General descriptives

## General

As a response to our recruitment call, 1141 researchers signed up to participate in our study. Out of these volunteers, 459 signed up to analyse at least one datasets and submitted their work by the deadline or an extended deadline.

Throughout the project, 509 re-analyses have been submitted. This number is higher than the number of co-analysts as some co-analysts volunteered to analyse more than one dataset.

Out of the submitted analyses, \_\_\_ were withdrawn, and \_\_\_ were omitted from the summary analysis for the following reasons: 4 analyses failed the peer evaluation.

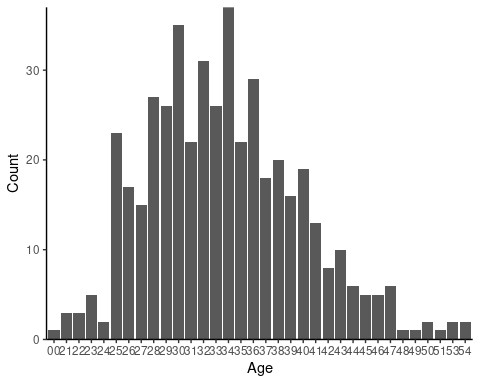
As a result, we ended up with 505 re-analyses, submitted by 458 co-analysts.

## Task 1 Survey results

Out of all the co-analysts who submitted their work by the deadline, there were 23 professors, 41 associate professors, 108 post-doctoral researchers, 122 doctoral students, 59 from other academic/research positions, and from other positions.

The gender distribution of the co-analysts is as follows: 117 female, 334 male, 1 other, and 7 didn’t want to respond to this question.

Warning in geom\_histogram(stat = "count"): Ignoring unknown parameters:  
`binwidth`, `bins`, and `pad`



The age distribution of the co-analysts is depicted in Fig X. 378 Young adults (-39 years); 81 middle-ages adults (40-59 years); and no old adults (60- years).

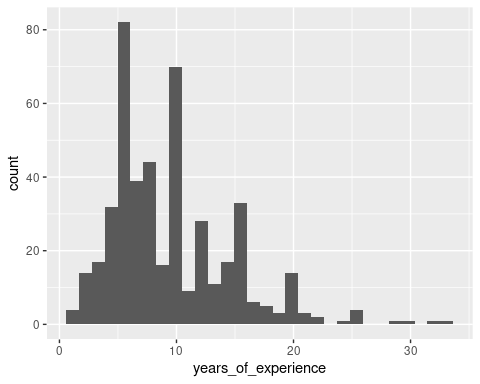
Regarding the highest level of education, co-analysts had Bachelor’s degree or equivalent, Master’s degree or equivalent, had Doctoral degree or equivalent; and reported other degree.

The country of residence of the co-analysts is shown on the map on Fig X. Regarding continents, \_\_ co-analysts were from Africa, \_\_\_ from Asia, \_\_\_ from Australia, \_\_\_ from Europe, \_\_\_ from North America, \_\_\_ from South America.

We asked the co-analysts which discipline is the closest to their research area. The following table (Table X) summarises the distribution of their disciplinary orientation. Co-analysts from \_\_\_ and \_\_\_ disciplines participated in the highest ratio in this study.

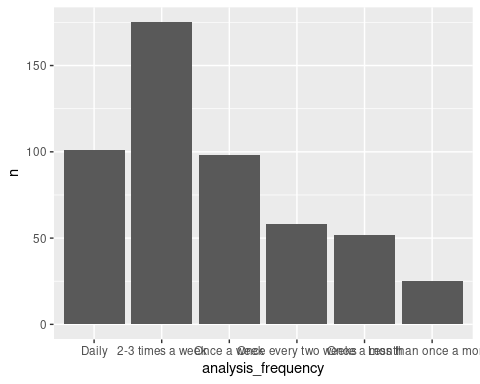
# A tibble: 62 × 4  
# Groups: analyst\_id [31]  
 analyst\_id years\_of\_experience n n\_per\_analyst  
 <chr> <dbl> <int> <int>  
 1 018OL 9 1 2  
 2 018OL 10 1 2  
 3 0XXW0 4 1 2  
 4 0XXW0 5 1 2  
 5 1HC64 5 1 2  
 6 1HC64 8 1 2  
 7 3AQDX 3 1 2  
 8 3AQDX 5 1 2  
 9 5X3KG 4 1 2  
10 5X3KG 10 1 2  
# ℹ 52 more rows

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

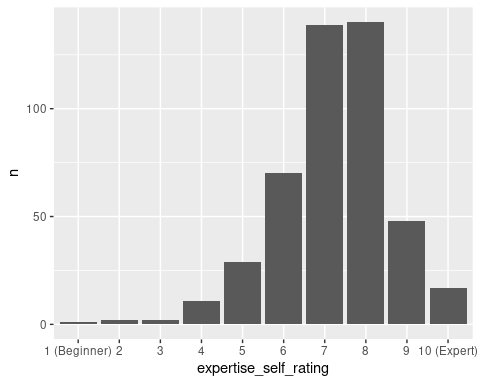


The distribution of the years of experience with data analysis is depicted on Fig X. The median time of experience with data analysis was X years among our co-analysts.

