Eine Einführung in R: Deskriptive Statistiken und Graphiken

Katja Nowick, Lydia Müller und Markus Kreuz

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig

http://www.bioinf.uni-leipzig.de/teaching/currentClasses/class211.html

17. November 2015

I. Ergänzungen zu Übung 1

Scope [Gültigkeitsbereich] von Variablen bei Funktionen

Es können drei Arten von Variablen in einer Funktion auftauchen:

- Formale Parameter:
  Werden beim Aufruf der Funktion angegeben
- Lokale Variablen:
  Werden beim Abarbeiten einer Funktion erzeugt
- Freie Variablen:
  Alle anderen

Frage: Wo sucht `R` nach freien Variablen?
Antwort: In der Umgebung der Variable

```
z <- 3
f <- function(x) {
y <- 2*x
print(z)
}
```

Ausgabe bei Aufruf der Funktion:

f(1)	f(60)
3	3

- *x*: Formaler Parameter
- *y*: Lokale Variable
- *z*: Freie Variable, die in diesem Bsp. von R außerhalb der Funktion gesucht wird

```
z <- 3
f <- function(x) {
y <- 2*x
z <- 5
print(z)
}
```

Ausgabe bei Aufruf der Funktion:

f(1)	f(60)
5	5

- *z* ist keine freie Variable mehr, da sie nun innerhalb der Funktion definiert ist (lokale Variable) und die freie Variable *z* außerhalb der Funktion verdeckt
- Zugriff auf verdeckte Variablen per `<<-` Befehl

Ermittlung der Rechenzeit

```
system.time(expr)
```

- `expr`: R-Befehl, dessen Rechenzeit ausgewertet werden soll

Beispiel: `colMeans` gegen `apply`

```
try<-matrix(1:4000000, nrow=4)
system.time(colMeans(try))
```

```
user system elapsed
0.02 0.00 0.01
```

```
system.time(apply(try, MARGIN=2, FUN=mean, na.rm=TRUE))
```

```
user system elapsed
32.16 0.00 32.20
```

Alternativ:

```
ptm <- proc.time()
exrps
proc.time()-ptm
```

Pakete und Hilfe

- Download unter *http://cran.r-project.org*
- R besteht aus einem Grundprogramm mit vielen Zusätzen den sogenannten *packages* oder Pakete
- Hilfe per `?<Name>` oder `help.search(suchbegriff)`
- Übersicht über die Hilfe `help.start()`
- Pakete speziell für Bioinformatik / Biostatistik: *http://bioconductor.org/*

Was sind Pakete?

- R bietet eine Vielzahl frei verfügbarer Pakete
- Ein Paket enthält unterschiedlichste, spezielle Funktionen
- Beim Start von R ist nur eine Grundausstattung geladen, alle anderen Pakete müssen zusätzlich geladen werden
- Jeder kann sein eigenes Paket schreiben
- Derzeit gibt es 7482 Pakete (Stand Oktober 2009: 2112 Pakete)
- Es besteht aber KEINE GARANTIE für richtige Funktionsweise!

Was sind Pakete?

- Überblick über die geladenen Pakete `sessionInfo( )`
- package laden `require(packagename)` oder `library(packagename)`
- package installieren `install.packages(packagename)`
- Repositories auswählen `setRepositories()`
- Wichtige Pakete:
  - `survival`: Überlebenszeitanalysen (Kaplan-Meier, Log-Rank-Tests Cox-Modelle)
  - `mvtnorm`: Multivariate Normalverteilung
  - `R2HTML`: R Ausgabe in HTML
- Mögliche Pakete:
  - `sendmailR`: send email from inside R
  - `twitteR`: R based Twitter client
  - `sudoku`: Sudoku Puzzle Generator and Solver

II. Diskrete Daten: Deskriptive Statistiken und Graphiken

Was sind diskrete Variablen?

Diskrete Variablen nehmen nur eine endliche Anzahl an Werten an:

- Kategorial: Es besteht keine Rangordnung der Kategorien
- Ordinal: Kategorien können geordnet werden

Kategoriale oder ordinale Variablen sollten in R als Faktoren definiert sein.

Mit einer Häufigkeitstabelle kann man ein kategoriales Objekt zusammenfassen:

- `table(object)`: Absolute Häufigkeiten
- `prop.table(table(object))`: Relative Häufigkeiten

Betrachten wir einen Faktor mit 4 Ausprägungen:

`DNA <- rep(c(“A”, “C”, “G”, “T”), 10)`

1	“A”
2	“C”
3	“G”
3	“T”
⋮	⋮

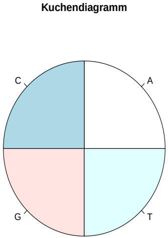
- `table(DNA)` ergibt:

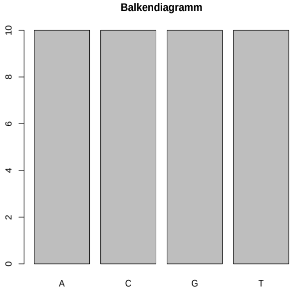
```
 A  C  G  T
10 10 10 10
```

- `prop.table(table(DNA))` ergibt:

```
   A    C    G    T
0.25 0.25 0.25 0.25
```

Kuchendiagramm und Balkendiagramm





Zu erzeugen mit:

`pie(table(DNA))` `barplot(table(DNA))`

III. Stetige Daten: Deskriptive Statistiken und Graphiken

Was sind stetige Variablen?

Stetige Variablen können (in der Theorie) eine unendliche Anzahl an Werten annehmen. Beispiele:

- Gewicht
- Größe
- Gehalt

> R speichert stetige Variablen als
> metrische Objekte (`numeric`) ab.

Häufigkeitstabelle sind für stetige Variablen meist nicht geeignet. Wichtiger sind:

- Maße für die Lage
- Maße für die Streuung

Maße für die Lage

Die Lage (*location*) gibt an, in welcher Größenordnung sich Daten bewegen.

