CAS Datenanalyse HS16 - DeskStat

Qualitative Daten

Qualitative Daten

- Als qualitative (nominale) Merkmale bezeichnet man Merkmale, bei denen sich die Merkmalsausprägungen zwar eindeutig in Kategorien unterscheiden lassen, diese Antworten jedoch keinen mathematischen Wert annehmen können.

Qualitative Daten

- Als qualitative (nominale) Merkmale bezeichnet man Merkmale, bei denen sich die Merkmalsausprägungen zwar eindeutig in Kategorien unterscheiden lassen, diese Antworten jedoch keinen mathematischen Wert annehmen können.
- Streng genommen zählen auch ordinale Merkmale zu den qualitativen Merkmalen. Bei ordinalen Merkmalen kann eine Hierarchie erstellt werden, eine genaue numerische Skalierung ist aber nicht möglich.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.
- **painters** enthält Informationen zu Malern des 18. Jahrhunderts.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.
- **painters** enthält Informationen zu Malern des 18. Jahrhunderts.

```
library(MASS)
head(painters,3)  # oder painters[1:3,]

##              Composition Drawing Colour Expression School
## Da Udine              10       8     16          3      A
## Da Vinci              15      16      4         14      A
## Del Piombo             8      13     16          7      A
```

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.
- Die Variable `school` ist damit qualitativ.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.
- Die Variable `school` ist damit qualitativ.

```
painters$School
```

```
##  [1] A A A A A A A A A A B B B B B B C C C C C C D D D D D D
## [29] D D D D E E E E E E E F F F F G G G G G G G H H H H
## Levels: A B C D E F G H
```

Häufigkeitsverteilung

Definition

Die Häufigkeitsverteilung gibt an, wie oft eine Merkmalsausprägung in einer Variable vorkommt.

Problem: Bestimmen Sie die Häufigkeitsverteilung der Variablen `School` aus `painters`.

Lösung:

```
library(MASS)                  # das MASS-Paket laden
school = painters$School       # die Schulen der Maler
school.freq = table(school)    # Anwenden der table-Funktion
```

Beispiel: Häufigkeitsverteilung

Antwort: Die Häufigkeitsverteilung der Variablen `School` ist:

```
school.freq

## school
##  A  B  C  D  E  F  G  H
## 10  6  6 10  7  4  7  4
```

Beispiel: Häufigkeitsverteilung

Erweiterte Antwort: Mit `cbind` stellen wir das Ergebnis in Spalten dar.

```
cbind(school.freq)
```

```
##   school.freq
## A          10
## B           6
## C           6
## D          10
## E           7
## F           4
## G           7
## H           4
```

Relative Häufigkeitsverteilung

Definition

Die relative Häufigkeitsverteilung gibt an, welchen Anteil die Merkmalsausprägungen einer Variable einnehmen.

Problem: Bestimmen Sie die relative Häufigkeitsverteilung der Variablen `School` aus `painters`.

Lösung:

```
library(MASS)                  # das MASS-Paket laden
school = painters$School       # die Schulen der Maler
school.freq = table(school)    # Anwenden der table-Funktion
school.relfreq = school.freq / nrow(painters)
```

Beispiel: Relative Häufigkeitsverteilung

Antwort: Die Häufigkeitsverteilung der Variablen `School` ist

```
school.relfreq
```

```
## school
##          A          B          C          D          E
## 0.18518519 0.11111111 0.11111111 0.18518519 0.12962963
##          F          G          H
## 0.07407407 0.12962963 0.07407407
```

Beispiel: Relative Häufigkeitsverteilung

Erweiterte Antwort: Wir drucken Spalten und weniger Stellen.

```
old=options(digits=3)
head(cbind(school.relfreq*100))

##    [,1]
## A 18.52
## B 11.11
## C 11.11
## D 18.52
## E 12.96
## F  7.41
```

Balkendiagramm

Definition

Ein Balkendiagramm stellt die Häufigkeitsverteilung von qualitativen Daten durch vertikale Balken graphisch dar.

Problem: Ein Balkendiagramm der Variable `school` von **painters** gibt mit vertikalen Balken die Anzahl der Maler pro Schule an.

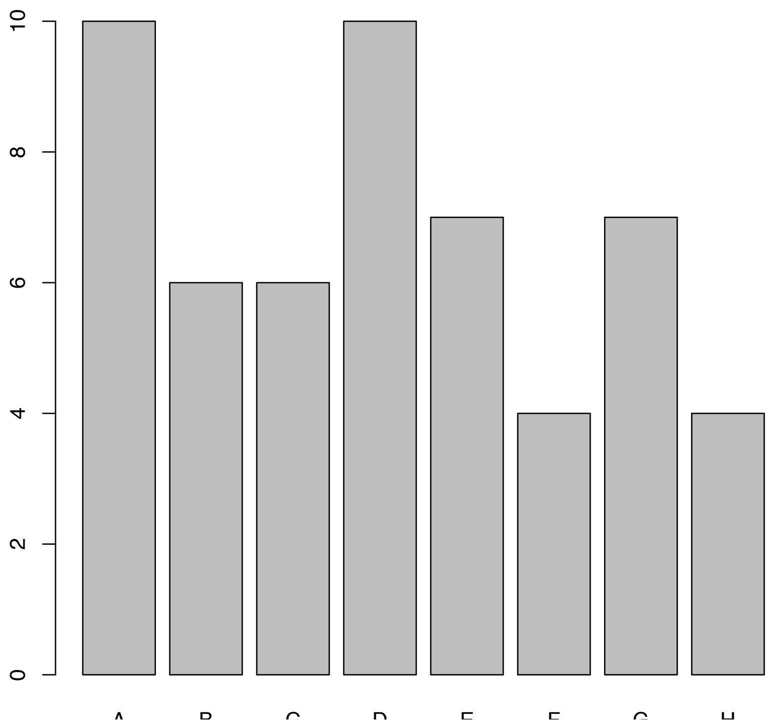
Lösung:

```
options(old)
school=painters$School
school.freq=table(school)
```

Beispiel: Balkendiagramm

Lösung:

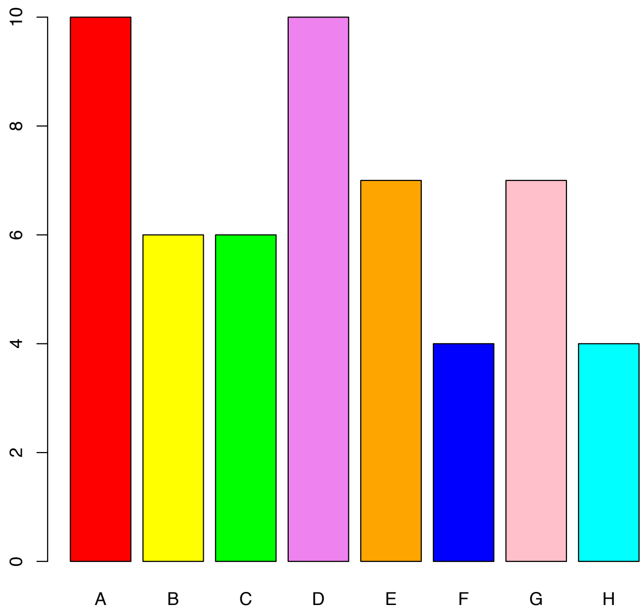
```
barplot(school.freq)
```



Beispiel: Balkendiagramm

Erweiterte Antwort:

```
farben=c("red","yellow","green","violet","orange","blue","pink","c
barplot(school.freq,col=farben)
```



Kuchendiagramm

Definition

Ein Kuchendiagramm stellt die Häufigkeitsverteilung von qualitativen Daten durch Pizzastücke graphisch dar.

Problem: Ein Kuchendiagramm der Variable `school` von **painters** gibt mit Pizzastücken die Anzahl der Maler pro Schule an.

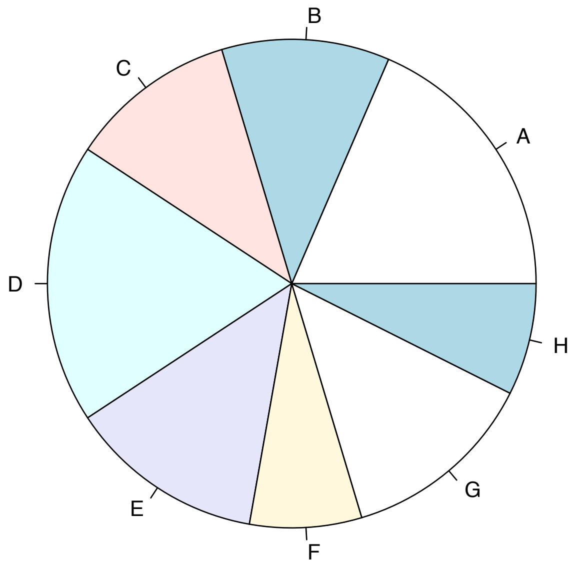
Lösung:

```
school=painters$School
school.freq=table(school)
```

Beispiel: Kuchendiagramm

Lösung:

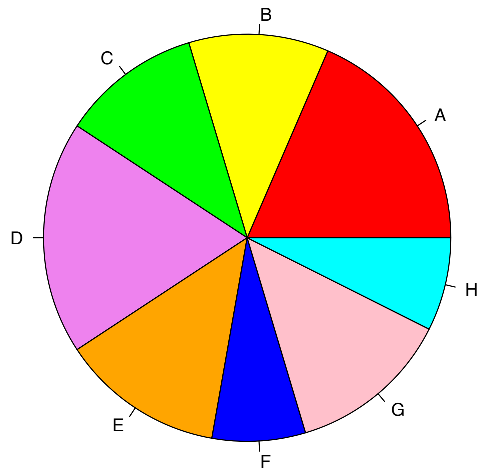
```
pie(school.freq)
```



Beispiel: Kuchendiagramm

Erweiterte Antwort:

```
farben=c("red","yellow","green","violet","orange","blue","pink","c
pie(school.freq,col=farben)
```



Gruppenstatistik

Problem: Bestimmen Sie den durchschnittlichen Wert von `Composition` in der Schule C.

Lösung:

```
school=painters$School
c_school= school=="C"
c_painters = painters[c_school, ]
mean(c_painters$Composition)

## [1] 13.16667
```

Gruppenstatistik

Erweiterte Antwort: Anstatt den Durchschnittswert von `Composition` jeder Schule manuell zu bestimmen, verwenden wir die Funktion `tapply`:

```
tapply(painters$Composition, painters$School, mean)

##        A        B        C        D        E        F
## 10.40000 12.16667 13.16667  9.10000 13.57143  7.25000
##        G        H
## 13.85714 14.00000
```

CAS Datenanalyse HS16 - DeskStat

Quantitative Daten

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.
- **faithful** zeigt die Eruptionsdauer und die Wartezeit zwischen den Eruptionen des Geysirs Old Faithful im Yellowstone Nationalpark.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.
- **faithful** zeigt die Eruptionsdauer und die Wartezeit zwischen den Eruptionen des Geysirs Old Faithful im Yellowstone Nationalpark.

```
head(faithful,3)

##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
```

Häufigkeitsverteilung quantitativer Daten

Definition

Die Häufigkeitsverteilung einer quantitativen Variablen gibt an, wie sich die Merkmalswerte über nicht-überlappende Intervalle verteilen.

Problem: Bestimmen Sie die Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Häufigkeitsverteilung quantitativer Daten

Lösung:

```
# Schritt 1: Spannweite bestimmen
duration = faithful$eruptions
range(duration)

## [1] 1.6 5.1

# Schritt 2: Spannweite in gleichlang, nichtüberlappende Interval…
# Runde Spannweite zu [1.5,5.5]
breaks = seq(1.5, 5.5, by=0.5)
breaks

## [1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

Häufigkeitsverteilung quantitativer Daten

Lösung:

```
# Schritt 3: Eruptionszeiten in Intervalle verteilen
duration.cut = cut(duration, breaks, right=FALSE)
# Schritt 4: Häufigkeit pro Intervall bestimmen
duration.freq = table(duration.cut)
```

Antwort: Die Häufigkeitsverteilung der Variablen `eruption` ist:

```
duration.freq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##      51      41       5       7      30      73      61
## [5,5.5)
##       4
```

Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Die stellen die Verteilung in einer Spalte dar.

```
cbind(duration.freq)

##           duration.freq
## [1.5,2)              51
## [2,2.5)              41
## [2.5,3)               5
## [3,3.5)               7
## [3.5,4)              30
## [4,4.5)              73
## [4.5,5)              61
## [5,5.5)               4
```

Histogramm

Definition

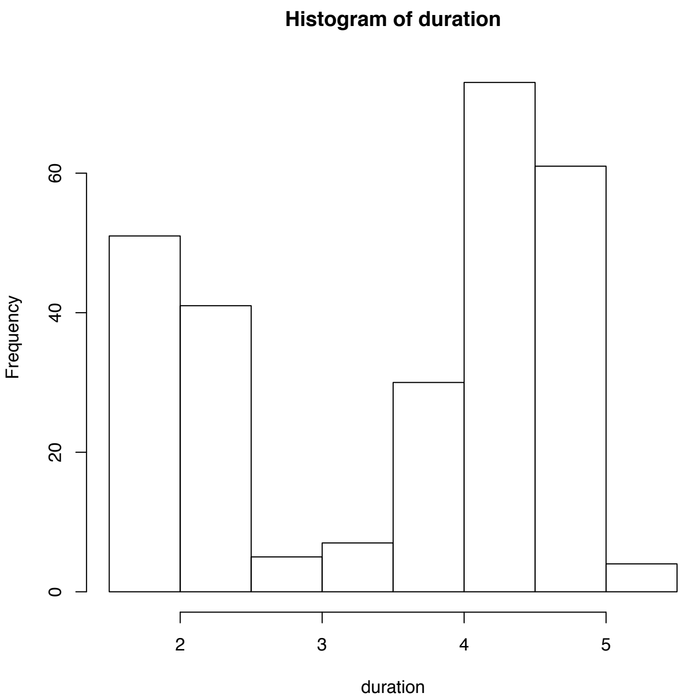
Ein Histogramm stellt die Häufigkeitsverteilung einer quantitativen Variablen graphisch dar.

Problem: Zeichnen Sie das Histogramm der Eruptionszeiten aus **faithful**.

Histogramm

Lösung:

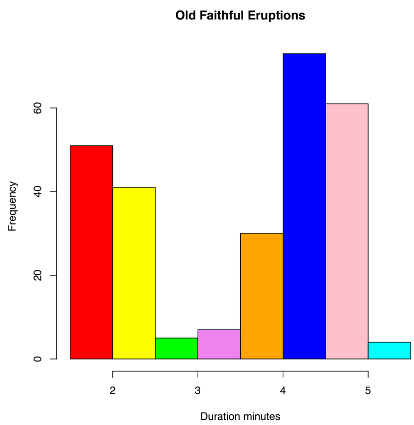
```
duration = faithful$eruptions
hist(duration, right=FALSE) # Intervalle sind rechts offen
```



Histogramm

Erweiterte Antwort: Wir verwenden Farben und fügen Titel sowie Achenbeschriftungen ein.

```
colors = c("red", "yellow", "green", "violet", "orange", "blue",
"pink", "cyan")
hist(duration, right=FALSE, col=colors,
main="Old Faithful Eruptions", xlab="Duration minutes")
```



Relative Häufigkeitsverteilung quantitativer Daten

Definition

Die relative Häufigkeitsverteilung einer quantitativen Variablen gibt an, wie sich die Anteile der Merkmalswerte über nicht-überlappende Intervalle verteilen.

Problem: Bestimmen Sie die relative Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Relative Häufigkeitsverteilung quantitativer Daten

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.relfreq = duration.freq/nrow(faithful)
duration.relfreq

## duration.cut
##     [1.5,2)     [2,2.5)     [2.5,3)     [3,3.5)     [3.5,4)
## 0.18750000 0.15073529 0.01838235 0.02573529 0.11029412
##     [4,4.5)     [4.5,5)     [5,5.5)
## 0.26838235 0.22426471 0.01470588
```

Relative Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Wir zeigen weniger Stellen an.

```
old = options(digits=1)
duration.relfreq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##    0.19    0.15    0.02    0.03    0.11    0.27    0.22
## [5,5.5)
##    0.01

options(old)  # alter Status
```

Relative Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Wir zeigen weniger Stellen an.

```
duration.percentage = duration.relfreq*100
old = options(digits=3)
head(cbind(duration.freq, duration.percentage),5)

##          duration.freq duration.percentage
## [1.5,2)             51               18.75
## [2,2.5)             41               15.07
## [2.5,3)              5                1.84
## [3,3.5)              7                2.57
## [3.5,4)             30               11.03

options(old)
```

Kumulierte Häufigkeitsverteilung

Definition

Die kumulierte Häufigkeitsverteilung einer quantitativen Variablen summiert die Anteile der Merkmalswerte über nicht-überlappende Intervalle.

Problem: Bestimmen Sie die kumulierte Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Kumulierte Häufigkeitsverteilung

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumfreq

## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##      51      92      97     104     134     207     268
## [5,5.5)
##     272
```

Kumulierte Häufigkeitsverteilung

Erweiterte Antwort: Wir präsentieren das Ergebnis als Spalte.

```
cbind(duration.cumfreq)
```

```
##          duration.cumfreq
## [1.5,2)                51
## [2,2.5)                92
## [2.5,3)                97
## [3,3.5)               104
## [3.5,4)               134
## [4,4.5)               207
## [4.5,5)               268
## [5,5.5)               272
```

Kumulierte Häufigkeitsverteilungskurve

Definition

Die kumulierte Häufigkeitsverteilungskurve einer quantitativen Variablen stellt die summierten Häufigkeiten der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte Häufigkeitsverteilungskurve der Eruptionszeiten aus **faithful**.

Kumulierte Häufigkeitsverteilungskurve

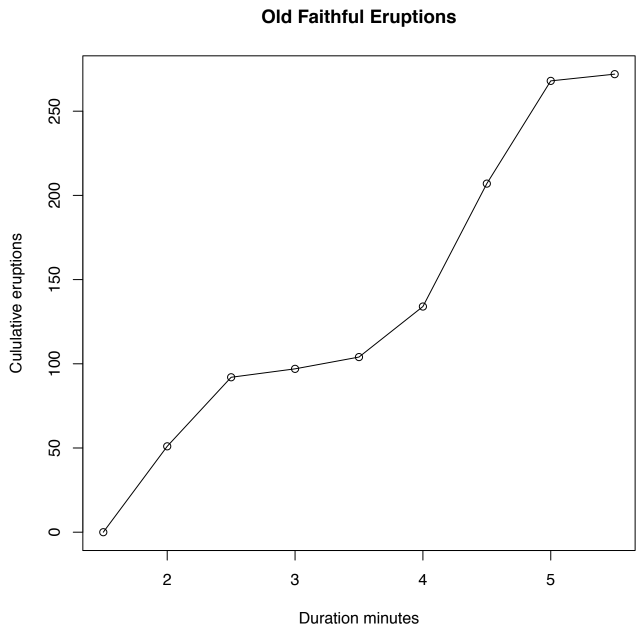
Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq0 = c(0,cumsum(duration.freq))
```

Kumulierte Häufigkeitsverteilungskurve

Lösung:

```
plot(breaks, duration.cumfreq0, main="Old Faithful Eruptions",
xlab="Duration minutes", ylab="Cululative eruptions")
lines(breaks, duration.cumfreq0)
```



Kumulierte relative Häufigkeitsverteilung

Definition

Die kumulierte Häufigkeitsverteilung einer quantitativen Variablen stellt die summierten Anteile der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte relative Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Kumulierte relative Häufigkeitsverteilung

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumrelfreq = duration.freq/nrow(faithful)
duration.cumrelfreq

## duration.cut
##    [1.5,2)    [2,2.5)    [2.5,3)    [3,3.5)    [3.5,4)
## 0.18750000 0.15073529 0.01838235 0.02573529 0.11029412
##    [4,4.5)    [4.5,5)    [5,5.5)
## 0.26838235 0.22426471 0.01470588
```

Kumulierte relative Häufigkeitsverteilung

Erweiterte Antwort: Wir drucken weniger Stellen.

