Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

t-Test in R

Bei obigem Experiment zur Tonhöhenwahrnehmung erhalten wir für eine weitere Versuchsperson bei $n = 10$ Reizdarbietungen folgende Einstellungen

$$1005, 993, 995, 1003, 996, 1004, 1010, 1001, 997, 992$$

Mit $\alpha = 0.05$ soll unter der Annahme unabhängiger Stichproben, sowie jeweils Normalverteilung mit identischen Varianzen, die Nullhypothese getestet werden, dass sich die Punkte subjektiver Gleichheit (PSEs) beider Versuchspersonen (also die Erwartungswerte) nicht unterscheiden

Die Funktion `t.test()` wird dann mit folgender Eingabe aufgerufen

```
x <- c(998, 1003, 1004, 993, 990, 997, 1001, 992, 1000, 995)
y <- c(1005, 993, 995, 1003, 996, 1004, 1010, 1001, 997, 992)
t.test(x, y, mu = 0, var.equal = TRUE, conf.level = 0.95)
```

t-Test in R

R liefert folgende Ausgabe

```
        Two Sample t-test

data:  x and y
t = -0.96009, df = 18, p-value = 0.3497
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.332964  2.732964
sample estimates:
mean of x mean of y
    997.3     999.6
```

Die Nullhypothese wird nicht abgelehnt, da der resultierende p-Wert *nicht kleiner als* die vorgegebene Irrtumswahrscheinlichkeit $\alpha = 0.05$ ist

Es lässt sich also kein Unterschied in den PSEs nachweisen

Bestimmung der Post-hoc-Teststärke

Für den mit den angegebenen Daten bei einem Stichprobenumfang von je $n = 10$ erhaltenen Effekt $\mu_y - \mu_x = \delta = 2.3\,\text{Hz}$ und der Schätzung $\hat{\sigma} = 5.36\,\text{Hz}$, sowie der Irrtumswahrscheinlichkeit $\alpha = 0.05$, bestimmt man die Power dann durch die Simulation

```
spool <- sqrt((var(x) + var(y)) / 2)     # gepoolte SD

pval <- replicate(2000, {
  x <- rnorm(n = 10, mean = 997.3, sd = spool)
  y <- rnorm(n = 10, mean = 999.6, sd = spool)
  t.test(x, y, mu = 0, var.equal = TRUE)$p.value
})
mean(pval < 0.05)
```

Die Power beträgt nur ca. 15% bzw. der Fehler 2. Art β ca. 85%

Vorsicht: Post-hoc-Power ist aber meist uninteressant und wird oft fehlinterpretiert!

Post-hoc-Teststärke

Eine, wie in obigem Beispiel, auf Grund eines beobachteten Effekts post-hoc berechnete Power ("observed power") liefert keine über den p-Wert hinaus gehende Information

Hoenig und Heisey (2001) zeigen, dass in diesem Fall eine eineindeutige Beziehung zwischen dem erhaltenen p-Wert und der Power besteht

"Advocates of observed power argue that there is evidence for the null hypothesis being true if statistical significance was not achieved despite the computed power being high at the observed effect size. (Usually, this is stated in terms of the evidence for the null hypothesis (no effect) being weak if observed power was low.) Observed power can never fulfill the goals of its advocates because the observed significance level of a test ("p value") also determines the observed power; for any test the observed power is a 1:1 function of the p value." (S. 20)

Der beobachtete Effekt und die Standardabweichung liefern lediglich Anhaltspunkte für die Planung zukünftiger Experimente

Betrachtung von Konfidenzintervallen als Alternative

"From a pedagogic point of view, it is interesting to compare the inference one would obtain from consideration of confidence intervals to that obtained from the power analysis approach. Confidence intervals have at least two interpretations. One interpretation is based on the equivalence of confidence intervals and hypothesis tests. That is, if a confidence interval does not cover a hypothesized parameter value, then the value is refuted by the observed data. Conversely, all values covered by the confidence interval could not be rejected; we refer to these as the set of nonrefuted values. If the nonrefuted states are clustered tightly about a specific null value, one has confidence that nature is near the null value. If the nonrefuted states range widely from the null, one must obviously be cautious about interpreting the nonrejection as an indication of a 'near-null' state. The more widely known interpretation is that confidence intervals cover the true value with some fixed level of probability. Using either interpretation, the breadth of the interval tells us how confident we can be of the true state of nature being close to the null." (Hoenig & Heisey, 2001, S. 22)

