Zürcher Hochschule
für Angewandte Wissenschaften

# Äquivalenztests zur Evaluation von Interventionen in der Lehre

Gekürzte Version

Treffen der Statistikinteressierten an der ZHAW
25. 6. 2019

Claude Müller Werder (muew, ZID)
Thoralf Mildenberger (mild, IDP)

Basiert teilweise auf Paper mit Maren Lübcke (HIS Institut für Hochschulentwicklung Hannover)

# Äquivalenz von Behandlungen

**Ziel:** Es soll gezeigt werden, dass zwei Populationen den selben Mittelwert haben

**t-Test** $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Nullhypothese kann nicht abgelehnt werden ($p > 0.05$).
**Häufiger Fehlschluss:** „Damit ist gezeigt, dass es keinen Unterschied gibt."

Warum ist das falsch?

- Hypothesentests bewerten nur Evidenz **gegen** die Nullhypothese, **nicht für** die Nullhypothese
- Nicht-Ablehnen von $H_0$ *kann* heissen, dass $H_0$ zutrifft.
- Nicht-Ablehnen von $H_0$ *kann* aber auch daher kommen, dass $H_0$ falsch ist, dies sich aber anhand der Daten nicht eindeutig zeigen lässt (z.B. zu kleine Stichprobe, zu grosse Varianz, generell zu tiefe Power des Tests).
- Analogie: Vor Gericht wird ein Angeklagter freigesprochen, wenn Ihm keine Schuld nachgewiesen werden kann. Vielleicht ist er unschuldig, vielleicht gibt es einfach keine Beweise!

# Äquivalenz von Behandlungen

**Was kann man dagegen machen?**

**Verschiedene Ansätze:**

- **Klassischer t-Test mit Analyse der Power**
  t-Test wird wie gehabt durchgeführt und Power gegen Effekte bestimmter Grössen analysiert. Ältester Ansatz.
- **Bayes-t-Test**
  Bayes-Ansätze behandeln $H_0$ und $H_1$ symmetrisch. Bayes-Faktor sogar unabhängig von Priors *für* Hypothesen (aber ein Prior *unter* $H_1$ wird benötigt!). Problem: Wir glauben meist gar nicht, dass die Differenz exakt Null ist, sondern wollen zeigen, dass sie klein ist.
- **Konfidenzintervalle**
  Für einfache Nullhypothesen praktisch äquivalent zum Testen, aber informativer. Hier extrem nützlich, da man schauen kann, welche Bereiche man einigermassen sicher ausschliessen kann. Nicht immer optimal.
- **Äquivalenztests**
  Eine Art Umkehrung der klassischen Tests. Fällt komplett unter klassisches Neyman-Pearson-Framework, ist in der Praxis aber wenig bekannt. Bayes-Versionen gibt es auch!

# Äquivalenztests

**t-Test** $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

**Verbreiteter Irrtum:** „Nullhypothese besagt immer, dass kein Effekt existiert bzw. eine Differenz Null ist.“

Definition von Null- und Alternativhypothese ergibt sich aus der Fragestellung:
Die Nullhypothese möchte man widerlegen („Beweislast“).

**t-Test umgekehrt** $H_0: \mu_1 \neq \mu_2$ vs. $H_1: \mu_1 = \mu_2$

**Probleme:**

- Es existiert keine gute Lösung des Problems. Kein Test kann höhere Power haben als der triviale Niveau-α-Test (Daten gar nicht benutzen und $H_0$ mit Wahrscheinlichkeit α verwerfen)
- Wir wollen eigentlich auch gar nicht zeigen, dass es gar keine Differenz gibt, sondern dass sie klein genug ist um praktisch irrelevant zu sein.

# Äquivalenztests

**Eigentlich interessierendes Testproblem:**

$H_0$: $\mu_1 - \mu_2 \leqq -\varepsilon$ *oder* $\mu_1 - \mu_2 \geqq +\varepsilon$ vs. $H_1$: $-\varepsilon < \mu_1 - \mu_2 < +\varepsilon$

**Nullhypothese:** Die Differenz ist grösser als ein Toleranzwert
**Alternative:** Die Differenz ist kleiner als ein Toleranzwert

Verwerfen von $H_0$ bedeutet, dass man Gleichheit bis auf ε nachgewiesen hat.
Nichtverwerfen von $H_0$ bedeutet nicht, dass man einen grossen Effekt nachgewiesen hat!

