Simulation-based power analysis

Nora Wickelmaier

Last modified: 2025-09-11

Help, my effect size is too large!

Help, my effect size is too large!

- Probably never heard anyone complain about this
- But it is a huge problem for the scientific integrity of our field
- Reported effect sizes in the literature are way too large

Help, my effect size is too large!

- Probably never heard anyone complain about this
- But it is a huge problem for the scientific integrity of our field
- Reported effect sizes in the literature are way too large

- We will look at an example from the literature:
  Decision biases from two-hand tapping

Refresher: Framing

- Tversky and Kahneman (1981)

  "Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed" (p. 453)

  If Program A is adopted **200** people will be **saved** [109]

  If Program B is adopted there is 1/3 probability that **600** people will be **saved**, and 2/3 probability that **no people** will be **saved** [43]

  If Program C is adopted **400** people will **die** [34]

  If Program D is adopted there is 1/3 probability that **nobody** will **die**, and 2/3 probability that **600** people will **die** [121]

- Odds ratio (OR) = 9.0

Decision biases from two-hand tapping

- McElroy and Seta (2004), $n = 48$

  "a behavioral task of finger tapping was used to induce asymmetrical activation of the respective hemispheres ... Framing effects were found when the right hemisphere was selectively activated whereas they were not observed when the left hemisphere was selectively activated" (p. 572)

	right-hand tapping		left-hand tapping		ratio of odds
	safe	risky	safe	risky	ratios (ROR)
gain	8	4	12	1	
loss	7	4	3	9	
OR		1.1		36	31.5

- Our replication (see Gelman, 2020), $n = 332$

gain	52	31	56	27	
loss	26	57	30	53	
OR		3.7		3.7	1.0

Large effects from subtle manipulations?

There is a simple explanation for the seemingly large effects published all over the psychological literature

- that works without any real large effects
- but assumes that they are statistical artifacts based on a combination of

$$\frac{\text{low power} \\ \wedge \\ \text{selection by significance}}{\Rightarrow \text{inflated effect}}$$

(type M error; Gelman & Carlin, 2014)

Classical inference in a nutshell

- Deciding between two hypotheses about parameter of data-generating model (Neyman & Pearson, 1933)
- Null hypothesis (specific), alternative hypothesis (logical opposite)
  – Example: Binomial model, H_0: $\pi = 0.5$, H_1: $\pi \neq 0.5$
- Possible decision errors

	Decision for H_0	Decision for H_1
H_0 true	correct	type I error, α
H_1 true	type II error, β	correct

Conventions

- $\alpha = 0.05$
- $\beta < 0.2$

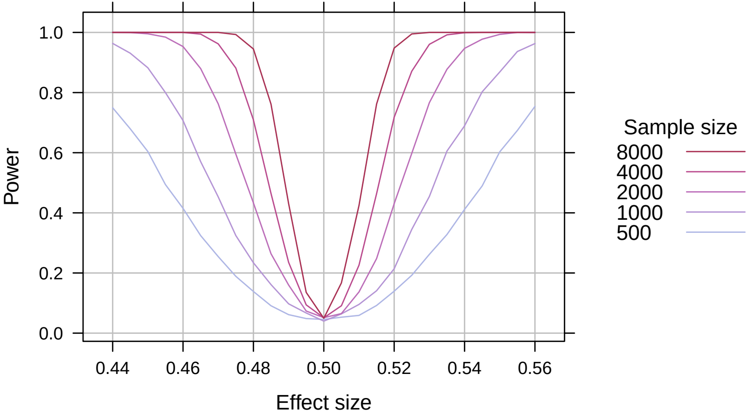
- Decision based on data (p-value)
  – If $p < \alpha$, choose H_1; else retain H_0
- Power $= 1 - \beta$
  – Probability of test to detect an effect of a given size

