

Statistik für Digital Humanities

Vergleich zweier Mittelwerte

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

31. Mai 2021

[Letzte Aktualisierung: 26/07/2021, 11:39]

Arten Verschiedener Messungen

- Gruppendesign
 - Verschiedene Probanden in Gruppen
 - Gleichzeitige Messung möglich
 - Unabhängiges Design
- Messwiederholungsdesign
 - Gleiche Probanden
 - Wiederholte Messung
 - Abhängiges Design

Arten Verschiedener Messungen

Beispiel Arachnophobie

- Gruppendesign

| Proband | Gruppe | Angst |
|---------|--------|-------|
| 1 | Real | 30 |
| 2 | Real | 35 |
| 3 | Real | 45 |
| 4 | Bild | 40 |
| 5 | Bild | 35 |
| 6 | Bild | 50 |

- Messwiederholungsdesign

| Proband | Messung 1 | Messung 2 |
|---------|-----------|-----------|
| 1 | 30 | 40 |
| 2 | 35 | 35 |
| 3 | 45 | 50 |

Konfidenzintervalle bei Messwiederholungsdesign

- Gruppendesign

- Individuelle Veranlagung Teil der Fehlerquelle und/oder Untersuchungsgegenstand

- Messwiederholungsdesign

- Zeitliche Entwicklung Untersuchungsgegenstand
 - Individuelle Veranlagung verzerrt Fehler
 - Eigentlich spezifischer als Gruppendesigns

→ Rechnerisch besteht kein Unterschied zwischen beiden Designs, obwohl einer bestehen sollte, deshalb...

Adjustierung

... bei Messwiederholungsdesigns die Messwerte mit der Veranlagung individueller Probanden normalisieren.

- 1. $pMean$ = Mittelwert pro Proband
- 2. Grand Mean berechnen (Pooled Mean) = Gesamtdurchschnitt
- 3. Pro Proband Adjustierungsfaktor $adj = Grand\ Mean - pMean$
- 4. Pro Proband: Adjustierter Wert = Wert + adj

Dadurch werden die Konfidenzintervalle der Messwiederholungen enger, das Modell also präziser bei gleichbleibenden Mittelwerten und Differenzen der Gruppen

Beispiel Adjustierung

- 1. $pMean$ = Mittelwert pro Proband
- 2. Grand Mean berechnen (Pooled Mean) = Gesamtdurchschnitt
- 3. Pro Proband Adjustierungsfaktor adj = Grand Mean - $pMean$
- 4. Pro Proband: Adjustierter Wert = Wert + adj

| Proband | Messung 1 | Messung 2 | $pMean$ | adj | Adj M1 | Adj M2 |
|---------|-----------|-----------|---------|-------|--------|--------|
| 1 | 30 | 40 | 35 | 4.167 | 34.167 | 44.167 |
| 2 | 35 | 35 | 35 | 4.167 | 39.167 | 39.167 |
| 3 | 45 | 50 | 47.5 | -8.5 | 36.5 | 41.5 |

$$\text{Grand Mean} = \frac{30+35+45+40+35+50}{6} = 39.167$$

Ergebnis der Adjustierung

- Differenzen zwischen Gruppen und deren Mittelwerte bleiben gleich
- Verschiebung pro Proband wird herausgerechnet
- → Konfidenzintervalle der Gruppen werden kleiner
- → höhere Spezifität
- Verringerter Standardfehler bei Signifikanztests

t-Test

Bereits verwendet:

- Korrelationskoeffizient signifikant ungleich 0?
- Regressionskoeffizient signifikant ungleich 0?

Jetzt:

- 2 Mittelwerte signifikant verschieden?

t-Test

Experiment mit 1 Prädiktor und 1 Outcome, Messung in 2 verschiedenen Gruppen (bzw. 1 Manipulation und 1 Kontrollgruppe)

- Ist der Film Scream 1 gruseliger als der Film Scream 2?
- Verbessert sich die Arbeit durch Musik?
- Verbessert sich die Arbeit durch Lieblingsmusik?