- (Empirische) Mittelwert

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \ldots + x_n).$$

- In `R`: `mean()`

Maße für die Lage II

- x%-Quantile, trennen die Daten in zwei Teile.
  So liegen x% der Daten unter dem x%-Quantile
  und $100 - x$% darüber.
  - Median $x_{0.5}$ entspricht dem 50%-Quantil
  - In R: `median()`
  - 25%-Quantil $x_{0.25}$ (das erste Quartil)
  - In R: `quantile(x,0.25)`
  - 75%-Quantil $x_{0.75}$ (das dritte Quartil)
  - In R: `quantile(x,0.75)`
- Der Median ist robuster gegen Ausreißer als der Erwartungswert
- Oder gleich in R: `summary()`

Maße für die Streuung

Die Streuung (*scale*) gibt an, wie stark die verschiedenen Werte voneinander abweichen.

- Die (empirische) Varianz

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left((x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\right).$$

- Spannbreite:
  Differenz vom größten zum kleinsten Wert
- Interquartilsabstand:

$$\mathrm{IQR} = x_{0.75} - x_{0.25}$$

Beispiel: *oecd*-Daten

Betrachten wir das durchnittliche, frei verfügbare Einkommen einer Familie [pro Kind, in tausend US-Dollar].

- Einen Überblick erhält man durch:

```
summary(Einkommen)
Min. 1st Qu. Median Mean 3rd Qu. Max.
5.10 16.60 21.10 19.18 22.65 34.20
```

- Die Varianz bzw. Standardabweichung

```
var(Einkommen)
[1] 50.75937
```

`sd(Einkommen)` (alternativ `sqrt(var(Einkommen))`)

```
[1] 7.124561
```

Beispiel: *oecd*-Daten II

- Den Interquartilsabstand erhält man durch:

```
IQR(Einkommen)
[1] 6.05
```

- Die Spannweite mit

```
max(Einkommen)-min(Einkommen)
[1] 29.1
```

Bei der Variable *Alkohol* (Prozentsatz der 13-15 jährigen Kinder, die mindestens zweimal betrunken waren) bestehen fehlende Werte.

- Mittelwertsberechnung über

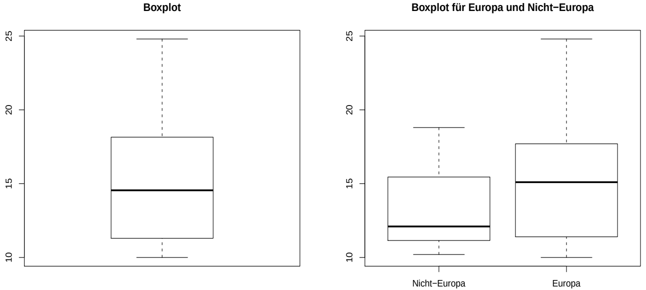
```
mean(Alkohol,na.rm=TRUE)
[1] 15.225
```

Was ist ein Boxplot?

Der Boxplot ist eine Graphik zur Darstellung stetiger Variablen.
Er enthält:

- Minimum und Maximum
- 25%-Quantil und 75%-Quantil
- Median
- In R: `boxplot(variable)`
- Um Variablen getrennt nach Faktorstufen zu untersuchen, bietet sich an: `boxplot(variable ~ factor)`
- Einschub: Ein Label für den Faktor *Geo*
  ```
  factor(Geo,levels=c("R","E"),
  labels=c("Nicht-Europa","Europa"))
  ```

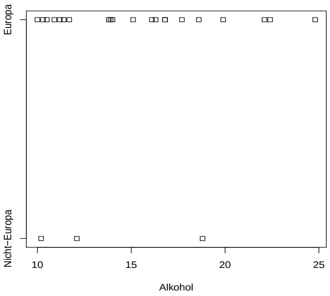
Boxplot: *Alkohol*



Zu erzeugen mit:

`boxplot(Alkohol)` `boxplot(Alkohol∼ Geo)`

Eine Alternative zum Boxplot bei wenigen Beobachtungen ist der Stripchart:



Zu erzeugen mit:

```
stripchart(Alkohol~Geo)
```

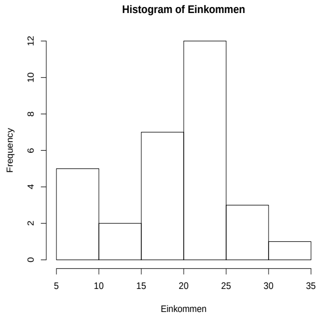
Was ist ein Histogramm?

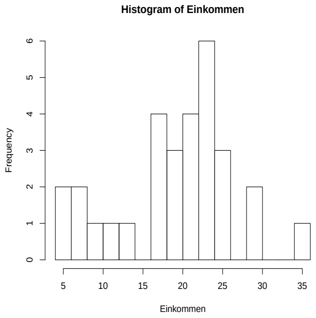
- Zur Erstellung eines Histogramms teilt man die Daten in homogene Teilintervalle ein und plottet dann die absolute Häufigkeit pro Teilintervall
- Dieses Verfahren gibt einen ersten Überblick über die Verteilung der Daten
  (=> Ermitteln der "empirischen Dichte" möglich)

```
hist(x, breaks = "AnzahlBins", freq = NULL )
```

- `x`: Daten
- `breaks = "AnzahlBins"`: Steuerung der Teilintervalle
- `freq=TRUE`: absolute Häufigkeiten
- `freq=FALSE`: relative Häufigkeiten ("empirische Dichte")

Histogramme des Einkommens mit verschiedenen Binstärken





Zu erzeugen mit:

`hist(Einkommen)` `hist(Einkommen, breaks=15)`

Aufgabenkomplex 1

IV. Graphiken in R: Grundaufbau und Parameter

Graphiken in R

R kennt einen Standardbefehl für einfache Graphiken (`plot()`), aber auch viele spezielle Befehle, wie `hist()` oder `pie()`.

```
plot(x, y, type, main, par (...) )
```

- `x`: Daten der x-Achse
- `y`: Daten der y-Achse
- `type=“l”`: Darstellung durch eine Linie
- `type=“p”`: Darstellung durch Punkte
- `main`: Überschrift der Graphik
- `par (...)`: Zusätzlich können sehr viele Parametereinstellungen geändert werden

Parameter für Graphiken in R

```
par(cex, col, lty, mfrow, pch, x/yaxs)
```

- `cex`: Skalierung von Graphikelementen
- `col`: Farbe (`colors()` zeigt die vordefinierten Farben an)
- `lty`: Linienart
- `mfrow`: Anordnen von mehreren Graphiken nebeneinander
- `pch`: Andere Punkte oder Symbole
- `x/yaxs`: Stil der x- bzw. y-Achse

Einen Überblick über die Parameter erhält man mit `?par`.
`par()` kann entweder im `plot()` -Befehl gesetzt werden oder als eigene Funktion vor einem oder mehreren `plot()`-Befehlen.