**Methoden:**
- TOST (Two One-sided t-Tests) für nicht-standardisierte Mittelwertdifferenz
- UMPI-Test für standardisierte Mittelwertdifferenz (Cohen‘s d, Effektstärke) vgl. Wellek (2010)

# Äquivalenztests: TOST

**Two One-Sided t-Tests:**

$H_{0,1}$: $\mu_1 - \mu_2 \leqq -\varepsilon$ vs. $H_{1,1}$: $-\varepsilon < \mu_1 - \mu_2$

$H_{0,2}$: $\mu_1 - \mu_2 \geqq +\varepsilon$ vs. $H_{1,2}$: $\mu_1 - \mu_2 < +\varepsilon$

Wir führen also **zwei** Tests durch:

- Ablehnen von $H_{0,1}$ bedeutet, die Differenz ist nicht kleiner als $-\varepsilon$
- Ablehnen von $H_{0,2}$ bedeutet, die Differenz ist nicht grösser als $+\varepsilon$

Ablehnen von $H_{0,1}$ **und** $H_{0,2}$ bedeutet also, $-\varepsilon < \mu_1 - \mu_2 < +\varepsilon$.

Die einzelnen Tests sind normale t-Tests, alle Varianten möglich

- Unverbundene Stichproben (gleiche oder ungleiche Varianz)
- Verbundene Stichprobe
- Eine Stichprobe

Zu welchen Niveau müssen wir die beiden Tests durchführen?

Müssen wir eine **Korrektur für multiples Testen** machen?

# Äquivalenztests: TOST

Wir wollen zum Niveau α testen, d.h. wenn keine Äquivalenz vorliegt, wollen wir nur mit Wahrscheinlichkeit α Äquivalenz behaupten.

**Es reicht, beide Tests zum Niveau α durchzuführen!**

Warum?

Sei dazu $H_0$ richtig. Wir machen einen Fehler 1. Art, wenn $H_{0,1}$ und $H_{0,2}$ ablehnen.

Da nur **entweder** $\mu_1 - \mu_2 \leqq -\varepsilon$ **oder** $\mu_1 - \mu_2 \geqq +\varepsilon$ sein kann,
ist entweder $H_{0,1}$ oder $H_{0,2}$ richtig, aber **nicht beide**.

Sei o.E. $H_{0,1}$ richtig.

$P_{H0,1}$(Fehler 1. Art) = $P_{H0,1}$ ($H_{0,1}$ ablehnen und $H_{0,2}$ ablehnen) $\leqq P_{H0,1}$ ($H_{0,1}$ ablehnen) $\leqq \alpha$

Analog für den Fall $H_{0,2}$ richtig.

Wir müssen also keine **Korrektur für multiples Testen** durchführen!

