husson, Sébastien Lê, and Jérôme Pagès. *Exploratory Multivariate Analysis by Example Using R*. CRC Press, Boca Raton, FL, 2nd edition, 2017.
- [LPM06] Ludovic Lebart, Marie Piron, and Alain Morineau. *Statistique exploratoire multidimensionnelle: Visualisation et inférence en fouille de données*. Sciences Sup. Dunod, Paris, France, 4ème édition edition, 2006. Cours et exercices corrigés, Masters, Écoles d’ingénieurs.