Teststärke und Versuchsplanung

Power dient der Planung zukünftiger Experimente

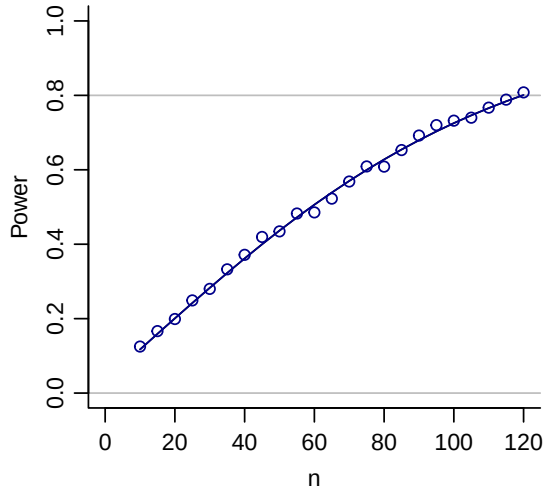
Bestimmung des Stichprobenumfangs n pro Gruppe, um mit einer Power von ca. 80% einen Effekt $\delta = 2\,\text{Hz}$ bei $\sigma = 5.5\,\text{Hz}$ und Irrtumswahrscheinlichkeit $\alpha = 0.05$ zeigen zu können

```
n <- 120                    # Stichprobenumfang pro Gruppe

pval <- replicate(2000, {
  x <- rnorm(n, mean = 997 + 2, sd = 5.5)
  y <- rnorm(n, mean = 997,     sd = 5.5)
  t.test(x, y, mu = 0, var.equal = TRUE)$p.value
})
mean(pval < 0.05)
```

Teststärke und Versuchsplanung

Es ergibt sich für den betrachteten Fall folgende Abhängigkeit der Power von der Stichprobengröße n pro Gruppe



Vergleich von Erwartungswerten (Welch-t-Test bei zwei unabhängigen Stichproben)

Annahmen

- Unabhängige Stichproben $X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ und $Y_1, \ldots, Y_m \sim N(\mu_y, \sigma_y^2)$, Varianzen σ_x^2, σ_y^2 unbekannt (keine Varianzhomogenität gefordert)

Hypothesen

1. $H_0\colon\ \mu_x - \mu_y = \delta_0 \qquad H_1\colon\ \mu_x - \mu_y \neq \delta_0$
2. $H_0\colon\ \mu_x - \mu_y \geq \delta_0 \qquad H_1\colon\ \mu_x - \mu_y < \delta_0$
3. $H_0\colon\ \mu_x - \mu_y \leq \delta_0 \qquad H_1\colon\ \mu_x - \mu_y > \delta_0$

Teststatistik

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

Welch-t-Test bei zwei unabhängigen Stichproben

Verteilung der Teststatistik unter $\mu_x - \mu_y = \delta_0$

- $t(k)$-verteilt mit

$$k = (S_x^2/n + S_y^2/m)^2 \; / \left(\frac{1}{n-1} \left(\frac{S_x^2}{n} \right)^2 + \frac{1}{m-1} \left(\frac{S_y^2}{m} \right)^2 \right)$$

k ist eine reelle Zahl, die zu einer ganzen Zahl abgerundet werden kann

Ablehnungsbereich

1. $|T| > t_{1-\alpha/2}(k)$
2. $T < t_\alpha(k) = -t_{1-\alpha}(k)$
3. $T > t_{1-\alpha}(k)$

Welch-t-Test in R

Wir betrachten die Daten der beiden Versuchspersonen aus dem Experiment zur Tonhöhenwahrnehmung

998,1003,1004,993,990,997,1001,992,1000,995
1005,993,995,1003,996,1004,1010,1001,997,992

Mit $\alpha = 0.05$ soll unter der Annahme unabhängiger, normalverteilter Stichproben (ohne Varianzgleichheit), die Nullhypothese getestet werden, dass sich die Punkte subjektiver Gleichheit beider Versuchspersonen nicht unterscheiden

Die Funktion `t.test()` wird dann mit folgender Eingabe aufgerufen

```
x <- c(998, 1003, 1004, 993, 990, 997, 1001, 992, 1000, 995)
y <- c(1005, 993, 995, 1003, 996, 1004, 1010, 1001, 997, 992)
t.test(x, y, mu = 0, var.equal = FALSE, conf.level = 0.95)
```