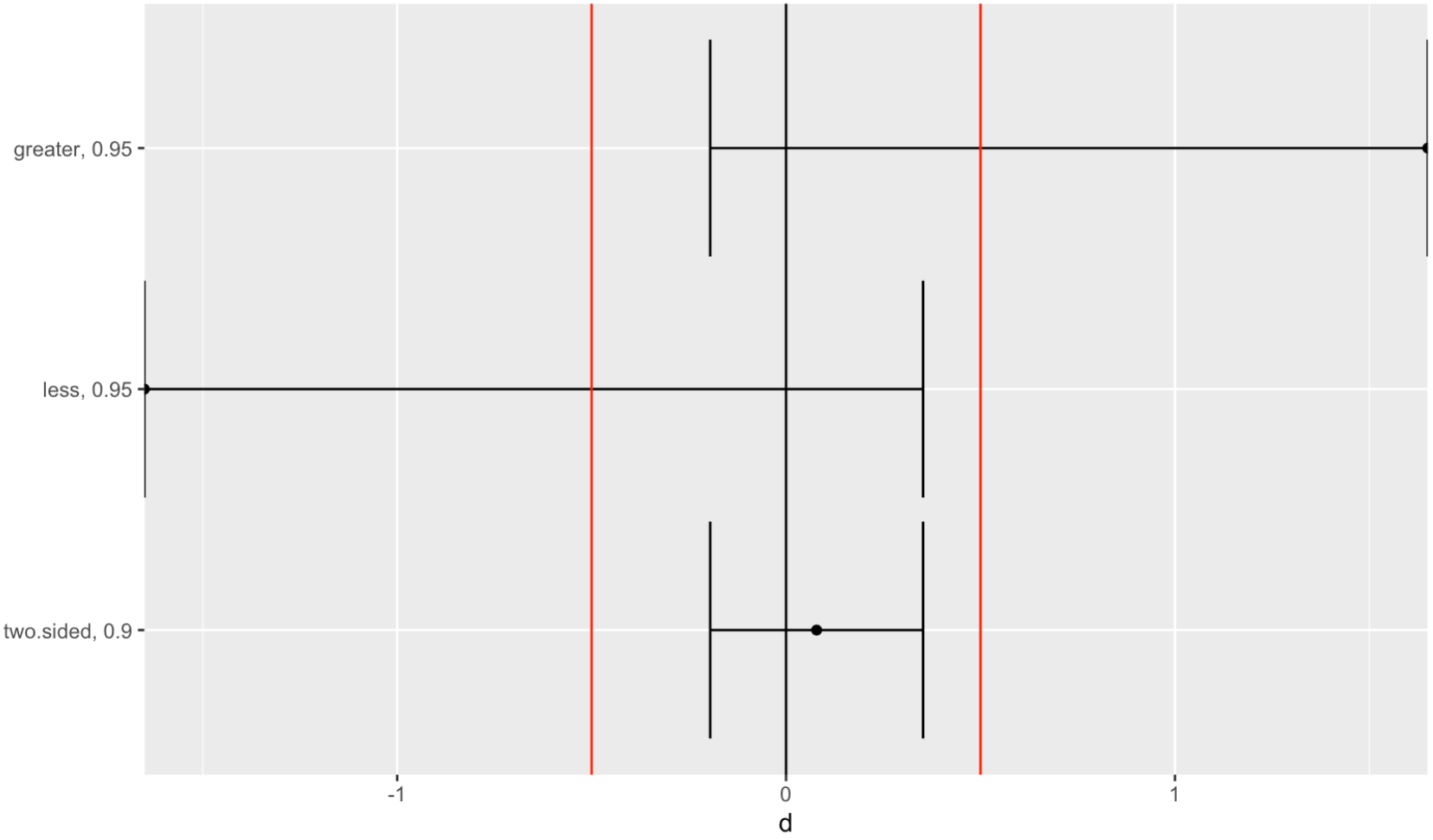
# Äquivalenztests: TOST

Dies hat eine weitere Konsequenz: Wir können den Test mit einem **(1-2α)-Konfidenzintervall** durchführen:

- $H_{0,1}$ wird genau dann abgelehnt, wenn das einseitige (1-α)-KI $[d_l,\infty)$ komplett rechts von $-\varepsilon$ liegt
- $H_{0,2}$ wird genau dann abgelehnt, wenn das einseitige (1-α)-KI $(\infty, d_u]$ komplett links von $+\varepsilon$ liegt

- Die Schnittmenge $[d_l, d_u] = [d_l,\infty) \cap (\infty, d_u]$ ist aber genau das (1-2α)-Konfidenzintervall für die Differenz der Erwartungswerte!

# Äquivalenztests: TOST



# Äquivalenztests: TOST

Wir können den TOST zum Niveau α also auf zwei Arten durchführen:

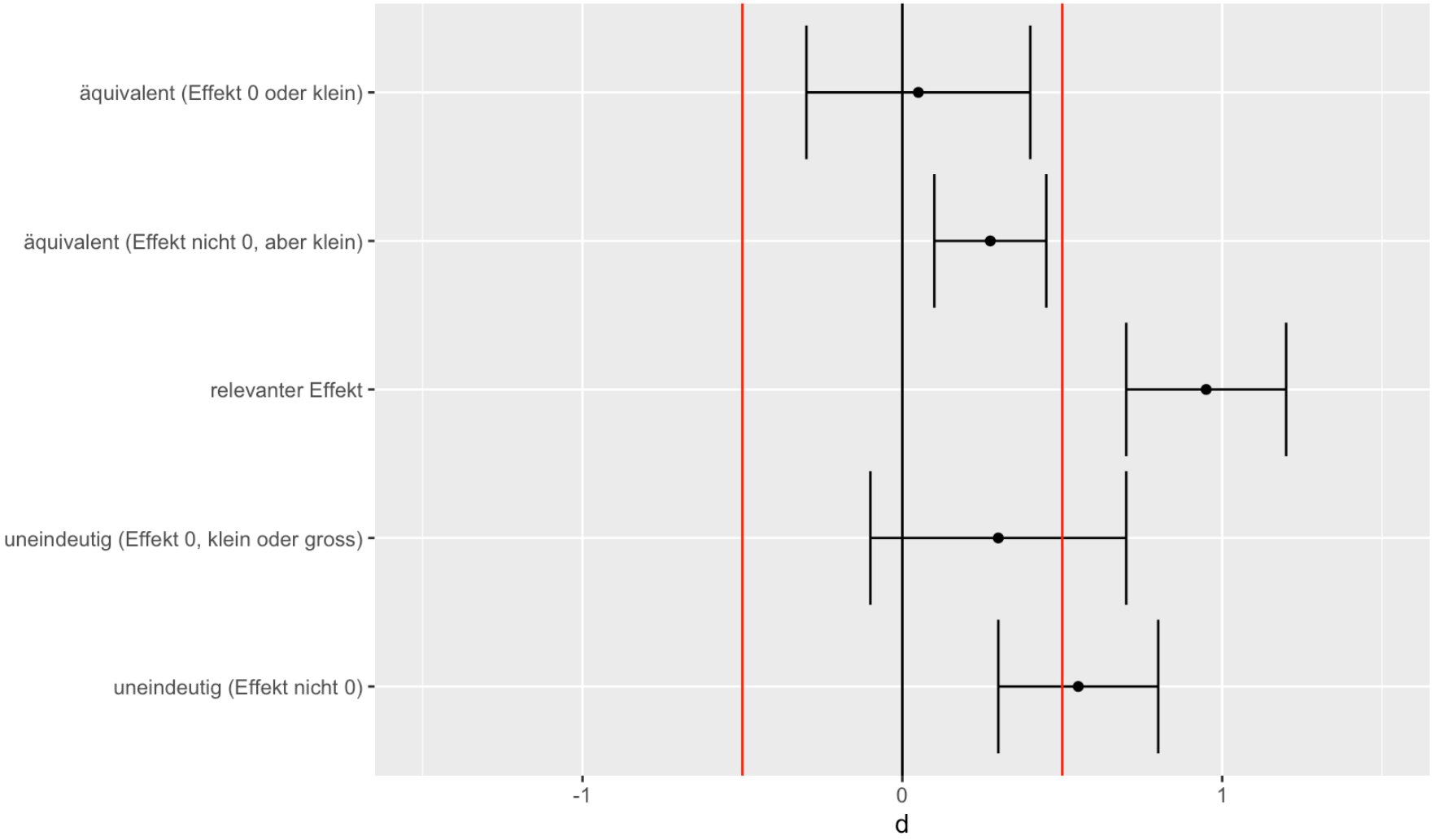
- Zwei einseitige t-Tests, $H_0$ ablehnen, wenn beide p-Werte kleiner als α
- (1-2α)-Konfidenzintervall für die Mittelwertdifferenz berechnen. Wir schliessen auf Äquivalenz, wenn das KI komplett in (-ε, + ε) enthalten ist.

