Principles in sample size estimation

Power versus precision

Statistik-Kolloquium, André Meichtry

Departement of Health Professions
Bern University of Applied Sciences

March 5, 2025

1. Quantity of interest
2. Power approach (Neyman-Pearson)
3. Precision approach

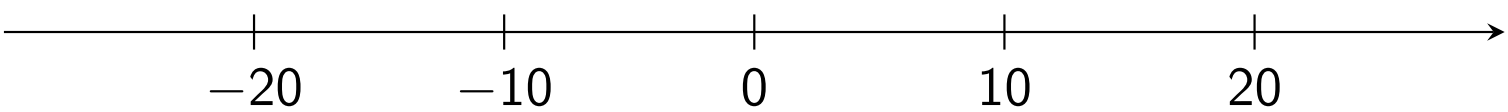
What is your quantity of interest θ?

- a true[1] slope in regression
- a true log hazard ratio
- a true within-subject change
- a true between-group difference
- a true sensitivity of a diagnostic test
- a true reliability measure (ICC, Kappa)
- a true risk ratio
- a true log odds ratio
- a true R^2
- etc. etc. etc

[1] "true": unknown value in the population from which we have sampled.

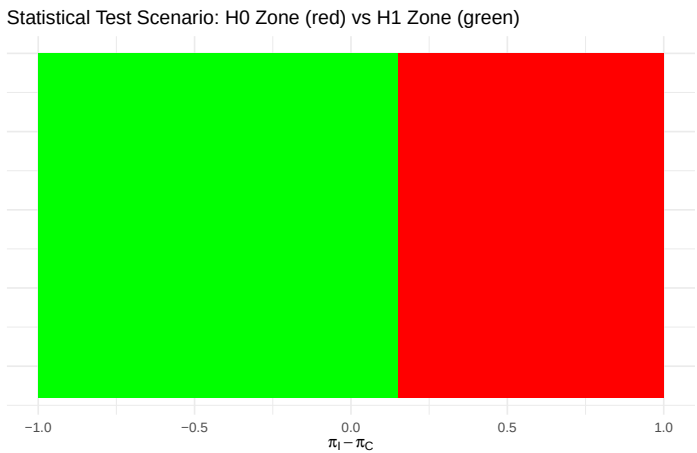
Sample size using power approach

- You need both the null and alternative hypothesis.
- You have a decision problem!
- Assume a quantity of interest θ with possible values on the real line, i.e log Odds Ratios (difference in logits)



Sample size using power approach

- Example: Non-inferiority study



Sample size using power approach

- With θ_0 as the superiority or non-inferiority margin.
- We have the following options for complementary H_0 and H_1:

$$
\begin{array}{rlcll}
\textit{clinical superiority}: & H_0 : \theta \leq \theta_0 & \textit{vs.} & H_1 : \theta > \theta_0, & (\theta_0 > 0) \\
\textit{statistical superiority}: & H_0 : \theta \leq 0 & \textit{vs.} & H_1 : \theta > 0 & \\
\textit{non-inferiority}: & H_0 : \theta \leq \theta_0 & \textit{vs.} & H_1 : \theta > \theta_0, & (\theta_0 < 0) \\
\textit{equivalence}: & H_0 : |\theta| \geq \theta_0 & \textit{vs.} & H_1 : |\theta| < \theta_0 & \\
\textit{equality}: & H_0 : \theta = 0 & \textit{vs.} & H_1 : \theta \neq 0 &
\end{array}
$$

θ_0 is very often set to 0, unfortunately! $\rightarrow$ "nil-null hypothesis"

Strawmen research

"There is a form of H_0 testing that has been used in astronomy and physics for centuries, what Meehl (1967) called the strong form, as advocated by Karl Popper (1959). Popper proposed that a scientific theory be tested by attempts to falsify it. In null hypothesis testing terms, one takes a central prediction of the theory, say, a point value of some crucial variable, sets it up as the H_0, and challenges the theory by attempting to reject it. This is certainly a valid procedure, potentially even more useful when used in confidence interval form. What I and my ilk decry is the weak form in which theories are confirmed by rejecting null hypotheses. ([3], p.999)."