```
old = options(digits=2)
duration.cumrelfreq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##   0.188   0.151   0.018   0.026   0.110   0.268   0.224
## [5,5.5)
##   0.015

options(old)
```

Kumulierte relative Häufigkeitsverteilungskurve

Definition

Die kumulierte relative Häufigkeitsverteilungskurve einer quantitativen Variablen stellt die summierten Anteile der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte relative Häufigkeitsverteilungskurve der Eruptionszeiten aus **faithful**.

Kumulierte relative Häufigkeitsverteilungskurve

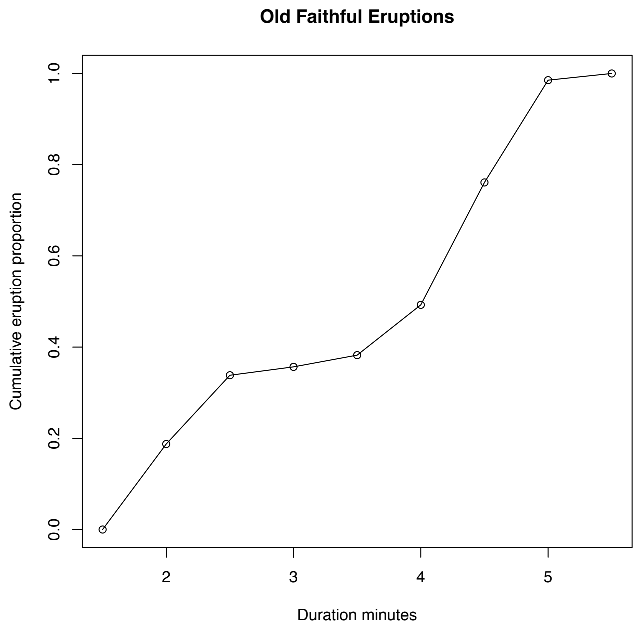
Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumrelfreq = duration.cumfreq/nrow(faithful)
duration.cumrelfreq0 = c(0, duration.cumrelfreq)
```

Kumulierte relative Häufigkeitsverteilungskurve

Lösung:

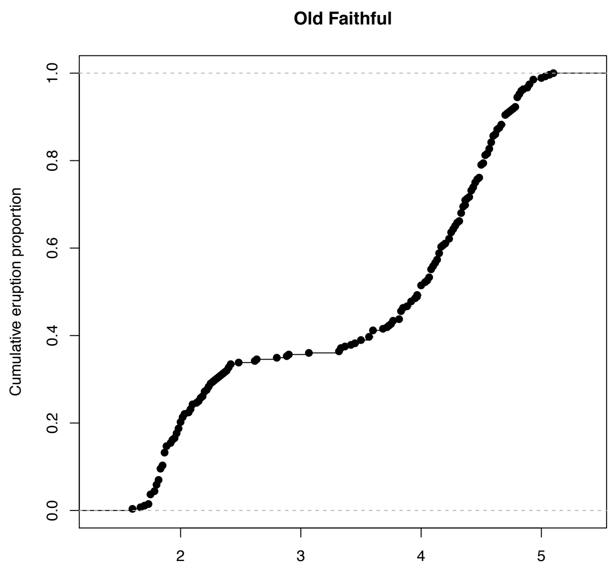
```
plot(breaks, duration.cumrelfreq0, main="Old Faithful Eruptions",
xlab="Duration minutes", ylab="Cumulative eruption proportion")
lines (breaks, duration.cumrelfreq0)
```



Kumulierte relative Häufigkeitsverteilungskurve

Erweiterte Antwort: Wir interpolieren die relative Häufigkeitsverteilung mit dem Befehl `ecdf`.

```
Fn = ecdf(duration)
plot(Fn, main="Old Faithful", xlab="Duration minutes",
ylab="Cumulative eruption proportion")
```



Streudiagramm

Definition

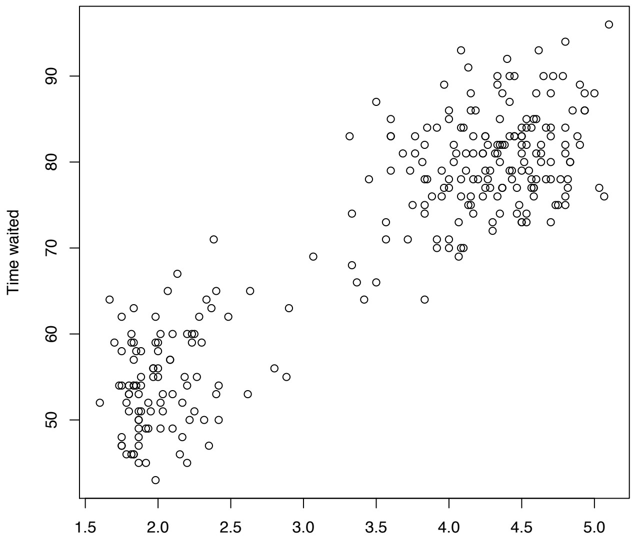
Ein Streudiagramm ist die graphische Darstellung von beobachteten Wertepaaren zweier statistischer Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt.

Problem: Bestimmen Sie das Streudiagramm der Eruptions- und Wartezeiten aus **faithful**.

Streudiagramm

Lösung:

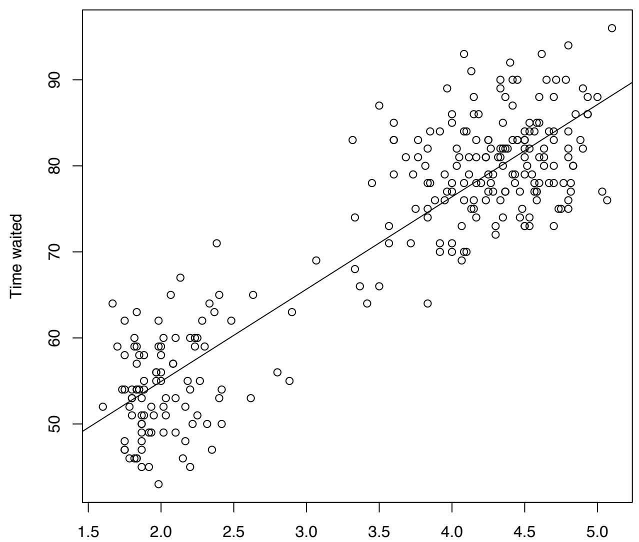
```
duration = faithful$eruptions
waiting = faithful$waiting
plot(duration, waiting, xlab="Erution duration",
ylab="Time waited")
```



Streudiagramm

Erweiterte Antwort: Wir berechnen mit `lm` ein lineares Model der beiden Variablen und fügen mit `ablilne` eine Trendline hinzu.

```
plot(duration, waiting, xlab="Eruption duration",
ylab="Time waited")
abline(lm(waiting ~ duration))
```



CAS Datenanalyse HS16 - DeskStat

Zweidimensionale Häufigkeitsverteilungen

Zweidimensionale Verteilung

Definition

Werden die Gesamtheit aller möglichen Kombinationen und deren Merkmalsausprägungen mit deren absoluten oder auch relativen Häufigkeiten in eine Tabelle eingetragen, so spricht man von der zweidimensionale Häufigkeitsverteilung.

Absolute Zweidimensionale Verteilung

Problem: Laden Sie die Tabelle **Daten_WachstumX**. Bestimmen Sie die absolute zweidimensionale Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche`.

Lösung:

```
load("U:/RFiles/Daten_WachstumX.RData")
# load("~/Documents/RScripts/Daten_WachstumX.RData")
tab = table(Daten_Wachstum$Geschlecht, Daten_Wachstum$Branche)
tab

##
##        Dienstleistung Industrie
##   Frau             26         9
##   Mann             40        25
```

Randverteilung

Problem: Fügen Sie die Randverteilungen zur zweidimensionalen Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche` hinzu.

Lösung:

```
attach(Daten_Wachstum)
tab = table(Geschlecht, Branche)
addmargins(tab)

##             Branche
## Geschlecht Dienstleistung Industrie  Sum
##       Frau             26         9   35
##       Mann             40        25   65
##       Sum              66        34  100
```

Relative Zweidimensionale Verteilung

Problem: Bestimmen Sie die relative zweidimensionale Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche`.

Lösung:

```
tab = table(Geschlecht, Branche)
prop.table(tab)

##             Branche
## Geschlecht Dienstleistung Industrie
##        Frau           0.26      0.09
##        Mann           0.40      0.25
```

Bedingte Verteilung

Problem: Wie verteilen sich die Tätigkeiten in den beiden Branchen innerhalb der Geschlechtergruppen?

Lösung:

```
tab = table(Geschlecht, Branche)
addmargins(prop.table(tab,1))

##             Branche
## Geschlecht Dienstleistung Industrie       Sum
##       Frau      0.7428571  0.2571429 1.0000000
##       Mann      0.6153846  0.3846154 1.0000000
##       Sum       1.3582418  0.6417582 2.0000000
```

Bedingte Verteilung

Problem: Wie verteilen sich die beiden Geschlechter auf die Branchen?

Lösung:

```
tab = table(Geschlecht, Branche)
addmargins(prop.table(tab,2))

##             Branche
## Geschlecht Dienstleistung Industrie       Sum
##       Frau      0.3939394  0.2647059 0.6586453
##       Mann      0.6060606  0.7352941 1.3413547
##       Sum       1.0000000  1.0000000 2.0000000

detach(Daten_Wachstum)
```

CAS Datenanalyse HS16 - DeskStat

Zweidimensionale Häufigkeitsverteilungen

Zweidimensionale Verteilung

Definition

Werden die Gesamtheit aller möglichen Kombinationen und deren Merkmalsausprägungen mit deren absoluten oder auch relativen Häufigkeiten in eine Tabelle eingetragen, so spricht man von der zweidimensionale Häufigkeitsverteilung.

Absolute Zweidimensionale Verteilung

Problem: Laden Sie die Tabelle **Daten_WachstumX**. Bestimmen Sie die absolute zweidimensionale Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche`.

Lösung:

```
load("U:/RFiles/Daten_WachstumX.RData")
# load("~/Documents/RScripts/Daten_WachstumX.RData")
tab = table(Daten_Wachstum$Geschlecht, Daten_Wachstum$Branche)
tab

##
##          Dienstleistung Industrie
##   Frau               26         9
##   Mann               40        25
```

Randverteilung

Problem: Fügen Sie die Randverteilungen zur zweidimensionalen Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche` hinzu.

Lösung:

```
attach(Daten_Wachstum)
tab = table(Geschlecht, Branche)
addmargins(tab)

##             Branche
## Geschlecht Dienstleistung Industrie  Sum
##       Frau             26         9   35
##       Mann             40        25   65
##       Sum              66        34  100
```

Relative Zweidimensionale Verteilung

Problem: Bestimmen Sie die relative zweidimensionale Häufigkeitsverteilung der Merkmale `Geschlecht` und `Branche`.

Lösung:

```
tab = table(Geschlecht, Branche)
prop.table(tab)

##             Branche
## Geschlecht Dienstleistung Industrie
##       Frau           0.26      0.09
##       Mann           0.40      0.25
```

Bedingte Verteilung

Problem: Wie verteilen sich die Tätigkeiten in den beiden Branchen innerhalb der Geschlechtergruppen?

Lösung:

```
tab = table(Geschlecht, Branche)
addmargins(prop.table(tab,1))

##             Branche
## Geschlecht Dienstleistung Industrie       Sum
##       Frau      0.7428571 0.2571429 1.0000000
##       Mann      0.6153846 0.3846154 1.0000000
##       Sum       1.3582418 0.6417582 2.0000000
```

Bedingte Verteilung

Problem: Wie verteilen sich die beiden Geschlechter auf die Branchen?

Lösung:

```
tab = table(Geschlecht, Branche)
addmargins(prop.table(tab,2))

##             Branche
## Geschlecht Dienstleistung Industrie       Sum
##       Frau      0.3939394  0.2647059 0.6586453
##       Mann      0.6060606  0.7352941 1.3413547
##       Sum       1.0000000  1.0000000 2.0000000

detach(Daten_Wachstum)
```

CAS Datenanalyse HS16 - DeskStat

Lorenzkurve und Ginikoeffizient

Lorenzkurve

Definition

Die Lorenzkurve stellt statistische Verteilungen grafisch dar und veranschaulicht dabei das Ausmaß an Ungleichheit respektive relativer Konzentration innerhalb der Verteilung. Grundlage dieser Berechnungen ist eine Liste der von links nach rechts aufsteigend sortierten Einzeleinkommen oder -vermögen.

Lorenzkurve

Problem: Bestimmen Sie die Lorenzkurve von Beispiel 12 des Foliensatzes „Folien Kapitel 1 Teil 2.pdf“.

Lorenzkurve

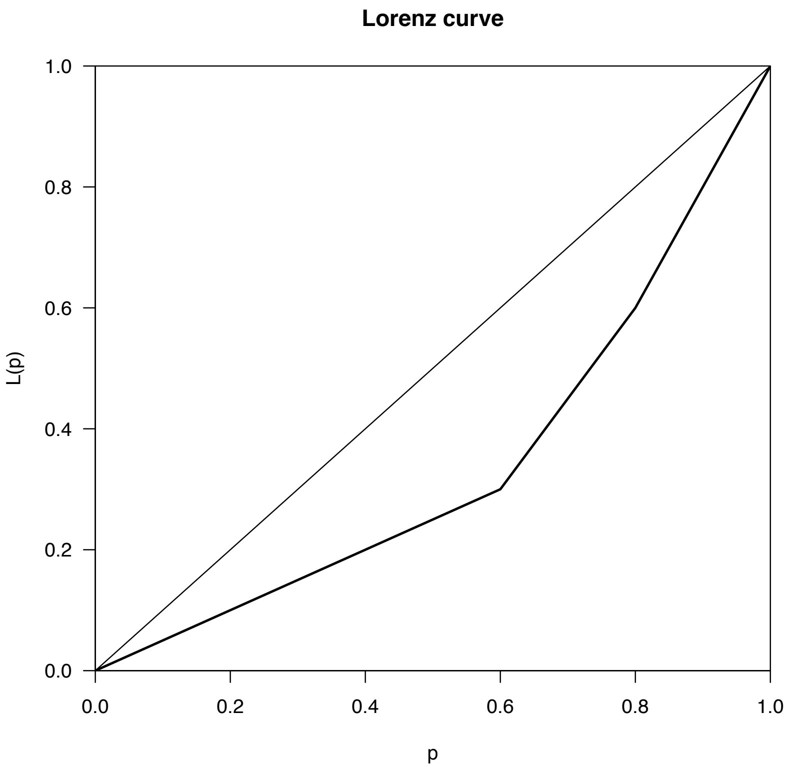
Lösung:

```
install.packages("ineq", repos="http://cran.rstudio.com/")

##
## The downloaded binary packages are in
##  /var/folders/8t/66zyqwx177q7xz30x8cbqgd80000gp/T//Rtmpl5xGmE/d

library("ineq")
einkommen = c(1000, 1000, 1000, 3000, 4000)
```

Lorenzkurve

Lösung:

```
Lc(einkommen, plot=TRUE)
```

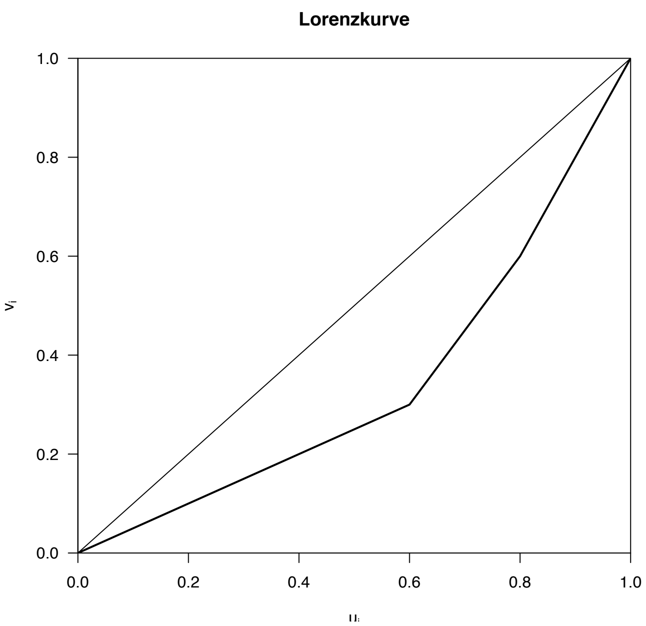


Lorenzkurve

Der Befehl `plot` bietet mehr Komfort.

Lösung:

```
Lcx = Lc(einkommen)

plot(Lcx, main="Lorenzkurve", xlab=expression(u[i]), ylab=expressi
```



Ginikoeffizient

Definition

Der Ginikoeffizient oder auch Gini-Index ist ein statistisches Mass zur Darstellung von Ungleichverteilungen. Der Gini-Koeffizient wird aus der Lorenz-Kurve abgeleitet und nimmt einen Wert zwischen 0 (bei einer gleichmäßigen Verteilung) und 1 (wenn nur eine Person das komplette Einkommen erhält, d.h. bei maximaler Ungleichverteilung) an. Er beträgt das Zweifache der Fläche zwischen der Lorenzkurve und der Geraden $y = x$.

Ginikoeffizient

Problem: Bestimmen Sie den Ginikoeffizienten sowie den normierten Ginikoeffizienten von Beispiel 12 des Foliensatzes „Folien Kapitel 1 Teil 2.pdf“.

Ginikoeffizient

Lösung:

```
Gini(einkommen)

## [1] 0.32

GiniKorrigiert = function(x){ifelse(length(x)==1, NA,
 Gini(x)/(1-1/length(x)))}
GiniKorrigiert(einkommen)

## [1] 0.4
```

CAS Datenanalyse HS16 - DeskStat

Numerische Masszahlen

Arithmetischer Mittelwert

Definition

Das arithmetische Mittel (auch Durchschnitt) ist derjenige Mittelwert, der als Quotient aus der Summe der betrachteten Zahlen und ihrer Anzahl berechnet ist

Problem: Bestimmen Sie die durchschnittliche Eruptionsdauer aus **faithful**.

Lösung:

```
duration = faithful$eruptions
mean(duration)

## [1] 3.487783
```

Median (Zentralwert)

Definition

Der Median (oder Zentralwert) einer Auflistung von Zahlenwerten ist der Wert, der an der mittleren (zentralen) Stelle steht, wenn man die Werte der Größe nach sortiert.

Problem: Bestimmen Sie den Median der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
median(duration)

## [1] 4
```

Quartile

Definition

Quartile zerlegen eine sortierte Datenreihe von Beobachtungen in vier (annähernd) gleich grosse Abschnitte oder Klassen.

- Das erste Quartil teilt die geordnete Datenreihe in das untere Viertel und das obere Dreiviertel. Das erste Quartil wird auch unteres Quartil genannt (abgekürzt Q_1).
- Das zweite Quartil ist der Median.
- Das dritte Quartil teilt die geordnete Datenreihe in das untere Dreiviertel und das obere Viertel. Das dritte Quartil wird auch oberes Quartil genannt (abgekürzt Q_3).

Quartile

Problem: Bestimmen Sie die Quartile der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
quantile(duration)   # Achtung: quaNtile!

##       0%      25%      50%      75%     100%
## 1.60000 2.16275 4.00000 4.45425 5.10000
```

Quantile (Perzentile)

Definition

Quantile (genauer p-Quantile) sind Werte, die eine Menge von Daten in zwei Teile spalten, und zwar so, dass mindestens ein Anteil p kleiner oder gleich dem p-Quantil ist, und mindestens ein Anteil $1 - p$ grösser oder gleich dem p-Quantil.

Man bezeichnet Quantile entweder durch den Anteil p, oder durch eine Prozentzahl. So ist z.B. ein 0.2-Quantil ist dasselbe wie ein 20%-Quantil.

