A survey of the quality and transparency of statistical power analyses conducted using GPower

Robert T. Thibault & Emmanuel Zavalis
(Contributor roles are detailed before the references)

Address correspondence to robert.thibault@stanford.edu

2022-04-25

```
library(readr)
library(tidyverse) # for cleaner code
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.8
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr) # for knitting and kable function
library(kableExtra) # for kable table styling
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
# this chunk estimates the total number of published articles that use GPower for doing (1) a power ana

#fileEncoding variable so that the column headers don't have junk text
ex=read.csv('./queryHits.csv', header=T, fileEncoding="UTF-8-BOM")

# number of articles coded
denom <- nrow(df_gpower_orig)
# number of articles meeting our inclusion criteria
include <- nrow(df_gpower)
# number of articles used for a priori sample size calculations
apriori <- sum(grepl("A priori", df_gpower$typeof_powercalc) & grepl("Before", df_gpower$when_powercalc
# percentage of PMC articles that were hits
proportion <- ex$pmcHits/ex$pmcTotal

#initialize dataframe for point estimates and confidence intervals
cis <- data.frame(matrix(nrow=6, ncol=3))
```

```r
# create function to calculate the point estimates and confidence intervals
ciCalc <- function(raw, proportion, num, denom){
  cis <- c(raw * proportion * num / denom,
           raw * proportion * prop.test(x=num, n=denom, conf.level=.95)$conf.int[1],
           raw * proportion * prop.test(x=num, n=denom, conf.level=.95)$conf.int[2]
  )
  return(cis)
}

# run the function to calculate point estimates and confidence intervals
cis <- rbind(ciCalc(ex$pmcTotal, proportion, include, denom),
             ciCalc(ex$pmcTotal, proportion, apriori, denom),
             ciCalc(ex$pubmedTotal, proportion, include, denom),
             ciCalc(ex$pubmedTotal, proportion, apriori, denom),
             ciCalc(ex$dimensionsTotal, proportion, include, denom),
             ciCalc(ex$dimensionsTotal, proportion, apriori, denom)
)

# assign row names and column names
rownames(cis) <- c("pmcInclude",
                   "pmcApriori",
                   "pubmedInclude",
                   "pubmedApriori",
                   "dimensionsInclude",
                   "dimensionsApriori"
                   )
colnames(cis) <- c("point",
                   "ci.lb",
                   "ci.ub"
                   )


cis <- cis %>% round(0)

#supplementary table 1
st1 <- cis
rownames(st1) <- c("PubMed Central included",
                   "PubMed Central a priori",
                   "PubMed included",
                   "PubMed, a priori",
                   "Dimensions included",
                   "Dimensions a priori"
                   )
colnames(st1) <- c("Point estimate",
                   "95% confidence Interval, lower bound",
                   "95% confidence Interval, upper bound"
                   )
```

```r
#initialize dataframe for type and timing of power analyses
type <- data.frame(matrix(nrow=3, ncol=2))
rownames(type) <- c("Sample size",
                    "Power",
                    "Effect size"
```

Supplementary Table 1. Estimates of the number of published articles that use GPower.

	Point estimate	95% confidence Interval, lower bound	95% confidence Interval, upper bound
PubMed Central included	18279	12315	20597
PubMed Central a priori	11249	5784	16393
PubMed included	41227	27775	46455
PubMed, a priori	25370	13045	36973
Dimensions included	179393	120860	202140
Dimensions a priori	110395	56764	160882

* The total number of article in each database since 2017 is: PubMed Central 3,166,809; PubMed 7,142,566; dimensions.ai 31,079,708.

Supplementary Table 2. Type and timing of power calculations

	Before study	After study
Sample size	8	0
Power	0	1
Effect size	0	1

```
)
colnames(type) <- c("Before study",
                    "After study"
)

type[1,1] <- apriori
type[1,2] <- sum(grepl("A priori", df_gpower$typeof_powercalc) & grepl("After", df_gpower$when_powercalc
type[2,1] <- sum(grepl("Post hoc", df_gpower$typeof_powercalc) & grepl("Before", df_gpower$when_powercal
type[2,2] <- sum(grepl("Post hoc", df_gpower$typeof_powercalc) & grepl("After", df_gpower$when_powercalc
type[3,1] <- sum(grepl("effect size", df_gpower$typeof_powercalc) & grepl("Before", df_gpower$when_power
type[3,2] <- sum(grepl("effect size", df_gpower$typeof_powercalc) & grepl("After", df_gpower$when_power
```

```
## Warning in add_indent(x, c(2, 3, 5, 6, 7, 9, 10, 11, 13, 14, 15, 19, 20, :
## Please specify format in kable. kableExtra can customize either HTML or LaTeX
## outputs. See https://haozhu233.github.io/kableExtra/ for details.
```

Table 2: Power calculation prerequisites reporting

	% Reporting
Alpha	92.3
0.05	91.7
Other level of significance	8.3
Power	76.9
80%	30
95%	60
Other level of power	70
Effect size	100
d	30.8
f	15.4
Other	53.8
Statistical test	
ANOVA	0
T-test	0

	% Reporting
Other test	100
Sample size	76.9
Median (IQR)	53 (83.5)
Reproducible	15.4
Reproducible with assumptions	15.4
Reproducible w/o assumptions	0
Irreproducible	84.6

[Robby: Introduction]

Including the Daniel Lakens paper, and the commentary that you authored.