Aufbau von Graphiken in R

1. `plot()`: Bildet den Grundstein einer Graphik
2. Zusätzlich können weitere Elemente eingefügt werden wie:
   - `lines()`: Linien
   - `points()`: Punkte
   - `legend()`: Legende
   - `text()`: Text
3. `dev.off()`: schließt die Graphik

Einen Überblick erhält man mit der betreffenden Hilfefunktion, z.B. `?legend`.

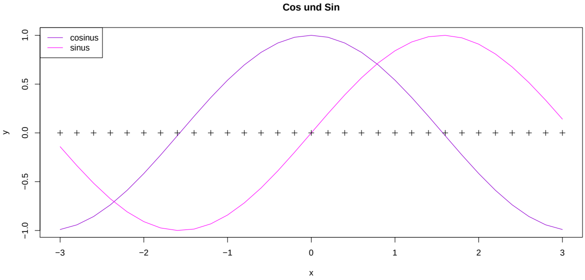
Abspeichern von Graphiken

Folgende Graphikformate können in R erzeugt werden:

- `pdf()`
- `ps()`
- `jpg()`

Beispiel:

```
pdf(file=“boxplot.pdf”, width=13, height=6)
par(mfrow=c(1,2))
boxplot(Alkohol, main=“Boxplot”)
boxplot(Alkohol~Geo, main=“Boxplot für ...”)
par(mfrow=c(1,1))
dev.off()
```




```
pdf(file="RGraphiken/beispiel.pdf", width=12, height=6)
plot(x,y, type="l", col="darkviolet", main="Cos und Sin")
lines(x,z, col="magenta")
points(x,null, pch=3)
legend("topleft", c("cosinus","sinus"), col=c("darkviolet",
"magenta"), lty=1)
dev.off()
```

V. Dichten und Verteilungsfunktionen in R

Einschub: Zufallsvariablen

Eine Variable oder Merkmal X, dessen Werte die Ergebnisse eines Zufallsvorganges sind, heißt Zufallsvariable.

Notation:

- X: Die Zufallsvariable
- x: Eine Realisierung oder Beobachtung der Zufallsvariable

Induktive (Schließende) Statistik:

Mittels einer Stichprobe wird versucht Aussagen bezüglich einer Grundgesamtheit zu treffen.

- Grundgesamtheit: Menge aller für die Fragestellung relevanten Objekte
- Stichprobe: Tatsächlich untersuchte Teilmenge der Grundgesamtheit

Die Aussagen beziehen sich auf Merkmale der Grundgesamtheit.

- Merkmal: Die interessierende Größe oder Variable
- Merkmalsausprägung: Der konkret gemessene Wert an einem Objekt der Stichprobe

Das Model: Theoretische Ebene

- Statistische Analysen beruhen auf Modellannahmen.
- Ziel: Formalisierung eines reellen Sachverhaltes
  - Stetige Variablen mit Erwartungswert und Varianz
  - Diskrete Variablen mit Gruppenzugehörigkeiten
- Parametrischer Ansatz: Verteilungsannahmen,
  wie eine Zufallsvariable X ist normalverteilt mit Erwartungswert μ und Varianz σ^2
- Non-Parametrischer Ansatz: Ohne Verteilungsannahmen

Die beobachteten Daten: Die empirische Ebene

- Erwartungswert und Varianz einer Grundgesamtheit können nicht in der Realität beobachtet werden, sondern müssen aus der Stichprobe geschätzt werden.
- Beobachtet werden n Realisierungen $x_1, ..., x_n$ einer Zufallsstichprobe X.
- Notation:
  - Erwartungswert μ
  - Schätzer für den Erwartungswert $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Gesetz der großen Zahlen: "Je mehr Realisierungen einer Zufallszahl beobachtet werden, desto besser approximiert der Mittelwert den Erwartungswert"
- Realisierungen einer Zufallsvariable folgen nicht exakt einer bestimmten Verteilung. Nur bei großer Stichprobenzahl nähert sich die empirische Dichte der theoretischen an.

Normalverteilung $N(\mu, \sigma)$

Die Normal- oder Gauß -Verteilung ist formalisiert durch Erwartungswert μ und Varianz σ^2:

$$f(x|\mu,\sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- Diese Funktion ist in R implementiert:
  `dnorm(x, mean=0, sd=1)`
  (Vorsicht: mean steht hier für den Erwartungswert)
- Erzeugen von n Realisierungen $x_1, ..., x_n$:
  `rnorm(n, mean=0, sd=1)`

Beispiel: Normalverteilung

- Darstellung: Gesetz der großen Zahlen

```
x10<-matrix(rnorm(100),nrow=10,ncol=10)
x1000<-matrix(rnorm(10000),nrow=10,ncol=1000)
apply(x10,MARGIN=1, mean)
```

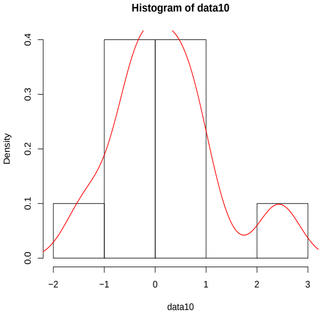
-0.392 -0.309 0.195 -0.727 -0.150 0.327 0.142 0.020 0.069 0.594

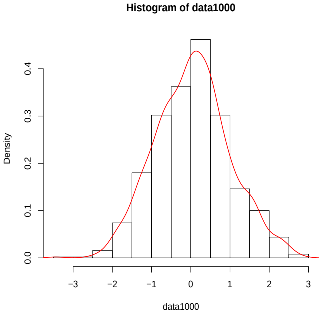
```
apply(x1000,MARGIN=1, mean)
```

-0.018 -0.011 0.007 -0.011 -0.021 -0.013 0.036 0.026 0.074 0.010

Beispiel: Normalverteilung

- Anpassung der empirischen an die theoretische Verteilung:





V.I Diskrete Daten

Eine Zufallsvariable heißt diskret, wenn sie endlich viele Werte $x_1, \dots, x_k$ annehmen kann.