Quantile (Perzentile)

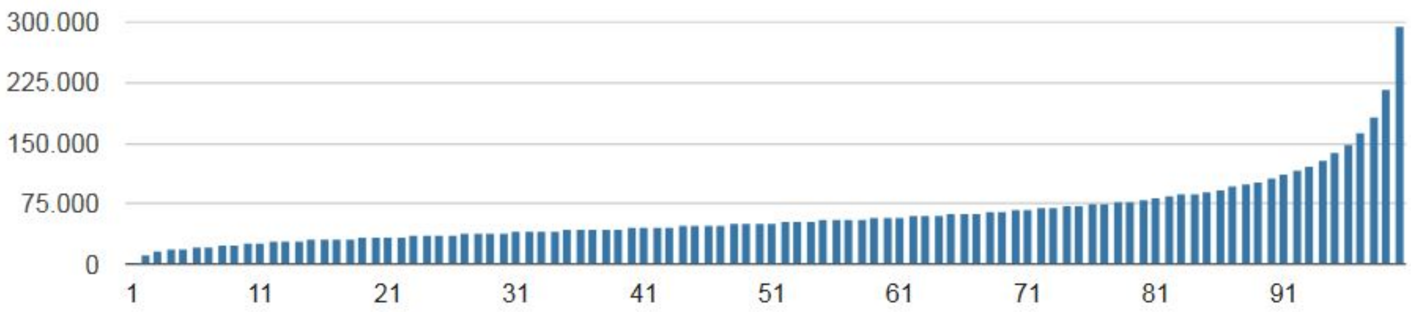
Problem: Bestimmen Sie das 0.32-Quantil, das 0.57-Quantil und das 98%-Quantil der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
quantile(duration, c(0.32, 0.57, 0.98))

##     32%     57%     98%
## 2.39524 4.13300 4.93300
```

Anwendung der Quantile: Einkommensverteilung

Die folgende Grafik zeigt die Einkommensverteilung in der Schweiz für jedes Einkommens-Perzentil der Bevölkerung. Diese Grafik wird auch als Pen Parade bezeichnet.



Quelle: BAKBASEL

Spannweite

Definition

Die Spannweite eines Datensatzes ist die Differenz aus dem grössten und dem kleinesten Wert:

Spannweite = Maximum - Minimum

Problem: Bestimmen Sie die Spannweite der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
max(duration)-min(duration)

## [1] 3.5
```

Interquartilsabstand

Definition

Der Interquarilsabstand eines Datensatzes ist die Differenz aus dem oberen und dem unteren Quartil:

$$\text{IQR} = Q_3 - Q_1$$

Problem: Bestimmen Sie den Interquartilsabstand der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
IQR(duration)

## [1] 2.2915
```

Boxplot

Definition

Ein Boxplot ist eine grafische Zusammenfassung der folgenden fünf Punkte:

Minium - 1.Quartil - Median - 3.Quartil - Maximum

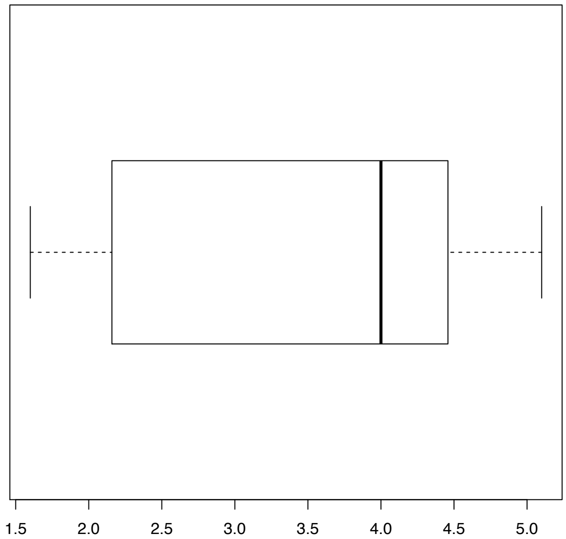
Bemerkung: Whiskers werden meistens nicht bis zum Minimum bzw. Maximum gezeichnet. Falls die Whiskers länger sind als $1.5 \cdot \text{IQR}$, werden sie nicht bis zum letzten Punkt gezeichnet, sondern nur bis zum letzten Punkt der weniger als das 1.5-fache des IQR von der Box entfernt ist. Alle Datenpunkte, die ausserhalb der Whiskers liegen, werden als Ausreisser separat eingezeichnet.

Boxplot

Problem: Bestimmen Sie den Boxplot der Eruptionsdauern.

Lösung:

```
duration = faithful$eruptions
boxplot(duration, horizontal=TRUE)
```



Varianz

Definition

Die Varianz ist numerisches Streumass für die Abweichung eines

Datensatz vom Mittelwert.

Stichprobenvarianz:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Populationsvarianz:

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Varianz

Problem: Bestimmen Sie die beidne Varianzen der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
# Stichprobenvarianz
var(duration)

## [1] 1.302728

# Populationsvarianz
var(duration)*(length(duration)-1)/length(duration)

## [1] 1.297939
```

Standardabweichung

Definition

Die Standardabweichung ist die Quadratwurzel aus der Varianz.

Stichprobenstandardabweichung:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Populationsstandardabweichung:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Standardabweichung

Problem: Bestimmen Sie die Standardabweichungen der Eruptionsdauern aus **faithful**.

Lösung:

```
duration = faithful$eruptions
# Stichprobenstandardabweichung
sd(duration)

## [1] 1.141371

# Populationsstandardabweichung
sqrt(var(duration)*(length(duration)-1)/length(duration))

## [1] 1.139271
```

Populationsvarianz und -standardabweichung

Bemerkung: Werden Populationsvarianz und -standardabweichung oft benötigt, empfiehlt es sich, diese als eigene Funktionen zu definieren.

```
varianz <- function(x) {n=length(x) ; var(x) * (n-1) / n}
stdabw <- function(x) {n=length(x) ; sqrt(var(x) * (n-1) / n)}
duration = faithful$eruptions
# Populationsvarianz
varianz(duration)

## [1] 1.297939

# Populationsstandardabweichung
stdabw(duration)

## [1] 1.139271
```

Kovarianz

Definition

Die Kovarianz ist eine nichtstandardisierte Masszahl für den linearen Zusammenhang zweier statistischer Variablen.

Stichprobenkovarianz:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Populationskovarianz:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Kovarianz

Problem: Bestimmen Sie die Kovarianz der Eruptionsdauern und Wartezeiten aus **faithful**.

Lösung:

```
duration = faithful$eruptions
waiting = faithful$waiting
# Stichprobenkovarianz und Populationskovariant
cov(duration, waiting)

## [1] 13.97781

cov(duration, waiting)*(length(duration)-1)/length(duration)

## [1] 13.92642
```

Korrelationskoeffizient

Definition

Der Korrelationskoeffizient ist ein dimensionsloses Mass für den Grad des linearen Zusammenhangs zwischen zwei Variablen. Er kann Werte zwischen -1 und +1 annehmen.

$$\text{Korrelationskoeffizient } r = \frac{s_{xy}}{s_x \cdot s_y}$$

Korrelationskoeffizient

Problem: Bestimmen Sie die Korrelation zwischen den Eruptionsdauern und den Wartezeiten aus **faithful**.

Lösung:

```
duration = faithful$eruptions
waiting = faithful$waiting
cor(duration, waiting)

## [1] 0.9008112
```

CAS Datenanalyse HS16 - DeskStat

Wahrscheinichkeitsverteilungen

Zufallsvariablen

Definition

Eine Variable X ist eine Zufallsvariable, wenn der Wert, den X annimmt, von dem Ausgang eines Zufallsexperiments abhängt. Eine Zufallsvariable ordnet jedem Ergebniss eines Zufallsexperiments einen numerischen Wert zu.

Zufallsvariablen werden meist mit Großbuchstaben geschrieben.

Zufallsvariablen

Bemerkung: Zufallsvariablen sind daher Funktionen, die jedem Ergebnis eine (reelle) Zahl zuordnen. Sie haben also nicht direkt etwas mit Zufall zu tun. Da nun Ergebnisse durch Zahlen repräsentiert werden, kann mit ihnen gerechnet werden.

Wahrscheinlichkeitsverteilungen

Definition

Eine Wahrscheinlichkeitsverteilung beschreibt, wie sich die Werte einer Zufallsvariablen verteilen.

Binomialverteilung

Definition

Die Binomialverteilung beschreibt die Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben („Erfolg“ oder „Misserfolg“). Solche Versuchsserien werden auch Bernoulli-Prozesse genannt.

Bezeichnet p die Wahrscheinlichkeit eines erfolgreichen Versuchs, so bestimmt sich die Wahrscheinlichkeit für x erfolgreiche Ergebnisse in n unabhängigen Versuchen folgendermassen:

$$B(x|p, n) = \binom{n}{x} p^x (1-p)^{n-x} \text{ für } x \in \mathbb{N}$$

Binomialverteilung

Problem: Eine Multiple-Choise-Prüfungen besteht aus 12 Fragen. Jede Fragen gibt 5 verschiedenen Antworten, von denen aber nur jeweils eine Antwort richtig ist. Ein Student löst die Aufgaben nach dem Zufallsprinzip. Bestimmen Sie die Wahrscheinlichkeit dafür, dass der Student maximal vier korrekte Antworten gibt.

Binomialverteilung

Antwort: Für eine korrekten Antwort gilt $p = 0.2$. Die Wahrscheinlichkeit für genau 4 richtige Antworten finden wir mit:

```
dbinom(4, size=12, prob=0.2)

## [1] 0.1328756
```

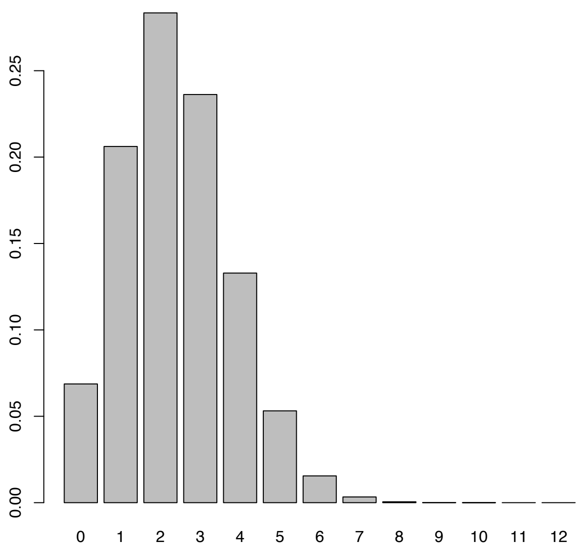
Die Wahrscheinlichkeit für maximal 4 korrekte Antworten ist somit:

```
dbinom(4, size=12, prob=0.2) +
+ dbinom(3, size=12, prob=0.2) +
+ dbinom(2, size=12, prob=0.2) +
+ dbinom(1, size=12, prob=0.2) +
+ dbinom(0, size=12, prob=0.2)

## [1] 0.9274445
```

Binomialverteilung

Erweiterte Antwort:

```
yprob <- dbinom(0:12,size=length(0:12)-1,prob = 1/5)
names(yprob) <- 0:12
barplot(yprob)
```



Binomialverteilung

Erweiterte Antwort: Alternativ können wir die kummulierte Wahrscheinlichkeit direkt berechnen mit:

```
pbinom(4, size=12, prob=0.2)

## [1] 0.9274445
```

Die Wahrscheinlichkeit für vier oder weniger korrekte Antworten beträgt damit 92.7%.

Hypergeometrische Verteilung

Definition

Die Hypergeometrische Verteilung beschreibt eine Stichprobe, die ohne Zurücklegen gezogen wird. Die einzelnen Versuche sind dann nicht unabhängig.

Sei N die Anzahl der Elemente in der Grundgesamtheit; M die Anzahl der Elemente, die für uns günstig sind; n sei die Grösse der Stichprobe; k die Anzahl der Elemente aus M, die in n enthalten sind; $\binom{n}{k}$ ist der Binomialkoeffizient.

$$Hyper(k|M, N, n) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

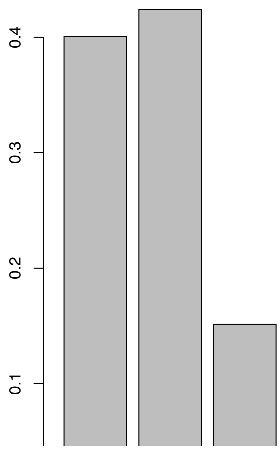
Hypergeometrische Verteilung

Problem: Beim Schweizer Zahlenlotto sind 6 Zahlen aus 42 zu ziehen. Wir bezeichnen mit x die Anzahl der richtig angekreutzten Zahlen. Bestimmen Sie die Wahrscheinlichkeitsverteilung und stellen Sie diese grafisch dar.

Hypergeometrische Verteilung

Antwort:

```
ylotto <- dhyper(0:6,m=6,n=39,k=6)

names(ylotto) <- 0:6

barplot(ylotto)
```



Poissonverteilung

Definition

Die Poissonverteilung ist eine diskrete Verteilung, mit der man die Anzahl von Ereignissen in einem gegebenen Zeitintervall modelliert. Ihr einziger Parameter λ bezeichnet die durchschnittlich zu erwartende Anzahl an Ereignissen.

$$Pois(x|\lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \text{ mit } x \in \mathbb{N}$$

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.
- Die Anzahl der SMS, die Handynutzer während eines Tages verschicken.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.
- Die Anzahl der SMS, die Handynutzer während eines Tages verschicken.
- Die Anzahl der Gäste, die ein Restaurant zwischen 20 Uhr und 22 Uhr besuchen.

Poissonverteilung

Problem: Eine Brücke wird durchschnittlich von 12 Autos pro Minute passiert. Wie gross ist die Wahrscheinlichkeit, dass sich in einer Minute mehr als 17 Autos auf der Brücke befinden?

Poissonverteilung

Antwort: Die Wahrscheinlichkeit für weniger als 16 Autos auf der Brücke finden wir mit der Funktion `ppoiss`.

```
ppois(16, lambda=12) # lower tail
```

```
## [1] 0.898709
```

Die Wahrscheinlichkeit für 17 und mehr Autos ist somit:

```
1-ppois(16, lambda=12) # oder
```

```
## [1] 0.101291
```

```
ppois(16, lambda=12, lower=FALSE)
```

```
## [1] 0.101291
```

Stetige Gleichverteilung

Definition

Die stetige Gleichverteilung ist eine Verallgemeinerung der diskreten Gleichverteilung. Während bei der diskreten Gleichverteilung jede ganze Zahl zwischen a und b möglich ist (beim Würfelwurf ist z.B. $a = 1$ und $b = 6$), so ist bei der stetigen Gleichverteilung nun jede reelle Zahl im Intervall von a bis b ein mögliches Ergebnis. Ihre Dichtefunktion lautet:

$$Uni(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{für } a \leqslant x \leqslant b \\ 0 & \text{für } x < a \text{ oder } x > b \end{cases}$$

Stetige Gleichverteilung

Beispiel:

- Zufallszahlen.

Stetige Gleichverteilung

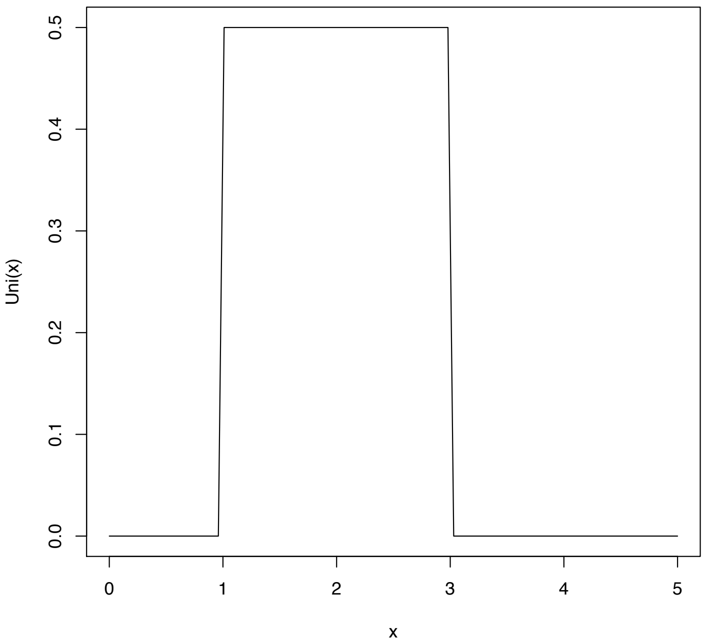
Beispiel:

- Zufallszahlen.
- Wartezeiten auf den Bus.

Stetige Gleichverteilung

Beispiel:

```
xv <- seq(0,5,length=100)
plot(xv, dunif(xv,1,3), type = "l", ylab = "Uni(x)", xlab = "x")
```



Stetige Gleichverteilung

Problem: Bestimmen Sie 10 Zufallszahlen zwischen 1 und 3.

Stetige Gleichverteilung

Antwort: Wir verwenden die Zufallszahlfunktion `runif` der stetigen Gleichverteilung.

```
runif(10, min=1, max=3)
```

```
##  [1] 2.115256 1.553116 2.338847 1.638142 2.121753 2.945975
##  [7] 2.722053 1.858388 2.954451 1.650964
```

Exponentialverteilung

Definition

Die Exponentialverteilung beschreibt die Dauer zwischen zufällig auftretenden Ereignissen. Der einzige Parameter λ steht für die Zahl der erwarteten Ereignisse pro Einheitsintervall. Ihre Dichtefunktion lautet:

$$Exp(x|\lambda) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geqslant 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.
- Lebensdauer von Bauteilen, Maschinen und Geräten, wenn Alterungserscheinungen nicht betrachtet werden müssen.

Exponentialverteilung

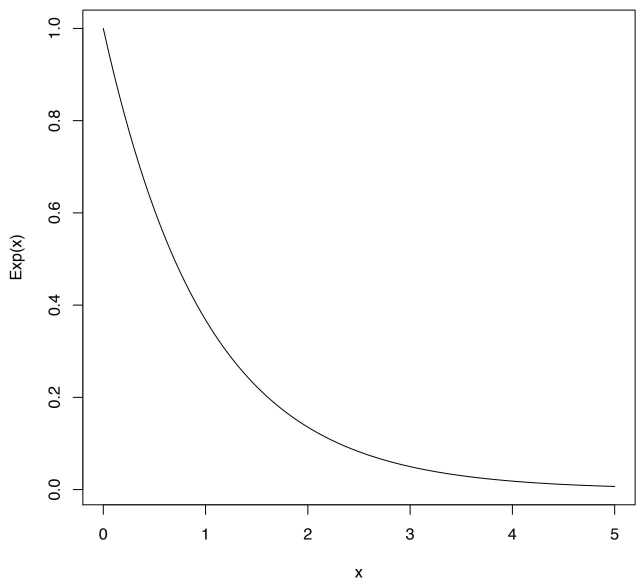
Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.
- Lebensdauer von Bauteilen, Maschinen und Geräten, wenn Alterungserscheinungen nicht betrachtet werden müssen.
- als grobes Modell für kleine und mittlere Schäden in Hausrat, Kraftfahrzeug-Haftpflicht, Kasko in der Versicherungsmathematik.

Exponentialverteilung

Beispiel:

```
xv <- seq(0,5,length=100)
plot(xv, dexp(xv, rate=1), type = "l", ylab = "Exp(x)",
xlab = "x")
```



Exponentialverteilung

Problem: Die durchschnittliche Abfertigungszeit an der Kasse eines Supermarktes betrage 3 Minuten. Mit welcher Wahrscheinlichkeit wird ein Kunde in weniger als 2 Minuten bedient?

Exponentialverteilung

Antwort: Die durchschnittliche Anzahl Kunden, die pro Minute bedient werden, beträgt $\lambda = \frac{1}{3}$.

```
pexp(2, rate=1/3)
```

```
## [1] 0.4865829
```

Der Kunde wird mit einer Wahrscheinlichkeit von 48.7% innerhalb von 2 Minuten bedient.

Normalverteilung

Definition

Die Normalverteilung ist wohl die wichtigste Verteilung in der Statistik.

Sie besitzt zwei Parameter, den Mittelwert μ und die

Standardabweichung σ. Ihre Dichtefunktion lautet:

$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

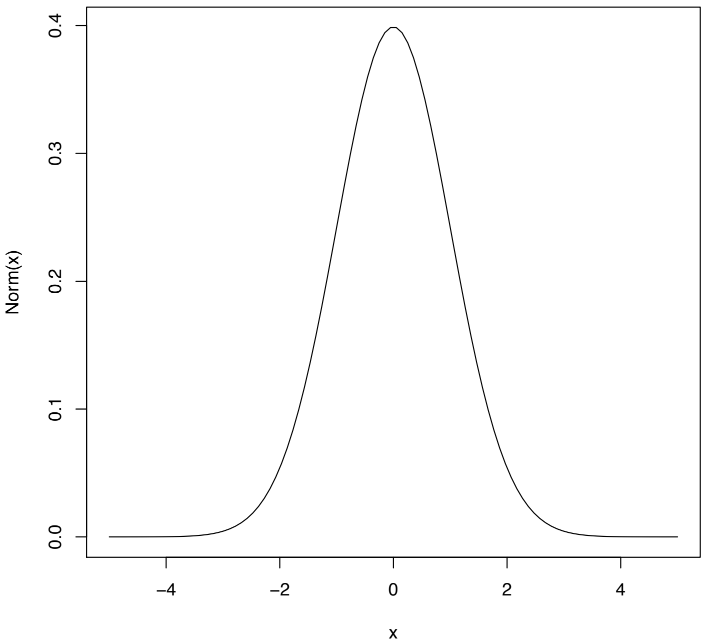
Für die Standardnormalverteilung gilt $\mu = 0$ und $\sigma = 1$,

d.h. $Z \sim N(0,1)$.