**Bemerkungen**

- Konfidenzintervall ist evtl. informativer, da man damit auch Äquivalenz ausschliessen kann
- Der Test zum 5%-Niveau entspricht einem 90%-Konfidenzintervall
- **In R:** z.B. Paket TOSTER

Zürcher Hochschule
für Angewandte Wissenschaften

# Äquivalenztests: TOST / Konfidenzintervall



# Äquivalenztests: UMPI-Test

Optimaler Äquivalenztest für standardisierte Mittelwertdifferenzen

Implementierungen in R-Paketen:

**EQUINONINF**:
- R-Paket zum Buch von Wellek (2010)
- Berechnet Ablehnbereiche, nicht p-Werte
- Kann mit asymmetrischen Äquivalenzbereichen umgehen
- Paket enthält auch andere Äquivalenztests

**equivUMP**:
- Selbst implementiert, seit April 2019 auf CRAN
- Syntax und Output soweit wie möglich analog zu t.test()
- Berechnet p-Werte
- Kann (bis jetzt) nicht mit asymmetrischen Äquivalenzbereichen umgehen
- Ein- und Zweistichprobenversionen (gepaart, ungepaart), ausserdem auch einseitige Versionen (non-inferiority testing, non-superiority testing)

# Äquivalenztests: UMPI-Test

```
> equiv.test(grade ~ group, data = dat, alternative = "two.sided",
eps = 0.5, mu = 0, paired = FALSE)

        Two sample equivalence test

data:  grade by group
t = 0.45781, df = 98.0000, ncp = 2.2913,
p-value = 0.03034
alternative hypothesis: equivalence
null values:
     lower upper
[1,]  -Inf  -0.5
[2,]   0.5   Inf
sample estimates:
         d
0.09990274
```

# Literatur

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference, 26*(2), 231-245. doi: https://doi.org/10.1016/j.foodqual.2012.05.003

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton: CRC Press.

# R-Pakete

- **BayesFactor:** Computation of Bayes Factors for Common Designs
  *Kann benutzt werden, um einen Bayesianischen Äquivalenztest durchzuführen*
- **BEST:** Bayesian Estimation Supersedes the t-Test
  *Bayesianischer Ansatz auf Basis der Region of Practical Equivalence (ROPE)*
- **Equivalence:** Provides Tests and Graphics for Assessing Tests of Equivalence
  *TOST und weitere Tests*
- **EquivalenceTest:** Equivalence Test for the Means of Two Normal Distributions
  *Einige neuere spezielle Tests*
- **EQUIVNONINF:** Testing for Equivalence and Noninferiority
  *zu Wellek (2010), optimale Tests*
- **equivUMP:** Uniformly Most Powerful Invariant Tests of Equivalence
  *UMPI-Test*
- **TOSTER:** Two One-Sided Tests (TOST) Equivalence Testing
  *TOST mit Varianten*