# A tibble: 53 × 4  
# Groups: analyst\_id [26]  
 analyst\_id analysis\_frequency n n\_per\_analyst  
 <chr> <chr> <int> <int>  
 1 7ZZIN Daily 1 3  
 2 7ZZIN Once a week 1 3  
 3 7ZZIN Once every two weeks 1 3  
 4 018OL 2-3 times a week 1 2  
 5 018OL Daily 1 2  
 6 0GFYM 2-3 times a week 1 2  
 7 0GFYM Once a week 1 2  
 8 0XXW0 2-3 times a week 1 2  
 9 0XXW0 Once a week 1 2  
10 1HC64 2-3 times a week 1 2  
# ℹ 43 more rows



We asked our co-analysts of how regularly they perform data analysis. Fig X. shows that the most frequent category was 2-3 times a week.



We also asked them how they rated their level of expertise in data analysis between Beginner (1) and Expert (10). The distribution on Fig X shows that the most prevalent answer was 8 .

All together, 8.25 % (42 out of 509) co-analysts indicated that they were familiar with the paper that the provided dataset belongs to before beginning their work on the project.

No co-analysts reported that they communicated about the details of their analysis with other co-analysts working with the same dataset.

We asked the co-analysts what programming language/software/tool they used in their data analysis. The following figure indicates that \_\_ (X%), \_\_ (Y%), and \_\_ (Z%) were the most popular responses. FIGURE

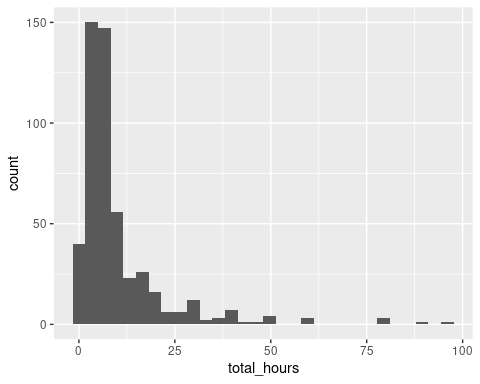
## Task 2 Survey results

In Task 2, when we asked the co-analysts to present one main statistical result, 97.25% of them (495 our of 509) based their conclusion on p-value and 2.75% of them (14 our of 509) used Bayes Factor.

A difference in Task 2 compared to Task 1 was that the co-analysts received some constraints for their analysis in order to make them linkable to a single result in the original study. 47.35 % (241 out of 509) the co-analysts reported that they had to make additional calculations in the second task. In 52.65% (268 out of 509) the co-analysts indicated that despite the limitations in the instructions, they received the same result in Task 2 and Task 1.

In Task 2, 12.77% of the results (65 our of 509) were in the opposite direction as claimed by the original study, disregarding whether the effect was conclusive/significant.

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The co-analysts were asked to estimate the time they spent to perform Task 1 and Task 2 together. The median value of their response is 6 hours (Fig X).

## Peer evaluation

### Peer evaluators

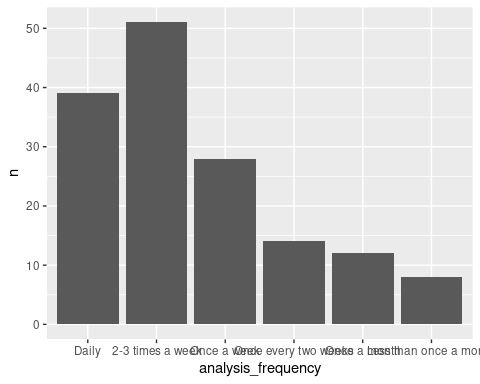
Basic demographic info.

Experience with conducting statistical analysis:

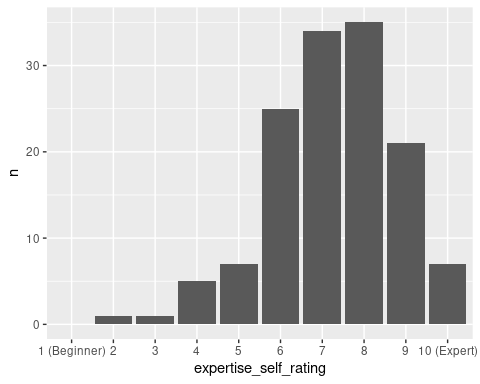
# A tibble: 20 × 4  
# Groups: evaluator\_id [10]  
 evaluator\_id years\_of\_experience n n\_per\_analyst  
 <chr> <dbl> <int> <int>  
 1 018OL 9 1 2  
 2 018OL 10 1 2  
 3 5X3KG 4 1 2  
 4 5X3KG 10 1 2  
 5 AHW5W 20 2 2  
 6 AHW5W 22 1 2  
 7 CGTZS 19 1 2  
 8 CGTZS 23 1 2  
 9 EZI7J 3 1 2  
10 EZI7J 5 1 2  
11 N8P2J 27 1 2  
12 N8P2J 30 1 2  
13 NSDML 5 1 2  
14 NSDML 7 1 2  
15 WMHM7 11 1 2  
16 WMHM7 13 1 2  
17 XF5GJ 14 1 2  
18 XF5GJ 15 1 2  
19 Z5YKQ 6 1 2  
20 Z5YKQ 8 1 2