— *Jacob Cohen*

Power function

Power is a function of the specified alternative θ_A

$$\text{Power}(\theta_A) = \Pr(\text{reject } H_0 \mid \theta_A). \tag{1}$$

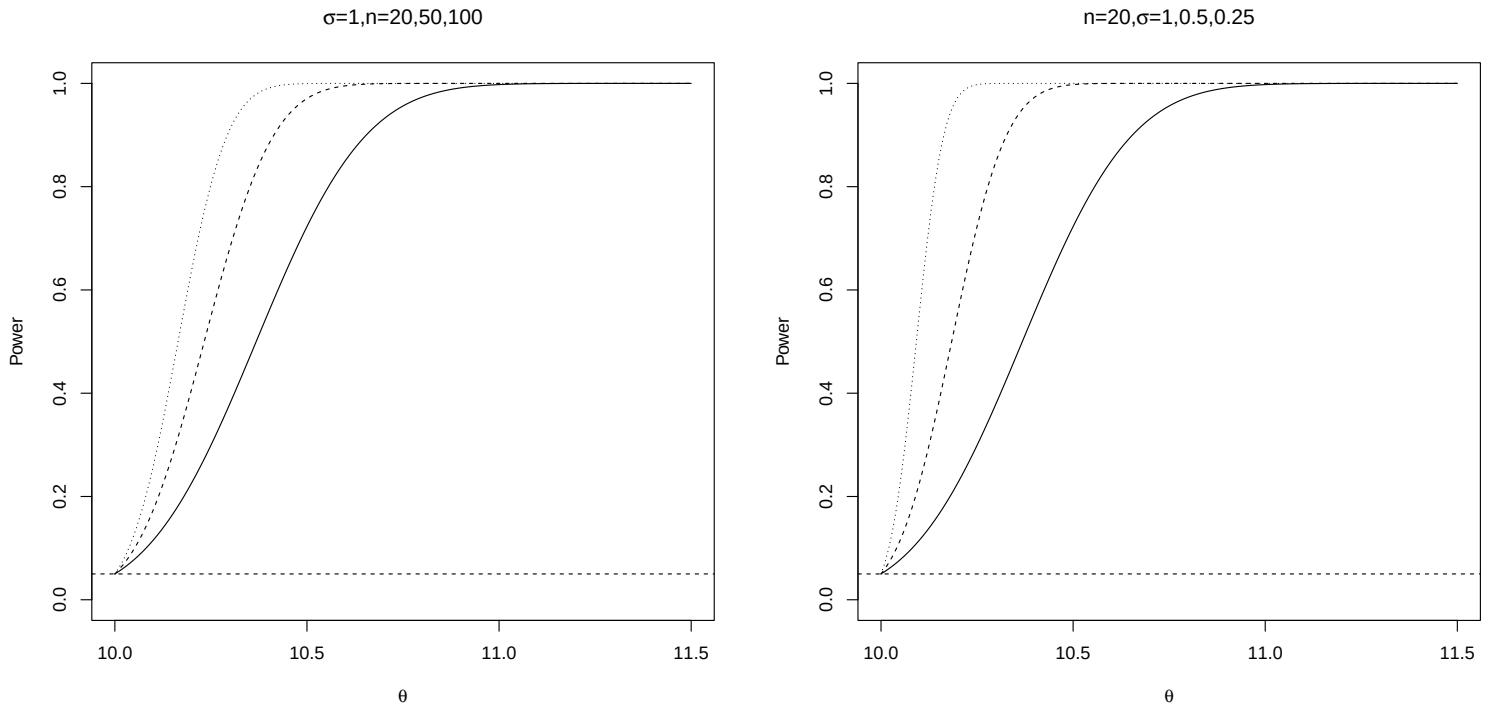
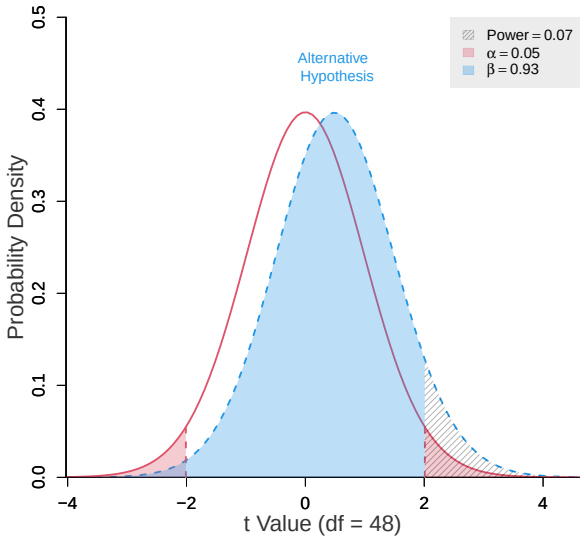
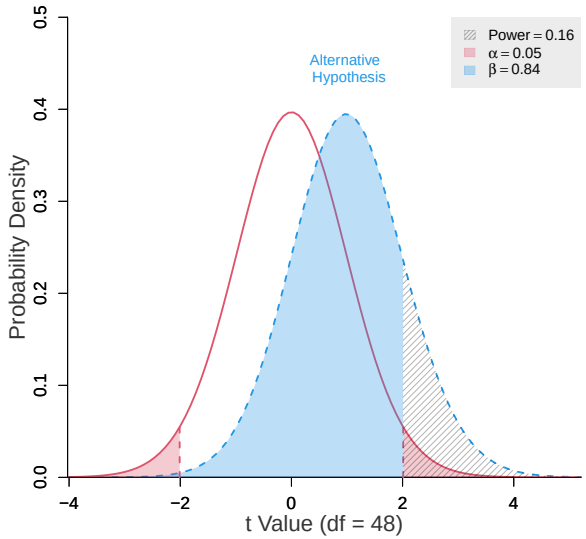
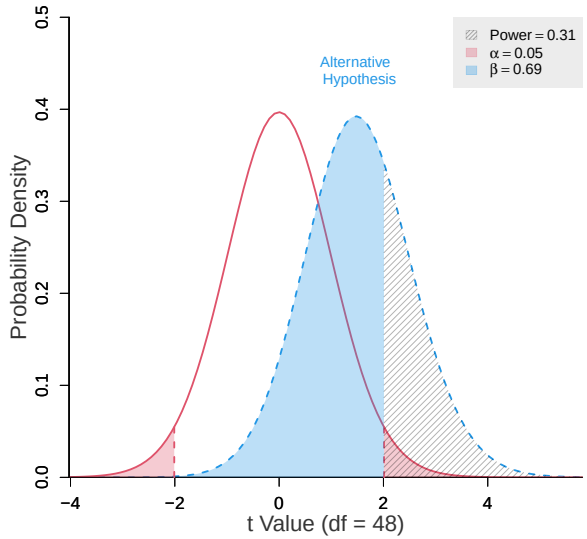


