Code ▾

Hypothesentest t-Test für unabhängige Stichproben

1) Grundlegende Konzepte: Was ist t-Test für unabhängige Stichproben?

Der t-Test für unhängige Stichproben testet, ob die Mittelwerte zweier unahängiger Stichproben verschieden sind.

Von unahängigen Stichproben wird gesprochen, wenn ein Messwert in einer Stichprobe und ein bestimmter Messwert in einer anderen Stichprobe sich gegenseitig nicht beeinflussen. Der t-Test eignet sich damit zur Untersuchung einer Unterschiedshypothese zwischen zwei unabhängigen Stichproben. Die zu testende Variable sollte dabei intervallskaliert und normalverteilt sein. Im Gegensatz zum t-Test für verbundene Stichproben müssen jedoch die Stichproben nicht gleich gross sein.

Der t-Test für zwei unabhängige Stichproben findet dann Verwendung, wenn der Frage nachgegangen werden soll, ob sich zwei Gruppen nach einem bestimmten Merkmal (Konsumverhalten, soziale Einbindung, Stressresistenz, Einstellung gegenüber Sachfragen etc.) unterscheiden oder ob eine Experimentalgruppe gegenüber einer Kontrollgruppe andere Ergebnisse (Behandlungserfolg, Auswirkungen bestimmter Tests etc.) aufweist. So eignet sich dieser t-Test für Fragen wie „Unterscheidet sich das Konsumverhalten von Männer und Frauen?“ oder “Reagieren die Experimentalgruppe und die Kontrollgruppe unterschiedlich auf die Behandlungsmethode?“

2) Aufgabenstellung - Beschreibung

Es wurden 1753 Absolventinnen und Absolventen, die sich in Ihrem ersten Jahr nach Ihren Abschluss in einer festen vollzeit Anstellung befinden bezüglich Ihrem jähr. Bruttogehalt in Euro befragt. Ziel dieser Studie ist es herauszufinden, ob es einen signifikanten Unterschied in den jährl. Bruttogehälter zwischen Absolventinnen und Absolventen gibt.

3) Hypothese

H0: Es gibt keinen signifikanten Unterschied in den Mittelwerten zwischen den Gehältern von Absolventinnen und Absolventen.

$$M_M = M_F$$

H1: Es gibt einen signifikanten Unterschied in den Mittelwerten zwischen den Gehältern von Absolventinnen und Absolventen.

$$M_M \neq M_F$$

Variable 1 = Gender -> Geschlecht UV

Variable 2 = Salary -> Gehalt AV

4) Voraussetzungen des t-Tests für unabhängige Stichproben

Die abhängige Variable ist mindestens intervallskaliert -> Erfüllt -> Gehalt ist ratioskalier.

Es liegt eine unabhängige Variable vor, mittels der die beiden zu vergleichenden Gruppen gebildet werden-> Erfüllt -> Geschlecht ist Dichotome Variable, welche aus zwei Ausprägungen besteht (männlich und weiblich).

Die einzelnen beobachteten Messwerte (Gehalt) innerhalb und zwischen den Gruppen sind voneinander unabhängig.

- Unabhängigkeit innerhalb der Gruppen: Die Absolventen und Ablsoventinnen in der jeweiligen Gruppe, haben untereinander keinen Einfluss aufeinander um so die Messswerte zu verzerren.–> Erfüllt -> Es wurde sichergestellt, dass die einzelnen Absolventinnen und Absolventen innerhalb der jeweiligen Gruppen keinen Kontakt zueinander hatten.

- Unabhängigkeit zwischen den Gruppen: Die beiden Gruppen “Absolventen” und “Ablsoventinnen” sind unabhängig von einander befragt wurden. D.h. ein Absolvent(männlich) wird explizit in die männlich-Gruppe eingeordnet und taucht nicht nocheinmal in der weiblich-Gruppe auf.

Befinden sich signifikante Ausreißer in den Daten: Signifikante Ausreißer können die Ergebnisse und statistiken verfälschen und somit zu falschen Schlussfolgerungen führen. -> Es gibt keine extremen Ausreißer -> siehe Ausreißer Sektion.

- Die abhängige Variable (Gehalt) ist in den Grundgesamtheiten der beiden Gruppen normalverteilt -> Erfüllt. Für weitere Erklärung, bitte siehe Histogramm.

