Power Analysis & Sample Size Calculation

(Require Substance-Matter Knowledge)

Florian Wickelmaier

Department of Psychology
University of Tübingen

Overview

- Large effects from subtle manipulations?
- Inference and power
- Power analysis by simulation
- Do it yourself

Help, my effect size is too large!

Examples

- Decision biases from two-hand tapping
- Beautiful parents have more daughters

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

Beautiful parents have more daughters

- Kanazawa (2007)

  "Very attractive individuals are 26% less likely to have a son" (p. 133)

  – $n_{total} = 2970$
  – $n_{v.att.} < 400$

- Gelman and Weakliem (2009)

  "the noise is stronger than the signal" (p. 314)

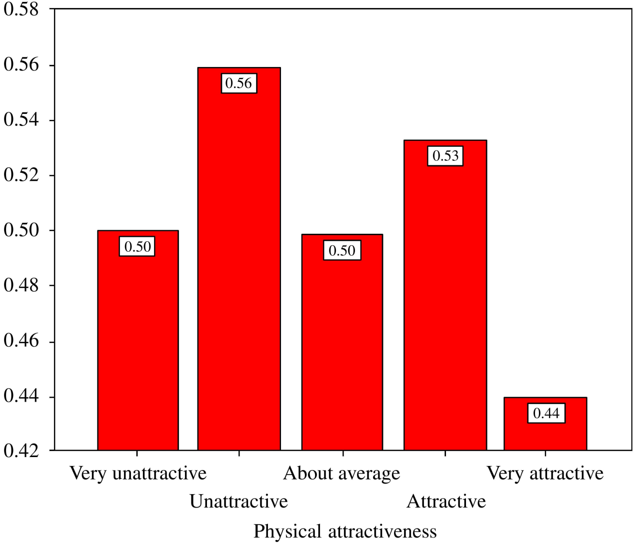


Fig. 1. Proportion of boys among the first child, by parent's physical attractiveness.

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

$$\frac{\begin{array}{c}\text{low power}\\ \wedge\\ \text{selection by significance}\end{array}}{\Rightarrow \text{inflated effect}}$$

(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

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

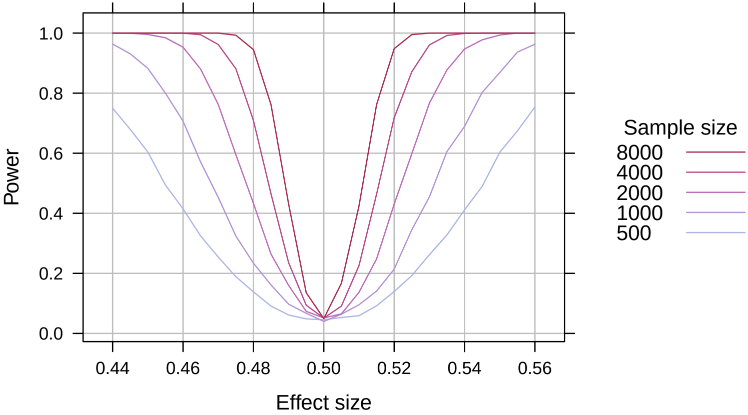
- Decision based on data (p-value)
  – If $p < \alpha$, choose H_1; else retain H_0
- Power $= 1 - \beta$
  – Probability of test to detect an effect of a given size

Power function

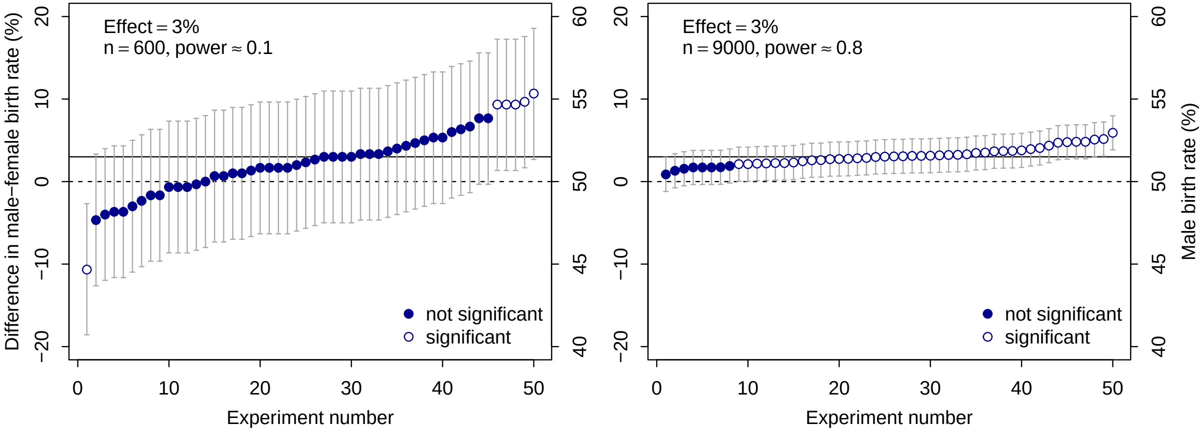
Power of a test depends on

- 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)