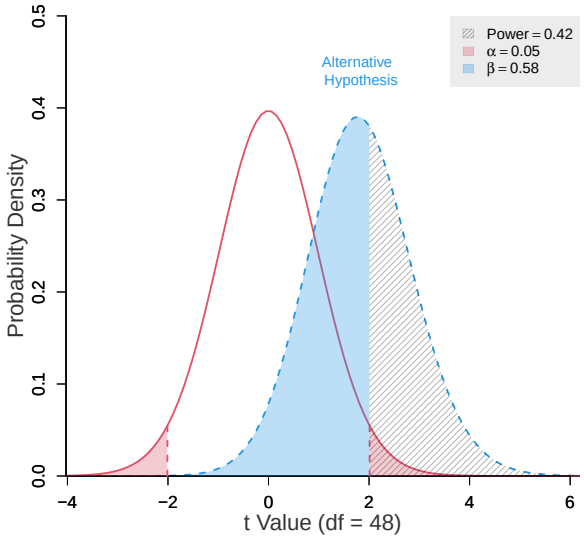
Figure: Power function one-sided z-test, with $H_0 : \theta \leq 10$ versus $H_1 : \theta > 10$. Left: $n = 20(—), 50(--), 100(\cdots)$ und $\sigma = 1$. Right: $n = 20$ und $\sigma = 1(—), 0.5(--), 0.25(\cdots)$.

