

Einführung in R

Clemens Brunner

18.-19.10.2019

Mittelwertsvergleich

Nicht nur die Beziehungen zwischen zwei (oder mehreren) Variablen sind in Studien interessant (welche man beispielsweise mit Korrelationen beschreiben kann). Oft ist man auch an mittleren Unterschieden zwischen zwei (oder mehreren) Gruppen interessiert.

Im Gegensatz zu korrelativen Studiendesigns kann man mit experimentellen Studiendesigns auch auf Kausalität rückschließen. In experimentellen Designs ist man daher oft an Mittelwertsunterschieden interessiert, welche auf eine konkrete Behandlung zurückzuführen sind.

Prinzipiell unterscheidet man bei experimentellen Designs zwischen abhängigen und unabhängigen Designs. In unabhängigen Designs sind die Individuen in den einzelnen Gruppen voneinander unabhängig (z.B. verschiedene Personen). In abhängigen Designs sind die Individuen zwar innerhalb der Gruppen unabhängig, aber zwischen den Gruppen besteht eine Abhängigkeit (z.B. befinden sich dieselben Personen in mehreren Gruppen). Abhängige Designs berücksichtigen individuelle Unterschiede, die für die Behandlung nicht relevant sind, und können daher kleinere Unterschiede besser detektieren als unabhängige Designs. Dies lässt sich an einem Beispiel veranschaulichen.

In einem (fiktiven) Versuch wurde die Angst vor Spinnen auf einer Skala von 0 bis 100 gemessen (größere Werte stehen für höhere Angst). Es wurden 24 Personen in zwei Gruppen untersucht (also 12 Personen pro Gruppe). Einer Gruppe wurden Fotos von Spinnen gezeigt, der anderen Gruppe wurden echte Spinnen gezeigt. Wir beginnen mit dem Laden der Daten:

```
library(readr)
spider <- read_tsv("spider.dat")
spider
```

```
# A tibble: 24 x 2
  Group Anxiety
  <chr>   <dbl>
1 Picture    30
2 Picture    35
3 Picture    45
4 Picture    40
5 Picture    50
6 Picture    35
7 Picture    55
8 Picture    25
9 Picture    30
10 Picture   45
# ... with 14 more rows
```