Exercise: First steps in simulation

- Generate data from a binomial model using the `rbinom()` function in R;
  try out different values of

  – n (10, 500, 2000)
  – the parameter π (0.5, 0.8, 0.44, 0.515)

  and see how this affects the output
- With these data, test different null hypotheses using `binom.test()`;
  these may or may not coincide with the values of π used for data generation
- If you repeat data generation and testing, can you usually reject H_0?

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

The steps in general

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

Power is estimated from the proportion of significant test results

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
- Plausible values from the literature (beware of significance filter)

(3) Specify the effect of interest

- *Not* the true effect (else no need to run the study!)
- *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)

Power simulation and sample size

The steps in 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

- 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 ...)

Examples and exercises

Selected examples

- Birth rates (with or without beautiful parents)
- Temporal value asymmetry
- Anchoring and adjustment
- How to fix the two-hand tapping study?

More examples

- Wickelmaier (2022) includes power simulation examples and R code for many classical statistical tests

Example: Birth rates

- Fisher's principle states that the male-female sex ratio is about 1:1
- Plan a study and calculate the sample size necessary to
  – detect a deviation from Fisher's principle of 106:100
  – with about 80% power
- Check your setup
  – Set the effect size to zero; what "power" estimate do you expect to get?

```r
n <- ...                                    # adjust sample size
pval <- replicate(5000, {                   # replications of experiment
  x <- rbinom(1, size = n,                  # data-generating model with
              prob = 106/(106 + 100))       #   minimum relevant effect
  binom.test(x, n = n, p = 1/2)$p.value     # p-value of test against H0
})
mean(pval < 0.05)                           # simulated power at alpha = 0.05
```

Exercise: Birth rates

- Kanazawa (2007) claims that beautiful parents have more daughters
- Plan a study and calculate the sample size necessary to
  – detect a deviation from the global 106:100 male-female sex ratio
  – with about 80% power
- Wanted: Substance-matter knowledge
  – What would be a minimum relevant deviation (effect)?
  – Considering the literature on birth rates, what would be a realistic deviation?
- Some background
  – `https://en.wikipedia.org/wiki/Human_sex_ratio`
  – Literature cited there (e. g., Davis et al., 1998; Mathews & Hamilton, 2005)

Exercise: Birth rates

```
## Fisher's principle

n <- 9500
pval <- replicate(5000, {
  x <- rbinom(1, size = n, prob = 106/(106 + 100))
  binom.test(x, n = n, p = 1/2)$p.value
})
mean(pval < 0.05)
```

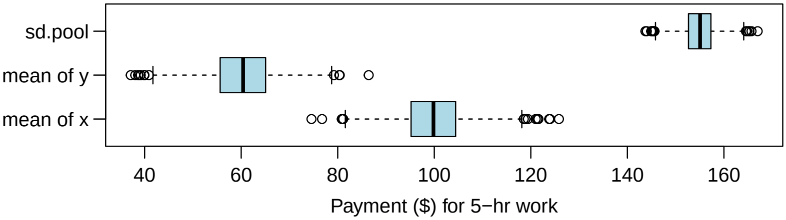
```
## Beautiful parents ###

n <- 22000
pval <- replicate(5000, {
  x <- rbinom(1, size = n, prob = 102/(102 + 100))
  binom.test(x, n = n, p = 106/(106 + 100))$p.value
})
mean(pval < 0.05)
```

Exercise: Temporal value asymmetry

- Caruso et al. (2008)
  "participants ... were asked to imagine that they had agreed to spend 5 hr entering data into a computer and to indicate how much money it would be fair for them to receive. Some participants imagined that they had completed the work 1 month previously, and others imagined that they would complete the work 1 month in the future ... Participants believed that they should receive 101% more money for work they would do 1 month later ($M = \$125.04$) than for identical work that they had done 1 month previously ($M = \$62.20$), $t(119) = 2.22$, $p = .03$, $d = 0.41$" (p. 797)
- Plan a direct replication of the study
  – What is a plausible standard deviation? Hint: $d = (M_1 - M_2)/SD$
  – What is an interesting minimal effect size (in $)?
- Parameter recovery
  – Re-estimate the parameters (μ_1, μ_2, σ) from the simulated responses
- Calculate total n necessary for 80% power

