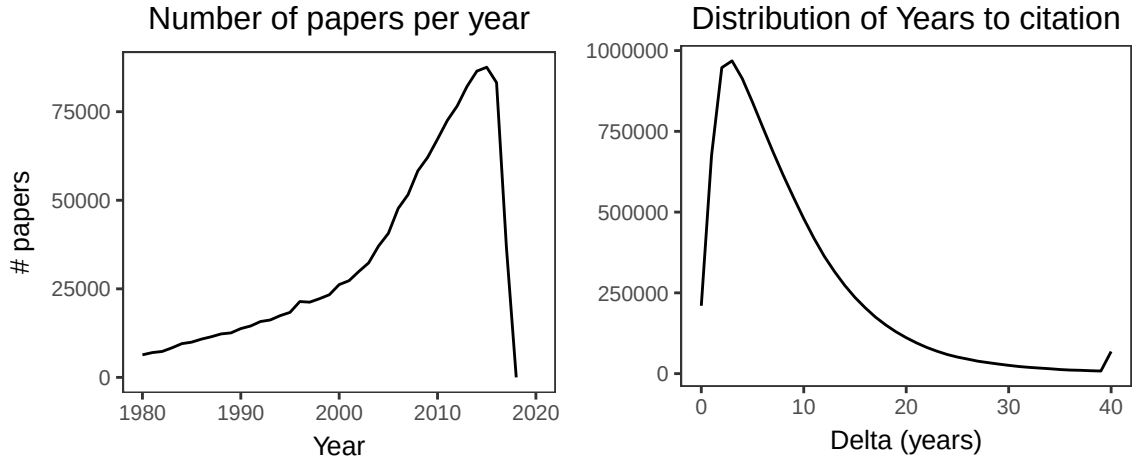
MAG - 'Psychiatry'

Jan Overgoor

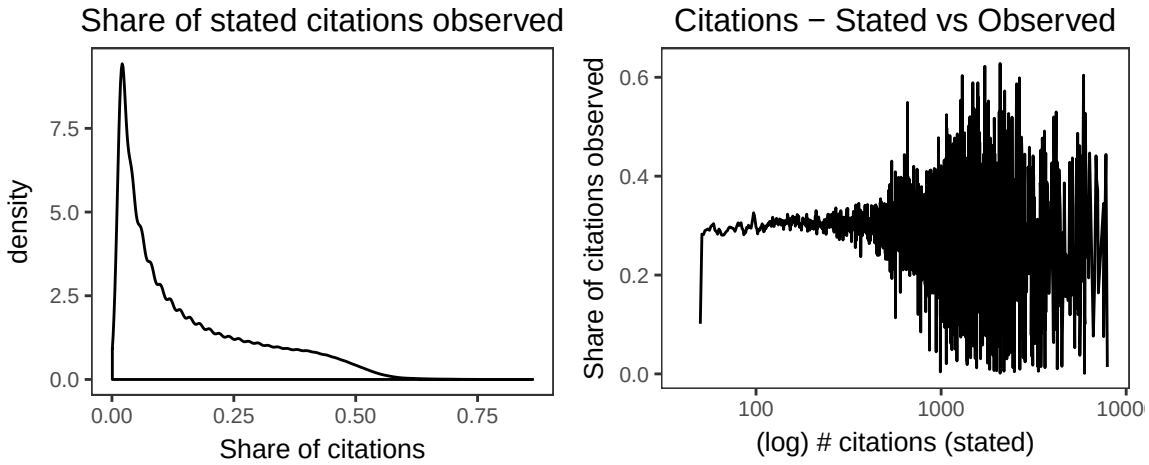
```
to build from command-line, run with:
R -e "rmarkdown::render('data_mag_psy.Rmd', output_file='data_mag_psy.pdf')"
```

- total number of papers (=nodes): 1388536
- total number of references: 25558327
- total number of references to known nodes (=edges): 10805458
- share of references to known nodes: 0.4227764



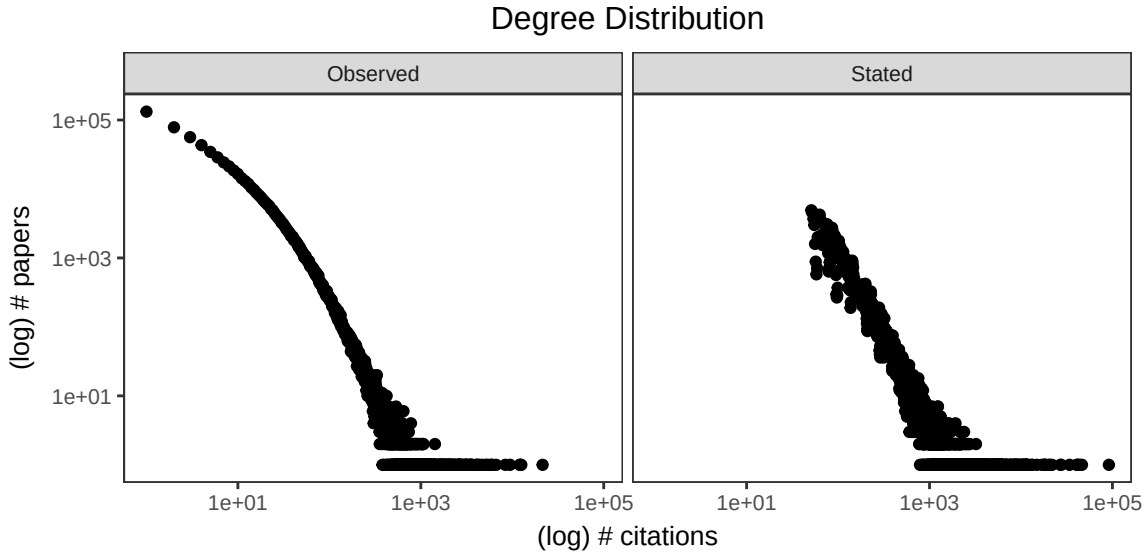
- Left: number of papers per year, linear increase since 2000, drop for recent years
- Right: distribution of years between publication and getting cited. Most citations happen within 2-3 years of publication.