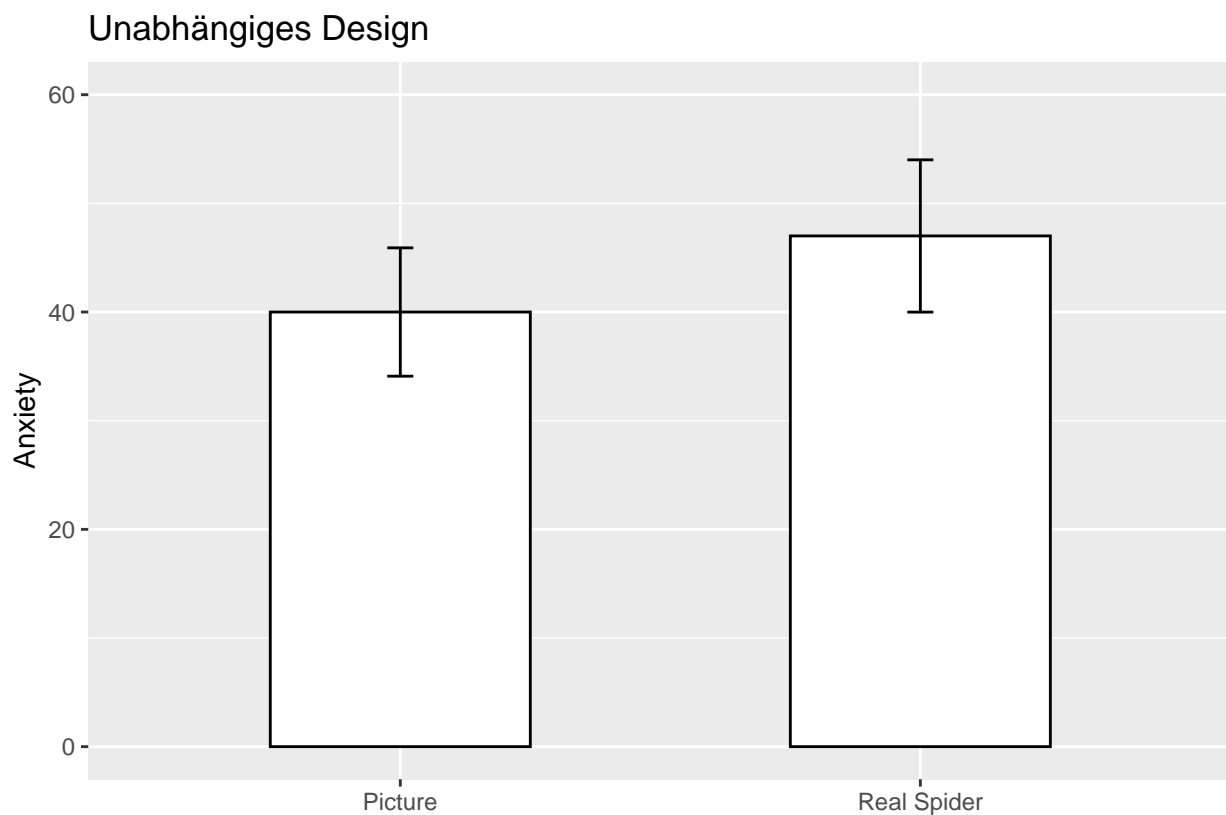
Unabhängige Stichproben

Die Daten in `spider` liegen im Long-Format vor, d.h. dieses Format würde gut für ein unabhängiges Design passen, da sich jede Person in einer eigenen Zeile befindet. Stellen wir nun die Mittelwerte beider Gruppen in einer Balkengrafik gegenüber. Die Fehlerbalken stellen die 95%-Konfidenzintervalle dar. Der nachfolgende Code verwendet das Paket `dplyr` aus dem Tidyverse, welches sich auszeichnet zur Datenmanipulation

eignet. Leider fehlt die Zeit in dieser Veranstaltung, dieses Paket genauer zu besprechen. Wichtig ist hier aber ohnehin die resultierende Grafik, die man auch ohne Hintergründe zum Code interpretieren kann.

```
library(dplyr)
spider.stats <- spider %>%
  group_by(Group) %>%
  summarize(N=length(Anxiety),
            mean=mean(Anxiety),
            sd=sd(Anxiety),
            se=sd/sqrt(N),
            ci=se*qt(0.975, N-1))

library(ggplot2)
ggplot(spider.stats, aes(Group, mean)) +
  geom_bar(stat="identity", width=0.5, fill="white", color="black") +
  geom_errorbar(aes(ymin=mean-ci, ymax=mean+ci), width=0.05) +
  xlab("") +
  ylab("Anxiety") +
  scale_y_continuous(limits=c(0, 60)) +
  ggtitle("Unabhängiges Design")
```



Man erkennt, dass die Konfidenzintervalle stark überlappen, d.h. der Unterschied zwischen den beiden Mittelwerten ist wahrscheinlich nicht signifikant.

Abhängige Stichproben

Wenn die beiden Gruppen abhängig wären (also beispielsweise aus einem Design mit Messwiederholung, in welchem jede der 12 Personen in beiden Gruppen war), dann bringt man die Daten zuerst am besten ins Wide-Format.

```
library(tidyr)
spider_w <- spread(cbind(id=rep(1:12), spider), Group, Anxiety)
spider_w$id <- NULL
names(spider_w) <- c("picture", "real")
spider_w <- as_tibble(spider_w)
spider_w
```

```
# A tibble: 12 x 2
  picture real
  <dbl> <dbl>
1      30    40
2      35    35
3      45    50
4      40    55
5      50    65
6      35    55
7      55    50
8      25    35
9      30    30
10     45    50
11     40    60
12     50    39
```

