Untersuchungen zur koronaren Herzkrankheit (t-tests)

Untersuchungen zur koronaren Herzkrankheit

In diesem Abschnitt sollen Daten von Probanden bzw. Patienten auf das Risiko für koronare Herzkrankheit untersucht werden. Dies ist eine Erkrankung der Herzkranzgefäße (Koronararterien), die sich durch Ablagerungen in den Gefäßwänden verengen. Der Original-Herz-Datensatz ist unter

- https://archive.ics.uci.edu/ml/datasets/Heart+Disease

beschrieben. Wir nutzen eine konsolidierte CSV-Datei, die bereits Header enthält. Download unter:

- https://oc.informatik.hs-mannheim.de/s/wyzFq34K9HiNjXR/download

Die Datei enthält 13 Merkmale, die einen Einfluss auf eine koronare Herzkrankheit haben können. Das 14. Merkmal `goal` (im Original auch `num`) ist die Diagnose (Klassifizierung). Der Wert ist 0, falls keine krankhafte Verengung der Gefäße vorliegt, oder 1, 2, 3 oder 4, falls – je nach Stärke – eine krankhafte Verengung der Gefäße vorliegt. Wir unterscheiden im Folgenden nur die Zustände "gesund" (0) und "krank" (1, 2, 3, 4). Unter

- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

finden Sie eine Beschreibung aller Attribute. Hier ist eine Zusammenfassung. Wir benötigen insbesondere die Merkmale `sex`, `trestbps`, `chol` und `goal`.

Feld	Bedeutung
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	resting systolic blood pressure (in mmHg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest.[1]
slope	slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
goal	diagnosis of heart disease (0: < 50% diameter narrowing ; 1, 2, 3, 4: > 50% diameter narrowing)

[1] ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline.

Einlesen der Herz-Daten

Lesen Sie die Datei aus der URL als Data Frame zur weiteren Bearbeitung ein. Überlegen Sie, ob sie Faktoren sinnvoll einsetzen können. Geben Sie die ersten drei Zeilen und fünf Spalten aus[2]:

```
heart_df = read.csv("https://oc.informatik.hs-mannheim.de/s/wyzFq34K9HiNjXR/download")
# Kategoriale Spalten in Faktoren
# Geschlecht in Faktoren mit sprechendem Namen umwandeln
heart_df$sex <- factor(heart_df$sex, levels=c(1,0), labels=c("male", "female"))
heart_df$cp = as.factor(heart_df$cp)
heart_df$restecg = as.factor(heart_df$restecg)
heart_df$exang = as.factor(heart_df$exang)
heart_df$slope = as.factor(heart_df$slope)
heart_df$thal = as.factor(heart_df$thal)
heart_df[1:3,1:5] # erste 3 Zeilen und 5 Spalten
```

```
##   age  sex cp trestbps chol
## 1  63 male  1      145  233
## 2  67 male  4      160  286
## 3  67 male  4      120  229
```