Next, we compare the stated number of citations to the amount we can actually find in the data.



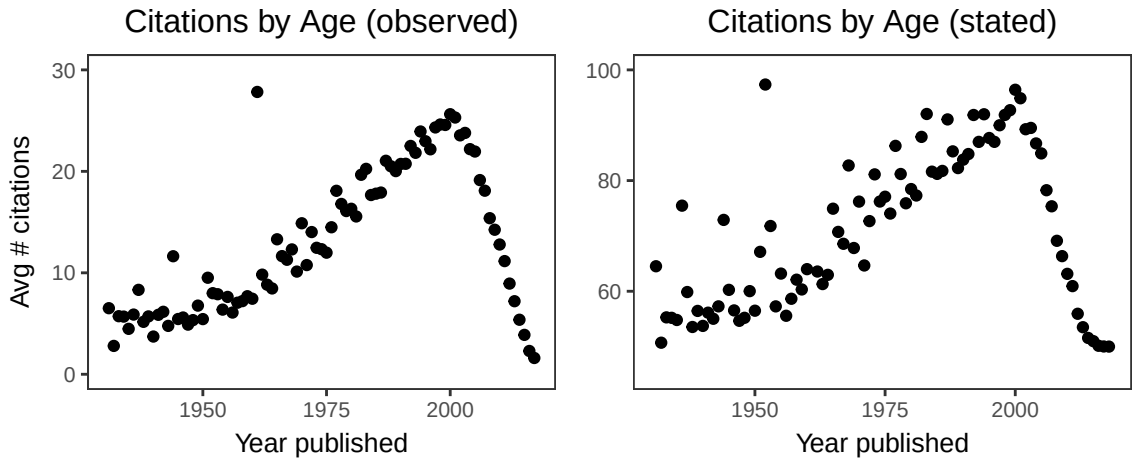
- Left: distribution of "share of stated observations observed", for mostly papers this is <25%. Since I look at the graph filtered by field of study, citing papers might not be included. The whole graph is hard to work with, so this is what we got.
- Right: per "stated number of citations", what is the average "share of citations observed"? Very stable by x, with a much higher variance for the highly cited papers (as there are fewer of them).

Are the degree distributions similar for the stated and observed citation counts?



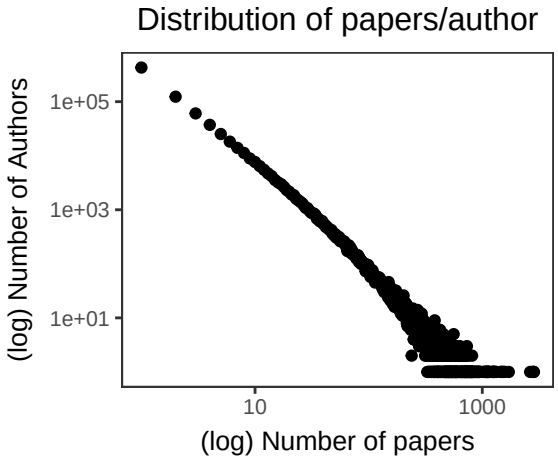
Yes, and this shows the censoring at 50 very clearly.

TODO: fit α?



- Left: average number of citations by year of publishing (as observed). linear increase until 2000 (newer papers more cited), but then drops off
- Right: same, but as stated. The trend is the same, but the numbers are inflated by about 50.

Here is the distribution of papers/author:



Very heavy-tailed as well.

What are the top keywords?

keywords	n
human	177931
homme	176260
hombre	171859
humans	90864
depression	86816
safety	80952
suicide prevention	79559
human factors	79256
falls	79070
ergonomics	78939

TODO: here, language might play a role..

Model

Data construction process:

- sample 1000 citations from after 2011
- for each actual citation, sample 24 non-cited papers (from before publication date)
- for each of the (paper,option) pairs, compute features (n citations, years since, has same author)

```
##
## ===============================================================================
##                                         y
##                      (1)         (2)         (3)         (4)         (5)
## -------------------------------------------------------------------------------
## log Citations     0.868***    1.257***    0.922***    1.288***    1.274***
##                   (0.023)     (0.033)     (0.025)     (0.036)     (0.036)
##
## log Age                      -1.322***               -1.261***   -1.231***
##                               (0.055)                 (0.059)     (0.060)
##
## Has same author                           7.431***    6.851***    6.868***
##                                           (0.598)     (0.628)     (0.645)
##
## # same keywords                                                   0.157***
##                                                                  (0.015)
##
## -------------------------------------------------------------------------------
## Observations       1,000       1,000       1,000       1,000       1,000
## Log Likelihood  -2,292.067 -1,918.289 -1,958.921 -1,665.612 -1,608.969
## ===============================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```