Frequency of data analysis:

# A tibble: 20 × 4  
# Groups: evaluator\_id [10]  
 evaluator\_id analysis\_frequency n n\_per\_analyst  
 <chr> <chr> <int> <int>  
 1 018OL 2-3 times a week 1 2  
 2 018OL Daily 1 2  
 3 AHW5W 2-3 times a week 1 2  
 4 AHW5W Daily 2 2  
 5 CC3UD 2-3 times a week 1 2  
 6 CC3UD Daily 1 2  
 7 DAILV Less than once a month 1 2  
 8 DAILV Once a month 1 2  
 9 EZI7J Daily 1 2  
10 EZI7J Once every two weeks 1 2  
11 N8P2J 2-3 times a week 1 2  
12 N8P2J Daily 1 2  
13 NSDML 2-3 times a week 1 2  
14 NSDML Daily 1 2  
15 WMHM7 2-3 times a week 1 2  
16 WMHM7 Once a week 1 2  
17 XF5GJ 2-3 times a week 1 2  
18 XF5GJ Once a week 1 2  
19 Z5YKQ 2-3 times a week 1 2  
20 Z5YKQ Once a week 1 2



Self-reported expertise in data-analysis:



### Peer evaluations

Nr. of peer evaluations:

Descriptives of peer evaluations.

For those analyses where there were more than one peer evaluations, for 49.51% (101 out of 204) of the analysis the evaluators disagreed on the analytical pipeline for task 1, and 58.82% (120 out of 204) for task 2.

% (x out of y) of acceptable analysis pipelines (Task 1) - the outcome of the procedure, + % of peer evaluations we need to adjust,

% (x out of y) of acceptable analysis pipelines (Task 2) - the outcome of the procedure, + % of peer evaluations we need to adjust,

For those analyses where there were more than one peer evaluator, 10.78% (22 out of 204) of evaluators disagreed on the adequacy of the conclusions.

% (x out of y) of adequate conclusions (Task 1) - outcome of the procedure, + % of peer evaluations we need to adjust,

% (x out of y) of cases where the correction of the self-categorization of the conclusion was necessary

Nr. of analytical reproducibility checks:

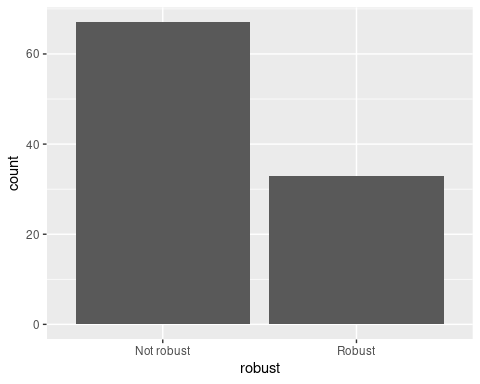
74.3% (185 out of 249 of the analytical reproducibility checks were successful

## How robust are conclusions published in social sciences to analytical choices?

Do different analysts arrive at the same conclusions as the analysts of the original study?

### Task 1 Survey results

In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive to a single conclusion. Out of 100 re-analysed studies, the conclusions of 33 (33%) remained robust to independent re-analysis, so that all assigned co-analysts arrived at the same conclusion as reported in the article of the original study.



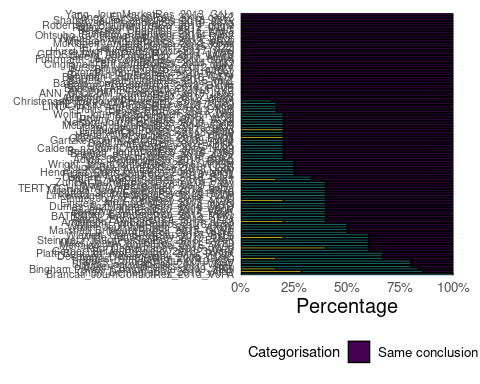


Fig X shows the histotrophic display of the different and identical conclusions resulting from the re-analysis of each of the studies.

Across all the studies, 73.28 % (373 out of 509) of the re-analyses arrived to the same conclusion; 24.36% (124 out of 509) to no effects, and 2.36% (12 out of 509) to opposite effect compared to the original conclusion.

### Robustness

#### Robustness by field

We were interested to see whether these results show a different pattern when inspecting them in different fields. The following figure shows that for the major fields (>=10 studies) the pattern were…, FIGURE

#### Robustness by study type (observational, experimental)

Here, we were interested to see whether these results show a different pattern when separating them by study type. The following figure shows that for that…,

FIGURE

#### Robustness by expertise (self-reported expertise in data analysis)

Here, we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts. The following figure shows that for that…,

FIGURE

#### Robustness by prior familiarity with the dataset

Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. The following figure shows that for that…,

FIGURE

#### Robustness by the suitability of their self-judged analysis

FIGURE

#### Robustness by the sample size

Here, we were interested to see whether these results show a different pattern when considering sample size. The following figure shows that for that…,

FIGURE