2 Arten von t-Tests

- Unabhängiger t-Test → Gruppendesign (verschiedene Probanden) Auch Independent-Measures oder Independent Samples t-Test
- Abhängiger t-Test → Messwiederholungsdesign (gleiche Probanden) Auch Matched Pairs oder Paired Samples t-Test

Grundprinzip

Student (1908): *The Probable Error of a Mean*

- 2 Stichproben gesammelt und Mittelwerte berechnet
- $H_0 = 2$ Mittelwerte sind gleich / sehr ähnlich
- Ähnlichkeitstoleranz der Mittelwerte durch Variabilität der Werte bestimmt → bei großem Standardfehler sind große Unterschiede zwischen Mittelwerten typisch
- Je kleiner der Standardfehler und je größer die Unterschiede der Mittelwerte, desto sicherer ist H_0 falsch
- $t = \frac{\text{model}}{\text{error}} = \frac{\text{meandiff}_{\text{observed}} - \text{meandiff}_{\text{expected}}}{\text{standardfehler}}$

Unabhängiger t-Test (Gruppensdesign, verschiedene Probanden)

- $t = \frac{\text{model error}}{\text{standardfehler}} = \frac{\text{meandiff}_{\text{observed}} - \text{meandiff}_{\text{expected}}}{\text{standardfehler}}$
- $\rightarrow t = \frac{(\bar{x}_1 - \bar{x}_2) - (\bar{\mu}_1 - \bar{\mu}_2)}{\text{standardfehler}}$ // Unter H_0 gilt $\bar{\mu}_1 - \bar{\mu}_2 = 0$
- $= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (bei gleicher Gruppengröße)
- Bei unterschiedlich großen Gruppen muss s^2 gepooled werden (kann eigentlich immer getan werden, schadet nicht)
- $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ mit $s_p^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$
- $= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 * (\frac{1}{n_1} + \frac{1}{n_2})}}$
- t_{kr} folgt aus Tabelle mit $df = n_1 + n_2 - 2$
- $abs(t) > t_{kr} \rightarrow H_0$ wird abgewiesen \rightarrow Mittelwerte sind signifikant verschieden
- Achtung: R berechnet mit Welch's t-Test df anders (kommt ohne Homogenität der Varianzen aus), was zu Abweichungen führen kann

Beispiel

Zeichenlänge des Dokumententitels pro Autor

| Dokument | Autor 1 | Dokument | Autor 2 |
|----------|---------|----------|---------|
| 1 | 30 | 1 | 40 |
| 2 | 35 | 2 | 35 |
| 3 | 45 | 3 | 50 |

$$\overline{Autor1} = 36.667, \overline{Autor2} = 41.667$$

$$s_{Autor1}^2 = 58.333, s_{Autor2}^2 = 58.333$$

$$s_p^2 = \frac{(n_1-1)*s_1^2 + (n_2-1)*s_2^2}{n_1+n_2-2} = \frac{(2)*58.333 + (2)*58.333}{4} = 58.333$$

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2 * (\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{36.667 - 41.667}{\sqrt{58.333 * (\frac{1}{3} + \frac{1}{3})}} = -0.802$$

$abs(t) = 0.802 < t_{kr95}(df = n_1 + n_2 - 2 = 4) = 2.78 \rightarrow H_0$ kann nicht abgelehnt werden
 \rightarrow Mittelwertschwankung bei gegebener Wertevarianz zu erwarten, nicht statistisch signifikant

Unabhängiger t-Test in R

```
autor1<-c(30,35,45)
autor2<-c(40,35,50)
t.test(autor1,autor2)
t.test(autor1,autor2, alternative = c("two.sided", "less", "greater"), mu = 0,
       paired = FALSE, var.equal = FALSE, conf.level = 0.95,
       na.action = na.exclude)
```

Welch Two Sample t-test

```
data: autor1 and autor2
t = -0.80178, df = 4, p-value = 0.4676
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.31418  12.31418
sample estimates:
mean of x mean of y
 36.66667  41.66667
```