Anschließend berechnen wir angepasste Werte, welche die individuellen Unterschiede der Personen berücksichtigen. Dazu fügen wir zunächst eine Spalte mit der mittleren Angst jeder Person hinzu.

```
spider_w$mean <- rowMeans(spider_w)
spider_w
```

```
# A tibble: 12 x 3
  picture real mean
  <dbl> <dbl> <dbl>
1      30    40   35
2      35    35   35
3      45    50  47.5
4      40    55  47.5
5      50    65  57.5
6      35    55   45
7      55    50  52.5
8      25    35   30
9      30    30   30
10     45    50  47.5
11     40    60   50
12     50    39  44.5
```

Nun berechnen wir die Differenz der Personenmittelwerte zum Gesamtmittelwert aller Daten - dies ist der Korrekturfaktor für Designs mit Messwiederholung.

```
spider_w$adj <- mean(c(spider_w$picture, spider_w$real)) - spider_w$mean
spider_w
```

```
# A tibble: 12 x 4
```

	picture	real	mean	adj
	<dbl>	<dbl>	<dbl>	<dbl>
1	30	40	35	8.5
2	35	35	35	8.5
3	45	50	47.5	-4
4	40	55	47.5	-4
5	50	65	57.5	-14
6	35	55	45	-1.5
7	55	50	52.5	-9
8	25	35	30	13.5
9	30	30	30	13.5
10	45	50	47.5	-4
11	40	60	50	-6.5
12	50	39	44.5	-1

Damit können wir die einzelnen Angstwerte pro Person korrigieren.

```
spider_w$picture.adj <- spider_w$picture + spider_w$adj
spider_w$real.adj <- spider_w$real + spider_w$adj
```

Dadurch haben wir erreicht, dass die korrigierten Werte nun für alle Personen denselben Mittelwert ergeben, d.h. personenspezifische Unterschiede werden berücksichtigt und nur der Unterschied zwischen den Gruppen wird untersucht:

```
rowMeans(spider_w[, c("picture.adj", "real.adj")])
```

```
[1] 43.5 43.5 43.5 43.5 43.5 43.5 43.5 43.5 43.5 43.5 43.5 43.5
```