Varianzhomogenität zwischen den beiden Gruppen sind annähernd identisch-> Das ist nicht gegeben, daher wurde ein Welch-Test durchgeführt.

Libraries:

Hide

```
library(dplyr) #->  Gruppierung der Daten
library(ggplot2) #-> Diagramm
library(car) #-> Prüfung auf Varianzhomogenität
library(effsize)# -> Cohen´s d
```

DatenImport:

Hide

```
library(readxl)
StudentenGehalt <- read_excel("C:/Data_Science_Projects/Repository_2/Aufgabe4/StudentenGehal
t.xlsx")
View(StudentenGehalt)
```

Hide

```


library(dplyr) #-> Rename columns
StudentenGehalt <- StudentenGehalt %>% #pipe
           rename(Geschlecht = 'gender',
                  Gehalt   = 'salary')
```

Hide

```
StudentenGehalt <- StudentenGehalt %>%
  mutate(Geschlecht   = factor(Geschlecht, levels = c("Female", "Male"), labels = c("weiblic
h", "männlich")))
```

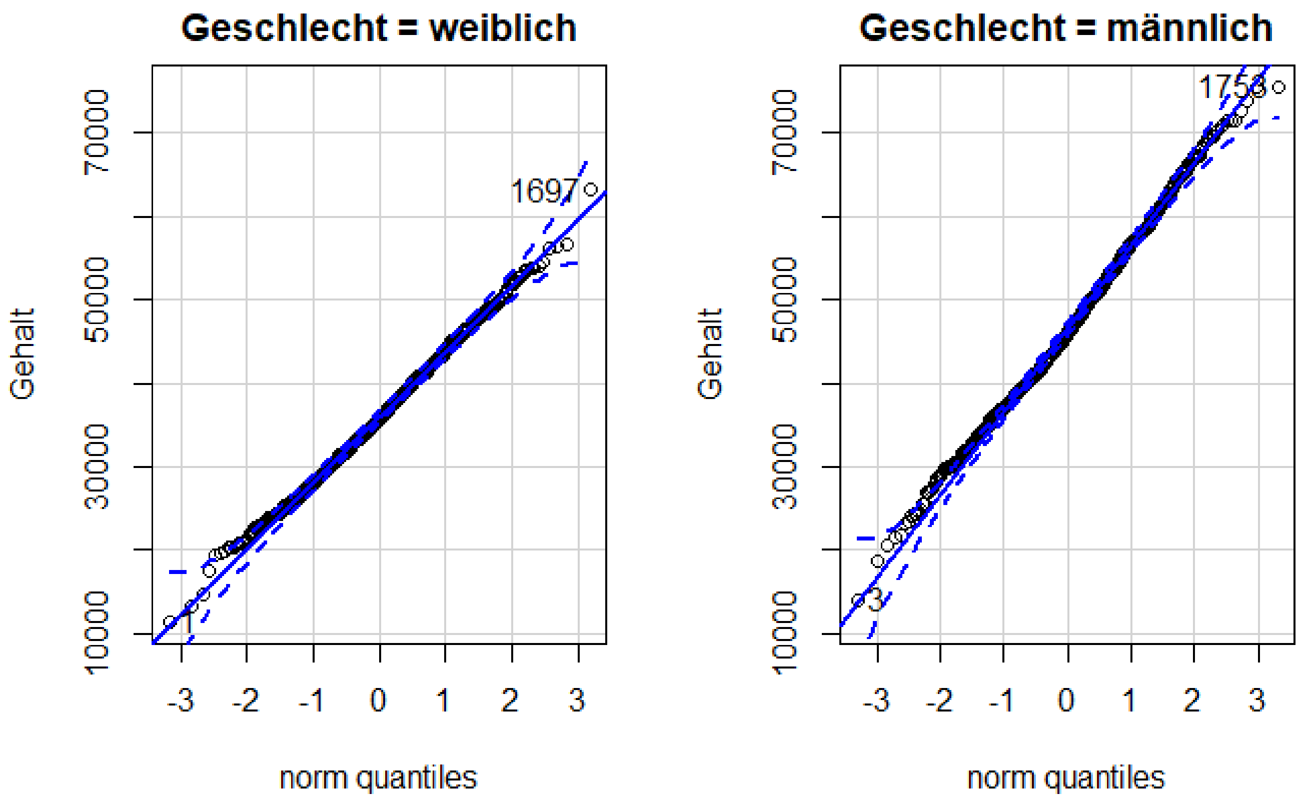
Voraussetzungsprüfung

Prüfung der Normalverteilung der abhängigen Variable(Gehalt)