Welch-t-Test in R

R liefert folgende Ausgabe

```
    Welch Two Sample t-test

data:  x and y
t = -0.9601, df = 17.242, p-value = 0.3503
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.348872  2.748872
sample estimates:
mean of x mean of y
    997.3     999.6
```

Die Nullhypothese kann nicht abgelehnt werden

Es lässt sich kein Unterschied in den PSEs nachweisen, wegen der geringen Power (siehe oben) lässt sich aber ein großer Bereich von (teilweise unplausiblen) Werten nicht verwerfen

Approximative Tests

Im Falle zweier unabhängiger Stichproben kann, wie bei Ein-Stichproben-Tests, auf die Annahme einer unterliegenden Normalverteilung verzichtet werden, wenn die Stichprobenumfänge n und m hinreichend groß sind

- $X_1, \ldots, X_n$ sowie $Y_1, \ldots, Y_m$ seien unabhängig und identisch verteilt bei $n \geq 30$ und $m \geq 30$, Varianzen σ_x^2, σ_y^2 unbekannt

Dann ist die oben definierte Teststatistik T des Zwei-Stichproben-t-Tests bei Varianzhomogenität approximativ $t(n+m-2)$-verteilt und bei fehlender Varianzhomogenität approximativ $t(k)$-verteilt (Bestimmung der Anzahl der Freiheitsgrade k wie oben)

Vergleich von Erwartungswerten (t-Test bei zwei abhängigen Stichproben)

Im Gegensatz zu den bisher betrachteten unabhängigen Stichproben bestehen bei *abhängigen Stichproben* (oder *verbundenen Stichproben*) die Daten aus Paaren von Merkmalen

$$\begin{array}{c} X_1,\ Y_1 \\ \vdots \\ X_i,\ Y_i \\ \vdots \\ X_n,\ Y_n \end{array}$$

wobei die verschiedenen Paare (X_i, Y_i), für $i = 1, \ldots, n$, voneinander stochastisch unabhängig sind

Kennzeichnend für verbundene Stichproben ist, dass eine eineindeutige Zuordnung zwischen den Werten der beiden Stichproben existiert

Verbundene Stichproben

Derartige eineindeutige Zuordnungen resultieren beispielsweise aus Within-Subjects-Designs (Messwiederholungs-Designs)

Beispiele

- Versuchsperson 1 liefert unter einer experimentellen Bedingung A eine Realisierung der Zufallsvariablen X_1 und unter einer zweiten Bedingung B eine Realisierung von Y_1, Versuchsperson 2 liefert unter Bedingung A eine Realisierung von X_2 und unter Bedingung B eine Realisierung von Y_2, ...
- X_i und Y_i sind Test- und Retest-Wert des Probanden i
- "Matched Samples"

Die Zuordnung wird durch eine Untersuchungseinheit vermittelt, die etwa eine Versuchsperson oder ein Reiz sein kann

Differenzwerte

Auch in diesem Fall interessieren wir uns für den Vergleich der Erwartungswerte $E(X_1) = \ldots E(X_n) = \mu_x$ und $E(Y_1) = \ldots E(Y_n) = \mu_y$

Hierzu betrachten wir die Differenzen

$$D_i = X_i - Y_i$$

für alle $i = 1, \ldots, n$, die stochastisch unabhängig sind

Mit der Betrachtung der Differenzen $D_1, \ldots, D_n$ sind wir in einer Situation, die der einer einzigen Stichprobe entspricht

- Es können verschiedene Annahmen hinsichtlich der den Differenzen unterliegenden Verteilung gemacht werden (z. B. normalverteilt mit (un-)bekannter Varianz)

t-Test bei zwei abhängigen Stichproben

Annahmen

- $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$ Stichproben mit $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ unabhängig und identisch $N(\mu_D, \sigma_D^2)$-verteilt mit unbekannter Varianz σ_D^2

Folgerungen

$$\hat{\mu}_D = \bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{X} - \bar{Y}$$

und

$$\begin{aligned}\hat{\sigma}_D^2 = S_D^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left((X_i - Y_i) - (\bar{X} - \bar{Y})\right)^2 \\ &= S_X^2 + S_Y^2 - 2 \cdot S_{XY}\end{aligned}$$

t-Test bei zwei abhängigen Stichproben

Hypothesen

1. $H_0\colon\ \mu_D = \delta_0 \qquad H_1\colon\ \mu_D \neq \delta_0$
2. $H_0\colon\ \mu_D \geq \delta_0 \qquad H_1\colon\ \mu_D < \delta_0$
3. $H_0\colon\ \mu_D \leq \delta_0 \qquad H_1\colon\ \mu_D > \delta_0$