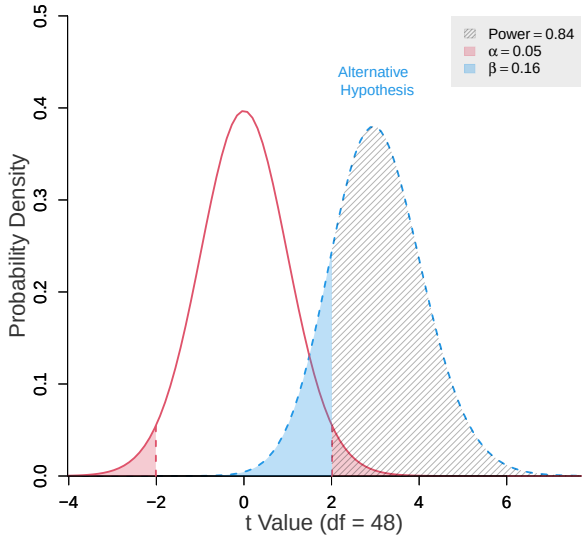
Which specific alternative?

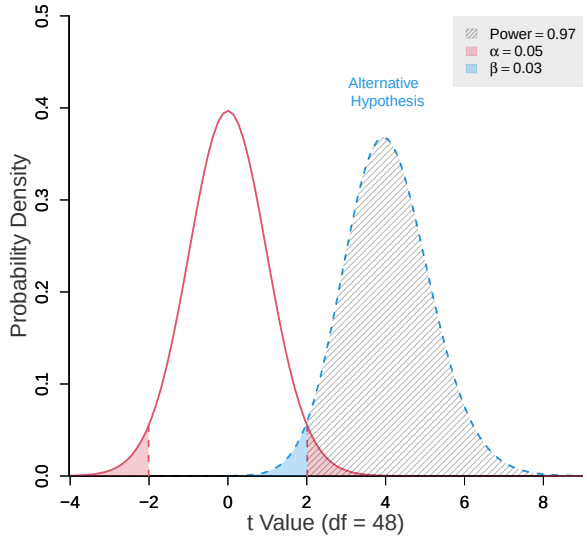












Which specific alternative?

To be able to compute sample size, you have to specify the alternative (to specify a distance with respect to H_0) (in order to control the type II error...)

- Different Interpretations
  - "minimal relevant difference"
  - "worthwhile difference"
  - "realistic difference, thought likely to occur"

  These ideas tend to conflate the demands made (i.e. of the new treatment) and the expectations of its benefit.
- Combined role of "realistic and important"

Analysis stage

The specified alternative has no role a posteriori (in the analysis stage). You test H_0 against $\neg H_0$, that's all! The successful rejection of a null does not give any support for a specific alternative, unless we have ruled out any other alternative (which would be an infinite number, too).

Simulation versus analytic approach

Power for t-test of $H_0 : \mu \leq 10$ versus $H_1 : \mu > 10$ for specified alternative $\mu_A = 10.5$ with $n = 20$, $\sigma = 1$, $\alpha = 0.05$:

- Simulation:

```
R <- 10000  #number of simulations
n <- 20  #sample size
X <- matrix(0, n, R)  #matrix for R sims with n data
alpha <- 0.05  #Type I error
sigma <- 1  #SD from pilot study
mu <- 10.5  #Truth under H1
mu0 <- 10  #H0
reject <- c()
for (i in 1:R) {
    # simulate data from assumed truth (specified Alternative)
    X[, i] <- rnorm(n = n, mean = 10.5, sd = 1)
    # reject or not
    reject[i] <- t.test(X[, i], mu = mu0, type = "one.sample", alternative = "greater")$p.value < alpha
}
proportions(table(reject))[2]  #power
```

```
##  TRUE
## 0.688
```

- Analytical: $Power_{\mu}(\alpha) = \Phi\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha}\right)$, implemented in

```
stats::power.t.test(n = 20, delta = 0.5, sd = sigma, sig.level = alpha, type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 20
##           delta = 0.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.695
##     alternative = one.sided
```