Cholesterin im Vergleich Männer/Frauen

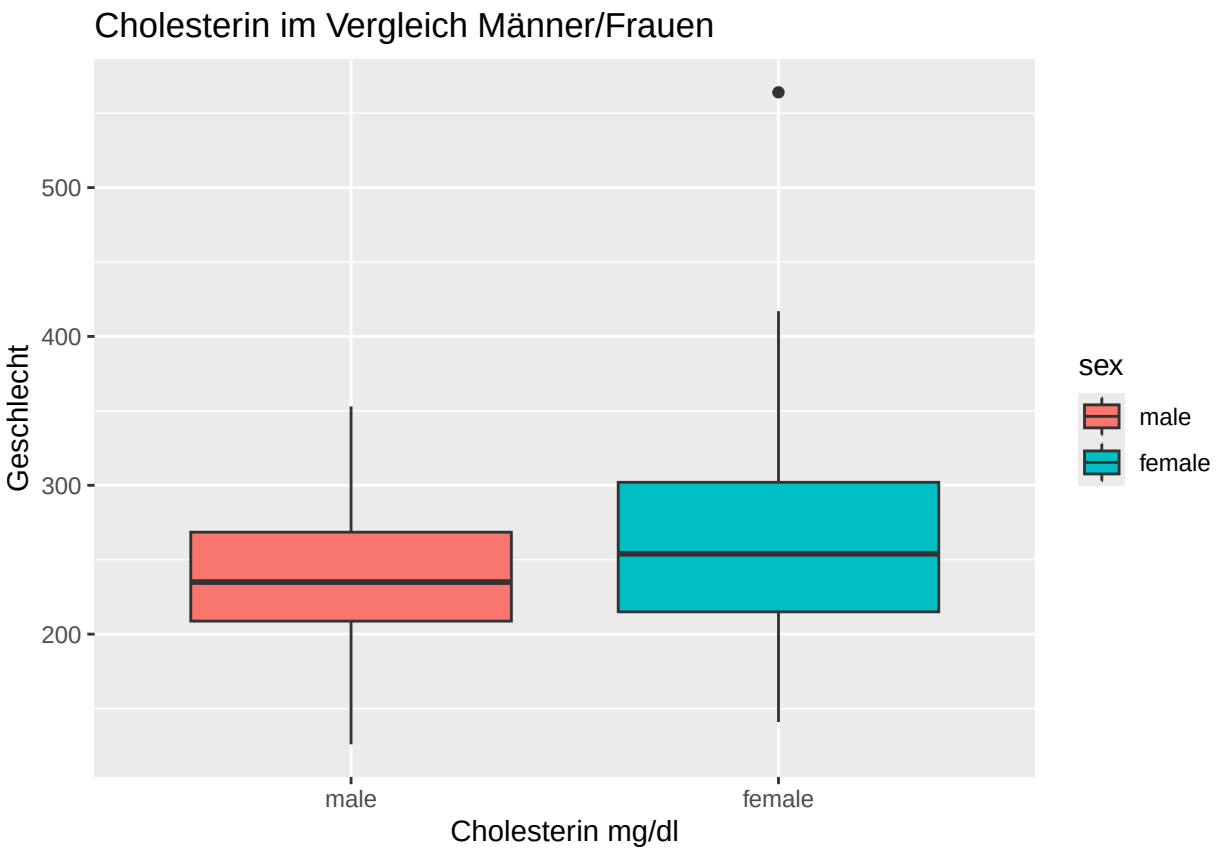
Nun sollen die Cholesterin-Werte untersucht werden – zunächst im Vergleich Männer zu Frauen.

Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für das Cholesterin gruppiert nach dem Geschlecht plotten.

```
ggplot(heart_df)+
  geom_boxplot(aes(x=sex, y=chol, fill=sex))+
  labs(title="Cholesterin im Vergleich Männer/Frauen", x="Cholesterin mg/dl", y="Geschlecht")
```

[2] Möglicherweise kommt es zu einem Fehler beim Einlesen des ersten Attributs (`age`). Manuelles Umbenennen hilft.



Konfidenz-Intervall

Berechnen Sie das Konfidenz-Intervall (Niveau 95%) für den Cholesterin-Level jeweils für Männer und Frauen.

Tabelle Geben Sie das Ergebnis als `kable`-Tabelle aus:

```
library(kableExtra)
# Ausführen der beiden zweiseitigen t-tests getrennt nach Geschlecht
test_male = t.test(heart_df$chol[heart_df$sex=="male"])
test_female = t.test(heart_df$chol[heart_df$sex=="female"])
# Zusammenbauen der Tabelle:
result <- data.frame(Geschlecht=c("Mann", "Frau"),
                     Mittelwert=c(test_male$estimate, test_female$estimate),
                     Untere_Grenze=c(test_male$conf.int[1],test_female$conf.int[1]),
                     Obere_Grenze=c(test_male$conf.int[2],test_female$conf.int[2]))

kable(result, digits=2, caption="Konfidenzintervalle für Cholesterin nach Geschlecht") |>
  kable_styling(full_width = FALSE) # Für schönere Darstellung der Tabelle
```

Table 1: Konfidenzintervalle für Cholesterin nach Geschlecht

Geschlecht	Mittelwert	Untere_Grenze	Obere_Grenze
Mann	239.60	233.74	245.46
Frau	261.75	248.67	274.83

Überlappung? Überlappen sich die Bereiche? -> Nein

Cholesterin im Vergleich zur Erkrankung

Nun sollen die Cholesterin-Werte in Abhängigkeit der Diagnose untersucht werden.

Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie einen Boxplot für das Cholesterin gruppiert nach der Diagnose plotten.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

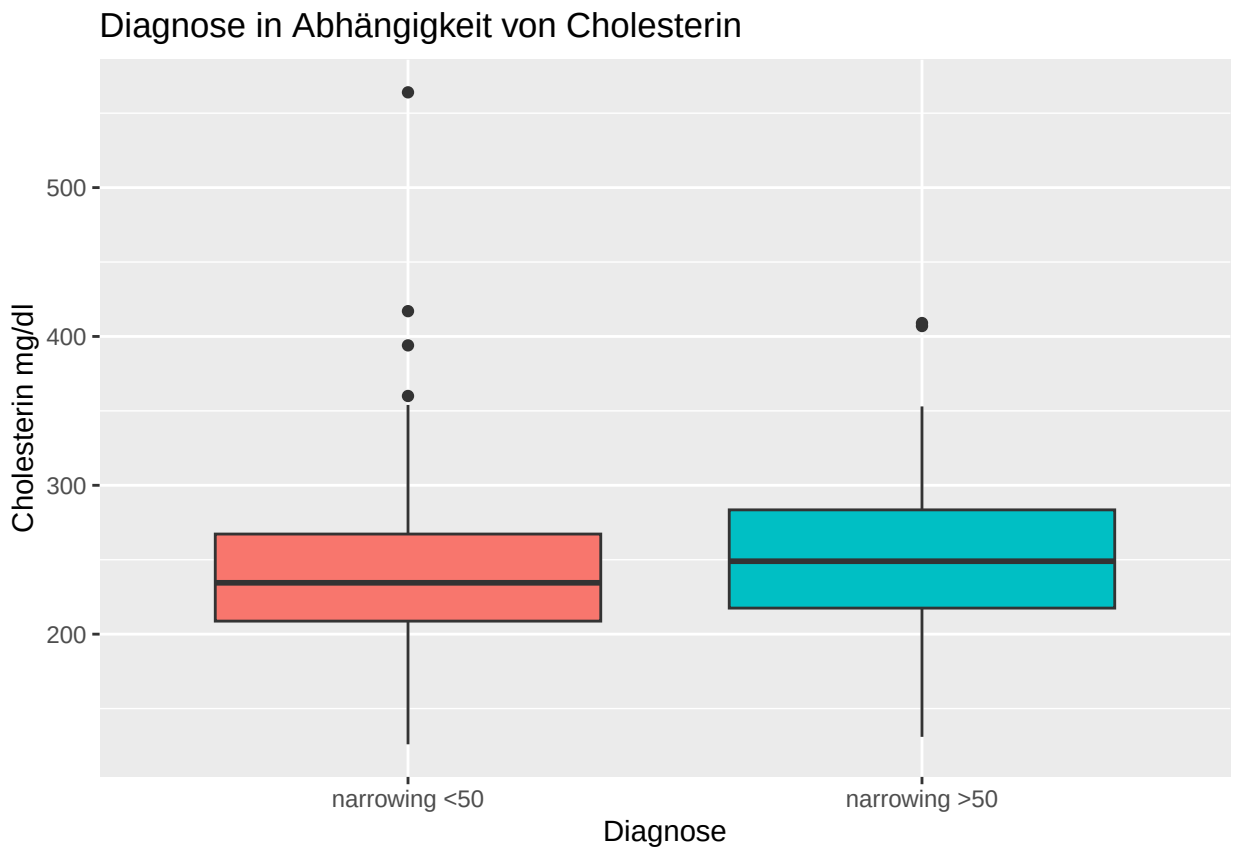
```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Narrowing größer 50 steht für die kranken Patienten, wohingegen eine Verengung
# kleiner 50 die Gesunden repräsentiert.
# Im Datensatz ist alles größer 0 krank
heart_df <- mutate(heart_df,diagnosis=ifelse(goal>0,"narrowing >50","narrowing <50"))

ggplot(heart_df)+
  geom_boxplot(aes(x=diagnosis, fill=diagnosis,y=chol))+
  labs(title="Diagnose in Abhängigkeit von Cholesterin",
       x="Diagnose",
       y="Cholesterin mg/dl") +
  theme(legend.position="none")
```



Konfidenz-Intervall

Berechnen Sie die Konfidenzintervalle für beide Gruppen und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
narrow_greater_50 = heart_df |> filter(diagnosis=="narrowing >50")
narrow_less_50 = heart_df |> filter(diagnosis=="narrowing <50")
test_greater_50 = t.test(narrow_greater_50$chol,conf.level=0.95)
test_less_50 = t.test(narrow_less_50$chol,conf.level=0.95)
result <- data.frame(Gruppe=c("Verengung >50","Verengung <50"),
                     "Mittelwert_der_Stichprobe"=c(test_greater_50$estimate,
                                                   test_less_50$estimate),
                     Untere_Grenze=c(test_greater_50$conf.int[1],test_less_50$conf.int[1]),
                     Obere_Grenze=c(test_greater_50$conf.int[2],test_less_50$conf.int[2]))

kable(result) |> kable_styling()
```

Gruppe	Mittelwert_der_Stichprobe	Untere_Grenze	Obere_Grenze
Verengung >50	251.4748	243.1752	259.7744
Verengung <50	242.6402	234.3977	250.8828