Normalverteilung

Beispiel:

```
xv <- seq(-5,5,length=100)
plot(xv, dnorm(xv), type = "l", ylab = "Norm(x)", xlab = "x")
```



Normalverteilung

Problem: Die Ergebnisse eines Abschlusstestes folgen einer Normalverteilung mit $\mu = 72$ und $\sigma = 15.2$. Welcher Anteil der Studierenden erreicht mindestens 84 Punkte?

Normalverteilung

Antwort:

```r
pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)

## [1] 0.2149176
```

Der Anteil der Studierenden, die mindestens 84 Punkte erzielen, beträgt 21.5%.

Chi-Quadrat-Verteilung

Definition

Die Chi-Quadrat-Verteilung wird in Zusammenhang mit Hypothesentest zu Kontingenztabellen und Verteilungsformen verwendet. Sie ist eine stetige Wahrscheinlichkeitsverteilung über der Menge der nicht-negativen reellen Zahlen. Der einzige Parameter ist die Anzahl der Freiheitsgrade *df*. Ist eine Zufallsvariable X chi-quadrat-verteilt, so gilt:

$$X \sim \chi^2(df)$$

Chi-Quadrat-Verteilung

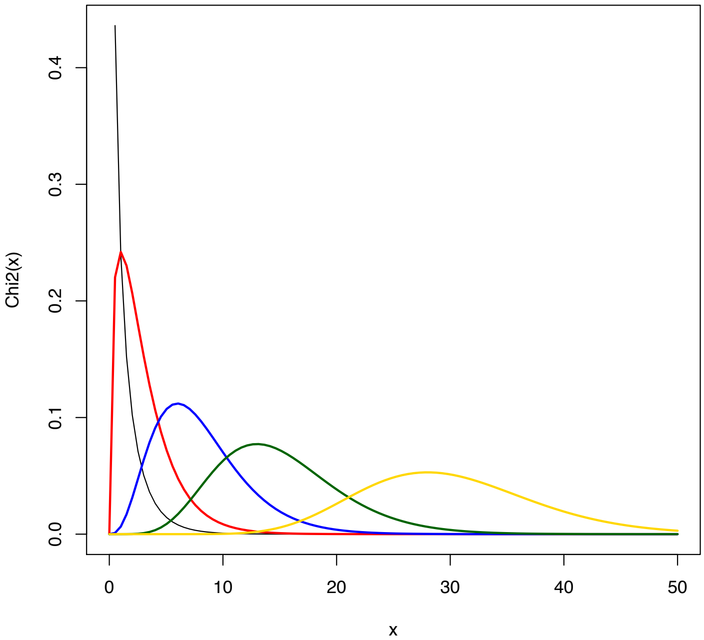
Beispiel:

```
xv <- seq(0,50,length=100)
degf <- c(3,8,15,30)
colors <- c("red", "blue", "darkgreen", "gold")
```

Chi-Quadrat-Verteilung

Beispiel:

```
plot(xv, dchisq(xv, df=1), type = "l", ylab = "Chi2(x)", xlab = "x")
for (i in 1:4){lines(xv, dchisq(xv,degf[i]), lwd=2, col=colors[i])}
```



Chi-Quadrat-Verteilung

Problem: Bestimmen Sie das 95%-Perzentil der χ^2-Verteilung mit Freiheitsgrad 7.

Chi-Quadrat-Verteilung

Antwort:

```
qchisq(.95, df=7)

## [1] 14.06714
```

Das 95%-Perzentil der χ^2-Verteilung mit *df* $= 7$ ist 14.067.

Studentsche t-Verteilung

Motivation

Wenn die Standardabweichung σ der Grundgesamtheit unbekannt ist, benutzt man die t-Verteilung (anstatt der Normalverteilung), vorausgesetzt die nötigen Bedingungen sind erfüllt. Die Variable X ist dann t-verteilt mit dem Freiheitsgrad $n - 1$.

$$X \sim t(df)$$

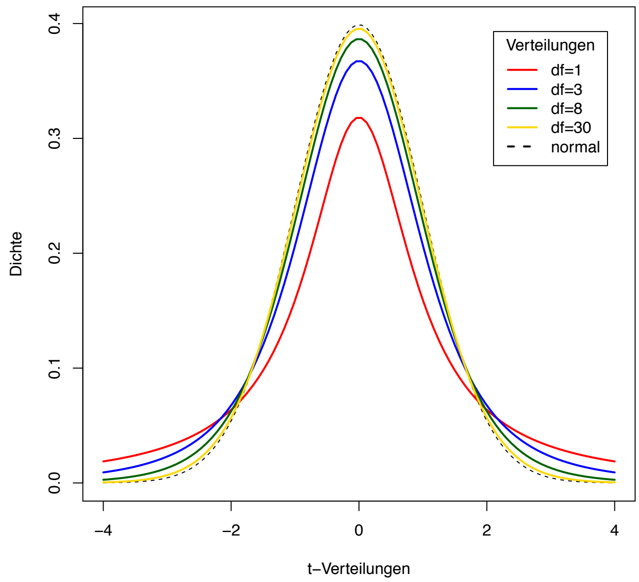
Studentsche t-Verteilung

Beispiel:

```r
x <- seq(-4, 4, length=100)
hx <- dnorm(x)
degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")
```

Studentsche t-Verteilung

```
plot(x, hx, type="l", lty=2, xlab="t-Verteilungen", ylab="Dichte")
for (i in 1:4){lines(x, dt(x,degf[i]), lwd=2, col=colors[i])}
legend("topright", inset=.05, title="Verteilungen",
labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```



Studentsche t-Verteilung

Problem: Bestimmen Sie das 2.5%- und das 97.5%-Perzentil der Studentschen t-Verteilung mit Freiheitsgrad 5.

CAS Datenanalyse HS16 - DeskStat

Konfidenzintervalle

Konfidenzintervalle

- Ausgehend von einer Zufallsstichprobe versuchen wir, den Wert eines Parameters zu schätzen.
- Diese Schätzung gelingt mit Konfidenzintervallen.
- Wir verwenden den von R mitgelieferten data frame **survey**.
- **survey** enthält die Ergebnisse einer Studentenumfrage in Australien.

Konfidenzintervalle

```
library(MASS)
head(survey, 3)

##       Sex Wr.Hnd NW.Hnd W.Hnd   Fold Pulse    Clap Exer
## 1 Female   18.5   18.0 Right R on L    92    Left Some
## 2   Male   19.5   20.5  Left R on L   104    Left None
## 3   Male   18.0   13.3 Right L on R    87 Neither None
##   Smoke Height      M.I    Age
## 1 Never  173.0   Metric 18.250
## 2 Regul  177.8 Imperial 17.583
## 3 Occas     NA     <NA> 16.917

# weitere Informationen: help(survey)
```

Punktschätzung eines Mittelwertes μ

- Der Mittelwert der Stichprobe $\bar{x}$ ist ein guter Schätzwert für den Mittelwert der Population μ.
- Der Durchschnitt aller aus mehreren Stichproben geschätzen Mittelwerte $\bar{x}_i$ ist gleich μ.
- Man sagt: die Zufallsvariable $\overline{X}$, welche die Stichprobenmittelwerte darstellt, ist erwartungstreu.
- Mathematisch formuliert: $E(\overline{X}_i) = \mu$.

Punktschätzung eines Mittelwertes μ

Problem: Bestimmen Sie einen Schätzwert für die Durchschnittsgrösse der Studierenden aufgrund der Daten aus **survey**.

Punktschätzung eines Mittelwertes μ

Antwort:

```
height.survey <- survey$Height
mean(height.survey, na.rm=TRUE)

## [1] 172.3809
```

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

- Ausgehend von der Punktschätzung bestimmen wir einen Vertrauensbereich, welcher den wahren Parameter μ mit grosser Wahrscheinlichkeit enthält.
- Das Konfidenzniveau $100(1-\alpha)$ wird zu Beginn festgelegt, mit Signifikanzniveau α.
- Wir nehmen an, dass die Varianz σ^2 bekannt ist.
- Die Endpunkte des Konfidenzintervalls sind gegeben durch

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

Problem: Für die Standardabweichung der Körpergrössen der Studierenden gelte $\sigma = 9.48$. Bestimmen Sie den Fehlerbereich und die Intervallschätzung der durchschnittlichen Körgergrösse bei einem Konfidenzniveau von 95%.

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

Antwort:

```
height.response <- na.omit(survey$Height)
n <- length(height.response)
sigma <- 9.48
# Standardfehler des Mittelwertes
sem <- sigma/sqrt(n)
sem

## [1] 0.6557453
```

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

Antwort:

```
# Fehlerbereich
ME <- qnorm(0.975)*sem
ME

## [1] 1.285237

xbar <- mean(height.response)
xbar + c(-ME,ME)

## [1] 171.0956 173.6661
```

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

Antwort: Bei einer Standardabweichung $\sigma = 9.48$ und einem Konfidenzniveau von 95% beträgt der Fehlerbereich 1.2852 cm. Der wahre durchschnittliche Körpergrösse wird vom Konfidenzintervall $[171.10; 173.67]$ mit einer Wahrscheinlichkeit von 95% überdeckt.

Intervallschätzung eines Mittelwertes μ, σ^2 bekannt

Erweiterte Antwort:

```
library(TeachingDemos)
z.test(height.response, sd=sigma)

##
##  One Sample z-test
##
## data:  height.response
## z = 262.88, n = 209.00000, Std. Dev. = 9.48000, Std.
## Dev. of the sample mean = 0.65575, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  171.0956 173.6661
## sample estimates:
## mean of height.response
##                172.3809
```

Intervallschätzung eines Mittelwertes μ, σ^2 unbekannt

- Die Populationsvarianz σ^2 ist meist unbekannt. Sie muss durch die Standardabweichung der Stichprobe s geschätzt werden.
- Die natürliche Schätzung s^2

$$s^2_{Pop} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  ist aber nicht erwartungstreu.
- Mathematische Herleitung führt zur erwartungstreuen Schätzung

$$s^2_{Samp} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Intervallschätzung eines Mittelwertes μ, σ^2 unbekannt

- Die Endpunkte des Konfidenzintervalls sind dann gegeben durch

$$\bar{x} \pm t_{\alpha/2} \frac{s_{Samp}}{\sqrt{n}}$$

- `R` berechnet mit dem Befehl `sd` automatisch die korrigierte Stichprobenstandardabweichung.
- Wir bezeichnen daher die Standardabweichung einfach mit s.

Intervallschätzung eines Mittelwertes μ, σ^2 unbekannt

Problem: Bestimmen Sie für die durchschnittliche Körpergrösse den Fehlerbereich und die Intervallschätzung der durchschnittlichen Körgergrösse bei einem Konfidenzniveau von 95%.

Intervallschätzung eines Mittelwertes μ, σ^2 unbekannt

Antwort:

```
library(MASS)
height.response <- na.omit(survey$Height)
n <- length(height.response)
s <- sd(height.response)
SE <- s/sqrt(n)
E <- qt(.975, df=n-1)*SE
xbar <- mean(height.response)
xbar + c(-E,E)

## [1] 171.0380 173.7237
```

Intervallschätzung eines Mittelwertes μ, σ^2 unbekannt

Erweiterte Antwort:

```
t.test(height.response)

##
##  One Sample t-test
##
## data:  height.response
## t = 253.07, df = 208, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  171.0380 173.7237
## sample estimates:
## mean of x
##  172.3809
```

Stichprobengrösse beim Stichprobenmittelwert

- Die Genauigkeit des Konfidenzintervalls wird durch ein Erhöhen der Stichprobengrösse verbessert.
- Die nachfolgende Formel bringt die beteiligten Grössen in Beziehung:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

- Unbekannte Parameter müssen aus der Stichprobe geschätzt werden.

Stichprobengrösse beim Stichprobenmittelwert

Problem: Bestimmen Sie benötigte Stichprobengrösse für die durchschnittliche Körpergrösse bei einem Fehlerbereich von 1.2 cm und einem Konfidenzniveau von 95%.

Stichprobengrösse beim Stichprobenmittelwert

Antwort:

```r
library(MASS)
height.response <- na.omit(survey$Height)
zstar <- qnorm(0.975)
s <- sd(height.response)
E <- 1.2
zstar^2*s^2/E^2

## [1] 258.695
```

Punktschätzung eines Populationsanteils p

- Der Anteil in der Stichprobe $\hat{p}$ ist ein guter Schätzwert für den Anteil in der Population p.
- Problem: Bestimmen Sie eine Punktschätzung für den Anteil der weiblichen Studierenden in **survey**.

Punktschätzung eines Populationsanteils p

Antwort:

```
library(MASS)
gender.response <- na.omit(survey$Sex)
n <- length(gender.response)
k <- sum(gender.response == "Female")
pbar <- k/n
pbar

## [1] 0.5
```

Intervallschätzung eines Populationsanteils p

- Ausgehend von der Punktschätzung bestimmen wir einen Vertrauensbereich, welcher den wahren Parameter p mit grosser Wahrscheinlichkeit enthält.
- Das Konfidenzniveau $100(1-\alpha)$ wird zu Beginn festgelegt, mit Signifikanzniveau α.
- Für die Stichprobengrösse gilt $np \geqslant 10$ und $n(1-p) \geqslant 10$.
- Die Endpunkte des Konfidenzintervalls sind gegeben durch

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Intervallschätzung eines Populationsanteils p

Problem: Bestimmen Sie den Fehlerbereich und die Intervallschätzung für den Anteil der weiblichen Studierenden aus **survey** bei einem Konfidenzniveau von 95%.

Intervallschätzung eines Populationsanteils p

Antwort:

```
library(MASS)
gender.response <- na.omit(survey$Sex)
n <- length(gender.response)
k <- sum(gender.response == "Female")
pbar <- k/n
pbar

## [1] 0.5
```

Intervallschätzung eines Populationsanteils p

Antwort:

```
SE <- sqrt(pbar*(1-pbar)/n)
SE

## [1] 0.03254723

E <- qnorm(.975)*SE
E

## [1] 0.06379139

pbar + c(-E,E)

## [1] 0.4362086 0.5637914
```

Intervallschätzung eines Populationsanteils p

Erweiterte Antwort:

```
prop.test(k, n)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  k out of n, null probability 0.5
## X-squared = 0, df = 1, p-value = 1
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4367215 0.5632785
## sample estimates:
##   p
## 0.5
```

Stichprobengrösse beim Populationsanteil p

- Die Genauigkeit des Konfidenzintervalls wird durch ein Erhöhen der Stichprobengrösse verbessert.
- Die nachfolgende Formel bringt die beteiligten Grössen in Beziehung:
$$n = \frac{(z_{\alpha/2})^2 p(1-p)}{E^2}$$
- Der Anteil p muss aus früheren Umfragen geschätzt werden.
- Fehlen diese Umfragen, gilt im Worst-Case $p = 0.5$.

Stichprobengrösse beim Populationsanteil p

Problem: Bestimmen Sie die Stichprobengrösse einer Umfrage zur Bestimmung des Anteils der weiblichen Studierenden. Der Fehlerbereich soll 5% betragen. Sie vermuten aus früheren Umfragen eine Anteil in der Grösse von $p = 0.5$. Das Konfidenzniveau ist 95%.

Stichprobengrösse beim Populationsanteil p

Antwort:

```
zstar <- qnorm(0.975)
p <- 0.5
E <- 0.05
zstar^2 * p * (1-p) / E^2

## [1] 384.1459
```

CAS Datenanalyse HS16 - DeskStat

Statistische Tests

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.
- Typ-II-Fehler: Eine falsche Nullhypothese wird beibehalten.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Standardabweichung der Population beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```
xbar <- 9900    # Stichprobenmittelwert
mu0 <- 10000    # Wert der Nullhypothese
sigma <- 120    # Standardabweichung
n <- 30         # Stichprobengrösse
z <- (xbar-mu0)/(sigma/sqrt(n))
z               # Testgrösse

## [1] -4.564355
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05               # Stichprobenmittelwert
z.alpha <- qnorm(alpha)  # kritischer Wert
z.alpha

## [1] -1.644854

z < z.alpha                 # H0 wird verworfen

## [1] TRUE
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z)
pval                 # unterer p-Wert

## [1] 2.505166e-06

pval < alpha         # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Nehmen Sie eine Standardabweichung von 0.25 g an. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \leqslant 2$ g, H_a: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
sigma <- 0.25
n <- 35
z <- (xbar-mu0)/(sigma/sqrt(n))
z

## [1] 2.366432
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05
z.critical <- qnorm(1-alpha)
z.critical

## [1] 1.644854

z > z.critical    # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z, lower.tail=FALSE)
pval                  # oberer p-Wert

## [1] 0.008980239

pval < alpha          # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha/2} \text{ oder } z < -z_{1-\alpha/2}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Standardabweichung der Population beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
sigma = 2.5
n = 35
z = (xbar-mu0)/(sigma/sqrt(n))
z

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha = .05
z.alpha = qnorm(1-alpha/2)
c(-z.alpha, z.alpha)

## [1] -1.959964  1.959964
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pnorm(z)     # lower tail
pval                    # zweiseitiger p-Wert

## [1] 0.05833852

# automatisierter p-Wert
pval = 2*ifelse(z < 0, pnorm(z), pnorm(z, lower.tail=FALSE))
pval

## [1] 0.05833852
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn
$$t < t_{\alpha,n-1}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Stichprobenstandardabweichung beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```
xbar = 9900               # Stichprobenmittelwert
mu0 = 10000               # Wert der Nullhypothese
s = 125                   # Stichprobenstandardabweichung
n = 30                    # Stichprobengrösse
t.val = (xbar-mu0)/(s/sqrt(n))
t.val                     # Testgrösse

## [1] -4.38178
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
t.alpha = qt(1-alpha, df=n-1)
-t.alpha                          # kritischer Wert

## [1] -1.699127

# alternative Lösung
pval = pt(t.val, df=n-1)
pval                              # unterer p-Wert

## [1] 7.035026e-05
```

Lösung: Linksseitiger Test bei μ, σ unbekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
test <- t.test(x, mu=mu0, alternative="less")
test$p.value

## [1] 1.591783e-05
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha,n-1}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Die Stichprobenstandardabweichung betage 0.3 g. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \leqslant 2$ g, H_a: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
s <- 0.3
n <- 35
t.val <- (xbar-mu0)/(s/sqrt(n))
t.val

## [1] 1.972027
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha <- 0.05
t.alpha <- qt(1-alpha, df=n-1)
t.alpha

## [1] 1.690924
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pt(t.val, df=n-1, lower.tail=FALSE)
pval                # oberer p-Wert

## [1] 0.02839295

pval < alpha        # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n - 1$ darstellt.