Die Wahrscheinlichkeitsfunktion $f(x)$ einer diskreten Zufallsvariable X ist für $x \in \mathbb{R}$ definiert durch die Wahrscheinlichkeiten p_i:

$$f(x) = \begin{cases} P(X = x_i) = p_i & \text{falls } x = x_i \in \{x_1, \dots, x_k\} \\ 0 & \text{sonst} \end{cases}$$

Die Verteilungsfunktion $F(x)$ einer diskreten Zufallsvariable ist gegeben durch die Summe:

$$F(y) = P(X \leq y) = \sum_{i: x_i \leq y} f(x_i)$$

Eigenschaften

Für die Wahrscheinlichkeitsfunktion $f(x)$ gilt:

$$0 \leq f(x) \leq 1$$

$$\sum_{i \geq 1} p_i = 1$$

Für die Verteilungsfunktion $F(x)$ gilt:

$$F(x) = \begin{cases} 1 & x \geq max(x) \\ 0 & x \leq min(x) \end{cases}$$

F(x) ist monoton steigend mit Wertebereich 0 bis 1.

Bernoulli-Experiment

Binäre Zufallsvariable X: Tritt ein Ereignis A ein?

$$X = \begin{cases} 1 & \text{falls } A \text{ eintritt} \\ 0 & \text{falls } A \text{ nicht eintritt} \end{cases}$$

Das Ereignis A tritt mit einer bestimmten Wahrscheinlichkeit $0 < \pi < 1$ ein

$$P(X = 1) = \pi$$
$$P(X = 0) = 1 - \pi$$

Binomialverteilung

Die Binomialverteilung entspricht dem n-maligen Durchführen eines Bernoulli-Experimentes mit Wahrscheinlichkeit π

$$f(x) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x} & \text{falls } x = 0, 1, ..., n \\ 0 & \text{sonst} \end{cases}$$

Beispiel

Ein Schütze schießt $n = 10$ mal auf eine Torwand.
Wie groß ist die Wahrscheinlichkeit, dass er genau fünfmal trifft, wenn er eine Trefferwahrscheinlichkeit π von 25 % hat?

$$P(X = 5) = \binom{10}{5}0.25^5(1-0.25)^{10-5} = 0.058$$

Diskrete Gleichverteilung

Die diskrete Gleichverteilung charakterisiert die Situation, dass $x_1, \ldots, x_k$-verschiedene Werte mit gleicher Wahrscheinlichkeit angenommen werden.

$$f(x) = \begin{cases} \frac{1}{k} & \text{falls } x_i \text{ mit } i = 1, ..., k \\ 0 & \text{sonst} \end{cases}$$

Beispiel

Würfeln, jede Zahl hat die gleiche Wahrscheinlichkeit $\frac{1}{6}$

V.II Stetige Daten

Eine Zufallsvariable heißt stetig, wenn sie unendlich viele Werte $x_1, \ldots, x_k, \ldots$ annehmen kann, wie beispielsweise metrische Variablen.

Die Dichte $f(x)$ einer stetigen Zufallsvariable X ist für ein Intervall $[a, b]$ definiert als:

$$P(a \leq X \leq b) = \int_a^b f(x) \partial x$$

Die Verteilungsfunktion $F(y)$ einer stetigen Zufallsvariable ist gegeben durch das Integral:

$$F(y) = P(X \leq y) = \int_{-\infty}^{y} f(x) \partial x$$

Eigenschaften

Für die Dichte $f(x)$ gilt:

$$\int_{-\infty}^{+\infty} f(x) \partial x = 1$$

$$P(X = a) = \int_{a}^{a} f(x) \partial x = 0$$

Für die Verteilungsfunktion $F(x)$ gilt:

$$F(x) = \begin{cases} 1 & \text{für } x \geq \max(x) \\ 0 & \text{für } x \leq \min(x) \end{cases}$$

$$F'(x) = \frac{\partial F(X)}{\partial x} = f(x)$$

Normalverteilung $N(\mu, \sigma)$

Eine der wichtigsten Verteilungen ist die Normal- oder Gauß -Verteilung mit Erwartungswert μ und Varianz σ^2:

$$f(x|\mu,\sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- Symmetrisch um μ
- Nur abhängig von μ und σ
- Beispiele: Klausurnoten, das (logarithmierte) Einkommen, Messfehler, Größe und Gewicht

Stetige Gleichverteilung $U(a, b)$

Gegeben: ein Intervall, definiert durch reelle Zahlen a und b mit $a < b$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

Die stetige Gleichverteilung spielt eine wichtige Rolle bei statistischen Tests.

Hat man $x_1, \ldots, x_n$ Realisierungen einer Variablen X mit Verteilungsfunktion F, so gilt:

$$F(x_1), \ldots, F(x_n) \sim U(0, 1)$$

Aufgabenkomplex 2

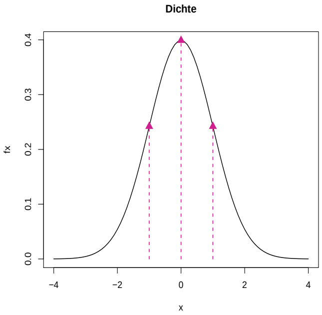
V.III Umgang mit Zufallszahlen

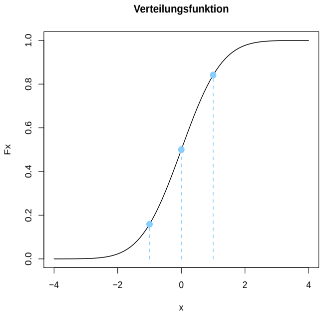
R ermöglicht den Umgang mit Zufallszahlen.
Beispiel: (Standard)Normalverteilung

1. Ziehen von n Zufallszahlen: `rnorm(n, mean=0, sd=1)`
2. Dichte im Wert x: `dnorm(x, mean=0, sd=1)`
   Beispiel: `dnorm(c(-1,0,1))`
   0.24197 0.39894 0.24197
3. Verteilungsfunktion im Wert x:
   `pnorm(x, mean=0, sd=1)`
   Beispiel: `pnorm(c(-1,0,1))`
   0.15866 0.50000 0.84134
4. Quantil für Wahrscheinlichkeit p:
   `qnorm(p, mean=0, sd=1)`
   Beispiel: `qnorm(c(0.25,0.5,0.75))`
   -0.67449 0.00000 0.67449