Hide

```
library(car)

qqPlot(Gehalt ~ Geschlecht, data=StudentenGehalt,
       layout=c(1, 2))
```



Anhand der beiden QQ Plots sieht man, dass sich die empirische Verteilung (schwarze Datepunkte) bei weiblich und maennlich nahe an der theoretischen Normalverteilung (die gerade Linie (blau) liegen. Abgesehen von ein paar Ausreißern am Ende der Verteilung kann man eine sehr gute Normalverteilung des abhängigen Variables(Gehalt) annehmen.

Histrogramme

Hide

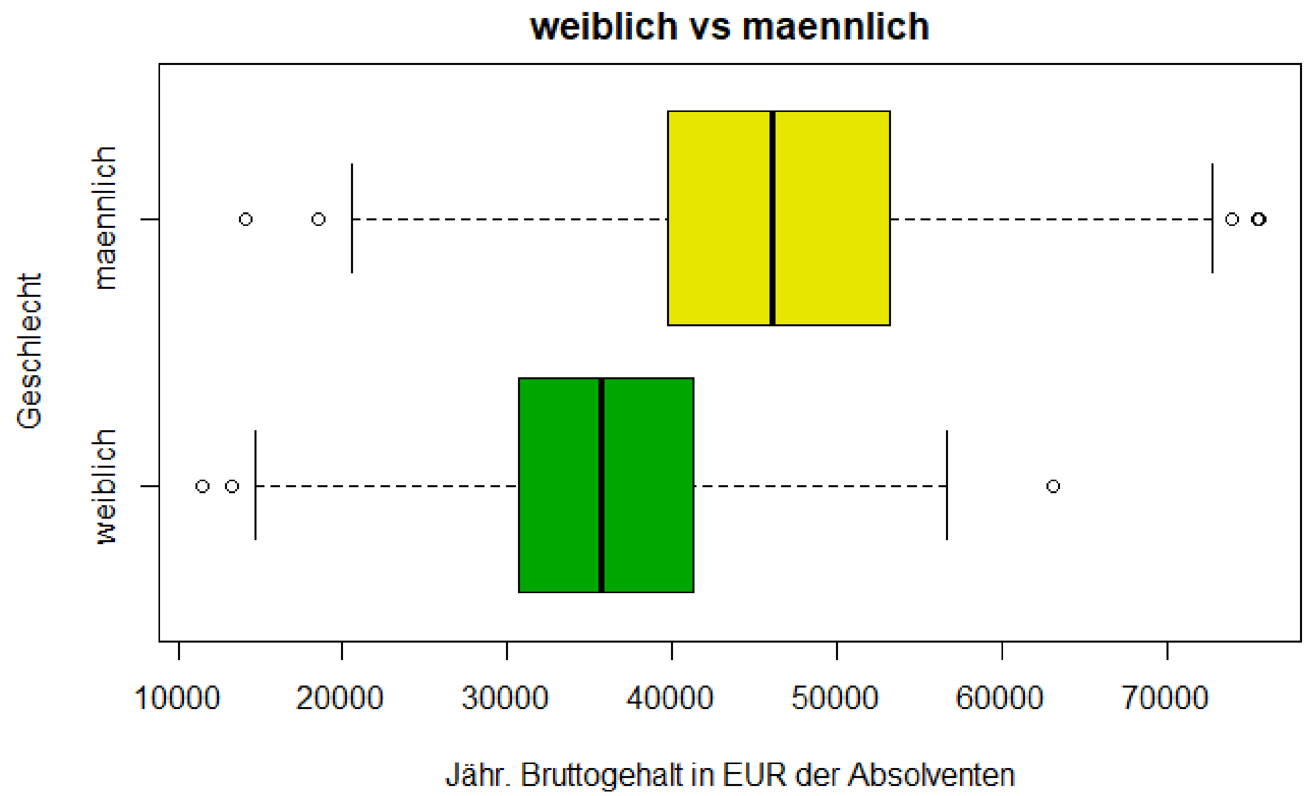
```
library(ggplot2)
StudentenGehalt %>%
  group_by(Geschlecht) %>%
  ggplot(aes(Gehalt, color=Geschlecht)) +
  geom_histogram(aes(fill = Geschlecht), bins = 20) +
  facet_wrap(~Geschlecht) +
  theme_grey()+
  labs(x= "Gehalt",y = "Anzahl" )
```

Auch anhand der Histrogramme für weiblich und männlich wird unsere Annahme einer normalverteilten abhängigen Variable bestätigt, da wir hier eine gute symmetrische Verteilung, mit abfallenden Häufigkeiten zu den Rändern.