Teststatistik

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S_D}\sqrt{n}$$

Verteilung der Teststatistik unter $\mu_D = \delta_0$

- $t(n-1)$-Verteilung

t-Test bei zwei abhängigen Stichproben

Ablehnungsbereich

1. $|T| > t_{1-\alpha/2}(n-1)$
2. $T < t_\alpha(n-1) = -t_{1-\alpha}(n-1)$
3. $T > t_{1-\alpha}(n-1)$

mit dem p-Quantil $t_p(n-1)$ der $t(n-1)$-Verteilung

Bemerkung

- Entsprechend verfährt man bei Vorliegen der alternativen, für den Ein-Stichproben-Fall besprochenen Verteilungsannahmen bezüglich der Differenzen $D_i = X_i - Y_i$, $i = 1, \ldots, n$

Beispiel

Es soll geprüft werden, ob zwei Formen X und Y eines Intelligenztests "parallel" sind

Die IQ-Werte werden als (bivariat) normalverteilt angenommen mit $\sigma_x = \sigma_y$ und einer Korrelation (Reliabilität) von $\rho = 0.9$

Allgemein gilt für die Differenzen

$$\sigma_D^2 = \sigma_x^2 + \sigma_y^2 - 2 \cdot \rho \cdot \sigma_x \cdot \sigma_y$$

Zunächst wird der Stichprobenumfang n bestimmt, mit dem ein Unterschied $\mu_x - \mu_y = \mu_D$ von 7 IQ-Punkten mit einer Power von mindestens 80% gezeigt werden kann

Dies kann per Simulation auf zwei verschiedene Weisen geschehen

Beispiel: Teststärke und Versuchsplanung

Bestimmung des Stichprobenumfangs n, um mit einer Power von mindestens 80% einen Effekt von $\mu_D = 7$ IQ-Punkten bei $\sigma_x = \sigma_y = 15$ und $\alpha = 0.1$ zeigen zu können, $H_0\colon\ \mu_D = 0$

Alternative 1: Bivariate Normalverteilung von X_i und Y_i

```
n <- 8                                          # Stichprobenumfang
s <- 15; r <- 0.9; sxy <- r*s*s

pval <- replicate(2000, {
  x <- MASS::mvrnorm(n,
            mu = c(100 + 7, 100),               # Erwartungswertvektor
         Sigma = matrix(c(s^2, sxy,             # Kovarianzmatrix
                          sxy, s^2), 2, 2)
  )
  t.test(x[, 1], x[, 2], mu = 0, paired = TRUE)$p.value
})
mean(pval < 0.1)
```

Beispiel: Teststärke und Versuchsplanung

Bestimmung des Stichprobenumfangs n, um mit einer Power von mindestens 80% einen Effekt von $\mu_D = 7$ IQ-Punkten bei $\sigma_x = \sigma_y = 15$ und $\alpha = 0.1$ zeigen zu können, H_0: $\mu_D = 0$

Alternative 2: Normalverteilung der Differenzen $D_i = X_i - Y_i$

```
n <- 8                                          # Stichprobenumfang
s <- 15; r <- 0.9; sxy <- r*s*s

pval <- replicate(2000, {
  d <- rnorm(n, mean = 7, sd = sqrt(s^2 + s^2 - 2*sxy))
  t.test(d, mu = 0)$p.value
})
mean(pval < 0.1)
```

Beispiel

Anschließend bearbeitet eine Stichprobe von insgesamt acht Probanden die beiden "parallelen" Formen X und Y des Intelligenztests

Man erhält folgende IQ-Werte

x	97	87	116	137	88	95	111	87
y	97	85	114	134	102	103	108	92

Für die Differenzwerte $d_i = x_i - y_i$ für alle $i = 1, \ldots, 8$ gilt dann

$$0, 2, 2, 3, -14, -8, 3, -5$$

Geprüft werden soll für $\alpha = 0.10$ die Hypothese, dass die beiden Testformen denselben IQ-Wert liefern

Beispiel

Man berechnet

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{8}(0+2+2+3-14-8+3-5) = -2.125$$

und

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{7}\cdot 274.875 = 39.268$$

Für die Teststatistik erhält man dann mit $\delta_0 = 0$

$$t = \frac{\bar{d}}{s_d}\sqrt{n} = \frac{-2.125}{\sqrt{39.268}}\sqrt{8} = -0.959$$