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n - 1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha/2,n-1} \text{ oder } t < -t_{1-\alpha/2,n-1}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Stichprobenstandardabweichung beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
s = 2.5
n = 35
t.val = (xbar-mu0)/(s/sqrt(n))
t.val

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
ta = qt(1-alpha/2, df=n-1)
c(-ta, ta)

## [1] -2.032245  2.032245
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pt(t.val, df=n-1)
pval                              # zweiseitiger p-Wert

## [1] 0.06687552

# automatisierter p-Wert
pval = 2*ifelse(t.val < 0, pt(t.val, df=n-1), pt(t.val, df=n-1, lo
pval

## [1] 0.06687552
```

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Populationsanteils p

Problem: Die Wahlbeteiligung an den letzten Wahlen betrug 60%. Eine telefonische Umfrage ergab, dass 85 von 148 Befragten angaben, an den kommenden Wahlen teilzunehmen. Lässt sich die Hypothese, dass die kommende Wahlbeteiligung über 60% liegt, bei einem Signifikanzniveau von 5% verwerfen?

Linksseitiger Test des Populationsanteils p

Antwort:

H_0: $p \geqslant 60\%$, H_a: $p < 60\%$

```
pbar <- 85/148    # Stichprobenmittelwert
p0 <- 0.6         # Wert der Nullhypothese
n <- 148          # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                 # Testgrösse

## [1] -0.6375983
```

Linksseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                # Stichprobenmittelwert
z.alpha <- qnorm(alpha)      # kritischer Wert
z.alpha                      # H0 wird nicht verworfen

## [1] -1.644854

pval <- pnorm(z)
pval

## [1] 0.2618676
```

Linksseitiger Test des Populationsanteils *p*

Erweiterte Antwort:

```
prop.test(85, 148, p=0.6, alt="less", correct=FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  85 out of 148, null probability 0.6
## X-squared = 0.40653, df = 1, p-value = 0.2619
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.6392527
## sample estimates:
##         p
## 0.5743243
```

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Populationsanteils p

Problem: Die Apfelernte im letzten Jahr enthielt 12% faule Äpfel. Im aktuellen Jahr zeigte eine Zufallsstichprobe 30 verfaulte Äpfel auf insgesamt 214 Äpfeln. Lässt sich die Hypothese, dass in diesem Jahr der Anteil verfaulter Äpfel weniger als 12% beträgt, bei einem Signifikanzniveau von 5% verwerfen?

Rechtsseitiger Test des Populationsanteils p

Antwort:

H_0: $p \leqslant 12\%$, $\quad H_a$: $p > 12\%$

```
pbar <- 30/214    # Stichprobenmittelwert
p0 <- 0.12        # Wert der Nullhypothese
n <- 214          # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                 # Testgrösse

## [1] 0.908751
```

Rechtsseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                    # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha)        # kritischer Wert
z.alpha                          # H0 wird nicht verworfen

## [1] 1.644854

pval <- pnorm(z)
pval

## [1] 0.8182592
```

Rechtsseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(30, 214, p=0.12, alt="greater", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  30 out of 214, null probability 0.12
## X-squared = 0.82583, df = 1, p-value = 0.1817
## alternative hypothesis: true p is greater than 0.12
## 95 percent confidence interval:
##  0.1056274 1.0000000
## sample estimates:
##         p
## 0.1401869
```

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < -z_{1-\alpha/2} \text{ oder } z > z_{1-\alpha/2}$$

Zweiseitiger Test des Populationsanteils p

Problem: Nach 20 Würfen zeigt eine Münze 12 Kopf. Lässt sich bei einem Signifikanzniveau von 5% die Behauptung verwerfen, dass es sich um eine faire Münze handelt?

Zweiseitiger Test des Populationsanteils p

Antwort:

H_0: $p = 50\%$, $\quad H_a$: $p \neq 50\%$

```
pbar <- 12/20      # Stichprobenmittelwert
p0 <- 0.5          # Wert der Nullhypothese
n <- 20            # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                  # Testgrösse

## [1] 0.8944272
```

Zweiseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                          # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha/2)            # kritischer Wert
c(-z.alpha, z.alpha)                   # H0 wird nicht verworfen

## [1] -1.959964  1.959964

pval <- 2*pnorm(z, lower.tail=FALSE)
pval

## [1] 0.3710934
```

Zweiseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(12, 20, p=0.5, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  12 out of 20, null probability 0.5
## X-squared = 0.8, df = 1, p-value = 0.3711
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3865815 0.7811935
## sample estimates:
##   p
## 0.6
```

CAS Datenanalyse HS16 - DeskStat

Statistische Tests

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.
- Typ-II-Fehler: Eine falsche Nullhypothese wird beibehalten.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Standardabweichung der Population beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```
xbar <- 9900    # Stichprobenmittelwert
mu0 <- 10000    # Wert der Nullhypothese
sigma <- 120    # Standardabweichung
n <- 30         # Stichprobengrösse
z <- (xbar-mu0)/(sigma/sqrt(n))
z               # Testgrösse

## [1] -4.564355
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05              # Stichprobenmittelwert
z.alpha <- qnorm(alpha)   # kritischer Wert
z.alpha

## [1] -1.644854

z < z.alpha                # H0 wird verworfen

## [1] TRUE
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z)
pval                # unterer p-Wert

## [1] 2.505166e-06

pval < alpha        # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Nehmen Sie eine Standardabweichung von 0.25 g an. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \leqslant 2$ g, H_a: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
sigma <- 0.25
n <- 35
z <- (xbar-mu0)/(sigma/sqrt(n))
z

## [1] 2.366432
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05
z.critical <- qnorm(1-alpha)
z.critical

## [1] 1.644854

z > z.critical    # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z, lower.tail=FALSE)
pval              # oberer p-Wert

## [1] 0.008980239

pval < alpha      # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha/2} \text{ oder } z < -z_{1-\alpha/2}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Standardabweichung der Population beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
sigma = 2.5
n = 35
z = (xbar-mu0)/(sigma/sqrt(n))
z

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha = .05
z.alpha = qnorm(1-alpha/2)
c(-z.alpha, z.alpha)

## [1] -1.959964  1.959964
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pnorm(z)     # lower tail
pval                    # zweiseitiger p-Wert

## [1] 0.05833852

# automatisierter p-Wert
pval = 2*ifelse(z < 0, pnorm(z), pnorm(z, lower.tail=FALSE))
pval

## [1] 0.05833852
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t < t_{\alpha,n-1}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Stichprobenstandardabweichung beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```r
xbar = 9900                # Stichprobenmittelwert
mu0 = 10000                # Wert der Nullhypothese
s = 125                    # Stichprobenstandardabweichung
n = 30                     # Stichprobengrösse
t.val = (xbar-mu0)/(s/sqrt(n))
t.val                      # Testgrösse

## [1] -4.38178
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
t.alpha = qt(1-alpha, df=n-1)
-t.alpha                     # kritischer Wert

## [1] -1.699127

# alternative Lösung
pval = pt(t.val, df=n-1)
pval                         # unterer p-Wert

## [1] 7.035026e-05
```

Lösung: Linksseitiger Test bei μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```r
test <- t.test(x, mu=mu0, alternative="less")
test$p.value

## [1] 1.591783e-05
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha,n-1}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Die Stichprobenstandardabweichung betage 0.3 g. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \leqslant 2$ g, H_a: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
s <- 0.3
n <- 35
t.val <- (xbar-mu0)/(s/sqrt(n))
t.val

## [1] 1.972027
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha <- 0.05
t.alpha <- qt(1-alpha, df=n-1)
t.alpha

## [1] 1.690924
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pt(t.val, df=n-1, lower.tail=FALSE)
pval                  # oberer p-Wert

## [1] 0.02839295

pval < alpha          # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1-\frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1-\frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha/2,n-1} \text{ oder } t < -t_{1-\alpha/2,n-1}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Stichprobenstandardabweichung beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
s = 2.5
n = 35
t.val = (xbar-mu0)/(s/sqrt(n))
t.val

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
ta = qt(1-alpha/2, df=n-1)
c(-ta, ta)

## [1] -2.032245  2.032245
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pt(t.val, df=n-1)
pval                                 # zweiseitiger p-Wert

## [1] 0.06687552

# automatisierter p-Wert
pval = 2*ifelse(t.val < 0, pt(t.val, df=n-1), pt(t.val, df=n-1, lo
pval

## [1] 0.06687552
```

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Populationsanteils p

Problem: Die Wahlbeteiligung an den letzten Wahlen betrug 60%. Eine telefonische Umfrage ergab, dass 85 von 148 Befragten angaben, an den kommenden Wahlen teilzunehmen. Lässt sich die Hypothese, dass die kommende Wahlbeteiligung über 60% liegt, bei einem Signifikanzniveau von 5% verwerfen?

Linksseitiger Test des Populationsanteils p

Antwort:

H_0: $p \geqslant 60\%$, $\quad H_a$: $p < 60\%$

```
pbar <- 85/148    # Stichprobenmittelwert
p0 <- 0.6         # Wert der Nullhypothese
n <- 148          # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                 # Testgrösse

## [1] -0.6375983
```

Linksseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                # Stichprobenmittelwert
z.alpha <- qnorm(alpha)      # kritischer Wert
z.alpha                      # H0 wird nicht verworfen

## [1] -1.644854

pval <- pnorm(z)
pval

## [1] 0.2618676
```

Linksseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(85, 148, p=0.6, alt="less", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  85 out of 148, null probability 0.6
## X-squared = 0.40653, df = 1, p-value = 0.2619
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.6392527
## sample estimates:
##         p
## 0.5743243
```

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1 - \alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Populationsanteils p

Problem: Die Apfelernte im letzten Jahr enthielt 12% faule Äpfel. Im aktuellen Jahr zeigte eine Zufallsstichprobe 30 verfaulte Äpfel auf insgesamt 214 Äpfeln. Lässt sich die Hypothese, dass in diesem Jahr der Anteil verfaulter Äpfel weniger als 12% beträgt, bei einem Signifikanzniveau von 5% verwerfen?

Rechtsseitiger Test des Populationsanteils p

Antwort:

H_0: $p \leqslant 12\%$, $\quad H_a$: $p > 12\%$

```
pbar <- 30/214    # Stichprobenmittelwert
p0 <- 0.12        # Wert der Nullhypothese
n <- 214          # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                 # Testgrösse

## [1] 0.908751
```

Rechtsseitiger Test des Populationsanteils p

Antwort:

```r
alpha <- 0.05                  # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha)      # kritischer Wert
z.alpha                        # H0 wird nicht verworfen

## [1] 1.644854

pval <- pnorm(z)
pval

## [1] 0.8182592
```

Rechtsseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(30, 214, p=0.12, alt="greater", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  30 out of 214, null probability 0.12
## X-squared = 0.82583, df = 1, p-value = 0.1817
## alternative hypothesis: true p is greater than 0.12
## 95 percent confidence interval:
##  0.1056274 1.0000000
## sample estimates:
##         p
## 0.1401869
```

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < -z_{1-\alpha/2} \text{ oder } z > z_{1-\alpha/2}$$

Zweiseitiger Test des Populationsanteils p

Problem: Nach 20 Würfen zeigt eine Münze 12 Kopf. Lässt sich bei einem Signifikanzniveau von 5% die Behauptung verwerfen, dass es sich um eine faire Münze handelt?

Zweiseitiger Test des Populationsanteils p

Antwort:

H_0: $p = 50\%$, H_a: $p \neq 50\%$

```
pbar <- 12/20      # Stichprobenmittelwert
p0 <- 0.5          # Wert der Nullhypothese
n <- 20            # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                  # Testgrösse

## [1] 0.8944272
```

Zweiseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                        # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha/2)          # kritischer Wert
c(-z.alpha, z.alpha)                 # H0 wird nicht verworfen

## [1] -1.959964  1.959964

pval <- 2*pnorm(z, lower.tail=FALSE)
pval

## [1] 0.3710934
```

Zweiseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(12, 20, p=0.5, correct=FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  12 out of 20, null probability 0.5
## X-squared = 0.8, df = 1, p-value = 0.3711
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3865815 0.7811935
## sample estimates:
##   p
## 0.6
```

CAS Datenanalyse HS16 - DeskStat

Anpassungs- und Unabhängigkeitstests

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich.
- H_a: die beobachteten und die erwarteten Häufigkeiten sind verschieden.

Anpassungstests

- Die Abweichung zwischen den beiden Häufigkeiten wird die dem χ^2-Wert gemessen:

$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

Anpassungstests

- Die Abweichung zwischen den beiden Häufigkeiten wird die dem χ^2-Wert gemessen:

$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

- Mit dem zugehörigen p-Wert der χ^2-Verteilung finden wir die Testentscheidung.

Anpassungstests

Problem: Die Datenmenge `survey` enthält auch Informationen zum Rauchverhalten der australischen Studierenden aus Adelaide.

```
library(MASS)
levels(survey$Smoke)

## [1] "Heavy" "Never" "Occas" "Regul"

smoke.freq <- table(survey$Smoke)
smoke.freq

##
## Heavy Never Occas Regul
##    11   189    19    17
```

Anpassungstests

Problem: Aufgrund einer früheren Vollumfrage kennt die Unileitung die Rauchstatistiken.

Heavy	Never	Occassionaly	Regular
4.5%	79.5%	8.5%	7.5%

Entscheiden Sie, ob die Stichprobe aus **survey** die Behauptung der Unileitung stützt. Arbeiten Sie mit einem Signifikanzniveau von 5%.

Anpassungstests

Antwort:

```
smoke.prob = c(.045, .795, .085, .075)
chisq.test(smoke.freq, p=smoke.prob)

##
##  Chi-squared test for given probabilities
##
## data:  smoke.freq
## X-squared = 0.10744, df = 3, p-value = 0.9909
```

Anpassungstests

Antwort: Der p-Wert ist deutlich grösser als 5%. Die Nullhypothese H_0 wird daher nicht verworfen. Die Stichprobe verträgt sich mit der Behauptung der Unileitung.

Aufgabe: Anpassungstests

Problem: Die Unileitung vermutet folgendes Rauchverhalten ihrer Studierenden.

Heavy	Never	Occassionaly	Regular
4.5%	79.5%	8.5%	7.5%

Prüfen Sie, ob die Stichprobe aus **survey** sich mit dieser Behauptung verträgt. Bestimmen Sie den *p*-Wert, ohne auf die Funktion `chisq.test` zurückzugreifen.

Lösung: Anpassungstests

```
f = table(survey$Smoke)
e = smoke.prob*length(survey$Smoke)
e

## [1]  10.665 188.415  20.145  17.775

d = f-e
chi = sum(d*d/e)
chi

## [1] 0.1112089

df = length(f)-1
pchisq(chi, df=df, lower=FALSE)

## [1] 0.9904592
```

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich. Die beiden Zufallsvariablen sind unabhängig.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich. Die beiden Zufallsvariablen sind unabhängig.
- H_a: die beobachteten und die erwarteten Häufigkeiten sind verschieden. Die beiden Zufallsvariaben sind abhängig.

Unabhängigkeitstests

Problem: Untersuchen Sie, ob das Rauch- und Sportverhalten der Studierenden aus **survey** unabhängig sind. Die entsprechenden Variablen sind `smoke` und `Exer`. Arbeiten Sie mit einem Signifikanzniveau von 5%.

Unabhängigkeitstests

Antwort:

```
library(MASS)
tbl = table(survey$Smoke, survey$Exer)
tbl

##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

Unabhängigkeitstests

Antwort:

```
chisq.test(tbl)

## Warning in chisq.test(tbl): Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

H_0 wird nicht verworfen.

Unabhängigkeitstests

Erweiterte Antwort: Die Warnung beim `chisq.test` erscheint, weil gewisse Zelleneinträge der Tabelle `tbl` zu gering sind (<5). Wir fassen daher die zweite und dritte Spalte zu einer neuen Spalte zusammen.

```
ctbl = cbind(tbl[,"Freq"],tbl[,"None"] + tbl[,"Some"])
ctbl

##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8
```

Unabhängigkeitstests

Erweiterte Antwort:

```
chisq.test(ctbl)

##
##  Pearson's Chi-squared test
##
## data:  ctbl
## X-squared = 3.2328, df = 3, p-value = 0.3571
```

CAS Datenanalyse HS16 - DeskStat

Statistische Tests: Hypothesen über zwei Populationen

Aussagen über zwei Populationen

- Es ist oft notwendig, zwei Populationen miteinander zu vergleichen.

Aussagen über zwei Populationen

- Es ist oft notwendig, zwei Populationen miteinander zu vergleichen.
- Wir schätzen die folgende Parameter:

Aussagen über zwei Populationen

- Es ist oft notwendig, zwei Populationen miteinander zu vergleichen.
- Wir schätzen die folgende Parameter:
  - die Differenz zwischen zwei Populationsmittelwerten

Aussagen über zwei Populationen

- Es ist oft notwendig, zwei Populationen miteinander zu vergleichen.
- Wir schätzen die folgende Parameter:
  - die Differenz zwischen zwei Populationsmittelwerten
  - die Differenz zwischen zwei Populationsanteilen

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

- Zwei Stichproben sind verbunden, wenn sie durch wiederholte Messung desselben Objektes entstehen.

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

- Zwei Stichproben sind verbunden, wenn sie durch wiederholte Messung desselben Objektes entstehen.
- Wir nehmen an, dass die betrachteten Merkmale in der Population normalverteilt sind.

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

- Zwei Stichproben sind verbunden, wenn sie durch wiederholte Messung desselben Objektes entstehen.
- Wir nehmen an, dass die betrachteten Merkmale in der Population normalverteilt sind.
- Mit dem gepaarten t-Test schätzen wir die Differenz zwischen den beiden Populationsmittelwerten.

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

Problem: Das Dataframe **immer** zeigt die Gerstenernte von sechs unterschiedlichen Feldern in den Jahren 1931 bis 1932.

Wir nehmen an, dass die Erntemengen normalverteilt sind. Schätzen Sie mit einem 95%-Konfidenzintervall die Differenz zwischen den beiden Jahresdurchschnitten.

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

Antwort:

```r
library(MASS)
head(immer)

##   Loc Var     Y1    Y2
## 1  UF   M   81.0  80.7
## 2  UF   S  105.4  82.3
## 3  UF   V  119.7  80.4
## 4  UF   T  109.7  87.2
## 5  UF   P   98.3  84.2
## 6   W   M  146.6 100.4
```

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

Antwort:

```r
t.test(immer$Y1, immer$Y2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  immer$Y1 and immer$Y2
## t = 3.324, df = 29, p-value = 0.002413
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.121954 25.704713
## sample estimates:
## mean of the differences
##                15.91333
```

Vergleich von μ_1 und μ_2 bei verbundenen Stichproben

Antwort:

Das 95%-Konfidenzintervall für die Differenz der Durchschnittsernten in den Jahren 1931 und 1932 liegt zwischen6.122 und 25.705.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

- Zwei Stichproben sind unabhängig, wenn sie von unverbundenen Populationen stammen und sich nicht gegenseitig beeinflussen.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

- Zwei Stichproben sind unabhängig, wenn sie von unverbundenen Populationen stammen und sich nicht gegenseitig beeinflussen.
- Wir nehmen an, dass die betrachteten Merkmale in der Population normalverteilt sind.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

- Zwei Stichproben sind unabhängig, wenn sie von unverbundenen Populationen stammen und sich nicht gegenseitig beeinflussen.
- Wir nehmen an, dass die betrachteten Merkmale in der Population normalverteilt sind.
- Mit dem ungepaarten t-Test schätzen wir die Differenz zwischen den beiden Populationsmittelwerten.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Problem: Das Dataframe **mtcars** zeigt einige Aspekte von 32 Automodellen aus den Jahren 1973 bis 1974.

Die Variable `mpg` misst den Verbrauch durch die Angabe der zurückgelegten Meilen pro Gallone Treibstoff.

Die Variable `am` zeigt den Getriebetyp des Wagens (0: automatisches Getriebe, 1: manuelles Getriebe).

Der Bezinverbrauch sei normalverteilt. Bestimmen Sie ein 95%-Konfidenzintervall für die Differenz zwischen dem durchschnittlichen Verbrauch bei automatischen und manuellen Getrieben.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Antwort:

```
library(MASS)
head(mtcars, 3)

##                mpg cyl disp  hp drat    wt  qsec vs am gear
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4
##               carb
## Mazda RX4        4
## Mazda RX4 Wag    4
## Datsun 710       1
```

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Antwort:

```
L <- mtcars$am == 0
mpg.auto <- mtcars[L,]$mpg
mpg.auto

##  [1] 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2
## [12] 10.4 10.4 14.7 21.5 15.5 15.2 13.3 19.2

mpg.manual <- mtcars[!L,]$mpg
mpg.manual

##  [1] 21.0 21.0 22.8 32.4 30.4 33.9 27.3 26.0 30.4 15.8 19.7
## [12] 15.0 21.4
```

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Antwort:

```
t.test(mpg.auto, mpg.manual)

##
##  Welch Two Sample t-test
##
## data:  mpg.auto and mpg.manual
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Antwort: Der durchschnittliche Verbrauch liegt bei Autos mit Automatikgetriebe bei 17.15 mpg, bei Autos mit manuellem Getriebe bei 24.39 mpg. Das 95%-Konfidenzintervall für die Differenz der beiden Mittelwerte liegt zwischen 3.21 mpg und 11.28 mpg.