5) Deskriptive Statistiken:

Hide

```
boxplot(StudentenGehalt$Gehalt~StudentenGehalt$Geschlecht,
        col=terrain.colors(4),
        horizontal = T,
        names = c('weiblich', 'maennlich'),
        ylab = 'Geschlecht',
        xlab = 'Jähr. Bruttogehalt in EUR der Absolventen',
        main = "weiblich vs maennlich")
```



Bei dem Boxplot ist es ersichtlich, dass wir bei den maennlich sowie weiblich ein paar Ausreißer haben, die jedoch nicht schwerwiegend ins Gewicht fallen und somit nicht zu einer signifikanten Verzerrung unserer Messergebnisse führen sollten. Darüber hinaus wird anhand des Boxplots erkennbar, dass der Median der Absolventen (maennlich) größer ist als der der Absolventinnen (weiblich) bezüglich des jährlichen Bruttogehalts.

Hide

```
library(rstatix)
```

```
Paket 惏柂rstatix惏柞 wurde unter R Version 4.0.4 erstellt
Attache Paket: 惏柂rstatix惏柞

The following object is masked from 惏柂package:stats惏柞:

    filter
```

Hide

```
StudentenGehalt %>%
  group_by(Geschlecht) %>%
  identify_outliers(Gehalt)
```

Geschlecht <fctr>	Gehalt <dbl>	is.outlier <lgl>	is.extreme <lgl>
weiblich	11444.14	TRUE	FALSE
weiblich	13300.11	TRUE	FALSE
weiblich	63154.33	TRUE	FALSE
männlich	14081.10	TRUE	FALSE

Geschlecht <fctr>	Gehalt <dbl>	is.outlier <lgl>	is.extreme <lgl>
männlich	18571.24	TRUE	FALSE
männlich	73980.21	TRUE	FALSE
männlich	75530.39	TRUE	FALSE
männlich	75596.79	TRUE	FALSE

8 rows

Definition:

Q3: 2. Quantil Q1: 1. Quantil IQR: interquartile range (IQR = Q3 - Q1)

is.outlier: Werte die ober der "Q3 + 1.5xIQR" oder unter der "Q1 - 1.5xIQR" Quantil liegen, werden als Ausreißer betrachtet.

is.extreme: Werte die ober der "Q3 + 3xIQR" oder unter der "Q1 - 3xIQR" Quantil liegen, werden als Ausreißer betrachtet.

Auch bei der Ausreißerprüfung wird es deutlich, dass wir keine extreme Ausreißer in den beiden Gruppen haben.

Das Mindestgehalt einer Absolventin liegt bei 11444,14 EUR, wobei das höchste Gehalt eines Absolventen 75596,79 EUR beträgt.

Hide

```
library(dplyr) -> Gruppierung
StudentenGehalt %>%
group_by(Geschlecht) %>%
  summarize(Anzahl = n(), Mittelwert = mean(Gehalt), Median = median(Gehalt), Standardabweich
ung = sd(Gehalt)) %>%
  mutate_if(is.numeric, round, 2)
```

	Geschlecht <fctr>	Anzahl <dbl>	Mittelwert <dbl>	Median <dbl>	Standardabweichung <dbl>
1	weiblich	674	36018.46	35694.83	7729.23
2	männlich	1079	46584.63	46008.15	9657.67

2 rows

Es zeigt sich für diese Fragestellung einen Mittelwertsunterschied. Das jährliche Bruttogehalt bei den Absolventen (maennlich) ist höher (M=46584.63 ,SD=9657.67 , n=1079) als bei den Absolventinnen(weiblich) (M = 36018.46, SD = 7729.23, n = 674). Dieser Mittelwertsunterschied zw. Absolventinnen und Absolventen bezüglich des jährl. Bruttogehalts, deckt sich mit den Erkenntnissen, die wir in den grafischen Darstellungen (z.b. Boxplot, Histrogramm) erhalten haben.