```
# result <- heart_df |>
# group_by(Gruppe = diagnosis) |>
# summarise(test = list(t.test(chol)),
# Mittelwert = mean(chol),
# Untere_Grenze=test[[1]]$conf.int[1],
```

```
#   Obere_Grenze=test[[1]]$conf.int[2]) |>
# select(-test)
#kable(result, digits=2) |> kable_styling()
```

Test

Es sieht so aus, als ob der Cholesterin-Wert bei den erkrankten Patienten höher ist als bei den nicht erkrankten. Überprüfen Sie das mit einem Hypothesen-Test.

Wie lauten die Hypothesen? Formulieren Sie die Hypothesen (H_0 und H_1).

H_0: Die Cholesterin-Werte der erkrankten Patienten sind niedriger oder gleich der Gesunden.

H_1: Die Cholesterin-Werte der erkrankten Patienten sind höher als die der nicht erkrankten

Testanwendung Wenden Sie den Test mit R an. Was ist das Ergebnis?

```
t.test(narrow_greater_50$chol, narrow_less_50$chol, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  narrow_greater_50$chol and narrow_less_50$chol
## t = 1.4924, df = 298.64, p-value = 0.06832
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9327489        Inf
## sample estimates:
## mean of x mean of y
##  251.4748  242.6402
```

```
# Dieser Test prüft ob der Mittelwert der ersten Gruppe (krank) höher als der der Zweiten ist
```

Da der P-Wert größer 0.05 bzw 5% ist, kann die Nullhypothese nicht verworfen und die Alternativhypothese nicht angenommen werden.

Grund dafür könnten die Ausreißer in der gesunden Gruppe sein. Durch sie wird der Mittelwertsunterschied kleiner.

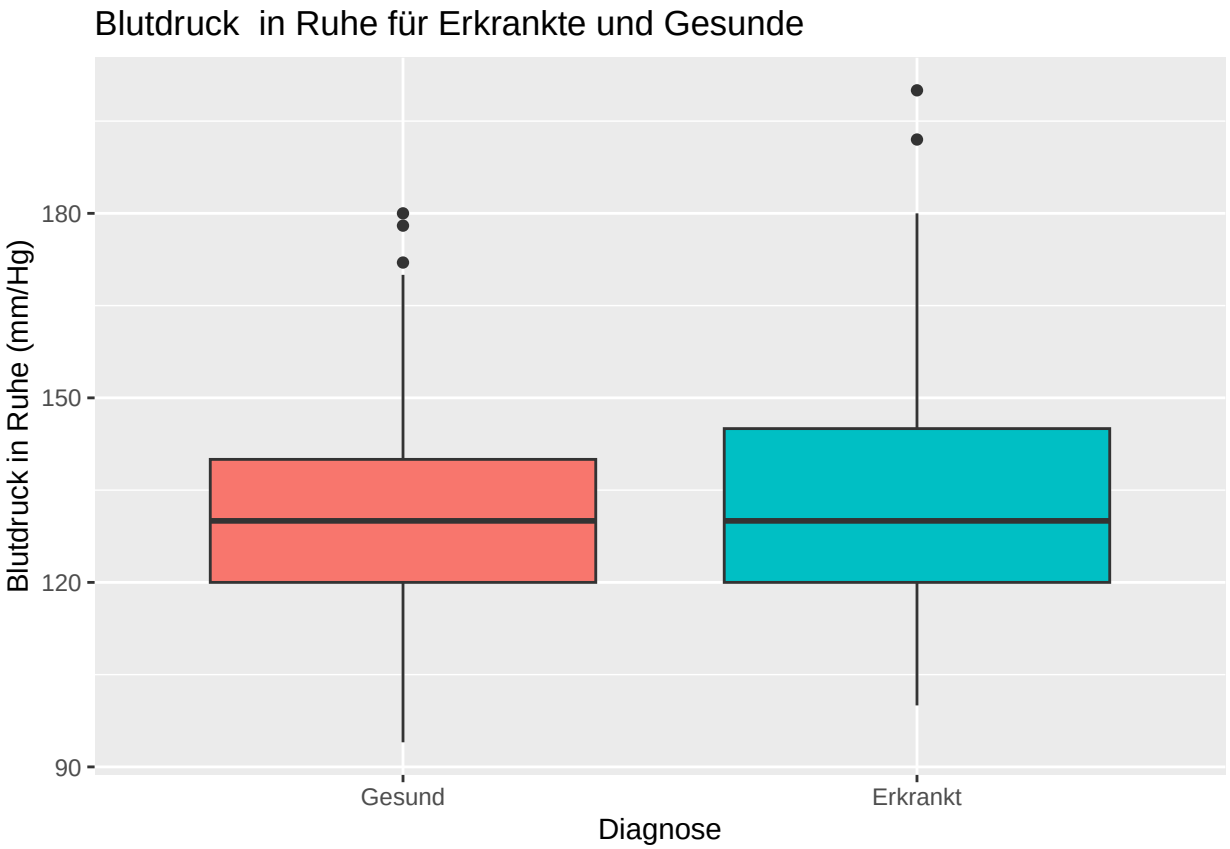
Systolischer Ruheblutdruck

Der systolische Blutdruck liegt beim gesunden Menschen bei ca. 120 mmHg.

Überblick über Blutdruck

Plot Verschaffen Sie sich einen Überblick, indem Sie einen Boxplot für den Blutdruck in Ruhe gruppiert nach der Diagnose plotten.