Vergleich von μ_1 und μ_2, unabhängige Stichproben

Erweiterte Antwort: Wir modellieren die abhängige Variable `mpg` durch den Prädiktor `am` und wenden anschliessend den t-Test an.

```
t.test(mpg ~ am, data=mtcars)

##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Lösung: Vergleiche μ_1 und μ_2, unabhängige Samples

Der durchschnittliche Gewichtsverlust liegt bei der Atkins-Diät bei 15.42 kg, bei der konventionallen Diät bei 7.01 kg. Das 95%-Konfidenzintervall für die Differenz der beiden Mittelwerte liegt zwischen 2.79 kg und 14.05 kg.

Vergleich von zwei Populationsanteilen

- Wird eine Umfrage in zwei unterschiedlichen Populationen durchgeführt, so werden die Resutate oft voneinander abweichen.

Vergleich von zwei Populationsanteilen

- Wird eine Umfrage in zwei unterschiedlichen Populationen durchgeführt, so werden die Resutate oft voneinander abweichen.
- Es besteht daher der Bedarf, diese beiden Resultaten miteinander zu vergleichen.

Vergleich von zwei Populationsanteilen

- Wird eine Umfrage in zwei unterschiedlichen Populationen durchgeführt, so werden die Resutate oft voneinander abweichen.
- Es besteht daher der Bedarf, diese beiden Resultaten miteinander zu vergleichen.
- Wir nehmen an, dass die betrachteten Merkmale in der Population normalverteilt sind.

Vergleich von zwei Populationsanteilen

Problem: Der Datensatz `quine` enthält Informationen zu Kindern einer australischen Kleinstadt.

```
head(quine, 6)
```

```
##   Eth Sex Age Lrn Days
## 1   A   M  F0  SL    2
## 2   A   M  F0  SL   11
## 3   A   M  F0  SL   14
## 4   A   M  F0  AL    5
## 5   A   M  F0  AL    5
## 6   A   M  F0  AL   13
```

Vergleich von zwei Populationsanteilen

Problem: Bestimmen Sie ein 95%-Konfidenzintervall für die Differenz des Frauenanteils unter den aboriginal und nicht-aboriginal Kindern.

```
table(quine$Eth, quine$Sex)

##
##      F  M
##   A 38 31
##   N 42 35
```

Vergleich von zwei Populationsanteilen

Antwort:

```
prop.test(table(quine$Eth, quine$Sex), correct=FALSE)

##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  table(quine$Eth, quine$Sex)
## X-squared = 0.0040803, df = 1, p-value = 0.9491
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1564218  0.1669620
## sample estimates:
##    prop 1    prop 2
## 0.5507246 0.5454545
```

Vergleich von zwei Populationsanteilen

Antwort: Das 95%-Konfidenzintervall der Differenz des Frauenanteils zwischen den beiden Gruppen liegt zwischen -15.64% und 16.70%.

CAS Datenanalyse HS16 - DeskStat

Lineare Regression

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

- Die beiden Parameter β_1 und β_2 sind unbekannt und sollen durch b_1 und b_2 geschätzt werden.

Lineare Regression

- Das einfache lineare Regressionsmodell beschreibt eine abhängige Variable als lineare Funktion einer unabhängigen Variablen.

$$y = \beta_1 \cdot x + \beta_2 + \epsilon$$

- Die beiden Parameter β_1 und β_2 sind unbekannt und sollen durch b_1 und b_2 geschätzt werden.
- Zum Beispiel:

$$\texttt{eruptions} = \beta_1 \cdot \texttt{waiting} + \beta_2 + \epsilon$$

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.
- Die Streuung der Residuen bleibt konstant.

Lineare Regression: Fehlerterme

- Die einzelnen Fehler pro Datenpunkt (Fehlerterm, Residuum) sind unabhängig.
- Der Erwartungswert der Residuen ist 0.
- Die Streuung der Residuen bleibt konstant.
- Die Residuen sind normalverteilt.

Lineare Regression: Schätzen eines y-Wertes

Problem: Wir modelieren den Zusammenhang zwischen den Eruptionsdauern und den Wartezeiten aus `faithful` mit einem lineare Modell. Wie lange dauert die nächste Eruptions im Schnitt, wenn die Wartezeit 80 Minuten beträgt?

Lineare Regression: Schätzen eines y-Wertes

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
coeffs <- coefficients(eruption.lm)
coeffs

## (Intercept)     waiting
## -1.87401599  0.07562795

waiting <- 80
duration <- coeffs[1] + coeffs[2]*waiting
duration

## (Intercept)
##     4.17622
```

Lineare Regression: Schätzen eines y-Wertes

Erweiterte Antwort:

```
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata)

##        1
## 4.17622
```

Wir erwarten eine Eruptionsdauer von ungefähr 4 Minuten.

Lineare Regression: Bestimmtheitsmass r^2

- Das Bestimmtheitsmass r^2 gibt an, welcher Anteil der Streuung, die in den Daten `eruptions` steckt, durch das Model erklärt werden kann.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Lineare Regression: Bestimmtheitsmass r^2

- Das Bestimmtheitsmass r^2 gibt an, welcher Anteil der Streuung, die in den Daten `eruptions` steckt, durch das Model erklärt werden kann.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Bei der linearen Regression entspricht das Bestimmtheitsmass dem Quadrat des Korrelationskoeffizienten.

Lineare Regression: Bestimmtheitsmass r^2

Problem: Bestimmen Sie das Bestimmtheitsmass r^2 des linearen Modells zu `faithful`.

Lineare Regression: Bestimmtheitsmass r^2

Antwort:

```r
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
summary(eruption.lm)$r.squared

## [1] 0.8114608
```

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?
- Wir testen die Hypothesen

$$H_0 : \beta_1 = 0 \text{ und } H_1 : \beta_1 \neq 0$$

Lineare Regression: Signifikanztests

- Ist der Zusammenhang zwischen der abhängigen Variablen und der unabhängigen Variablen überhaupt signifikant oder kommt der Wert von b_1 bloss durch Zufall zustande?
- Wir testen die Hypothesen

$$H_0 : \beta_1 = 0 \text{ und } H_1 : \beta_1 \neq 0$$

- Ist $\beta_1 = 0$, dann ist auch der Korrelationskoeffizient $\rho = 0$. In diesem Fall besteht kein linearer Zusammenhang zwischen den beiden Grössen x und y.

Lineare Regression: Signifikanztest für β_1

Problem: Untersuchen Sie, ob zwischen den Grössen `eruptions` und `waiting` aus `faithful` ein signifikanter Zusammenhang besteht.

Lineare Regression: Signifikanztest für β_1

```r
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
summary(eruption.lm)
```

```
##
## Call:
## lm(formula = eruptions ~ waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Lineare Regression: Signifikanztest für β_1

Antwort: Der p-Wert ist nahezu gleich 0. Die Nullhypothese $\beta_1 = 0$ wird verworfen. Offenbar besteht ein signifikanter Zusammenhang zwischen der Wartezeit und den Eruptiondauer.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.
- Dieser Wert wurde aufgrund einer Stichprobe ermittelt. Der wahre Durchschnittswert wird von diesem Wert abweichen.

Lineare Regression: Konfidenzintervalle für y

- Gemäss dem errechneten Modell führt eine Wartezeit von $x = 80$ Minuten zu einer durchschnittlichen Eruptionsdauer von $y = 4$ Minuten.
- Dieser Wert wurde aufgrund einer Stichprobe ermittelt. Der wahre Durchschnittswert wird von diesem Wert abweichen.
- Wir schätzen den wahren Wert mit einem Konfidenzintvervall ab.

Lineare Regression: Konfidenzintervalle für y

Problem: Bestimmen Sie ein 95%-Konfidenzintervall für die durchschnittliche Eruptionsdauer bei einer Wartezeit von 80 Minuten.

Lineare Regression: Konfidenzintervalle für y

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata, interval="confidence")

##        fit      lwr      upr
## 1 4.17622 4.104848 4.247592
```

Die durchschnittliche Eruptionszeit beträgt bei einer Wartezeit von 80 Minuten zwischen 4.10 und 4.24 Minuten, bei einem Signifikanzniveau von 95%.

Lineare Regression: Prognoseintervalle für y

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.

Lineare Regression: Prognoseintervalle für y

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.
- Dieser Wertebereich ist wiederum abhängig von einem Konfidenzniveau α.

Lineare Regression: Prognoseintervalle für *y*

- Das Prognoseintervall liefert einen Wertebereich für die zu erwartenden Lage eines einzelnen vorhergesagten Wertes der abhängigen Variablen.
- Dieser Wertebereich ist wiederum abhängig von einem Konfidenzniveau α.
- Das Prognoseintervall ist wird einen grösseren Wertebereich als das Konfidenzintervall liefern.

Lineare Regression: Prognoseintervalle für y

Problem: Bestimmen Sie ein 95%-Prognoseintervall für die Eruptionsdauer bei einer Wartezeit von 80 Minuten.

Lineare Regression: Prognoseintervalle für y

Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
newdata <- data.frame(waiting=80)
predict(eruption.lm, newdata, interval="predict")

##        fit      lwr      upr
## 1 4.17622 3.196089 5.156351
```

Die Eruptionszeit beträgt bei einer Wartezeit von 80 Minuten zwischen 3.20 und 5.16 Minuten, bei einem Signifikanzniveau von 95%.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.
  - Die Residuen haben eine gleichbleibende Streuung.

Lineare Regression: Residuen-Plot

- Die Abweichung eines Datenpunktes von seinem Modellwert nennen wir Residuum.

$$\text{Residuum}_i = y_i - \bar{y}$$

- Voraussetzungen des lineare Regressionsmodells an die Residuen:
  - Der Erwartungswert der Residuen ist 0.
  - Die Residuen haben eine gleichbleibende Streuung.
  - Die Residuen sind normalverteilt und unabhängig.

Lineare Regression: Residuen-Plot

Problem: Stellen Sie die Residuen des linearen Modells zwischen der Eruptionsdauer und der Wartezeit aus `faithful` grafisch dar.

Lineare Regression: Residuen-Plot

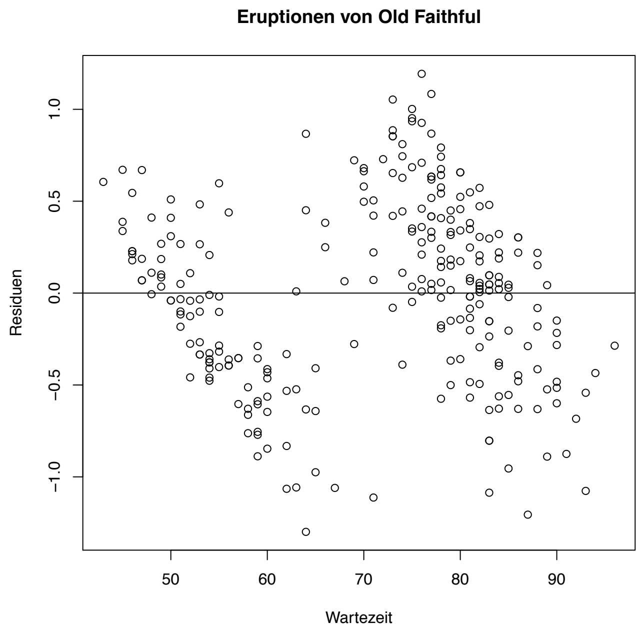
Antwort:

```
eruption.lm <- lm(eruptions ~ waiting, data=faithful)
eruption.res <- resid(eruption.lm)
```

Lineare Regression: Residuen-Plot

Antwort:

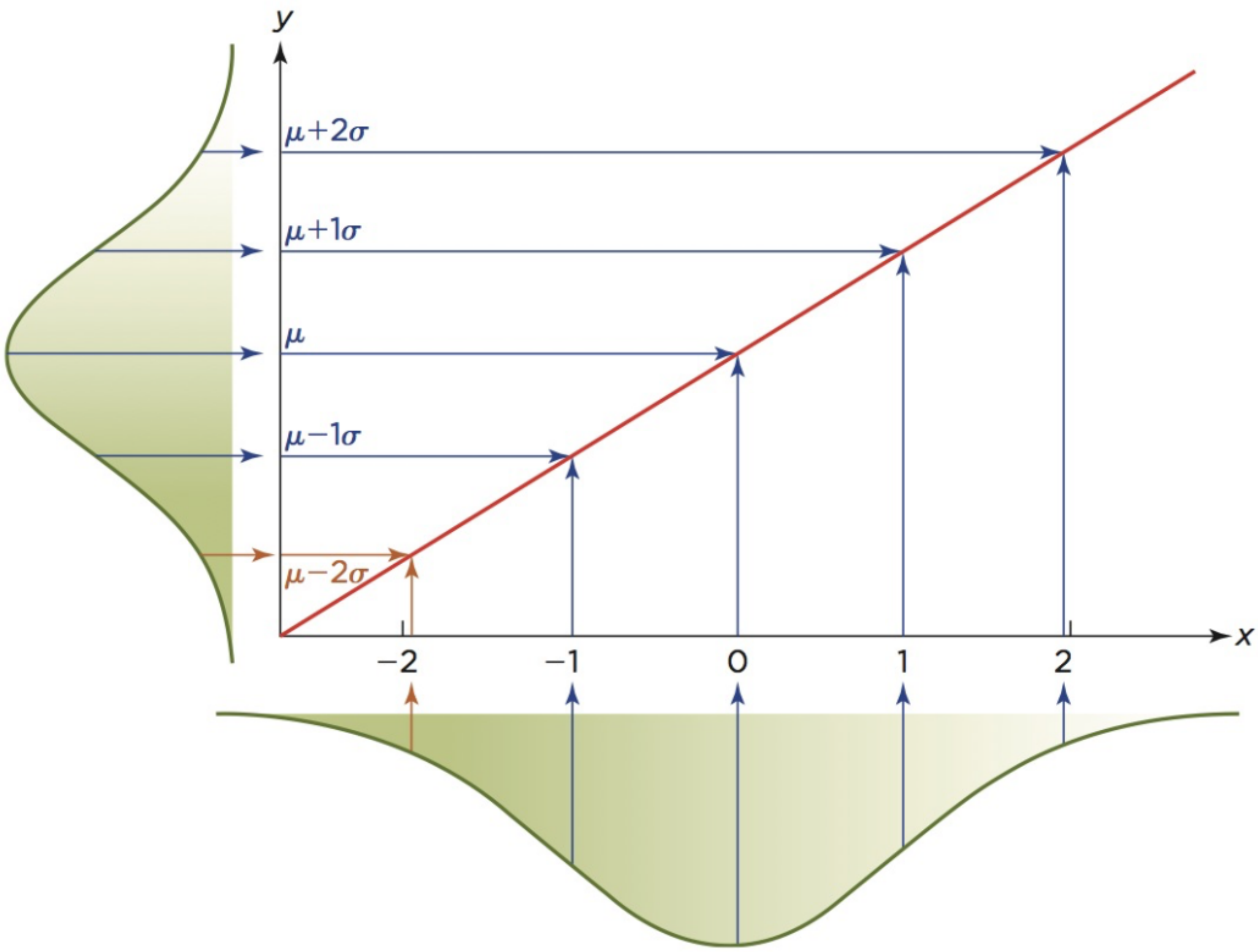
```
plot(faithful$waiting, eruption.res, ylab="Residuen",
 xlab="Wartezeit", main="Eruptionen von Old Faithful")
abline(0,0)
```



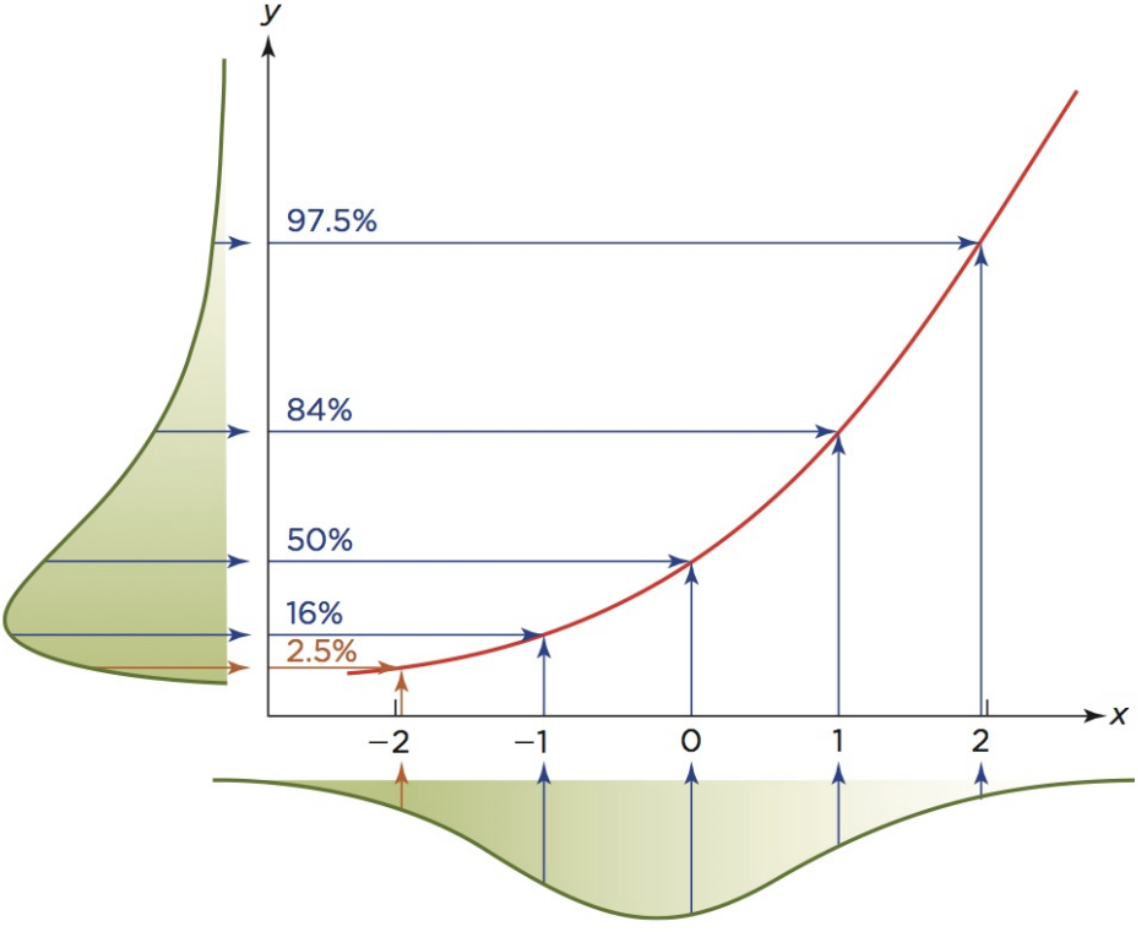
Lineare Regression: QQ-Plot

- Mit dem Normal-Wahrscheinlichkeits-Diagramm (auch Quantile-Quantile-Plot) der Residuen vergleichen wir die Residuen mit der Normalverteilung.

Lineare Regression: QQ-Plot



Lineare Regression: QQ-Plot



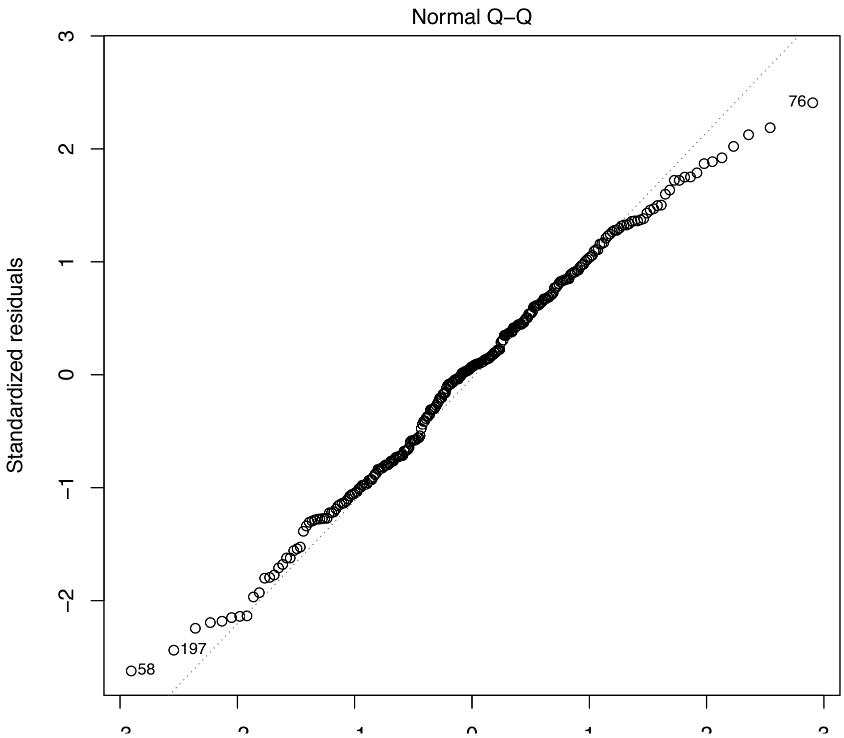
Lineare Regression: QQ-Plot

Problem: Erstellen Sie das Normal-Wahrscheinlichkeits-Diagramm der Residuen aus dem Datensatz `faithful`.

Lineare Regression: QQ-Plot

Antwort:

```
plot(eruption.lm, which=2)
```



CAS Datenanalyse HS16 - DeskStat

Kennzahlen des statistischen Zusammenhangs

Statistischer Zusammenhang

Bemerkung: Der statistische Zusammenhang zwischen Merkmalen hängt von deren Skalenniveau ab. Bei unterschiedlichen Niveaus bestimmt das tiefste Skalenniveau die Kennzahl.

Skalenniveau	Kennzahl
Nominal	Cramer's V
Ordinal	Spearman-Koeffizient
Metrisch	Pearson-Koeffizient

Statistischer Zusammenhang: Nominale Merkmale

Problem: Bestimmen Sie den Zusammenhang zwischen den Merkmalen `Studienrichtung` und `Geschlecht` aus Beispiel 13 des Foliensatzes „Folien Kapitel 1 Teil 2.pdf“.

Statistischer Zusammenhang: Nominale Merkmale

Lösung:

```
studis = matrix(c(110, 120, 20, 30, 20, 90, 60, 30, 10, 10),
nrow = 2, byrow = TRUE)
rownames(studis) = c("weiblich", "männlich")
colnames(studis) = c("BWL", "Soz", "VWL", "SoWi", "Stat")
studis
```

```
##           BWL Soz VWL SoWi Stat
## weiblich  110 120  20   30   20
## männlich   90  60  30   10   10
```

Statistischer Zusammenhang: Nominale Merkmale

```r
chisq.test(studis)

##
##  Pearson's Chi-squared test
##
## data:  studis
## X-squared = 18.056, df = 4, p-value = 0.001204

# Cramer's V
sqrt(chisq.test(studis)$statistic/(sum(studis)
*(min(dim(studis))-1)))

## X-squared
## 0.1900292
```

Statistischer Zusammenhang: Metrische Merkmale

Problem: Bestimmen Sie den Zusammenhang zwischen den Merkmalen `Einkommen` und `Alter` aus Beispiel 14 des Foliensatzes „Folien Kapitel 1 Teil 2.pdf“.

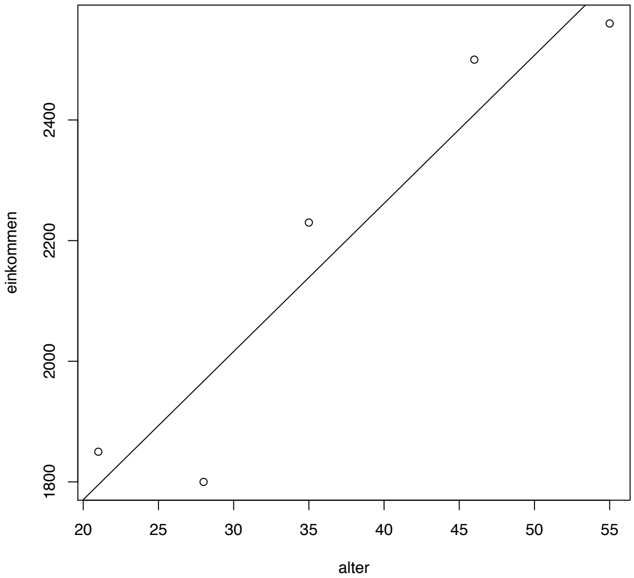
Statistischer Zusammenhang: Metrische Merkmale

Lösung:

```
einkommen <- c(1850, 2500, 2560, 2230, 1800)
alter <- c(21, 46, 55, 35, 28)
# Die passende Kennzahl ist der Korrelationskoeffizient
cor(einkommen, alter)

## [1] 0.9464183
```

Statistischer Zusammenhang: Metrische Merkmale

```r
# Streudiagramm
plot(alter, einkommen)
# Die Parameter im lm bestimmen und Gerade zeichnen
abline(lm(einkommen~alter))
```



Statistischer Zusammenhang: Ordinale Merkmale

Problem: Bestimmen Sie den Zusammenhang zwischen den Merkmalen `Mathematiknote` und `Statistiknote` aus Beispiel 16 des Foliensatzes „Folien Kapitel 1 Teil 2.pdf“.

Statistischer Zusammenhang: Ordinale Merkmale

Lösung:

```r
math.note <- c(1, 1, 5, 5, 4, 2)
stat.note <- c(2, 2, 5, 4, 4, 3)
# Die passende Kennzahl ist der
# Korrelationskoeffizient nach Spearman;
# R übernimmt die mühsame Rangbestimmung
cor(math.note, stat.note, method="spearman")

## [1] 0.9545455
```

CAS Datenanalyse HS16 - DeskStat

Wahrscheinichkeitsverteilungen

Zufallsvariablen

Definition

Eine Variable X ist eine Zufallsvariable, wenn der Wert, den X annimmt, von dem Ausgang eines Zufallsexperiments abhängt. Eine Zufallsvariable ordnet jedem Ergebniss eines Zufallsexperiments einen numerischen Wert zu.

Zufallsvariablen werden meist mit Großbuchstaben geschrieben.

Zufallsvariablen

Bemerkung: Zufallsvariablen sind daher Funktionen, die jedem Ergebnis eine (reelle) Zahl zuordnen. Sie haben also nicht direkt etwas mit Zufall zu tun. Da nun Ergebnisse durch Zahlen repräsentiert werden, kann mit ihnen gerechnet werden.

Wahrscheinlichkeitsverteilungen

Definition

Eine Wahrscheinlichkeitsverteilung beschreibt, wie sich die Werte einer Zufallsvariablen verteilen.

Binomialverteilung

Definition

Die Binomialverteilung beschreibt die Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben („Erfolg“ oder „Misserfolg“). Solche Versuchsserien werden auch Bernoulli-Prozesse genannt.

Bezeichnet p die Wahrscheinlichkeit eines erfolgreichen Versuchs, so bestimmt sich die Wahrscheinlichkeit für x erfolgreiche Ergebnisse in n unabhängigen Versuchen folgendermassen:

$$B(x|p,n) = \binom{n}{x} p^x (1-p)^{n-x} \text{ für } x \in \mathbb{N}$$

Binomialverteilung

Problem: Eine Multiple-Choise-Prüfungen besteht aus 12 Fragen. Jede Fragen gibt 5 verschiedenen Antworten, von denen aber nur jeweils eine Antwort richtig ist. Ein Student löst die Aufgaben nach dem Zufallsprinzip. Bestimmen Sie die Wahrscheinlichkeit dafür, dass der Student maximal vier korrekte Antworten gibt.

Binomialverteilung

Antwort: Für eine korrekten Antwort gilt $p = 0.2$. Die Wahrscheinlichkeit für genau 4 richtige Antworten finden wir mit:

```
dbinom(4, size=12, prob=0.2)

## [1] 0.1328756
```

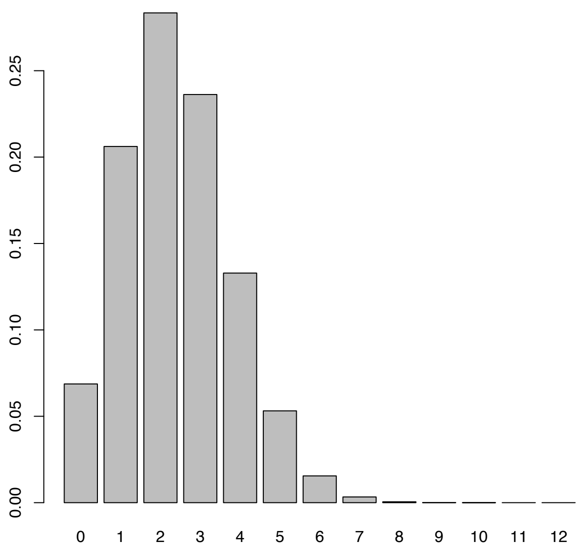
Die Wahrscheinlichkeit für maximal 4 korrekte Antworten ist somit:

```
dbinom(4, size=12, prob=0.2) +
+ dbinom(3, size=12, prob=0.2) +
+ dbinom(2, size=12, prob=0.2) +
+ dbinom(1, size=12, prob=0.2) +
+ dbinom(0, size=12, prob=0.2)

## [1] 0.9274445
```

Binomialverteilung

Erweiterte Antwort:

```
yprob <- dbinom(0:12,size=length(0:12)-1,prob = 1/5)
names(yprob) <- 0:12
barplot(yprob)
```



Binomialverteilung

Erweiterte Antwort: Alternativ können wir die kummulierte Wahrscheinlichkeit direkt berechnen mit:

```r
pbinom(4, size=12, prob=0.2)

## [1] 0.9274445
```

Die Wahrscheinlichkeit für vier oder weniger korrekte Antworten beträgt damit 92.7%.

Hypergeometrische Verteilung

Definition

Die Hypergeometrische Verteilung beschreibt eine Stichprobe, die ohne Zurücklegen gezogen wird. Die einzelnen Versuche sind dann nicht unabhängig.

Sei N die Anzahl der Elemente in der Grundgesamtheit; M die Anzahl der Elemente, die für uns günstig sind; n sei die Grösse der Stichprobe; k die Anzahl der Elemente aus M, die in n enthalten sind; $\binom{n}{k}$ ist der Binomialkoeffizient.

$$Hyper(k|M,N,n) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

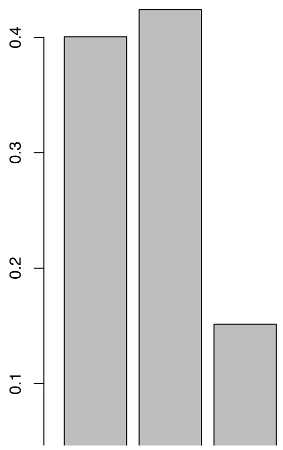
Hypergeometrische Verteilung

Problem: Beim Schweizer Zahlenlotto sind 6 Zahlen aus 42 zu ziehen. Wir bezeichnen mit x die Anzahl der richtig angekreutzten Zahlen. Bestimmen Sie die Wahrscheinlichkeitsverteilung und stellen Sie diese grafisch dar.

Hypergeometrische Verteilung

Antwort:

```
ylotto <- dhyper(0:6,m=6,n=39,k=6)

names(ylotto) <- 0:6

barplot(ylotto)
```



Poissonverteilung

Definition

Die Poissonverteilung ist eine diskrete Verteilung, mit der man die Anzahl von Ereignissen in einem gegebenen Zeitintervall modelliert. Ihr einziger Parameter λ bezeichnet die durchschnittlich zu erwartende Anzahl an Ereignissen.

$$Pois(x|\lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \text{ mit } x \in \mathbb{N}$$

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.
- Die Anzahl der SMS, die Handynutzer während eines Tages verschicken.

Poissonverteilung

Beispiel:

- Die Anzahl der Tore, die eine Fussballmannschaft während eines Spiels erzielt.
- Die Anzahl der Kunden, die während eines Tages am Postschalter auftauchen.
- Die Anzahl der SMS, die Handynutzer während eines Tages verschicken.
- Die Anzahl der Gäste, die ein Restaurant zwischen 20 Uhr und 22 Uhr besuchen.

Poissonverteilung

Problem: Eine Brücke wird durchschnittlich von 12 Autos pro Minute passiert. Wie gross ist die Wahrscheinlichkeit, dass sich in einer Minute mehr als 17 Autos auf der Brücke befinden?

Poissonverteilung

Antwort: Die Wahrscheinlichkeit für weniger als 16 Autos auf der Brücke finden wir mit der Funktion `ppoiss`.

```
ppois(16, lambda=12) # lower tail
```

```
## [1] 0.898709
```

Die Wahrscheinlichkeit für 17 und mehr Autos ist somit:

```
1-ppois(16, lambda=12) # oder
```

```
## [1] 0.101291
```

```
ppois(16, lambda=12, lower=FALSE)
```

```
## [1] 0.101291
```

Stetige Gleichverteilung

Definition

Die stetige Gleichverteilung ist eine Verallgemeinerung der diskreten Gleichverteilung. Während bei der diskreten Gleichverteilung jede ganze Zahl zwischen a und b möglich ist (beim Würfelwurf ist z.B. $a = 1$ und $b = 6$), so ist bei der stetigen Gleichverteilung nun jede reelle Zahl im Intervall von a bis b ein mögliches Ergebnis. Ihre Dichtefunktion lautet:

$$Uni(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{für } a \leqslant x \leqslant b \\ 0 & \text{für } x < a \text{ oder } x > b \end{cases}$$

Stetige Gleichverteilung

Beispiel:

- Zufallszahlen.

Stetige Gleichverteilung

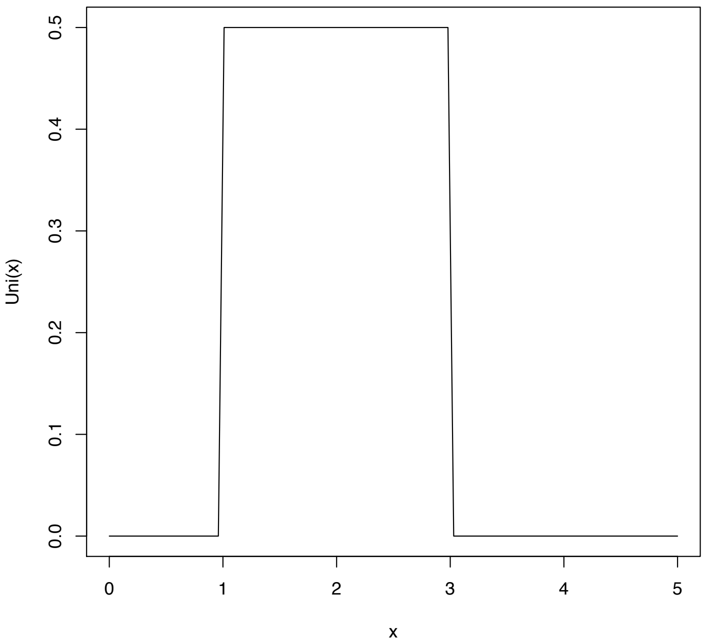
Beispiel:

- Zufallszahlen.
- Wartezeiten auf den Bus.

Stetige Gleichverteilung

Beispiel:

```
xv <- seq(0,5,length=100)
plot(xv, dunif(xv,1,3), type = "l", ylab = "Uni(x)", xlab = "x")
```



Stetige Gleichverteilung

Problem: Bestimmen Sie 10 Zufallszahlen zwischen 1 und 3.

Stetige Gleichverteilung

Antwort: Wir verwenden die Zufallszahlfunktion `runif` der stetigen Gleichverteilung.

```
runif(10, min=1, max=3)
```

```
##  [1] 2.115256 1.553116 2.338847 1.638142 2.121753 2.945975
##  [7] 2.722053 1.858388 2.954451 1.650964
```

Exponentialverteilung

Definition

Die Exponentialverteilung beschreibt die Dauer zwischen zufällig auftretenden Ereignissen. Der einzige Parameter λ steht für die Zahl der erwarteten Ereignisse pro Einheitsintervall. Ihre Dichtefunktion lautet:

$$Exp(x|\lambda) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geqslant 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.

Exponentialverteilung

Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.
- Lebensdauer von Bauteilen, Maschinen und Geräten, wenn Alterungserscheinungen nicht betrachtet werden müssen.

Exponentialverteilung

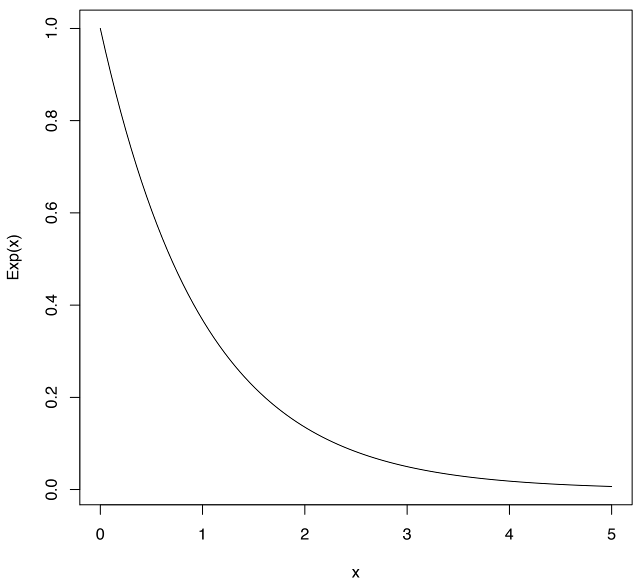
Beispiel:

- Zeit zwischen zwei Anrufen.
- Lebensdauer von Atomen beim radioaktiven Zerfall.
- Lebensdauer von Bauteilen, Maschinen und Geräten, wenn Alterungserscheinungen nicht betrachtet werden müssen.
- als grobes Modell für kleine und mittlere Schäden in Hausrat, Kraftfahrzeug-Haftpflicht, Kasko in der Versicherungsmathematik.

Exponentialverteilung

Beispiel:

```
xv <- seq(0,5,length=100)
plot(xv, dexp(xv, rate=1), type = "l", ylab = "Exp(x)",
xlab = "x")
```



Exponentialverteilung

Problem: Die durchschnittliche Abfertigungszeit an der Kasse eines Supermarktes betrage 3 Minuten. Mit welcher Wahrscheinlichkeit wird ein Kunde in weniger als 2 Minuten bedient?

Exponentialverteilung

Antwort: Die durchschnittliche Anzahl Kunden, die pro Minute bedient werden, beträgt $\lambda = \frac{1}{3}$.