Example complex problem*

Analytical Power for Stepped-Wedge-Design

```
swSS <- function(t = t0, m = m0, s = s0, theta = theta0, wpICC = wpICC0, CAC = CAC0, IAC = IAC0, beta = 0.2, alpha = 0.05, long = TRU
    # num<-2*(qnorm(1-alpha/2)+qnorm(1-beta))^2 Nparallel <- 2*(num/(theta/s)^2) ##Total N for parallel RCT
    Nparallel <- ceiling(power.t.test(delta = theta, sd = s, power = power)$n) * 2
    DFcluster <- function(m, wpICC) {
        1 + (m - 1) * wpICC
    }
    Rlong <- (m * wpICC * CAC + (1 - wpICC) * IAC)/(1 + (m - 1) * wpICC)
    Rcross <- (m * wpICC * CAC)/(1 + (m - 1) * wpICC)
    if (long == TRUE) {
        R <- Rlong
    } else {
        R <- Rcross
    }
    DFtime <- function(t, R) {
        (3 * t * (1 - R) * (1 + t * R))/((t^2 - 1) * (2 + t * R))
    }
    k <- (Nparallel * DFcluster(m, wpICC) * DFtime(t, R))/m
    if (long == TRUE) {
        Nsw = k * m
    } else {
        Nsw = k * m * (t0 + 1)
    }
    res <- data.frame(Nparallel = Nparallel, k = k, Nsw = Nsw, IAC = IAC, CAC = CAC, wpICC = wpICC)
    res
}
```

Sample size with precision approach

- There are many reasons for preferring to run estimation studies instead of hypothesis testing studies.
- Almost always more appropriated for our students.
- A null hypothesis may be irrelevant, and when there is adequate precision one can learn from a study regardless of the magnitude of a p-value.
- A universal property of precision estimates is that, all other things being equal, increasing the sample size by a factor of four improves the precision by a factor of two.

Sample size with precision approach

- Do not need to guess the true population value.
- Many studies are powered to detect a miracle and nothing less; if a miracle doesnt happen, the study provides no information.
- Planning on the basis of precision will allow the resulting study to be interpreted if the p-value is large, because the confidence interval will not be so wide as to include both clinically significant improvement and clinically significant worsening.

Example

Quantity of interest: $\mu_1 - \mu_2$. Question: n needed s.t. 95% confidence interval is on average of the form

$$\text{estimate} \pm \delta.$$

- n observations are i.i.d. normally distributed
- σ from literature or pilot study
- Two sided $(1-\alpha)$-confidence interval for $\mu_1 - \mu_2$:

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \times \sigma\sqrt{1/n_1 + 1/n_2}$$

- For $\alpha = 0.05$, the condition is:

$$1.96 \times \sigma\sqrt{1/n_1 + 1/n_2} \leq \delta$$

- Assume $n_1 = n_2$ and solve for n (per group):

$$n \geq 2 \times \frac{1.96^2}{(\delta/\sigma)^2}.$$

Implementations

- Quantity of interest: $\mu_1 - \mu_2$
- $\sigma = 4$
- Aim: $\text{estimate} \pm 2$.

- statpsych: https://search.r-project.org/CRAN/refmans/statpsych/html/00Index.html

```
statpsych::size.ci.mean2(alpha = 0.05, var = 16, w = 4, R = 1)
```

```
##  n1 n2
##  32 32
```

- presize: https://search.r-project.org/CRAN/refmans/presize/html/00Index.html

```
# n from precision
presize::prec_meandiff(delta = 3, sd1 = 4, sd2 = 4, r = 1, conf.width = 4, variance = "equal")
```

```
##
##      sample size for mean difference with equal variance
##
##   delta sd1 sd2 n1 n2 conf.width conf.level lwr upr
## 1     3   4   4 32 32          4       0.95   1   5
```

```
# precision from n
presize::prec_meandiff(delta = 3, sd1 = 4, sd2 = 4, r = 1, conf.width = NULL, n1 = 32, n2 = 32, variance = "equal")
```

```
##
##      precision for mean difference with equal variance
##
##   delta sd1 sd2 n1 n2 conf.width conf.level lwr upr
## 1     3   4   4 32 32          4       0.95   1   5
```

Complex Survey Design*

- Consider Design effects
- Collects the inflation of variance due to complex sampling design
- Sampling Designs
  - Probability sampling
  - Simple random sampling without replacement
  - Simple random sampling with replacement
  - Systematic sampling
  - Cluster sampling
  - Stratified random sampling
- `samplesize4survey`:
- `https://search.r-project.org/CRAN/refmans/samplesize4surveys/html/00Index.html`