Power function

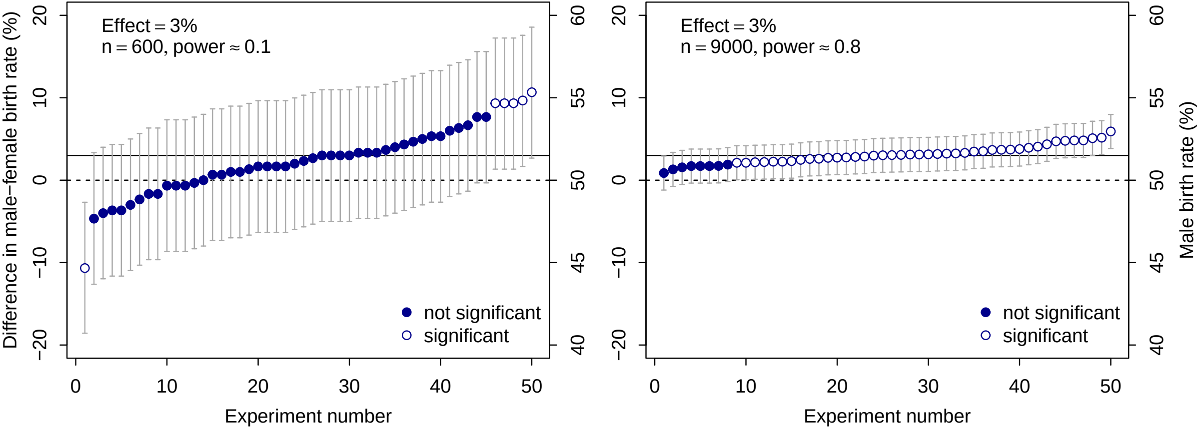
Power of a test depends on

- effect size
  (deviation from H_0)
- sample size n
- α

With effect size, power, and α fixed, we can calculate n



High power is a necessary condition for valid inference



"If power is low . . . every possible outcome under repeated sampling will be misleading: there will be a high proportion of inconclusive null results, and any significant effects will be due to mis-estimations of the true effect" (Vasishth & Gelman, 2021, p. 1317)

Power analysis by simulation

Why simulation?

- Simulation is at the heart of statistical inference
- Inference: Compare the data with the output of a statistical model
- If data look different from model output, reject model (or its assumptions)
- Simulation forces us to **specify a data model** and to attach meaning to its components
- Model should not be totally unrealistic for those aspects of the world we want to learn about

Power simulation

1. Specify the model including the effect of interest
2. Generate observations from the model
3. Test H_0
4. Repeat

Power corresponds to how often a test obtains a significant result for a given α level (e.g. 5%), given that an effect is truly present

$$\widehat{Power}_s = \frac{\sum_{i=1}^{niter} I(p\text{-}value_{is} < \alpha | H_1)}{niter}$$

Specify the model including the effect of interest

(1) Choose statistical model according to its assumptions

- Binomial test → binomial distribution → `rbinom()`
- t test → normal distribution → `rnorm()`
- ...

(2) Fix unknown quantities

- Standard deviations, correlations, ...

(3) Specify the effect of interest

- *Not* the effect one expects or hopes to find (size of effect is unknown!)
- *Never* an effect size taken from another study (significance filter!)
- *But* the biologically or clinically or psychologically "relevant effect one would regret missing" (Harrell, 2020)

Estimating power with simulation

Pseudo Code

```
Set sample size to n
replicate
{
  Draw sample from model with minimum relevant effect
  Test null hypothesis
}
Determine proportion of significant results
```

Sample size calculation

- Sample size n, minimal relevant effect and α must be predetermined
- Adjust n until desired power (0.8 or 0.95) is reached
- To be on the safe side, assume higher variation, less (or more) correlation, and smaller interesting effects (what results can we expect, if . . .)

Example: How to fix the two-hand tapping study?

1. Specify model

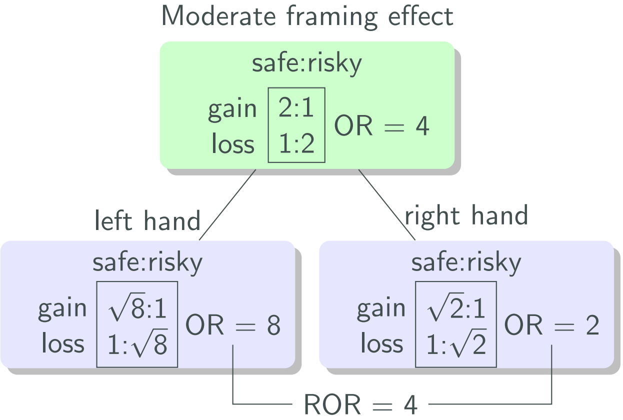
- Logit model with interaction

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \text{left hand} + \beta_2 \cdot \text{gain} + \beta_3 \cdot (\text{left hand} \times \text{gain})$$

- Suggest a minimum relevant effect
  - We can look at the original framing effect study and its many replications
  - Former study by McElroy and Seta (2003) found $ROR = 3.4$ for similar manipulation
  - Other studies investigating influencing factors (with RORs $\approx$ 2–3, e. g., foreign language effect, Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014; Wickelmaier, 2015)
- Underlying distribution: $X \sim Binom(n, p)$

Example: How to fix the two-hand tapping study?

1. Specify model



Translating into parameters

- $\exp(\beta_0) = \frac{1}{\sqrt{2}}$
  odds in reference categories: right and loss
- $\exp(\beta_1) = \frac{1}{2}$
  OR of switching to left hand
- $\exp(\beta_2) = 2$
  OR of switching to gain frame
- $\exp(\beta_3) = 4$
  ROR

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \text{left hand} + \beta_2 \cdot \text{gain} + \beta_3 \cdot (\text{left hand} \times \text{gain})$$

Example: How to fix the two-hand tapping study?

