**QCHECK**

# STATA PACKAGE FOR QUALITY CONTROL OF DATASETS

## Introduction

**qcheck** (shorthand for ‘quality check’) is a technical package for quality control of different data types. -qcheck- performs the following types of complementary types of analysis:

### - Static Analysis.

The static analysis allows checking for within the survey consistency. The static analysis of qcheck verifies the internal consistency of each variable and its relationship with other variables in the same dataset. It verifies that a variable is consistent with its definition (e.g., age is always a positive number) and checks the consistency with the other variables (e.g., 5 years old with graduate-level education). The user can create new tests, validations, and crosstabs to automate the assessment of variables across years, countries, and regions, among others.

### - BASIC Analysis.

The basic analysis is useful to compare within and across surveys, for example, evolution of a categorical variable over time or across countries or across regions within a country.

The basic analysis of qcheck store all the descriptive statistics provided by -sum var name, d- and the number of missing observations, number of non-missing observations, number of zeros, mean, standard deviation, maximum, minimum, skewness, kurtosis, and 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th ‘s percentiles. This output is store in a tabulated format.

### - CATEGORICAL Analysis.

Run tabulates for each category variable(s) specified by the user. The output is store in a tabulated long format, that in a visualization allow to compare the evolution of the categories over time, or within regions, or by population groups. These comparisons allow to identify anomalies or mistakes in the harmonization of the data.

## package components

### - Static Analysis.

For the **static** analysis, the qcheck package requires as input an Excel file. From this input file, the Stata command qcheck retrieves all the information needed to perform the assessment. The user must create and complete the input Excel with logic statements about the variables in the data. The ado-file is complemented with an example of an Excel file with a basic set of tests to check the quality of an example database. The user is free to modify such a file either by editing the tests or adding tests to it and running the qcheck analysis again to observe how the results change in response to the changes in the input file.

Static test workflow

Before performing the static analysis, qcheck revise the existence of the variables, and creates flags if the variable is not in the data, or if the variables was created all empty, either all zeros or missing. The diagram below indicates how the inspection is conducted.

### - BASIC and CATEGORIC Analysis.

Basic test workflow

Categorical test workflow

Unlike static analysis, the basic and categorical analysis does not require additional input other than the dataset to be analyzed. Once the Stata command qcheck has performed the assessment, results can be exported into long-formatted Excel files that can be read by Tableau/Power BI/R/Pivot tables Excel. We provide some examples, but the user may create their reports in the program of their preference or adapt the provided examples.

## Setup and Installation

The convention of the name of the Excel file is “qcheck\_NNN.xlsx” where NNN refers to a set of checks to be applied to a particular collection.

\*\*A word of caution here\*\*: it is expected that the suffix NNN of the “qcheck\_NNN.xlsx” file refers to the name of the collection to be tested. For example, the user may have the file “qcheck\_ABC.xlsx” to contain the check of the collection ABC. 1 step: Files location

## USE – STATIC ANALYSIS

### 1 step: Modify Excel file as needed (spreadsheet “test”)

The first step is to create the input Excel with the internal consistency logic statements. Before completing your input Excel, look at the example file “qcheck\_NNN.xlsx.” First, in the spreadsheet “TEST,” you can add, modify, or edit the set of quality checks of your database. Each row corresponds to a different check or logical statement, and each column corresponds to a particular check feature.

The first column contains the name of the variable to be checked. It may be the case that one variable has to be checked in relation to another variable so that both variables are checked jointly. It does not matter which variable name goes in the name as long as only one name is specified.

The second column, “Warning,” allows the user to specify the level of urgency. The purpose of this column is merely cosmetic. It allows the user to organize or filter the results easier in the Tableau dashboard or their own analyses.

The third and fourth columns are the checking code, but each has a particular function. The fourth column (iff) contains the logical statements that check the consistency of the variable. For instance, if you wanted to test that the variable corresponding to the person’s age does not have negative values, positive values above 100, or missing values, you may type something like this: **age < 0 | age > 100**. As you see, the logical test flags those observations that meet the criterion as inconsistent.[[1]](#footnote-2)

The third column (temporalvars), is for code lines that must be executed before the logical statement in column “iff.” Sometimes, it is needed to create a temporal variable with certain characteristics in order to check some inconsistencies. For instance, you may need to test whether the combination of household and person id is unique along the dataset. In order to do so, you can do the following:

**cap destring pid, replace**

**duplicates report hid pid**

**local n = r(unique\_value)**

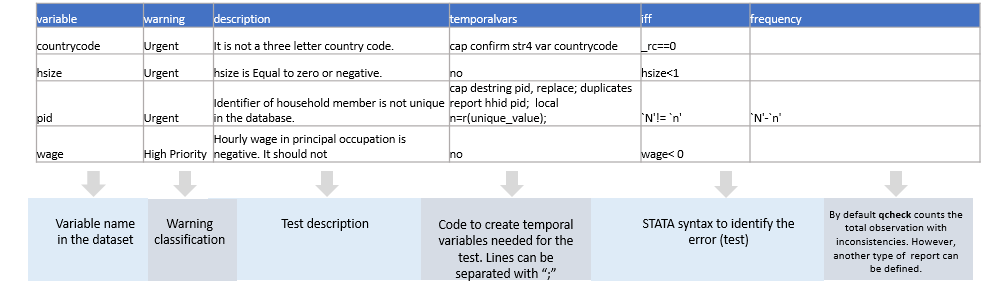
**count**

**count if r(N) != `n' // logical statement**

The first four lines of the code above create a temporal macro that counts the number of observations in the dataset that have a unique value for the combination hid and pid. If the dataset was constructed correctly, the number in local n should be the same as the number of observations in the dataset. Therefore, the last line of code is the logical test that verifies the aforementioned statement. Several things should be kept in mind.

* Given that there is only one cell for each check in column “temporalvars”, each line of code must be separated from the subsequent line with a semicolon (;) instead of a break of line.
* In the example above, the logical statement that goes in the corresponding cell of column “iif” is **r(N) != `n'**, rather than **count if r(N) != `n'**. Given that by design, all the consistency checks count the number of observations with problems, it is inefficient to ask the user to type “count if” for each cell. Instead, it is only necessary to type the logical statement of the code line.

See a small example below:



### 2 STEP: Modify Excel file as needed (spreadsheet “Variables”)

The dynamic assessment of qcheck performs different analyses depending on the variable type: welfare, categorical, and basic. Variables classified as ‘welfare’ are assumed to be continuous, and estimations of poverty and inequality are only performed with these variables. Categorical variables are numeric, but their values refer to a classification or characteristic of the observation rather than an ordinal correlation between its members. For instance, the variable of Labor Force Participation contains three numeric values: 1, 2, and 3. However, 1 means ‘employed,’ 2 means ‘unemployed,’ and 3 means ‘out of labor force.’ Finally, the basic classification of variables refers to variables that are either non-categorical or welfare aggregate.

## USE – DYNAMIC ANALYSIS

1. Notice that the test identifies those observations with problems, and not those that are fine. That is, the test should not be **inrange(age, 0, 100)**. [↑](#footnote-ref-2)