Beispiel: (Standard)Normalverteilung

1. Dichte im Wert x:
   `dnorm(c(-1,0,1))`
   `0.24197 0.39894 0.24197`
2. Verteilungsfunktion im Wert x:
   `pnorm(c(-1,0,1))`
   `0.15866 0.50000 0.84134`





R-Befehle für weitere Verteilungen

- `rnorm(n, mean=0, sd=1)` Normalverteilung mit Mittelwert `mean` und Standardabweichung `sd`
- `rexp(n, rate=1)` Exponentialverteilung mit Rate `rate`
- `rpois(n, lambda)` Poissonverteilung mit Rate `lambda`
- `rcauchy(n, location=0, scale=1)` Cauchyverteilung mit Lokations- und Skalenparameter
- `rt(n, df)`(Studen)t-verteilung mit Freiheitsgraden `df`
- `rbinom(n, size, prob)` Binomialverteilung vom Umfang `size` und Wahrscheinlichkeit `prob`
- `rgeom(n, prob)` Geometrische Verteilung mit Wahrscheinlichkeit `prob`
- `rhyper(nn, m, n, k)` Hypergeometrische Verteilung
- `runif(n, min=0, max=1)` Stetige Gleichverteilung im Intervall [`min`, `max`]

Darstellung: Histogramme und Kerndichteschätzer

1. Histogramme: Darstellung von stetigen und diskreten Verteilungen

   `hist(x, breaks = “AnzahlBins”, freq = NULL )`

   - `x`: Daten
   - `breaks = “AnzahlBins”`: Steuerung der Teilintervalle
   - `freq=TRUE`: absolute Häufigkeiten
   - `freq=FALSE`: relative Häufigkeiten (“empirische Dichte”)

2. Kerndichteschätzer: Darstellung von stetigen Verteilungen

   `plot(density(x, kernel=“gaussian”, bw))`

   - `density(x)`: Kerndichteschätzung der Daten
   - `kernel`: Option für spezielle Kerntypen
   - `bw`: Bandbreite

Darstellung: Kerndichteschätzer

Kerndichteschätzer sind aus dem Histogramm abgeleitete Verfahren zur Schätzung von stetigen Dichten

Hat man gegebene Daten $x_1, \ldots, x_n$ und eine konstante Bandbreite $h \in \mathbb{R}$ so ist der Kerndichteschätzer gegeben durch:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

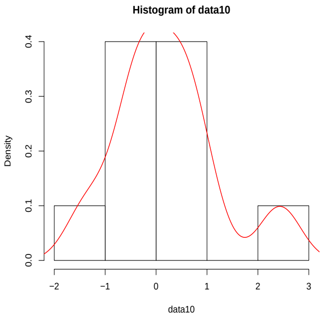
Typische Kerne sind:

- Bisquare Kern: $K(u) = \frac{15}{16}(1 - u^2)^2$ für $u \in [-1, 1]$ und 0 sonst
- Gauß Kern: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$ für $u \in \mathbb{R}$

Beispiel: Simulation aus der Normalverteilung

```
data10<-rnorm(10)
hist(data10, freq=FALSE)
lines(density(data10), col=2)
```

```
data1000<-rnorm(1000)
hist(data1000, freq=FALSE)
lines(density(data1000), col=2)
```

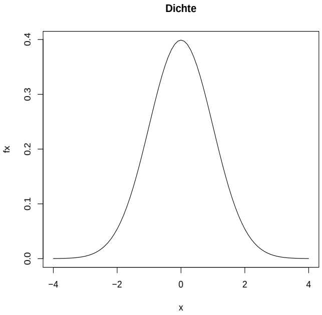


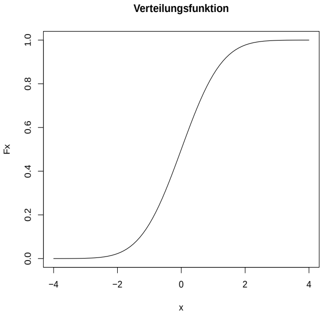
Beispiel: Wie plottet man die Normalverteilung?

```
x<-seq(from=-4, to=4, by=0.1)
```

```
♯ Dichte
fx<-dnorm(x)
plot(x,fx, type="l")
```

```
♯ Verteilungsfunktion
Fx<-pnorm(x)
plot(x,Fx, type="l")
```





Darstellung: Q-Q-Plot

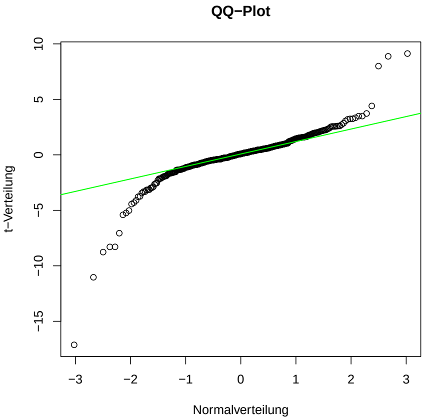
Quantil-Quantil-Plots tragen die Quantile (empirisch oder theoretisch) zweier Verteilungen gegeneinander ab. Somit können Verteilungen miteinander verglichen werden.

- `qqplot(x,y)`: Plottet die emp. Quantile von `x` gegen die emp. Quantile von `y`
- `qqnorm(y)`: Plottet die emp. Quantile von `y` gegen die theoretischen Quantile einer Standard-Normalverteilung
- `qqline(y)`: Fügt dem Quantilplot eine Gerade hinzu die durch das erste und dritte Quartil geht

Bsp: Vergleich von Normal- und t-Verteilung

```
data <- rt(400, df = 2)
qqnorm(data, main = “QQ-Plot”, xlab= “Normalverteilung”, ylab =
“t-Verteilung”)
qqline(data, col = “green”)
```

Darstellung: Q-Q-Plot



VI. Statistische Tests

VI.I Einführungsbeispiel

VI.I Einführungsbeispiel

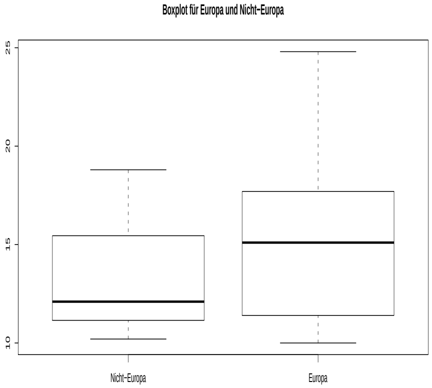
Fragestellung

Einführungsbeispiel: Trinkt die Jugend in Europa mehr Alkohol als im Rest der Welt?