Die Daten sehen nun so aus:

```
spider_w

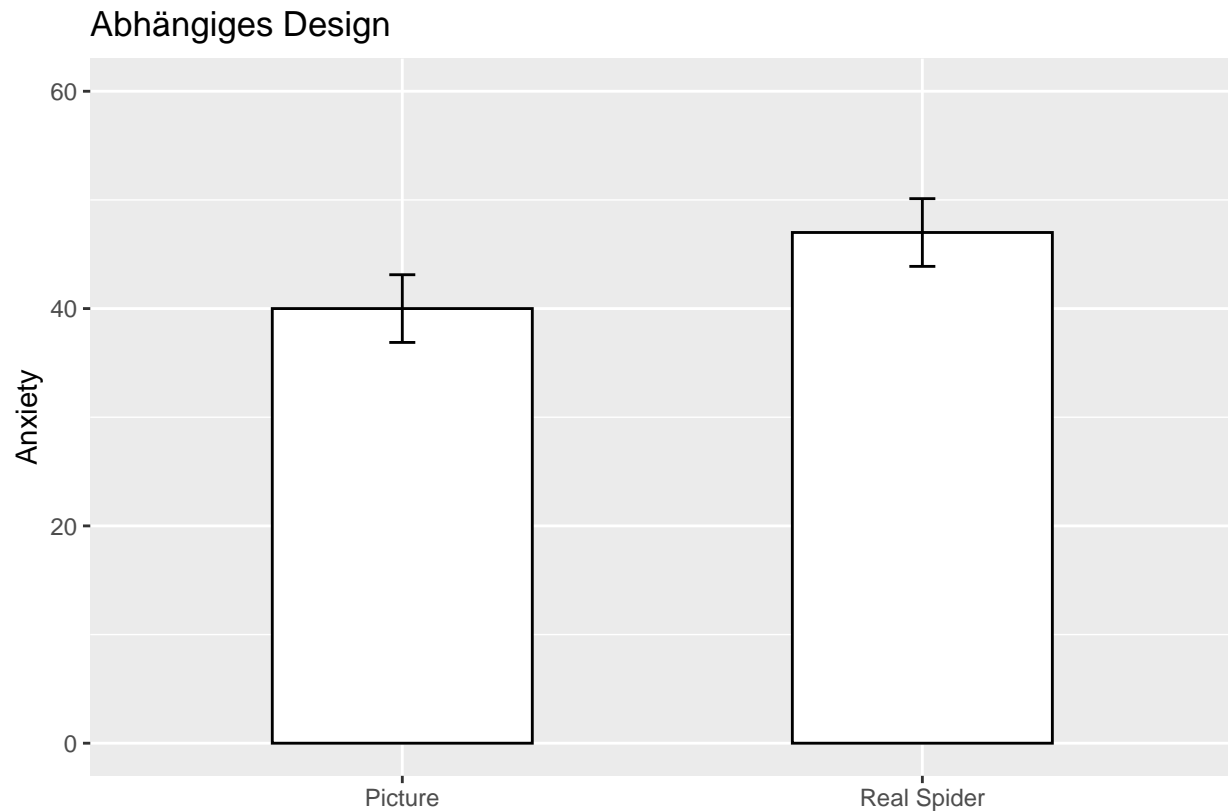
# A tibble: 12 x 6
  picture real mean adj picture.adj real.adj
  <dbl> <dbl> <dbl> <dbl>      <dbl>      <dbl>
1     30    40  35    8.5      38.5      48.5
2     35    35  35    8.5      43.5      43.5
3     45    50 47.5   -4      41        46
4     40    55 47.5   -4      36        51
5     50    65 57.5  -14      36        51
6     35    55  45   -1.5     33.5     53.5
7     55    50 52.5   -9      46        41
8     25    35  30   13.5     38.5     48.5
9     30    30  30   13.5     43.5     43.5
10    45    50 47.5   -4      41        46
11    40    60  50   -6.5     33.5     53.5
12    50    39 44.5   -1      49        38
```

Die angepassten Werte können wir wieder in einem Bar-Chart mit Konfidenzintervallen darstellen.

```
spider_w.stats <- data.frame(Group=c("Picture", "Real Spider"),
                             N=c(length(spider_w$picture.adj), length(spider_w$real.adj)),
                             mean=c(mean(spider_w$picture.adj), mean(spider_w$real.adj)),
                             sd=c(sd(spider_w$picture.adj), sd(spider_w$real.adj)))
spider_w.stats <- spider_w.stats %>%
  mutate(se=sd/sqrt(N), ci=qt(0.975, N-1)*se)

ggplot(spider_w.stats, aes(Group, mean)) +
```

```
geom_bar(stat="identity", width=0.5, fill="white", color="black") +
geom_errorbar(aes(ymin=mean-ci, ymax=mean+ci), width=0.05) +
xlab("") +
ylab("Anxiety") +
scale_y_continuous(limits=c(0, 60)) +
ggtitle("Abhängiges Design")
```



Aus der Grafik ist ersichtlich, dass die Mittelwerte (Höhe der Balken) gleich wie im unabhängigen Design sind. Die Konfidenzintervalle sind jedoch aufgrund des abhängigen Designs wesentlich kleiner geworden. In diesem Beispiel überlappen sie jetzt nicht mehr, was auf einen signifikanten Unterschied zwischen den Gruppenmittelwerten schließen lässt.