Methods

Search method and selection criteria

This is a pilot study that was conducted prior to the extraction of the actual sample that we will study in our paper. Using the search strategy 'g*power' in PubMed Central we searched for the papers that used the software GPower for their power analysis to examine their use of this tool and the handling of power calculations in general. The search was conducted on April 10th 2022 and retrieved 23927 publications.

From these a random sample of 15 articles were examined in detail using random seed 1453.

The search for the sample to find the articles and was used in PubMed Central was:

'PMC7273017 OR PMC5857919 OR PMC8005969 OR PMC5495109 OR PMC5629614 OR PMC7349576 OR PMC5609352 OR PMC6062074 OR PMC6542583 OR PMC7190572 OR PMC8244525 OR PMC8815471 OR PMC8952887 OR PMC7146211 OR PMC7547415'

We assessed the papers for eligibility first with the inclusion criteria being that the paper uses GPower to perform a power calculation. We also extracted meta-data that included what journal it was from, what impact factor the journal has, as well as the publication year. Finally we looked into whether the paper was a clinical trial defined as a study that includes human subjects and studies an intervention.

Data extraction

Then we proceeded to extract the full description of the power calculation. Continuing we looked if the calculation used an a priori (solving for sample size), post hoc(solving for power) or sensitivity (solving for effect size) type of power calculation. We then extracted the power of the study, the alpha level, the sample size as well as the effect size.

Furthermore, we investigated how many power calculations matched the main test, whether the BUG thingy of ANOVA sample sizes (where there is a difference in approaches used for the sample size calculation where results vary vastly). We also tabulated how many of the power analyses corrected for multiple comparisons.

Finally, we attempted to reproduce the power analysis. To assess the accuracy of our extraction we'll test for interrater agreement in x of our studies.

Precision analysis

For the actual study that will be performed after this pilot we will look at a sample of 41 from our PubMed Central search (the precision analysis using α=alpha expected proportion of p=0.5 and a two-sided margin of error of 20%; for precision of **WHAT?**).

Reproducibility of power analyses

	% Reporting
Alpha	92.3
0.05	91.7
Other level of significance	8.3
Power	76.9
80%	30
95%	60
Other level of power	70
Effect size	100
d	30.8
f	15.4
Other	53.8
Statistical test	
ANOVA	0
T-test	0
Other test	100
Sample size	76.9
Median (IQR)	53 (83.5)
Reproducible	15.4
Reproducible with assumptions	15.4
Reproducible w/o assumptions	0
Irreproducible	84.6

* Note that we have used placeholder variables for the rows statistical tests, these will be replaced with actual data

Table 1: ANOVA within-between/ ANOVA within main effect "bug"

Is this power calculation for within-between ANOVA or within main effect
No
Unsure what test the power calculation was for
Yes, but I cannot reasonably assume which option they used.
Yes, but the researchers use the default option without accounting for it (e.g., powering for a "medium" effect size by enter

Table 2: Multiple comparisons correction

Multiple comparisons correction
No, and the article contains multiple analyses with no clear indication of a sole primary analysis for which this power calcu
No, and there is no reason to account for multiple comparisons (e.g., there is only one analysis, or this analysis is clearly d
Unsure
Yes, reasonably so (i.e., the accounting for multiple comparisons matches the analyses conducted)

Results

Search results

Of the 15 papers, 13 were included in this study. 23.1 % of the papers were clinical trials.

Meta-data of the articles

The impact factors of the journals the studied papers was on average 4.3 (95% CI 2.4, 6.2). The papers were published from 2017 to 2022.

Extracted data

Of the sample of 15 papers using GPower for its power calculation 7 were missing one of the following parameters in reporting their power calculation:

I've called it ANOVA bug thingy throughout because I haven't been able to write something half intelligent about the difference in adjustments for the effect sizes so we can borrow from the introduction if you have something that would work?

- Papers missing α: 1
- Papers missing 1-β: 3
- Papers missing the type of effect size: 1

Of the ones that didn't report what type of effect size i.e. Cohen's d or f, reported a number but without specifying the type of effect size. In 7 papers the power calculation didn't match the statistical test for the primary outcome 7 could not be assessed due to missing information in the statistical plan. 2 studies had a completely reproducible power analysis.

The amount of papers that get the effect size that is input into the power calculation from prior research are 2, 0 cases we were based own own pilot studies 11 cases didn't base it on research but merely on either Cohen's rules of thumb or they did not contain a justification at all.

The discovered bug in ANOVA power calculations ....

Only 20% (1/5) of the power calculations that should adjust for multiple comparisons actually did use an α correction method (see Table 2).

Most power calculations (The ones found in 9 articles) where a priori power calculations (See Table 3)

If you extrapolate this across the 2.0736733×10^4 then about x y z will not be reproducible to.....

Table 3: Types of power calculation

The type of power calculation	Occurrences
A priori (i.e., solves for sample size)	9
Post hoc (i.e., solves for power)	2
Sensitivity (i.e.,solves for effect size)	1
Unsure	1