Untersucht wird die Variable Alkohol im oecd-Datensatz: Der Anteil an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren.

Erster Schritt: Deskriptive Analyse

1. Graphisch mit Boxplot: `boxplot(Alkohol~Geo)`



Zweiter Schritt: Kennzahlen

2. Kennzahlen:
   - Mittelwert:
     `mu<-tapply(Alkohol, Geo, FUN=mean, na.rm=TRUE)`

Nicht-Europa	**Europa**
13.700	15.443
   - Standardabweichung:
     `sigma<-tapply(Alkohol, Geo, FUN=sd, na.rm=TRUE)`

Nicht-Europa	**Europa**
4.518	4.341

Es ist zu erkennen, dass in Europa im Mittel ein höherer Anteil an Jugendlichen schon mindestens zweimal betrunken war als in nicht-europäischen Staaten.

Doch dies könnte auch ein Zufall sein! Denn die Beobachtungen beruhen auf Stichproben, sie sind Realisierungen einer Zufallsvariable.

Eigentliches Ziel:
Überprüfung von Annahmen über das Verhalten des interessierenden Merkmales in der Grundgesamtheit mittels Stichproben.

- **Annahme:** Jugendliche in Europa trinken mehr Alkohol als im Rest der Welt
- **Merkmal:** Alkoholkonsum der Jugend
- **Grundgesamtheit:** Jugendliche in Europa und im Rest der Welt
- **Stichprobe:** Die oecd-Daten

Für solche Fragestellungen mit gleichzeitiger Kontrolle der Fehlerwahrscheinlichkeit sind statistische Tests geeignet!

1. **Aufstellen von zwei komplementären Hypothesen:**
   - **Testhypothese** (H_0): Der Anteil in Europa ist kleiner dem im Rest der Welt $\mu_E \leq \mu_{NE}$
   - **Alternativhypothese** (H_1): Der Anteil in Europa größer als der im Rest der Welt $\mu_E > \mu_{NE}$
2. **Fehlerwahrscheinlichkeit festlegen:**
   H_0 soll mit einer W'keit von weniger als 5% abgelehnt werden, wenn H_0 wahr ist.

Also: Wenn der Anteil in Wahrheit kleiner oder gleich ist, soll der Test nur mit einer Wahrscheilichkeit von weniger als 5% zu dem (falschen) Ergebnis kommen, dass der Anteil größer ist.

3. **Beobachtete Daten: 2 Gruppen**

	$\hat{\mu}$	$\hat{\sigma}$	n
Nicht-Europa	13.700	4.518	3
Europa	15.443	4.341	21

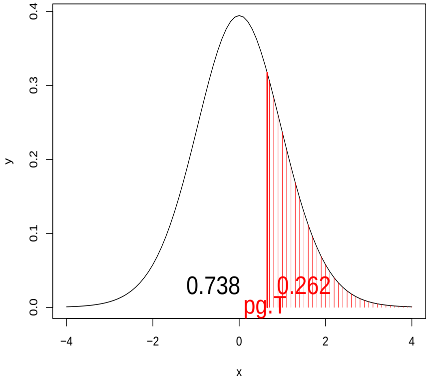
4. **(Weitere Annahmen, hier: Normalverteilung, Varianzgleichheit)**
5. **Berechnen der Prüfgröße T, einer Kennzahl, die zeigt, wie stark die Gruppenmittel voneinander abweichen:**
   - Mittelwertsdifferenz der beiden Gruppen
   - Standardisieren mit der entsprechenden Standardabweichung

$$T = (\hat{\mu_E} - \hat{\mu_{NE}}) / \sqrt{\left(\frac{1}{n_E} + \frac{1}{n_{NE}}\right) \frac{(n_E - 1)\hat{\sigma}_E^2 + (n_{NE} - 1)\hat{\sigma}_{NE}^2}{n_E + n_{NE} - 2}}$$

   - (Hypothetische Verteilung der Prüfgröße festlegen, hier t-Verteilung mit 3 + 21 - 2 = 22 Freiheitsgraden)

6. **Berechnung der Prüfgröße T in R:**
   - Mittelwertsdifferenz der beiden Gruppen
     ```
     m.diff<-mu[2]-mu[1]
     ```
   - Standardisieren mit der entsprechenden Standardabweichung
     ```
     diff.std2 <- sqrt((1/21+1/3)*
     (20*sigma[2]^2+2*sigma[1]^2)/(21+3-2))
     ```
   - Prüfgröße:
     ```
     pg.T <- m.diff/diff.std2
     0.648
     ```
7. **Wie wahrscheinlich ist es (unter der Nullhypothese), eine Prüfgröße T zu beobachten, die größer oder gleich** 0.648 **ist?**
   ```
   1-pt(pg.T, df=22)
   0.262
   ```

Statistisches Testen IV

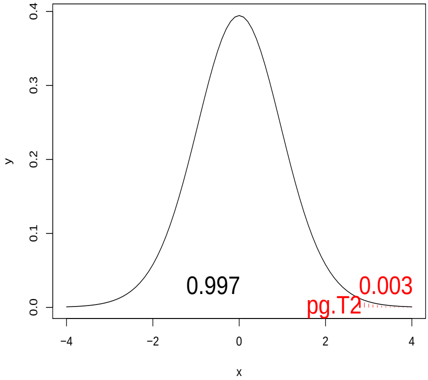


Mit hoher Wahrscheinlichkeit (26.2%) kann eine solche Prüfgröße $pg.T$ beobachtet werden, wenn der Mittelwert in Europa und kleiner als der in Nicht-Europa ist.

