Äquivalenztests, strenge Tests

Kolloquium für Statistik-Interessierte

Departement Gesundheit
Berner Fachhochschule

5. Juni 2024

Null-Hypothesen-Signifikanz-Test-Prozedur

- Sei δ eine Quantität von Interesse
- z.B. ein Zwischengruppenunterschied $\delta = \mu_1 - \mu_2$
- Sehr oft ist es nicht sinnvoll, Punkt-Nullhypothesen wie

$$H_0 : \delta = \delta_0 \quad \texttt{versus} \quad H_1 : \delta \neq \delta_0$$

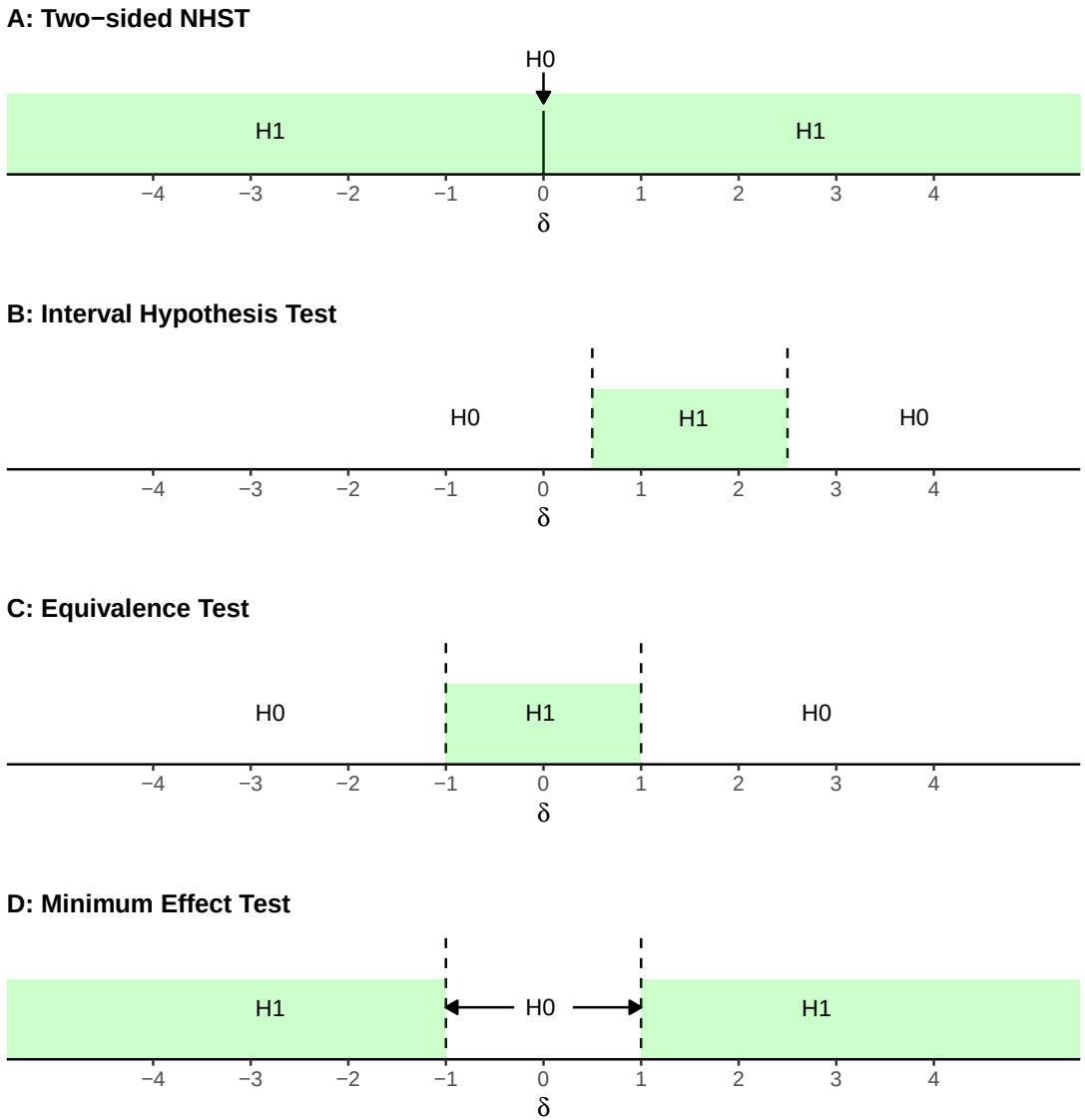
  zu testen.
- Oft wird zudem $\delta_0 = 0$ getestet.
- Wir können mit solchen Test "nur" etwas verwerfen, was sehr oft rein logisch (a priori!) schon falsch ist[1].