```
ggplot(heart_df) +
  geom_boxplot(aes(y=trestbps, x=diagnosis ,fill = diagnosis)) +
  labs(title="Blutdruck  in Ruhe für Erkrankte und Gesunde",
       x="Diagnose",
       y="Blutdruck in Ruhe (mm/Hg)")+
  scale_x_discrete(labels = c("narrowing <50" = "Gesund",
                              "narrowing >50" = "Erkrankt"))+
  theme(legend.position = "none") # entfernt die Legende
```



Normalverteilt? Kann überhaupt davon ausgegangen werden, dass die Daten normalverteilt sind? Da der Blutdruck von vielen Faktoren abhängt, könnte angenommen werden er sei normalverteilt. Denn viele voneinander unabhängige Einflüsse *addieren* sich (zentraler Grenzwertsatz). Aber Blutdruck-Werte sind auch nach unten begrenzt und nach oben ist viel Platz. Das kann für eine rechtsschiefe Verteilung sorgen.

Konfidenzintervalle nach Erkrankung

Berechnen Sie die Konfidenzintervalle für den Ruheblutdruck aufgeschlüsselt nach der Diagnose (erkrankt/nicht erkrankt) und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
result_bp = heart_df |>
  group_by(Gruppe = diagnosis) |>
  summarise(test = list(t.test(trestbps)),
            Mittelwert = mean(trestbps),
            Untere_Grenze = test[[1]]$conf.int[1],
            Obere_Grenze = test[[1]]$conf.int[2]) |>
  select(-test) # Das sorgt dafür, dass die zuvor erstellte Spalte Test,
# von der die anderen Spalten abgeleitet wurden, wieder entfernt wird

kable(result_bp, digits=2) |> kable_styling()
```

Gruppe	Mittelwert	Untere_Grenze	Obere_Grenze
narrowing <50	129.25	126.75	131.75
narrowing >50	134.57	131.42	137.72

Test, ob Kranke höheren Ruhe-Blutdruck haben

Überprüfen Sie mit einem Hypothesen-Test, ob Erkrankte einen höheren Ruhe-Blutdruck haben als gesunde Probanden.

Wie lauten die Hypothesen? Formulieren Sie die Hypothesen (H_0 und H_1).

H_0: Gesunde haben einen gleich großen oder höheren Ruhe-Blutdruck als Erkrankte. $\mu_{krank} \leq \mu_{gesund}$

H_1: Erkrankte haben einen höheren Ruhe-Blutdruck als Gesunde. $\mu_{krank} > \mu_{gesund}$

Testanwendung Wenden Sie den Test mit R an. Was ist das Ergebnis?

```r
# Testet die auf die Alternativhypothese, dass das erste Argument größer als das Zweite ist
restbps_diff = t.test(narrow_greater_50$trestbps,
                      narrow_less_50$trestbps,
                      alternative = "greater")
print(restbps_diff)
```

```
##
##  Welch Two Sample t-test
##
## data:  narrow_greater_50$trestbps and narrow_less_50$trestbps
## t = 2.6152, df = 274.64, p-value = 0.004705
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.962045      Inf
## sample estimates:
## mean of x mean of y
##  134.5683  129.2500
```

Der P-Wert ist mit 0.0047 niedriger als 0.05. Somit kann die Alternativhypothese angenommen werden und die Nullhypothese abgelehnt werden. Erkrankte haben also einen höheren Ruhe-Blutdruck als Gesunde.