Statistisches Testen V

8. **Entscheidung:** Aus diesen Daten kann nicht geschlossen werden, dass in Europa Jugendliche mehr Alkohol trinken als im Rest der Welt.
9. **Grund:** Zu geringe Fallzahl!
   Mit $nE = nNE = 101$ ergibt sich
   - Standardisieren mit der entsprechenden Standardabweichung
     ```
     diff.std <- sqrt((1/101+1/101)*
     (100*sigma[2]^2+100*sigma[1]^2)/(101+101-2))
     ```
   - Prüfgröße:
     pg.T2 <-m.diff/diff.std2
     2.796
   - Vergleich mit der t-Verteilung:
     1-pt(pg.T2, df=200)
     0.003

Statistisches Testen VI



Mit nur sehr geringer Wahrscheinlichkeit (0.003%) kann eine solche Prüfgröße $pg.T2$ beobachtet werden, wenn wenn der Mittelwert in Europa und kleiner als der in Nicht-Europa ist.

Fünf Schritte zum Testergebnis

I. Hypothesen aufstellen
II. Betrachtung der Daten
III. Aufstellen der Prüfgröße
IV. Durchführen des Tests
V. Testentscheidung

I. Hypothesen aufstellen

- **Was soll verglichen werden?**
  - Mittelwerte von unabhängigen Gruppen
  - Mittelwert gegen einen festen Wert
  - Gepaarte Messungen
- **Einseitige oder zweiseitige Fragestellung?**
  - Einseitige Fragestellung:
    $H_0 : \mu_1 \leq \mu_2$ gegen $H_1 : \mu_1 > \mu_2$
  - Zweiseitige Fragestellung:
    $H_0 : \mu_1 = \mu_2$ gegen $H_1 : \mu_1 \neq \mu_2$
- **Aufstellen der eigentlich interessierenden Alternativhypothese H_1 und der Nullhypothese H_0**
- **Signifikanzniveau α festlegen**

Fehler bei statistischen Tests

	Entscheidung H_0	Entscheidung H_1
H_0 wahr	richtig	Fehler erster Art α
H_1 wahr	Fehler zweiter Art (β)	richtig

- Fehler erster Art (α-Fehler):
  Obwohl H_0 wahr ist, entscheidet man sich für H_1
  (Falsch positives Testergebnis)
- Fehler zweiter Art (β-Fehler):
  Obwohl H_1 wahr ist, entscheidet man sich für H_0
  (Falsch negatives Testergebnis)

- Können Verteilungsannahmen getroffen werden?
  - Ja: Parametrische Tests
  - Nein: Nicht-Parametrische Tests
- Weitere Annahmen wie z.B. Varianzgleichheit in den Gruppen

Aus Schritt I. und II. folgt die Auswahl eines geeigneten Tests und alle weiteren Schritte!

III. Aufstellen der Prüfgröße

- Aus den Hypothesen ergibt sich die Form der Prüfgröße, z.B. die Mittelwertsdifferenz
- Standardisieren der Prüfgröße mit:
  - unter H_0 gültigen Erwartungswert
  - unter H_0 gültigen Standardabweichung
- Festlegen der Verteilung, die unter H_0 gültig ist

IV./V. Durchführen des Tests und Testentscheidung

Hier sind zwei Werte entscheidend:

- **Kritischer Wert** κ**:** Welchen Wert darf die Prüfgröße bei gegebenem Signifikanzniveau α maximal/minimal annehmen, wenn H_0 tatsächlich gültig ist
- p-**Wert:** Wahrscheinlichkeit, die vorliegenden Daten zu beobachten, wenn H_0 gültig ist

Entscheidung H_0 **ablehnen, falls:**

- die Prüfgröße größer als der kritische Wert ist (bzw. kleiner als der kritische Wert bei einigen nonparametrischen Tests)
- falls der p-Wert kleiner dem vorher festgelegten Signifikanzniveau α ist

t-Test - gegen festen Wert (Einstichproben-t-Test)

- Vergleich das emp. Populationsmittel $\overline{x}$ einer Population mit einem hypothetischen Mittelwert μ_0
- Voraussetzung: Normalverteilung der Stichprobe
- Varianz wird als unbekannt angenommen und aus den Daten geschätzt

Varianten für die Hypothesen:

1. **Einseitige Fragestellung 1:**
   $H_0 : \overline{x} \leq \mu_0$ gegen $H_1 : \overline{x} > \mu_0$
2. **Einseitige Fragestellung 2:**
   $H_0 : \overline{x} \geq \mu_0$ gegen $H_1 : \overline{x} < \mu_0$
3. **Zweiseitige Fragestellung:**
   $H_0 : \overline{x} = \mu_0$ gegen $H_1 : \overline{x} \neq \mu_0$

2. Teststatistik

- Teststatistik

$$T = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$$

- Schätzung der Standardabweichung σ durch:

$$s = \left[\frac{\sum_{i=1}^{n} (\overline{x} - x_i)^2}{n-1} \right]^{0.5}$$

3. Kritische Bereiche

1. **Einseitige Fragestellung 1:**
   $T > t_{1-\alpha}(df = n - 1)$

2. **Einseitige Fragestellung 2:**
   $T < t_{\alpha}(df = n - 1)$

3. **Zweiseitige Fragestellung:**
   $|T| > t_{1-\alpha/2}(df = n - 1)$

t-Test für unabhängige Stichproben (Zweistichproben-t-Test)

1. Ziel, Hypothesen und Voraussetzungen

- Vergleich das emp. Populationsmittel $\overline{x}_1$ und $\overline{x}_2$ miteinander
- Voraussetzung: Normalverteilung der Stichproben
- Varianz der Populationen unbekannt
- 2 Varianten: Varianzen der Populationen gleich oder ungleich

Varianten für die Hypothesen:

1. **Einseitige Fragestellung 1:**
   $H_0 : \overline{x}_1 \leq \overline{x}_2$ gegen $H_1 : \overline{x}_1 > \overline{x}_2$
2. **Einseitige Fragestellung 2:**
   $H_0 : \overline{x}_1 \geq \overline{x}_2$ gegen $H_1 : \overline{x}_1 < \overline{x}_2$
3. **Zweiseitige Fragestellung:**
   $H_0 : \overline{x}_1 = \overline{x}_2$ gegen $H_1 : \overline{x}_1 \neq \overline{x}_2$