Exercise: Temporal value asymmetry

```r
## Parameter recovery

n <- 1000                                        # total sample size
out <- replicate(2000, {
  x <- rnorm(n/2, mean = 60 + 40, sd = 155)
  y <- rnorm(n/2, mean = 60,      sd = 155)
  t <- t.test(x, y, mu = 0, var.equal = TRUE)
  c(t$estimate,
    sd.pool = sqrt(n)/2 * t$stderr)              # SE = 2/sqrt(n) * SD
})
boxplot(t(out))
```



Exercise: Temporal value asymmetry

```r
## Power analysis

n <- 480
pval <- replicate(2000, {
  x <- rnorm(n/2, mean = 60 + 40, sd = 155)    # SD = (Mx - My)/d
  y <- rnorm(n/2, mean = 60,      sd = 155)
  t.test(x, y, mu = 0, var.equal = TRUE)$p.value
})
mean(pval < 0.05)
```

Exercise: Anchoring and adjustment

- Items (see Jacowitz & Kahneman, 1995) and anchor values
  - How tall is the largest coast redwood in the world? [20, 168 m]
  - How many member states belong to the United Nations? [14, 127 members]
  - How much km/h is the maximum speed of a house cat? [11, 48 km/h]
- Research question
  - Does time pressure (respond within 7 s) increase the anchor effect?
- Suggest a minimum relevant effect
  - Go to `http://apps.mathpsy.uni-tuebingen.de/fw/pars2eta/`
  - Fix the parameters of the ANOVA model
- Some background
  - Open anchoring quest (Röseler et al., 2022, `https://osf.io/ygnvb/`)

Exercise: Anchoring and adjustment

Plan the study

- Pick one of the three items
- Parameter recovery
  - Make a data frame for the two-by-two design
  - With the parameter values determined before, simulate responses
  - Re-estimate the parameters
- Power simulation
  - Calculate the sample size necessary to detect the time-pressure effect

Bonus task

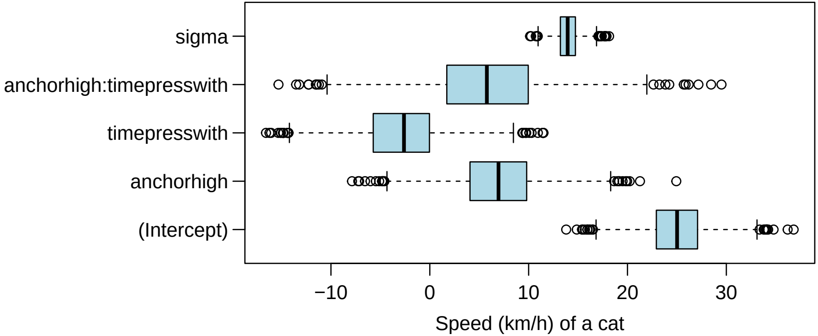
- Verify the plausibility of your model
  - Download the raw data from the open anchoring quest project
  - Estimate σ and compare it to your value

Exercise: Anchoring and adjustment

```r
## Parameter recovery

n <- 80
dat <- data.frame(
  anchor = factor(rep(1:2, each = n/2), labels = c("low", "high")),
  timepress = factor(rep(rep(1:2, each = n/4), 2), labels = c("w/o", "with"))
)
beta <- c(mu = 25, a2 = 7, b2 = -3, ab22 = 6)              # cat speed km/h
means <- model.matrix(~ anchor*timepress, dat) %*% beta

out <- replicate(2000, {
  y <- means + rnorm(n, sd = 14)
  m <- aov(y ~ anchor*timepress, dat)
  c(coef(m), sigma = sigma(m))
})
boxplot(t(out))
```