Beispiel

Als kritischen Wert erhält man aus einer Tabelle der t-Verteilung bei $df = 7$ Freiheitsgraden das Quantil $t_{1-\alpha/2} = t_{0.95}(7) = 1.89$

Wegen $|t| = 0.959 \leq 1.89$ wird die Nullhypothese beibehalten

Die vorgenommene Berechnung wird in R realisiert, indem in der Funktion `t.test()` die Option `paired = TRUE` gewählt wird

```
x <- c(97, 87, 116, 137, 88, 95, 111, 87)
y <- c(97, 85, 114, 134, 102, 103, 108, 92)
t.test(x, y, mu = 0, paired = TRUE, conf.level = 0.9)
```

Beispiel

Als Ausgabe in R erhält man

```
    Paired t-test

data:  x and y
t = -0.9591, df = 7, p-value = 0.3694
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -6.322457  2.072457
sample estimates:
mean of the differences
                 -2.125
```

Die Nullhypothese wird beibehalten, ein Unterschied zwischen den beiden Formen kann nicht nachgewiesen werden

Allerdings hatte die Studie für kleinere Effekte als $\mu_D = 7$ IQ-Punkte auch nicht genügend Power

Vergleich der Varianzen zweier Normalverteilungen

Annahmen

- Unabhängige Stichproben $X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ und $Y_1, \ldots, Y_m \sim N(\mu_y, \sigma_y^2)$, Varianzen σ_x^2, σ_y^2 unbekannt

Hypothesen

1. $H_0\colon\ \sigma_x^2 = \sigma_y^2 \qquad H_1\colon\ \sigma_x^2 \neq \sigma_y^2$
2. $H_0\colon\ \sigma_x^2 \geq \sigma_y^2 \qquad H_1\colon\ \sigma_x^2 < \sigma_y^2$
3. $H_0\colon\ \sigma_x^2 \leq \sigma_y^2 \qquad H_1\colon\ \sigma_x^2 > \sigma_y^2$

Teststatistik

$$T = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$$

Vergleich der Varianzen zweier Normalverteilungen

Verteilung der Teststatistik unter $\sigma_x^2 = \sigma_y^2$

- $F(n-1, m-1)$-Verteilung

Ablehnungsbereich

1. $T < f_{\alpha/2}(n-1, m-1)$ oder $T > f_{1-\alpha/2}(n-1, m-1)$
2. $T < f_{\alpha}(n-1, m-1)$
3. $T > f_{1-\alpha}(n-1, m-1)$

mit dem p-Quantil $f_p(n-1, m-1)$ der $F(n-1, m-1)$-Verteilung

Beispiel

Wir betrachten die Daten der beiden Versuchspersonen aus dem Experiment zur Tonhöhenwahrnehmung

998,1003,1004,993,990,997,1001,992,1000,995
1005,993,995,1003,996,1004,1010,1001,997,992

Mit $\alpha = 0.05$ soll unter der Annahme unabhängiger, normalverteilter Stichproben die Nullhypothese getestet werden, dass sich die Varianzen der Werte beider Versuchspersonen nicht unterscheiden (d. h. $\sigma_x^2/\sigma_y^2 = 1$)

Die Funktion `var.test()` wird dann mit folgender Eingabe aufgerufen

```
x <- c(998, 1003, 1004, 993, 990, 997, 1001, 992, 1000, 995)
y <- c(1005, 993, 995, 1003, 996, 1004, 1010, 1001, 997, 992)
var.test(x, y, ratio = 1, conf.level = 0.95)
```

Beispiel

Es resultiert folgende Ausgabe

```
    F test to compare two variances

data:  x and y
F = 0.6533, num df = 9, denom df = 9, p-value = 0.536
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1622777 2.6302990
sample estimates:
ratio of variances
          0.653329
```

Die Hypothese $\sigma_x^2 = \sigma_y^2$ (bzw. $\sigma_x^2/\sigma_y^2 = 1$) kann nicht abgelehnt werden

Ebenso wenig lassen sich aber Werte zwischen $\sigma_x^2/\sigma_y^2 = 1/6$ und $\sigma_x^2/\sigma_y^2 = 2.6$ zurückweisen

Vergleich der Varianzen zweier Normalverteilungen

Ablehnungsbereich