$p\text{-value} = 0.4676 > 0.05 \rightarrow H_0$ kann nicht abgewiesen werden

t-Test in R nach Andy Field

```
autor1<-c(30,35,45)
autor2<-c(40,35,50)

x1 <- mean(autor1)
x2 <- mean(autor2)
sd1 <- sd(autor1)
sd2 <- sd(autor2)
n1 <- length(autor1)
n2 <- length(autor2)

ttestfromMeans<-function(x1, x2, sd1, sd2, n1, n2)
{
  df<-n1 + n2 - 2
  poolvar <- (((n1-1)*sd1^2)+((n2-1)*sd2^2))/df
  t <- (x1-x2)/sqrt(poolvar*((1/n1)+(1/n2)))
  sig <- 2*(1-(pt(abs(t),df)))
  paste("t(df = ", df, ") = ", t, ", p = ", sig, sep = "")
}

ttestfromMeans(x1, x2, sd1, sd2, n1, n2)
```

Robuster Unabhängiger t-Test

- Welch's t-Test kann mit heterogenen Varianzen umgehen
Welch, B.L. (1947): *The Generalization of Student's Problem when Several Different Population Variances are Involved*
- Wilcox, R.R. (2005): *Robustness of Standard Tests*
→ R Funktionen *yuen()*, *yuenbt()*, *pb2gen()*

Abhängiger t-Test (Messwiederholungsdesign, gleiche Probanden)

- $t = \frac{\text{model}}{\text{error}} = \frac{\text{meandiff}_{\text{observed}} - \text{meandiff}_{\text{expected}}}{\text{standardfehler}}$
- $= \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$ // Unter H_0 gilt $\mu_D = 0$
- $= \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} = \frac{\text{Mittlere Differenz pro Proband}}{\frac{\text{Standardabweichung der Differenzen}}{\sqrt{\text{Gruppengröße}}}}$
- $df = n_{\text{group}} - 1$
- $\text{abs}(t) < t_{kr} \rightarrow H_0$ kann nicht abgelehnt werden.

Beispiel

Messung von Konzentrationsfähigkeit vor und nach Anhören von klassischer Musik

| Proband | Vor | Nach | Nach-Vor |
|---------|-----|------|----------|
| 1 | 30 | 40 | 10 |
| 2 | 35 | 35 | 0 |
| 3 | 45 | 50 | 5 |

$$\overline{Nach - Vor} = 5$$

$$s_{Nach-Vor} = 5$$

$$t = \frac{\overline{Nach - Vor}}{\frac{s_{Nach-Vor}}{\sqrt{n}}} = \frac{5}{\frac{5}{\sqrt{3}}} = 1.73205$$

$$t_{kr95}(df = n_{group} - 1 = 2) = 4.30 > t = 1.73205 \rightarrow H_0 \text{ kann nicht abgewiesen werden.}$$

Effektstärke und t-Test

- $r = \sqrt{\frac{t^2}{t^2+df}}$ Pearsons Korrelationskoeffizient

→ ± 0.5 : starker Einfluss

→ ± 0.3 : moderater Einfluss

→ ± 0.1 : schwacher Einfluss

Durchschnittlich zeigten die Probanden größere Angst vor echten Spinnen (mean=47.00, SE=3.18) als vor den Bildern von Spinnen (mean=40.00, SE=2.68). Dieser Unterschied war zwar nicht signifikant $t = -1.68, p > .05$, zeigte aber eine moderate Effektstärke mit $r = .34$.

Zusammenfassung

- Gruppendesign vs Messwiederholungsdesign
 - Messwiederholungsdesign hat stark reduzierte unsystematische Varianz (Fehler)
 - Adjustierung der Daten
- Unabhängiger t-Test für Gruppendesign
- Abhängiger t-Test für Messwiederholungsdesign
 - Berechnung
 - Interpretation
 - $t < t_{kr} \rightarrow H_0$ nicht widerlegt, Wertschwankung statistisch nicht signifikant
- Pearsons Korrelationskoeffizient als Effektstärke kann aus t bestimmt werden

DH - Beispiele

McKenna, A (2001): *Reflections on Form, Meaning, and Ideology in the Nausicaa Episode of Ulysses*

- Stylometrische Analyse als Literaturkritischer Input?
- t-Test zwischen den Episoden {1-11,15} und {12-14,16-17}

Lijffijt, J. et al (2016): *Significance testing of word frequencies in corpora*

- Metauntersuchung über die Eignung verschiedener Testverfahren (inkl. t-Test und einiger nichtparametrischer Verfahren) beim Vergleichen von Korpora