[1] Das wahre δ wird nie genau δ_0 sein (Das gilt vor allem bei Beobachtungsstudien).

Null-Hypothesen-Signifikanz-Test-Prozedur

- Die Wahrscheinlichkeit der Verwerfung dieser nil-null-Hypothese ist nur eine Frage der Präzision und nicht eine Frage der Gültigkeit der Hypothese.
- Dies stellt ein klägliches Unterfangen dar.
- Solche Strohmann-Nullhypothesen sind leicht zu verwerfen.
- Weak Tests münden in schlecht belastbare Wissenschaft und die berüchtigte Replikationskrise unserer Wissenschaften.
- Interessierte: Paul Meehl (2003)
  `https://meehl.umn.edu/sites/meehl.umn.edu/files/files/aumeehl2003sigtests-trimmed.mp3`

Parameterraum und Hypothesen (wichtigstes Slide)



Äquivalenztests

- Wenn eine Theorie aber nun einen Bereich für $\delta = \mu_1 - \mu_2$ vorhersagt, z.B. dass
- δ in einem Bereich liegt, der durch Grenzen $[-\epsilon, +\epsilon]$ festgelegt ist, haben wir folgende – viel stärkere – Testsituation:

$$H_0 : \delta \leq -\epsilon \text{ oder } \delta \geq +\epsilon \qquad H_1 : -\epsilon < \delta < \epsilon. \tag{1}$$

- Die Nullhypothese besagt, dass der wahre Parameter ausserhalb einer Toleranzregion liegt oder in der Region der Irrelevanz.
- Die Alternative besagt, dass der wahre Parameter im durch die Theorie vorhergesagten Region liegt, im Bereich von Relevanz.
- H_0 zu verwerfen heisst jetzt Irrelevanz zu verwerfen (definiert durch die Grenzen ϵ).

TOST

- Dieses Problem kann man lösen mit sogenannten TOST-Testverfahren ("Two One-sided t-Tests").
- Man macht zwei einseitige Tests auf α-Niveau:

$$H_{0a} : \delta \leq -\epsilon \qquad H_{1a} : \delta > -\epsilon \tag{2}$$

$$H_{0b} : \delta \geq +\epsilon \qquad H_{1b} : \delta < +\epsilon. \tag{3}$$

- Ablehnen von H_{0a} **und** H_{0b} bedeutet dann, $-\epsilon < \delta < \epsilon$.
- Die einzelnen Tests sind normale t-Tests, alle Varianten möglich (Unverbundene Stichproben (gleiche oder ungleiche Varianz), Verbundene Stichprobe, Eine Stichprobe).

Multiples Testen*

- Bei TOST-Testverfahren müssen wir keine Korrektur für multiples Testen durchführen, beide Tests werden zum Niveau α gemacht.
- Das hat zur Konsequenz, dass wir bei $\alpha = 0.05$ den Test auch mit einem 90% KI (statt einem 95% KI) (mit L, U also untere und obere Grenze) durchführen können.

Beweis.

- H_{0a} wird genau dann abgelehnt, wenn das einseitige $(1-\alpha)$-KI $[L, \infty)$ komplett rechts von $-\epsilon$ liegt
- H_{0b} wird genau dann abgelehnt, wenn das einseitige $(1-\alpha)$-KI $[-\infty, U)$ komplett links von $+\epsilon$ liegt
- Die Schnittmenge $[L, U] = [L, \infty) \cap (-\infty, U]$ ist aber genau das $(1-2\alpha)$-KI für die Differenz der Erwartungswerte.

□

Dualität Testen und Schätzen*

- Wir können den TOST zum Niveau α also auf zwei Arten durchführen:
  - Zwei einseitige t-Tests, H_0 ablehnen, wenn beide p-Werte kleiner als α
  - $(1-2\alpha)$-Konfidenzintervall für die Mittelwertdifferenz berechnen. Wir schliessen auf Äquivalenz, wenn das KI komplett in $(-\epsilon, +\epsilon)$ enthalten ist.

