# Paris Winter School: Exercises statcheck Day 2

Michèle B. Nuijten | m.b.nuijten@uvt.nl | http://mbnuijten.com

## Download statcheck Data

In recent research, we looked at the prevalence of reporting inconsistencies in 8 flagship Psychology journals, in articles from 1985 to 2013 (Nuijten, et al., 2016). The full (anonymized) data, and data organized per article, per journal, or per journal and year are available at the Open Science Framework (OSF): https://osf.io/e9qbp/.At the OSF page, go to the component "Data files" for all - you guessed it - data files.

At the OSF page you can also find the full analyses scripts that were used in our paper. For these scripts, go to the component "Analyses". It could be useful to take a look at these scripts for inspiration/instruction on how to perform a particular type of analysis. Finally, this page contains information about additional analyses that were not important enough to include in the paper, but important enough to be documenten somewhere.

### Exercises

1. Download the full statcheck data set: `150211FullFile_AllStatcheckData_Automatic1Tail.csv`.
2. Load the data in RStudio as follows:

```
# set your working directory to the folder where you saved the data
# for instance:
setwd("C:/MyComputer/Downloads/")

# load the data
data <-
read.csv2("150211FullFile_AllStatcheckData_Automatic1Tail.csv",header=TRUE)

# check if the data is loaded correctly by inspecting the dimensions of the
data frame:
dim(data)

## [1] 258105     19
```

If the data set you loaded doesn't show these dimensions, or when it gives an error when you try to load it, try `read.csv()` instead of `read.csv2()`:

```
data <-
read.csv("150211FullFile_AllStatcheckData_Automatic1Tail.csv",header=TRUE)
```

## Getting to Know the Data

To get a feel for the data, it is always smart to visualize it and look at descriptives. As a first step, look at the summary of the data:

```
summary(data)
```

With this function you get information about the distribution of all variables in the data frame. You can also immediately see if there are any missing values ("NA") In the example below is the output for the first four columns.

```
##        X              Source      Statistic          df1
##  Min.   :     1   Min.   :    1   Chi2: 24298   Min.   :    0.00
##  1st Qu.: 64527   1st Qu.: 5140   F   :146864   1st Qu.:    1.00
##  Median :129053   Median : 9465   r   : 10201   Median :    1.00
##  Mean   :129053   Mean   : 8644   t   : 66070   Mean   :    6.09
##  3rd Qu.:193579   3rd Qu.:11951   Z   : 10670   3rd Qu.:    2.00
##  Max.   :258105   Max.   :16695   NA's:     2   Max.   :22865.00
##                                                 NA's   :86941
```

Besides using the `summary()` function, it can also be useful to look at the data in more detail.

### Exercises

3. How many p-values were extracted?
4. What does the distribution of reported p-values look like? And that of computed p-values? (Tip: use `hist()`)
5. In how many different articles did statcheck find results? (Tip: use `unique()`)
6. Use the `table()` function to find out how many of each type of statistic were reported.
7. What percentage of all p-values was an inconsistency? And a gross inconsistency?

## Explorative Analyses

A data set as large as the statcheck data contains a lot of information, much of which has not been discussed in our paper. One of the main advantages of sharing data is that other people can answer different questions with the same data. The point of this exercise is to find the answer to previously unanswered (or unemphasized) questions.

Note: think about the level of the data to which your question pertains. If you want to know something about the p-values themselves (e.g., "how many results are an error?", or "is there a correlation between the degrees of freedom and the probability that a result is inconsisten?"), you can use the statcheck data that you already loaded.

However, you can also ask questions at a different level. For instance, if you want to know whether the prevalence of articles with at least one inconsistency has changed over the years, you need data at the article level. To make answering these kinds of questions at different levels easier, we also posted our data organized at the article level, year & journal level, and journal level.

You can find these data files at OSF again: https://osf.io/e9qbp/. These data files are saved as .txt files. To load these data, use `read.table()`, instead of `read.csv2()`.

If you want to organize the data at a different level than p-values yourself, or organize it at a level that isn't available on OSF yet (e.g., test statistic), you could use the function `by()`

(see the help files with `?by`). You can also look at the scripts we used to organize the data at different levels, you can find these under the "Analyses" component at the OSF page.

## Exercises

Choose one or more of the research questions below and try to answer them with the statcheck data. You can also come up with your own research question.

### Predicting (Gross) Inconsistencies

- Is there a relation between (gross) inconsistencies and...
  - ...year? In other words: what happens with the prevalence of inconsistencies over the years?
  - ...journal? Do some journals contain more inconsistencies than others?
  - ...test statistic? Which type of test statistic most often contains inconsistencies? And gross inconsistencies?
  - ...degrees of freedom? The second degree of freedom says something about the sample size. Test whether tests pertaining to smaller samples have a higher probability of being inconsistent than when the sample size is larger.

### Analyzing P-Values

- Can you see a bump just below p = .05 if you create a histogram of the computed p-values? And of the reported p-values? (NOTE: many p-values are reported in the format "p < .05". Take this into account when creating/interpreting the histogram.)
- In how many cases was a result inconsistent AND was the reported p-value lower than the computed p-value?
- It is likely that many inconsistencies arise because of random sloppiness. However, there are cases in which this might be less likely, for instance in cases where a computed p-value is .06, and the reported p-value < .05. Investigate how many of these cases there are in the data.
- How many inconsistencies are there in which the difference between the computed an reported p-value was larger than .01? And larger than .1? (Think about it: does this analysis work for p-values reported as p < ... as well?)

For more ideas about what you can exploratively analyze with these data, also check out the blog (incl. R code) on this topic by Daniel Lakens: http://daniellakens.blogspot.nl/2015/10/checking-your-stats-and-some-errors-we.html.

## References

Nuijten, M.B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, *48 (4)*, 1205-1226. DOI: 10.3758/s13428-015-0664-2