Der *t*-Test

Der *t*-Test wird verwendet, um die Mittelwerte zweier Gruppen miteinander zu vergleichen. Hier gibt es zwei Varianten, nämlich einen *t*-Test für abhängige Gruppen und einen für unabhängige Gruppen. Ersteren nennt man auch gepaarten oder abhängigen *t*-Test, letzteren nennt man unabhängigen *t*-Test.

Die *t*-Statistik ist wie viele Statistiken aufgebaut: sie setzt die Varianz, die erklärt werden kann in Beziehung zur Varianz, die nicht erklärt werden kann. Im Falle des *t*-Tests ist das statistische Modell der Unterschied der beiden Mittelwerte minus der erwarteten Differenz, und der Fehler wird durch den Standardfehler der Mittelwertsdifferenz geschätzt.

$$t = \frac{\text{Beobachtete Differenz} - \text{Erwartete Differenz}}{\text{Standardfehler der Differenz}}$$

Die Differenz bezieht sich immer auf die Differenz zwischen den Mittelwerten. Die erwartete Differenz (unter

der Annahme der Nullhypothese) ist in den meisten Fällen gleich Null. Die Vorgehensweise beim Testen ist also wie folgt:

- Zwei Stichproben werden erhoben und deren Mittelwerte berechnet. Die beiden Mittelwerte können sich wenig oder stark unterscheiden.
- Wenn die Stichproben aus derselben Population stammen, sollten die Mittelwerte ungefähr gleich sein. Dies ist die Annahme der Nullhypothese (es gibt keinen Unterschied). Große Differenzen können in seltenen Fällen zufällig auftreten.
- Wir vergleichen den beobachteten Unterschied mit dem erwarteten Unterschied, und wir verwenden den Standardfehler als Maß für die Variabilität der Stichprobenmittelwerte.
- Wenn der beobachtete Unterschied größer als der erwartete ist, kann das zwei Gründe haben:
 - Es gibt keinen Unterschied und der beobachtete Unterschied ist zufällig aufgetreten, d.h. die beiden Stichproben sind nicht repräsentativ für ihre Population.
 - Die beiden Stichproben kommen aus unterschiedlichen Populationen, d.h. beide Stichproben sind repräsentativ für ihre Population.
- Je größer der Unterschied bzw. die t -Statistik, desto wahrscheinlicher wird der zweite Fall (d.h. ein tatsächlicher Effekt).

Unabhängiger t -Test in R

Die t -Statistik lässt sich also folgendermaßen anschreiben:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{Standardfehler}}$$

Die Nullhypothese besagt dass $\mu_1 = \mu_2$, daher vereinfacht sich die Gleichung zu:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{Standardfehler}}$$

Der Standardfehler der Differenz beider Gruppen ist bei gleicher Gruppengröße wie folgt definiert:

$$\text{SE} = \sqrt{\frac{s_1^2}{N} + \frac{s_2^2}{N}}$$

Daher lautet die Gleichung für den t -Test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N} + \frac{s_2^2}{N}}}$$

Wenn die beiden Gruppen unterschiedlich viele Personen enthalten (also N_1 bzw. N_2), muss man den Standardfehler über die gepoolte Varianz berechnen.

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Daraus ergibt sich dann für den t -Test mit $N_1 + N_2 - 2$ Freiheitsgraden:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$

In R kann man t -Tests mit der Funktion `t.test` durchführen. Abhängig von der Form der Daten ist der Aufruf unterschiedlich. Liegen die Daten im Long-Format vor, dann ruft man die Funktion mit einer Formel als Argument auf:

```
model <- t.test(Anxiety ~ Group, data=spider)
model
```

Welch Two Sample t-test

```
data: Anxiety by Group
t = -1.6813, df = 21.385, p-value = 0.1072
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.648641  1.648641
sample estimates:
 mean in group Picture mean in group Real Spider
                        40                        47
```

Ausgegeben wird der Wert der t -Statistik, die Freiheitsgrade (standardmäßig korrigiert nach Welch, was die Voraussetzung der Varianzhomogenität überflüssig macht), sowie der p -Wert. Weiters gibt es noch das 95%-Konfidenzintervall für die t -Statistik sowie die Gruppenmittelwerte.