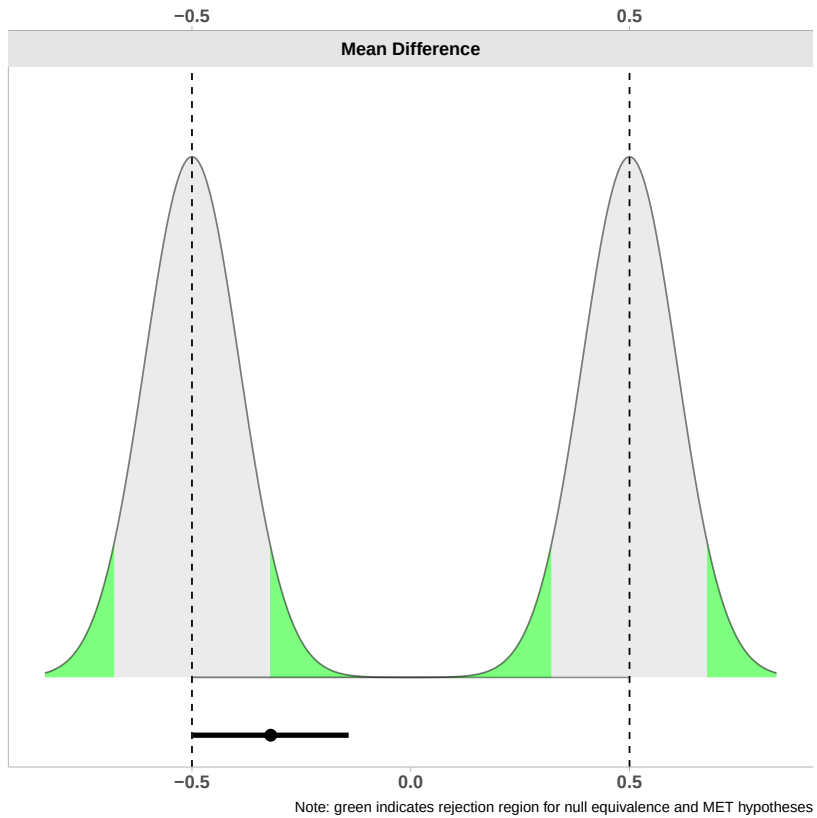
Implementation in R

- Packet `TOSTER`
- `https://cran.rstudio.com/web/packages/TOSTER/vignettes/IntroductionToTOSTER.html`

t-Test from TOSTER

```
res <- TOSTER::tsum_TOST(m1 = 4.55, m2 = 4.87, sd1 = 1.05, sd2 = 1.11,
                         n1 = 200, n2 = 200, low_eqbound = -0.5, high_eqbound = 0.5)

plot(res, type = "tnull",estimates="raw")
```