2. Generate observations

- Calculate logits for the model

```
n <- 100
dat <- read.table(header = TRUE, text = "
  hand frame
     r  gain
     r  loss
     l  gain
     l  loss")                                    # ref. cat.
dat$hand <- factor(dat$hand, levels = c("r", "l"))        #   right
dat$frame <- factor(dat$frame, levels = c("loss", "gain"))  #   loss

expbeta <- c(1/sqrt(2), 1/2, 2, 4)  # ROR = 4, linear on logit scale
logit <- model.matrix(~ hand * frame, dat) %*% log(expbeta)
```

Example: How to fix the two-hand tapping study?

2. Generate observations

- Simulate data from binomial distribution

```
y <- rbinom(4, size = n/4, prob = plogis(logit))
```

```
## Sim 1              Sim 2              ...
## hand frame  y      hand frame  y
##    r  gain  16        r  gain  15
##    r  loss  7         r  loss  13
##    l  gain  21        l  gain  19
##    l  loss  9         l  loss  7
```

Example: How to fix the two-hand tapping study?

3. Test H_0

- Fit null model to your generated observations – H_0: $\beta_3 = 0$

```
m1 <- glm(cbind(y, n/4 - y) ~ hand + frame, binomial, dat)
```

- Fit specified model to your generated observations – H_1: $\beta_3 \neq 0$

```
m2 <- glm(cbind(y, n/4 - y) ~ hand * frame, binomial, dat)
## ROR = 5.880952
```

- Do a likelihood ratio test to test interaction

```
anova(m1, m2, test = "LRT")
## Analysis of Deviance Table
##
## Model 1: cbind(y, n/4 - y) ~ hand + frame
## Model 2: cbind(y, n/4 - y) ~ hand * frame
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1     4.3436
## 2         0     0.0000  1   4.3436  0.03715
```

Example: How to fix the two-hand tapping study?

4. Repeat

- Do previous steps repeatedly

```r
## Power analysis ##
n <- 300

pval <- replicate(2000, {
  y <- rbinom(4, size = n/4, prob = plogis(logit))
  mm1 <- glm(cbind(y, n/4 - y) ~ hand + frame, binomial, dat)
  mm2 <- glm(cbind(y, n/4 - y) ~ hand * frame, binomial, dat)
  anova(mm1, mm2, test = "LRT")$"Pr(>Chi)"[2]
})

mean(pval < 0.05)
## 0.7945
```

Final thoughts

Statistical tests are no screening procedures

– Significance is not a substitute for relevance
– Nonsignificance does not imply absence of effect

- Often, data are rather uninformative and compatible with many models and hypotheses
- At the same time, "all models are wrong" (Box, 1976)
- Making data-based decisions using statistical inference requires a confirmatory setting where a-priori substantive knowledge goes into the power analysis
- When relying on statistical tests outside such a setting, all we do is descriptive statistics with p-values; this does more harm than good
- You can find a collection of worked out examples for power simulations in R in Wickelmaier (2022): `https://doi.org/10.48550/arXiv.2110.09836`

References I

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. doi: 10.1080/01621459.1976.10480949

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). "Piensa" twice: On the foreign language effect in decision making. *Cognition*, *130*, 236–254. doi: 10.1016/j.cognition.2013.11.010

Gelman, A. (2020, October 22). *An odds ratio of 30, which they (sensibly) don't believe.* Retrieved from `https://statmodeling.stat.columbia.edu/2020/10/22/an-odds-ratio-of-30-which-they-sensibly-dont-believe/`

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi: 10.1177/1745691614551642

Harrell, F. (2020, June 20). *Statistical problems to document and to avoid.* Retrieved from `https://discourse.datamethods.org/t/author-checklist/`

McElroy, T., & Seta, J. J. (2003). Framing effects: An analytic–holistic perspective. *Journal of Experimental Social Psychology*, *39*(6), 610–617. doi: 10.1016/S0022-1031(03)00036-2

References II

McElroy, T., & Seta, J. J. (2004). On the other hand am I rational? Hemispheric activation and the framing effect. *Brain and Cognition*, *55*(3), 572–580. doi: 10.1016/j.bandc.2004.04.002

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*(694–706), 289–337.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458. doi: 10.1126/science.7455683

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. doi: 10.1515/ling-2019-0051

Wickelmaier, F. (2015). On not testing the foreign-language effect: A comment on Costa, Foucart, Arnon, Aparici, and Apesteguia (2014). *ArXiv*. doi: 10.48550/arXiv.1506.07727

Wickelmaier, F. (2022). Simulating the power of statistical tests: A collection of R examples. *ArXiv*. doi: 10.48550/arXiv.2110.09836

P-value

> The p-value is the probability of obtaining a test statistic that signals a deviation from H_0 at least as extreme as that observed in the experiment, given H_0 is true and its underlying model holds

`http://apps.mathpsy.uni-tuebingen.de/fw/pvalbinom/`

On the role of power

- Vasishth and Gelman (2021)

  "the importance of power cannot be stressed enough. Power should be seen as the ball in a ball game; it is only a very small part of the sport, because there are many other important components. But the players would look pretty foolish if they arrive to play on the playing field without the ball. Of course, power is not the only thing to consider in an experiment; no amount of power will help if the design is confounded or introduces a bias in some way" (p. 1333)