Exercise: Anchoring and adjustment

Parameter recovery for two-by-two ANOVA model



Exercise: Anchoring and adjustment

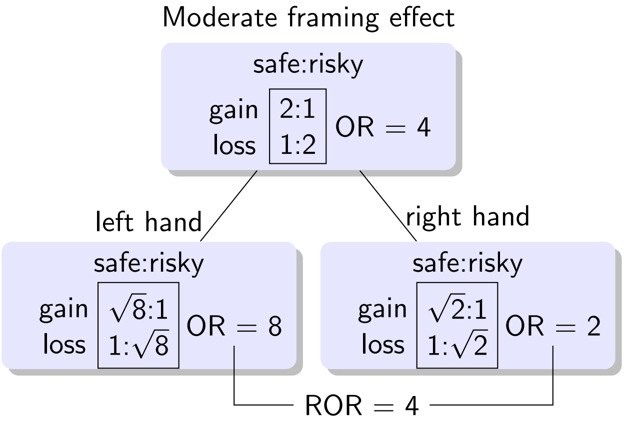
```r
## Power analysis

n <- 700
dat <- data.frame(
  anchor = factor(rep(1:2, each = n/2), labels = c("low", "high")),
  timepress = factor(rep(rep(1:2, each = n/4), 2), labels = c("w/o", "with"))
)
means <- model.matrix(~ anchor*timepress, dat) %*% beta

pval <- replicate(2000, {
  y <- means + rnorm(n, sd = 14)
  m <- aov(y ~ anchor*timepress, dat)
  summary(m)[[1]]$"Pr(>F)"[3]                          # test of interaction
})
mean(pval < 0.05)
```

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

Suggesting a minimum relevant effect

- Original framing effect, replication studies
- Factors (e. g., Costa et al., 2014; Wickelmaier, 2015, RORs $\approx$ 2–3)

How do these considerations translate into the parameters of this logit model?



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

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

Get a feel for model and data

- Analyze the original data (McElroy & Seta, 2004)
  - Mind the order of the factor levels
  - Formulate H_0 in terms of the parameters
  - Test the interaction

Plan a better study

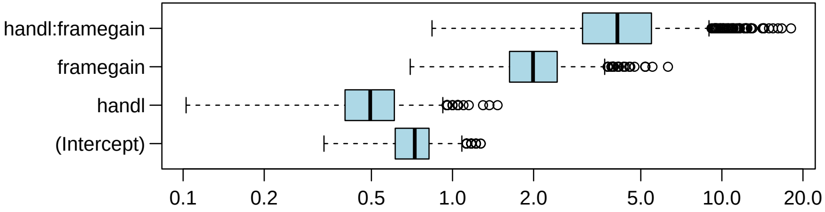
- Parameter recovery
  - Fix the parameters to the values determined before, simulate, recover
- Power simulation
  - Calculate the sample size necessary to detect the effect

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

```r
## Original data and analysis

dat <- read.table(header = TRUE, text = "
  hand frame safe risky
     r  gain    8     4
     r  loss    7     4
     l  gain   12     1
     l  loss    3     9
")                                                                  # ref. cat.
dat$hand <- factor(dat$hand, levels = c("r", "l"))                  #   right
dat$frame <- factor(dat$frame, levels = c("loss", "gain"))          #   loss

m1 <- glm(cbind(safe, risky) ~ hand + frame, binomial, dat)
m2 <- glm(cbind(safe, risky) ~ hand*frame, binomial, dat)
anova(m1, m2, test = "LRT")  # G(1) = 6.11, p = .013
```

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

```r
## Parameter recovery
n <- 400
beta <- c(1/sqrt(2), 1/2, 2, 4)  # ROR = 4, linear on logit scale
logit <- model.matrix(~ hand*frame, dat) %*% log(beta)
out <- replicate(2000, {
  y <- rbinom(4, size = n/4, prob = plogis(logit))
  mm2 <- glm(cbind(y, n/4 - y) ~ hand*frame, binomial, dat)
  exp(coef(mm2))
})
boxplot(t(out), log = "y")
```



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

```
## 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)
```

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

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

Caruso, E. M., Gilbert, D. T., & Wilson, T. D. (2008). A wrinkle in time: Asymmetric valuation of past and future events. *Psychological Science*, *19*(8), 796–801. doi: 10.1111/j.1467-9280.2008.02159.x

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

Davis, D. L., Gottlieb, M. B., & Stampnitzky, J. R. (1998). Reduced ratio of male to female births in several industrial countries: A sentinel health indicator? *Journal of the American Medical Association*, *279*(13), 1018–1023. doi: 10.1001/jama.279.13.1018

Gelman, A. (2020, October 22). *An odds ratio of 30, which they (sensibly) don't believe.* (`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

References II

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*(4), 310–316. doi: 10.1511/2009.79.310

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

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. doi: 10.1177/01461672952111004

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers–Willard hypothesis (gTWH). *Journal of Theoretical Biology*, *244*(1), 133–140. doi: 10.1016/j.jtbi.2006.07.017

Mathews, T. J., & Hamilton, B. E. (2005). Trend analysis of the sex ratio at birth in the United States. *National Vital Statistics Reports*, *53*(20), 1–20.

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.

Röseler, L., Weber, L., Helgerth, K. A. C., Stich, E., Günther, M., Tegethoff, P., . . . Schütz, A. (2022). *OpAQ: Open anchoring quest, version 1.1.42.95.* doi: 10.17605/OSF.IO/YGNVB

References III

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

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)