t-Test from TOSTER

```
res

##
## Welch Two Sample t-test
##
## The equivalence test was significant, t(396.78) = 1.666, p = 4.82e-02
## The null hypothesis test was significant, t(396.78) = -2.962, p = 3.24e-03
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: reject null equivalence hypothesis
##
## TOST Results
##                 t    df p.value
## t-test     -2.962 396.8   0.003
## TOST Lower  1.666 396.8   0.048
## TOST Upper -7.590 396.8 < 0.001
##
## Effect Sizes
##                Estimate    SE               C.I. Conf. Level
## Raw             -0.3200 0.108 [-0.4981, -0.1419]         0.9
## Hedges's g(av)  -0.2956 0.104 [-0.4605, -0.1304]         0.9
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

SESOI

How to specify the smallest effect size of interest (SESOI).

- Based on Theory
- Anchor based
- Cost-benefit
- **Small telescopes approach**
- Smallest detectable effect

Small telescopes approach

Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one.

Small telescopes approach

- Uri Simonsohn (2015) defines a small effect as one that would give 33% power to the original study.
- The effect size that would give the original study odds of 2:1 against observing a statistically significant result if there was an effect.
- Test if we can reject effects as large or larger than the effect the original study has 33% power to detect.
- It is a simple one-sided test, not against 0, but against a SESOI.
- A replication that obtains an effect size that is statistically significantly smaller than this small effect is inconsistent with the notion that the studied effect is large enough to have been detectable with the original sample size.

Example

- Eskine (2013) showed that participants who had been exposed to organic food were substantially harsher in their moral judgments relative to those exposed to control ($d = 0.81$,95% CI: $[0.19, 1.45]$)[2]
- A replication by Moery & Calin-Jageman (2016, Study 2) did not observe a significant effect (Control: $n = 95$, $M = 5.25$, $SD = 0.91$, Organic Food: $n = 89$, $M = 5.22$, $SD = 0.83$).

[2] d steht für die Effektgrösse nach Cohen, $d = \frac{\hat{\delta}}{s}$

Example

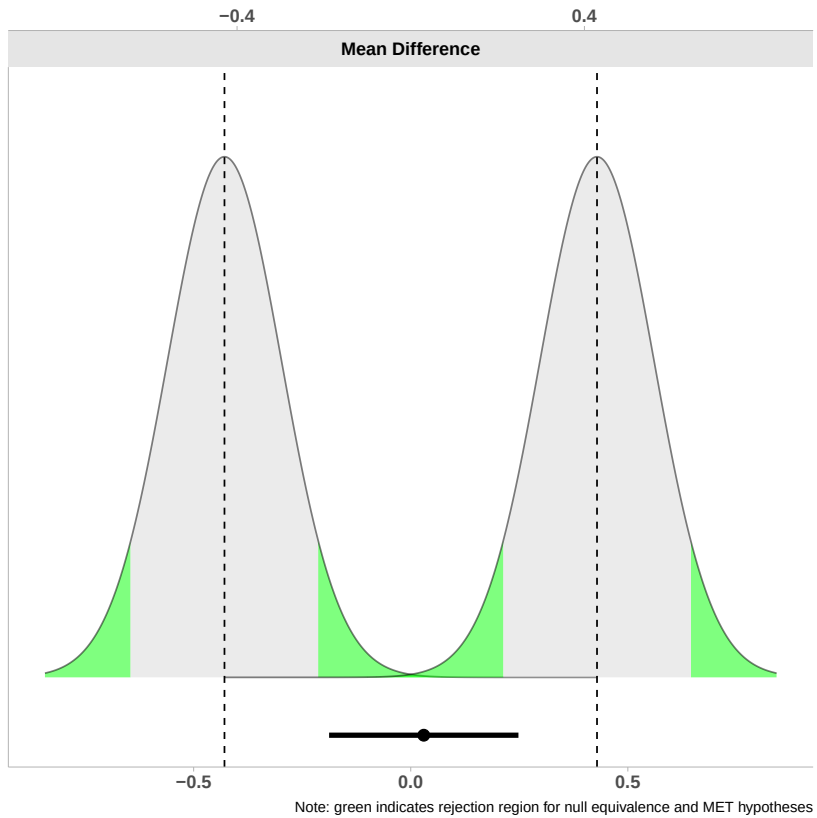
- Equivalence bound: Following Simonsohns (2015) recommendation the equivalence bound was set to the effect size the original study had 33% power to detect (with $n = 21$ in each condition, this means the equivalence bound is $d = 0.48$
- which equals a difference of 0.429 on a 7-point scale given the sample sizes and a pooled standard deviation of 0.894).

```
(d<-power.t.test(n=21,sd=1,power=0.33)$delta)
## [1] 0.481
spooled<-0.894
d*spooled
## [1] 0.43
```

Example

```
res2 <- TOSTER::tsum_TOST(m1=5.25,m2=5.22,sd1=0.95,sd2=0.83,n1=95,n2=89,
            low_eqbound=-0.429, high_eqbound=0.429, alpha = 0.05, var.equal=TRUE)

plot(res2, type = "tnull",estimates="raw")
```



Example

- Using a TOST equivalence test with $\alpha = 0.05$, assuming equal variances, and equivalence bounds of $d = -0.48$ and $d = 0.48$ is "significant", $t(182) = -3.03$, $p = 0.001$.
- We can reject effects more extreme than $d = 0.48$ (or a raw difference of 0.429 scalepoints).

Example

```
res2

##
## Two Sample t-test
##
## The equivalence test was significant, t(182) = -3.025, p = 1.42e-03
## The null hypothesis test was non-significant, t(182) = 0.227, p = 8.2e-01
## NHST: don't reject null significance hypothesis that the effect is equal to zero
## TOST: reject null equivalence hypothesis
##
## TOST Results
##                 t  df p.value
## t-test      0.2275 182    0.82
## TOST Lower  3.4804 182 < 0.001
## TOST Upper -3.0254 182   0.001
##
## Effect Sizes
##            Estimate     SE             C.I. Conf. Level
## Raw         0.03000 0.1319  [-0.188, 0.248]         0.9
## Hedges's g  0.03342 0.1469 [-0.2083, 0.275]         0.9
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```