Liegen die Daten im Wide-Format vor (oder in zwei getrennten Vektoren), dann ruft man die Funktion `t.test` so auf:

```
model <- t.test(spider_w$picture, spider_w$real)
model
```

Welch Two Sample t-test

```
data: spider_w$picture and spider_w$real
t = -1.6813, df = 21.385, p-value = 0.1072
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.648641  1.648641
sample estimates:
mean of x mean of y
      40      47
```

Das Ergebnis ist aber immer dasselbe, egal wie die Funktion aufgerufen wird. Der Unterschied zwischen den Gruppenmittelwerten ist nicht signifikant.

Die Effektgröße kann man aus dem Wert von t in r umrechnen:

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

Im Beispiel ist die Effektgröße r also:

```
t <- model$statistic[[1]]
df <- model$parameter[[1]]
r <- sqrt(t^2 / (t^2 + df))
round(r, 3)
```

```
[1] 0.342
```

Abhängiger t -Test in R

Der abhängige (oder gepaarte) t -Test funktioniert ebenso, sieht sich aber die Mittelwerte der einzelnen Differenzen an anstelle der Differenz der Mittelwerte:

$$t = \frac{\bar{D}}{s_D / \sqrt{N}}$$

Für den abhängigen t -Test verwendet man wieder die Funktion `t.test` und setzt übergibt das Argument `paired=TRUE`. Wieder gibt es zwei Möglichkeiten des Aufrufs, je nachdem, in welchem Format die Daten vorliegen.

```
model <- t.test(Anxiety ~ Group, data=spider, paired=TRUE)
model
```

Paired t-test

```
data: Anxiety by Group
t = -2.4725, df = 11, p-value = 0.03098
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.2312185 -0.7687815
sample estimates:
mean of the differences
                -7
```

```
model <- t.test(spider_w$picture, spider_w$real, paired=TRUE)
model
```

Paired t-test

```
data: spider_w$picture and spider_w$real
t = -2.4725, df = 11, p-value = 0.03098
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.2312185 -0.7687815
sample estimates:
mean of the differences
                -7
```

Das Ergebnis ist diesmal signifikant mit $p = 0.03098$. Dies entspricht unseren Überlegungen mit den unterschiedlich großen Konfidenzintervallen von abhängigen und unabhängigen Versuchsdesigns. Abhängige Designs können also kleinere Unterschiede detektieren.

Die Effektgröße kann man wieder aus t mit der Formel oben in r umwandeln.

```
t <- model$statistic[[1]]
df <- model$parameter[[1]]
r <- sqrt(t^2 / (t^2 + df))
round(r, 3)
```

```
[1] 0.598
```

Es handelt sich also in diesem Fall um einen großen Effekt.

Übungen

Übung 1

Vergleichen Sie im `iris`-Datensatz, ob sich die vier Merkmale zwischen den Spezies *Virginica* und *Versicolor* unterscheiden. Berichten Sie die relevanten Statistiken, Signifikanzen und Effektgrößen (Korrelationskoeffizient

r) für jeden der vier Tests.

Übung 2

Überprüfen Sie, ob sich die Anzahl der zusätzlichen Schlafzeit zwischen den beiden Gruppen im Datensatz `sleep` signifikant voneinander unterscheidet.

Übung 3

Laden Sie den Datensatz `immer` aus dem Paket `MASS`. Berechnen Sie ein 99%-Konfidenzintervall für den Unterschied des durchschnittlichen Gerstenertrages zwischen den beiden Jahren. Ist dieser Unterschied signifikant?



Diese Unterlagen sind lizenziert unter einer Creative Commons Namensnennung - Nicht-kommerziell - Weitergabe unter gleichen Bedingungen 4.0 International Lizenz.