6) Test auf Varianzhomogenität (Levene-Test)

Hide

```
library(car)# Prüfung auf Varianzhomogenität

leveneTest(StudentenGehalt$Gehalt, StudentenGehalt$Geschlecht, center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)    
group    1  35.239 3.508e-09 ***
      1751                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Da der t-Test für unabhängige Gruppen Varianzhomogenität (gleiche Varianzen) voraussetzt, wird anhand des Levene-Test bestimmt. Liegt Varianzheterogenität vor (also unterschiedliche Varianzen), so müssen unter anderem die Freiheitsgerade des t-Wertes angepasst werden. In solch einem Fall, wird der Welch-Test durchgeführt, die dann die fehlende Homogenität der Varianzen durch die Anpassung der Freiheitsgrade und des t-empirsch-Wertes korrigiert.

Der Levene-Test verwendet die Nullhypothese: “Die beiden Varianzen sind nicht unterschiedlich”. Alternativhypothese ist somit: “Die beiden Varianzen sind unterschiedlich”.

Anhand des p-values von da 3.508e-09, das kleiner ist als 0,05 ist der Levene-Test signifikant und es kann daher von Varianzheterogenität ausgegangen werden. Daher können wir nicht von gleichen Varianzen ausgehen($F(1,1751)=35.239$, $p=3.508e-09$).

Da Varianzheterogenität vorliegt, wird ein Welch t-test durchgeführt.

7) Welch-Korrektur - ungerichtete Hypothese

Hide

```
mit_welch_Ungerichtet<- t.test(StudentenGehalt$Gehalt~StudentenGehalt$Geschlecht, var.eq = F,
con= 0.95, alt = "two.sided")
mit_welch_Ungerichtet
```

```


	Welch Two Sample t-test

data:  StudentenGehalt$Gehalt by StudentenGehalt$Geschlecht
t = -25.252, df = 1647.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11386.870  -9745.475
sample estimates:
mean in group weiblich mean in group männlich 
              36018.46               46584.63 
```

Das ist ein zweiseitiges/ungerichtetes Welch-Test. Die Teststatistik beträgt $t = -25.252$ und der zugehörige Signifikanzwert $p < 2.2e-16$, das somit kleiner ist als 0,05. Damit ist der Unterschied signifikant: Die Mittelwerte zwischen Absolventen und Absolventinnen bezüglioch des jährlichen Gehaltes unterscheiden sich ($t(1647.5) = -25.252$, $p < 2.2e-16$).

8) Berechnung der Effektstärke bei ungleichgroßen Gruppen

Hide

```
library(effsize)
cohen.d(d = StudentenGehalt$Gehalt, f= StudentenGehalt$Geschlecht)
```

```


Cohen's d

d estimate: -1.178514 (large)
95 percent confidence interval:
    lower     upper
-1.282419 -1.074608
```

Interpretation von d nach Cohen (1988):

$$\begin{aligned} \text{Schwacher Effekt: } 0.20 &< ||d|| < 0.50 \\ \text{Schwacher bis mittlerer Effekt: } 0.50 &= ||d|| \\ \text{Mittlerer Effekt: } 0.50 &< ||d|| < 0.80 \\ \text{Mittlerer bis starker Effekt: } 0.80 &= ||d|| \\ \text{Starker Effekt: } 0.80 &< ||d|| \end{aligned}$$

Damit entspricht eine Effektstaerke von 1.178514 einem starken Effekt. Hier wird der Cohens.d verwendet, weil wir hier zwei verschiedene Gruppengrößen haben Absolventen(maennlich)=1079 und Absolventinnen(weiblich)=674.

9) Eine Aussage

Absolventen(maennlich) verdienen gemessen an dem jährlichen Bruttogehalt nach einem Jahr signfikant mehr (M=46584.63 ,SD=9657.67 , n=1079) als die Absolventinnen(weiblich) (M = 36018.46, SD = 7729.23, n = 674). Die Effektstärke liegt bei r = 1.178514 und entspricht damit einem starken Effekt nach Cohen.d (1988). H0 kann verworfen werden.