- Teststatistik

$$T = \frac{\overline{x}_1 - \overline{x}_2}{s} \cdot \sqrt{n}$$

- Schätzung der Standardabweichung σ durch:

$$s = \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \cdot \frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 1} \right]^{0.5}$$

wobei s_1 und s_2 die Standardvarianzschätzer für die Populationen sind

1. **Einseitige Fragestellung 1:**
   $T > t_{1-\alpha}(n_1 + n_2 - 2)$
2. **Einseitige Fragestellung 2:**
   $T < t_{\alpha}(n_1 + n_2 - 2)$
3. **Zweiseitige Fragestellung:**
   $|T| > t_{1-\alpha/2}(n_1 + n_2 - 2)$

t-Test für Paardifferenzen

- Teste die Differenz $\overline{d} = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} x_{1i} - x_{2i}$ miteinander gepaarter Stichproben (x_{1i}, x_{2i})
- Typisches Bsp.: Messen eines Blutwertes vor und nach einer med. Behandlung
- Voraussetzung: Normalverteilung der Stichproben

Varianten für die Hypothesen:

1. **Einseitige Fragestellung 1:**
   $H_0 : d \leq 0$ gegen $H_1 : d > 0$
2. **Einseitige Fragestellung 2:**
   $H_0 : d \geq 0$ gegen $H_1 : d < 0$
3. **Zweiseitige Fragestellung:**
   $H_0 : d = 0$ gegen $H_1 : d \neq 0$

- Teststatistik

$$T = \frac{\overline{d}}{s} \cdot \sqrt{n}$$

- Schätzung der Standardabweichung σ durch:

$$s = \left[\frac{\sum_{i=1}^{n} (\overline{d} - d_i)^2}{n-1} \right]^{0.5}$$

1. **Einseitige Fragestellung 1:**
   $T > t_{1-\alpha}(df = n-1)$
2. **Einseitige Fragestellung 2:**
   $T < t_{\alpha}(df = n-1)$
3. **Zweiseitige Fragestellung:**
   $|T| > t_{1-\alpha/2}(df = n-1)$

Der Wilcoxon-Rangsummen-Test

- Teste nicht-parametrisch, ob zwei Population den gleichen Median besitzen
- Zu verwenden, wenn Vor. für den t-Test nicht erfüllt sind
- Benötigt KEINE konkrete Verteilungsannahme
- Alternative für den t-Test

Varianten für die Hypothesen:

1. **Einseitige Fragestellung 1:**
   $H_0 : x_{1,med} \leq x_{2,med}$ gegen $H_1 : x_{1,med} > x_{2,med}$
2. **Einseitige Fragestellung 2:**
   $H_0 : x_{1,med} \geq x_{2,med}$ gegen $H_1 : x_{1,med} < x_{2,med}$
3. **Zweiseitige Fragestellung:**
   $H_0 : x_{1,med} = x_{2,med}$ gegen $H_1 : x_{1,med} \neq x_{2,med}$

- Bilde für sämtlichen Beobachtungen $x_{11}, \ldots x_{1n_1}, x_{21}, \ldots x_{2n_2}$ Ränge $rg(x_{11}), \ldots rg(x_{1n_1}), rg(x_{21}), \ldots rg(x_{2n_2})$
- Teststatistik:

$$R = \sum_{i=1}^{n_1} rg(x_{1i})$$

- Wertebereich: $\frac{n_1(n_1+1)}{2} < R < \frac{(n_1+n_2)(n_1+n_2+1)}{2} - \frac{n_1(n_1+)}{2}$
- Nullverteilung von R liegt tabelliert vor
- Approximation durch die Normalverteilung ab einer Stichprobengröße von ca. 20 möglich

1. **Einseitige Fragestellung 1:**
   $R > w_{1-\alpha}(n_1, n_2)$
2. **Einseitige Fragestellung 2:**
   $R < w_{\alpha}(n_1, n_2)$
3. **Zweiseitige Fragestellung:**
   $R > w_{1-\alpha/2}(n_1, n_2)$ oder $R < w_{\alpha/2}(n_1, n_2)$

t-Test und Wilcoxon-Rangsummen - Test in R - Praktische Durchführung

t-Test in R

`t.test(x, y, alternative, paired, var.equal)`

Erklärung der Parameter:

- `x,y = NULL`: Die Daten, beim *t*-Test für eine Population genügt es, *x* anzugeben
- `alternative = c("two.sided", "less", "greater")`: Varianten für die Alternativhypothese
- `var.equal = TRUE`: Gibt an, ob Varianzgleichheit bei den Populationen vorliegt
- `paired`: Gibt an, ob *x* und *y* als gepaarte Stichprobe anzusehen sind

Wilcoxon-Rangsummen - Test in R

`wilcox.test(x, y, alternative, paired, exact)`

Erklärung der Parameter:

- Im wesentlichen analog zum t-Test
- `exact`: Soll die Teststatistik exakt bestimmt werden, oder per Approximation an die Normalverteilung?

Beispiel:

- Nettokaltmieten pro m^2 für 1- (X) und 2-Raum (Y) Wohnungen
- Gibt es einen Unterschied zwischen beiden Gruppen?
- Wir untersuchen diese Frage per Wilcoxon- und t-Test

	1	2	3	4	5
X	8.70	11.28	13.24	8.37	12.16
Y	3.36	18.35	5.19	8.35	13.10

	6	7	8	9	10
X	11.04	10.47	11.16	4.28	19.54
Y	15.65	4.29	11.36	9.09	

t-Test

```
miete <- read.csv(‘‘Miete.csv‘‘)
attach(miete)
t.test(X,Y, var.equal = FALSE, paired = FALSE)
```

R-Ausgabe:

```
Welch Two Sample t-test
data: X and Y
t = 0.5471, df = 14.788, p-value = 0.5925
alternative hypothesis: true difference in means is not
equal to 0
```

$p > 0.05$, kein signifikanter Unterschied

wilcox.test(X,Y, exact = TRUE)

R-Ausgabe:

```
Wilcoxon rank sum test
data: X and Y
W = 51, p-value = 0.6607
alternative hypothesis: true location shift is not
equal to 0
```

$p > 0.05$, kein signifikanter Unterschied

Aufgabenkomplex 3