```
pexp(2, rate=1/3)

## [1] 0.4865829
```

Der Kunde wird mit einer Wahrscheinlichkeit von 48.7% innerhalb von 2 Minuten bedient.

Normalverteilung

Definition

Die Normalverteilung ist wohl die wichtigste Verteilung in der Statistik.

Sie besitzt zwei Parameter, den Mittelwert μ und die

Standardabweichung σ. Ihre Dichtefunktion lautet:

$$N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

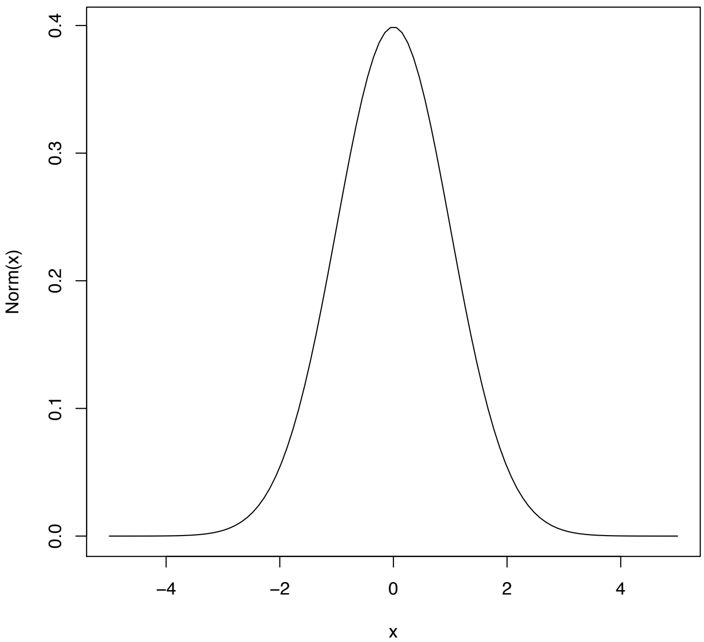
Für die Standardnormalverteilung gilt $\mu = 0$ und $\sigma = 1$,

d.h. $Z \sim N(0, 1)$.

Normalverteilung

Beispiel:

```
xv <- seq(-5,5,length=100)
plot(xv, dnorm(xv), type = "l", ylab = "Norm(x)", xlab = "x")
```



Normalverteilung

Problem: Die Ergebnisse eines Abschlusstestes folgen einer Normalverteilung mit $\mu = 72$ und $\sigma = 15.2$. Welcher Anteil der Studierenden erreicht mindestens 84 Punkte?

Normalverteilung

Antwort:

```r
pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)

## [1] 0.2149176
```

Der Anteil der Studierenden, die mindestens 84 Punkte erzielen, beträgt 21.5%.

Chi-Quadrat-Verteilung

Definition

Die Chi-Quadrat-Verteilung wird in Zusammenhang mit Hypothesentest zu Kontingenztabellen und Verteilungsformen verwendet. Sie ist eine stetige Wahrscheinlichkeitsverteilung über der Menge der nicht-negativen reellen Zahlen. Der einzige Parameter ist die Anzahl der Freiheitsgrade *df*. Ist eine Zufallsvariable X chi-quadrat-verteilt, so gilt:

$$X \sim \chi^2(df)$$

Chi-Quadrat-Verteilung

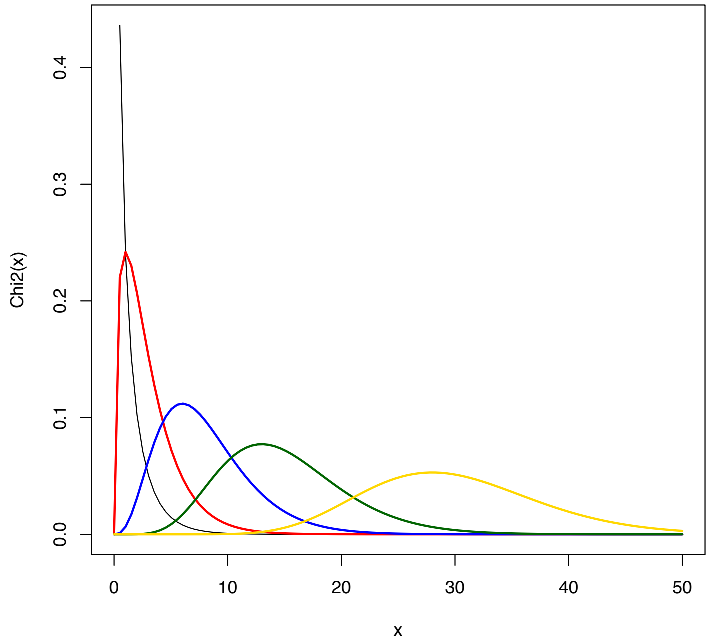
Beispiel:

```
xv <- seq(0,50,length=100)
degf <- c(3,8,15,30)
colors <- c("red", "blue", "darkgreen", "gold")
```

Chi-Quadrat-Verteilung

Beispiel:

```
plot(xv, dchisq(xv, df=1), type = "l", ylab = "Chi2(x)", xlab = "x")
for (i in 1:4){lines(xv, dchisq(xv,degf[i]), lwd=2, col=colors[i])}
```



Chi-Quadrat-Verteilung

Problem: Bestimmen Sie das 95%-Perzentil der χ^2-Verteilung mit Freiheitsgrad 7.

Chi-Quadrat-Verteilung

Antwort:

```
qchisq(.95, df=7)

## [1] 14.06714
```

Das 95%-Perzentil der χ^2-Verteilung mit $df = 7$ ist 14.067.

Studentsche t-Verteilung

Motivation

Wenn die Standardabweichung σ der Grundgesamtheit unbekannt ist, benutzt man die t-Verteilung (anstatt der Normalverteilung), vorausgesetzt die nötigen Bedingungen sind erfüllt. Die Variable X ist dann t-verteilt mit dem Freiheitsgrad $n-1$.

$$X \sim t(df)$$

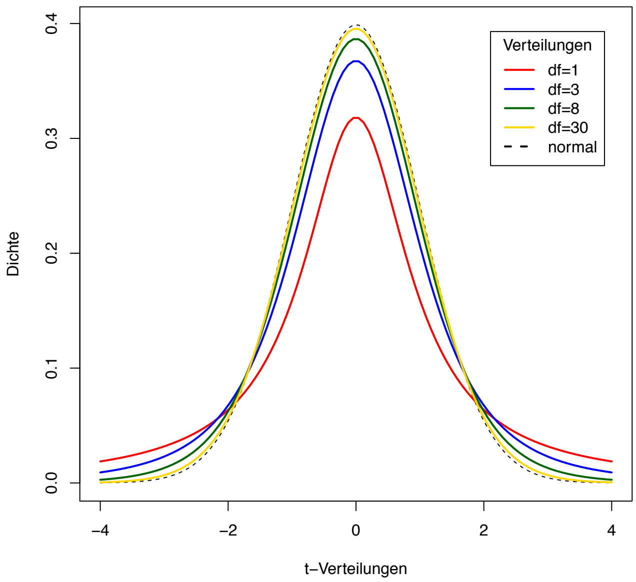
Studentsche t-Verteilung

Beispiel:

```
x <- seq(-4, 4, length=100)
hx <- dnorm(x)
degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")
```

Studentsche t-Verteilung

```
plot(x, hx, type="l", lty=2, xlab="t-Verteilungen", ylab="Dichte")
for (i in 1:4){lines(x, dt(x,degf[i]), lwd=2, col=colors[i])}
legend("topright", inset=.05, title="Verteilungen",
labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```



Studentsche t-Verteilung

Problem: Bestimmen Sie das 2.5%- und das 97.5%-Perzentil der Studentschen t-Verteilung mit Freiheitsgrad 5.

CAS Datenanalyse HS16 - DeskStat

Statistische Tests

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.

Hypothesentests

- Ausgehend von einer Zufallsstichprobe soll die Plausibilität einer Hypothese getestet werden.
- Die Nullhypothese H_0 wird verworfen, wenn ihr p-Wert unter dem Signifikanzniveau α liegt.
- Typ-I-Fehler: Eine wahre Nullhypothese wird verworfen.
- Typ-II-Fehler: Eine falsche Nullhypothese wird beibehalten.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \geqslant \mu_0 \text{ versus } H_a : \mu < \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Standardabweichung der Population beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```r
xbar <- 9900    # Stichprobenmittelwert
mu0 <- 10000    # Wert der Nullhypothese
sigma <- 120    # Standardabweichung
n <- 30         # Stichprobengrösse
z <- (xbar-mu0)/(sigma/sqrt(n))
z               # Testgrösse

## [1] -4.564355
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05              # Stichprobenmittelwert
z.alpha <- qnorm(alpha)    # kritischer Wert
z.alpha

## [1] -1.644854

z < z.alpha                # H0 wird verworfen

## [1] TRUE
```

Linksseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z)
pval                # unterer p-Wert

## [1] 2.505166e-06

pval < alpha        # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu \leqslant \mu_0 \text{ versus } H_a : \mu > \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Nehmen Sie eine Standardabweichung von 0.25 g an. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu \leqslant 2$ g, H_a: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
sigma <- 0.25
n <- 35
z <- (xbar-mu0)/(sigma/sqrt(n))
z

## [1] 2.366432
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha <- 0.05
z.critical <- qnorm(1-alpha)
z.critical

## [1] 1.644854

z > z.critical    # H0 wird verworfen

## [1] TRUE
```

Rechtsseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pnorm(z, lower.tail=FALSE)
pval                  # oberer p-Wert

## [1] 0.008980239

pval < alpha          # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1-\frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Mittelwerts μ, σ bekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha/2} \text{ oder } z < -z_{1-\alpha/2}$$

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Standardabweichung der Population beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
sigma = 2.5
n = 35
z = (xbar-mu0)/(sigma/sqrt(n))
z

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Antwort:

```
alpha = .05
z.alpha = qnorm(1-alpha/2)
c(-z.alpha, z.alpha)

## [1] -1.959964  1.959964
```

Zweiseitiger Test des Mittelwerts μ, σ bekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pnorm(z)     # lower tail
pval                    # zweiseitiger p-Wert

## [1] 0.05833852

# automatisierter p-Wert
pval = 2*ifelse(z < 0, pnorm(z), pnorm(z, lower.tail=FALSE))
pval

## [1] 0.05833852
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Linksseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \geqslant \mu_0$ versus $H_a : \mu < \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{\alpha,n-1}$, wobei $t_{\alpha,n-1}$ das 100α-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t < t_{\alpha,n-1}$$

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Hersteller von Glühbirnen behauptet eine Mindestlebensdauer von $10'000$ Stunden für seine Glühbirnen. Der Mittelwert einer Stichprobe aus 30 Glühbirnen ergab einen Stichprobenmittelwert von $9'900$ Stunden. Die Stichprobenstandardabweichung beträgt 120 Stunden. Können wir bei einem Signifikanzniveau von 5% die Behauptung des Herstellers verwerfen?

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \geqslant 10'000$ Stunden, H_a: $\mu < 10'000$ Stunden

```
xbar = 9900                # Stichprobenmittelwert
mu0 = 10000                # Wert der Nullhypothese
s = 125                    # Stichprobenstandardabweichung
n = 30                     # Stichprobengrösse
t.val = (xbar-mu0)/(s/sqrt(n))
t.val                      # Testgrösse

## [1] -4.38178
```

Linksseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
t.alpha = qt(1-alpha, df=n-1)
-t.alpha                          # kritischer Wert

## [1] -1.699127

# alternative Lösung
pval = pt(t.val, df=n-1)
pval                              # unterer p-Wert

## [1] 7.035026e-05
```

Lösung: Linksseitiger Test bei μ, σ unbekannt

Erweiterte Antwort:

Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
test <- t.test(x, mu=mu0, alternative="less")
test$p.value
```

```
## [1] 1.591783e-05
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

- Ist die Standardabweichung σ unbekannt, wird sie durch die Stichprobenstandardabweichung s geschätzt.
- $H_0 : \mu \leqslant \mu_0$ versus $H_a : \mu > \mu_0$
- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha,n-1}$, wobei $t_{1-\alpha,n-1}$ das $100(1-\alpha)$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha,n-1}$$

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Ein Produzent von Keksen behauptet, dass seine Produkte ein Höchstanteil an gesättigten Fettsäuren von 2 g pro Keks enthalten. In einer Stichprobe von 35 Keksen wurde ein Mittelwert von 2.1 g gemessen. Die Stichprobenstandardabweichung betage 0.3 g. Kann die Behauptung bei einem Signifikanzniveau von 5% verworfen werden?

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu \leqslant 2$ g, $\quad H_a$: $\mu > 2$ g

```
xbar <- 2.1
mu0 <- 2
s <- 0.3
n <- 35
t.val <- (xbar-mu0)/(s/sqrt(n))
t.val

## [1] 1.972027
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha <- 0.05
t.alpha <- qt(1-alpha, df=n-1)
t.alpha
```

```
## [1] 1.690924
```

Rechtsseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval <- pt(t.val, df=n-1, lower.tail=FALSE)
pval                  # oberer p-Wert

## [1] 0.02839295

pval < alpha          # H0 wird verworfen

## [1] TRUE
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1-\frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n-1$ darstellt.

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

- Wir formulieren die Hypothesen:

$$H_0 : \mu = \mu_0 \text{ versus } H_a : \mu \neq \mu_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Weiter bestimmen wir $t_{1-\alpha/2,n-1}$, wobei $t_{1-\alpha/2,n-1}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der t-Verteilung mit Freiheitsgrad $n - 1$ darstellt.
- H_0 wird verworfen, wenn

$$t > t_{1-\alpha/2,n-1} \text{ oder } t < -t_{1-\alpha/2,n-1}$$

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Problem: Das durchschnittliche Gewicht von antarktischen Königspinguinen einer bestimmten Kolonie betrug im letzten Jahr 15.4 kg. Eine Stichprobe von 35 Pinguinen derselben Kolonie zeigte ein Durchschnittsgewicht von 14.6 kg. Die Stichprobenstandardabweichung beträgt 2.5 kg. Lässt sich die Behauptung, dass sich das Durchschnittsgewicht nicht verändert hat, bei einem Signifikanzniveau von 5% verwerfen?

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

H_0: $\mu = 15.4$ kg, H_a: $\mu \neq 15.4$ g

```
xbar = 14.6
mu0 = 15.4
s = 2.5
n = 35
t.val = (xbar-mu0)/(s/sqrt(n))
t.val

## [1] -1.893146
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Antwort:

```
alpha = .05
ta = qt(1-alpha/2, df=n-1)
c(-ta, ta)

## [1] -2.032245  2.032245
```

Zweiseitiger Test des Mittelwerts μ, σ unbekannt

Erweiterte Antwort:
Anstatt den kritischen Wert zu berechnen, können wir den *p*-Wert bestimmen und diesen mit dem Signifikanzniveau α vergleichen.

```
pval = 2 * pt(t.val, df=n-1)
pval                               # zweiseitiger p-Wert

## [1] 0.06687552

# automatisierter p-Wert
pval = 2*ifelse(t.val < 0, pt(t.val, df=n-1), pt(t.val, df=n-1, lo
pval

## [1] 0.06687552
```

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.

Linksseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \geqslant p_0 \text{ versus } H_a : p < p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir z_α, wobei z_α das 100α-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < z_\alpha$$

Linksseitiger Test des Populationsanteils p

Problem: Die Wahlbeteiligung an den letzten Wahlen betrug 60%. Eine telefonische Umfrage ergab, dass 85 von 148 Befragten angaben, an den kommenden Wahlen teilzunehmen. Lässt sich die Hypothese, dass die kommende Wahlbeteiligung über 60% liegt, bei einem Signifikanzniveau von 5% verwerfen?

Linksseitiger Test des Populationsanteils p

Antwort:

H_0: $p \geqslant 60\%$, $\quad H_a$: $p < 60\%$

```
pbar <- 85/148   # Stichprobenmittelwert
p0 <- 0.6        # Wert der Nullhypothese
n <- 148         # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                # Testgrösse

## [1] -0.6375983
```

Linksseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                  # Stichprobenmittelwert
z.alpha <- qnorm(alpha)  # kritischer Wert
z.alpha                        # H0 wird nicht verworfen

## [1] -1.644854

pval <- pnorm(z)
pval

## [1] 0.2618676
```

Linksseitiger Test des Populationsanteils *p*

Erweiterte Antwort:

```
prop.test(85, 148, p=0.6, alt="less", correct=FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  85 out of 148, null probability 0.6
## X-squared = 0.40653, df = 1, p-value = 0.2619
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.0000000 0.6392527
## sample estimates:
##         p
## 0.5743243
```

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1 - \alpha)$-Perzentil der Standardnormalverteilung darstellt.

Rechtsseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p \leqslant p_0 \text{ versus } H_a : p > p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha}$, wobei $z_{1-\alpha}$ das $100(1-\alpha)$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z > z_{1-\alpha}$$

Rechtsseitiger Test des Populationsanteils p

Problem: Die Apfelernte im letzten Jahr enthielt 12% faule Äpfel. Im aktuellen Jahr zeigte eine Zufallsstichprobe 30 verfaulte Äpfel auf insgesamt 214 Äpfeln. Lässt sich die Hypothese, dass in diesem Jahr der Anteil verfaulter Äpfel weniger als 12% beträgt, bei einem Signifikanzniveau von 5% verwerfen?

Rechtsseitiger Test des Populationsanteils p

Antwort:

H_0: $p \leqslant 12\%$, $\quad H_a$: $p > 12\%$

```
pbar <- 30/214    # Stichprobenmittelwert
p0 <- 0.12        # Wert der Nullhypothese
n <- 214          # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                 # Testgrösse

## [1] 0.908751
```

Rechtsseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                   # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha)       # kritischer Wert
z.alpha                         # H0 wird nicht verworfen

## [1] 1.644854

pval <- pnorm(z)
pval

## [1] 0.8182592
```

Rechtsseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(30, 214, p=0.12, alt="greater", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  30 out of 214, null probability 0.12
## X-squared = 0.82583, df = 1, p-value = 0.1817
## alternative hypothesis: true p is greater than 0.12
## 95 percent confidence interval:
##  0.1056274 1.0000000
## sample estimates:
##         p
## 0.1401869
```

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.

Zweiseitiger Test des Populationsanteils p

- Wir formulieren die Hypothesen:

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Mit der Stichprobe berechnen wir die Testgrösse:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Weiter bestimmen wir $z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $100(1 - \frac{\alpha}{2})$-Perzentil der Standardnormalverteilung darstellt.
- H_0 wird verworfen, wenn

$$z < -z_{1-\alpha/2} \text{ oder } z > z_{1-\alpha/2}$$

Zweiseitiger Test des Populationsanteils p

Problem: Nach 20 Würfen zeigt eine Münze 12 Kopf. Lässt sich bei einem Signifikanzniveau von 5% die Behauptung verwerfen, dass es sich um eine faire Münze handelt?

Zweiseitiger Test des Populationsanteils p

Antwort:

H_0: $p = 50\%$, $\quad H_a$: $p \neq 50\%$

```
pbar <- 12/20       # Stichprobenmittelwert
p0 <- 0.5           # Wert der Nullhypothese
n <- 20             # Stichprobengrösse
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                   # Testgrösse

## [1] 0.8944272
```

Zweiseitiger Test des Populationsanteils p

Antwort:

```
alpha <- 0.05                  # Stichprobenmittelwert
z.alpha <- qnorm(1-alpha/2)    # kritischer Wert
c(-z.alpha, z.alpha)           # H0 wird nicht verworfen

## [1] -1.959964  1.959964

pval <- 2*pnorm(z, lower.tail=FALSE)
pval

## [1] 0.3710934
```

Zweiseitiger Test des Populationsanteils p

Erweiterte Antwort:

```
prop.test(12, 20, p=0.5, correct=FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  12 out of 20, null probability 0.5
## X-squared = 0.8, df = 1, p-value = 0.3711
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3865815 0.7811935
## sample estimates:
##   p
## 0.6
```

CAS Datenanalyse HS16 - DeskStat

Quantitative Daten

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.
- **faithful** zeigt die Eruptionsdauer und die Wartezeit zwischen den Eruptionen des Geysirs Old Faithful im Yellowstone Nationalpark.

Quantitative Daten

- Quantitative (stetige) Daten sind durch numerische Werte gegeben, die alle arithmetischen Operationen zulassen.
- Wir verwenden das data frame **faithful**.
- **faithful** zeigt die Eruptionsdauer und die Wartezeit zwischen den Eruptionen des Geysirs Old Faithful im Yellowstone Nationalpark.

```
head(faithful,3)

##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
```

Häufigkeitsverteilung quantitativer Daten

Definition

Die Häufigkeitsverteilung einer quantitativen Variablen gibt an, wie sich die Merkmalswerte über nicht-überlappende Intervalle verteilen.

Problem: Bestimmen Sie die Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Häufigkeitsverteilung quantitativer Daten

Lösung:

```
# Schritt 1: Spannweite bestimmen
duration = faithful$eruptions
range(duration)

## [1] 1.6 5.1

# Schritt 2: Spannweite in gleichlang, nichtüberlappende Interval
# Runde Spannweite zu [1.5,5.5]
breaks = seq(1.5, 5.5, by=0.5)
breaks

## [1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

Häufigkeitsverteilung quantitativer Daten

Lösung:

```
# Schritt 3: Eruptionszeiten in Intervalle verteilen
duration.cut = cut(duration, breaks, right=FALSE)
# Schritt 4: Häufigkeit pro Intervall bestimmen
duration.freq = table(duration.cut)
```

Antwort: Die Häufigkeitsverteilung der Variablen `eruption` ist:

```
duration.freq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##      51      41       5       7      30      73      61
## [5,5.5)
##       4
```

Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Die stellen die Verteilung in einer Spalte dar.

```
cbind(duration.freq)
```

```
##          duration.freq
## [1.5,2)             51
## [2,2.5)             41
## [2.5,3)              5
## [3,3.5)              7
## [3.5,4)             30
## [4,4.5)             73
## [4.5,5)             61
## [5,5.5)              4
```

Histogramm

Definition

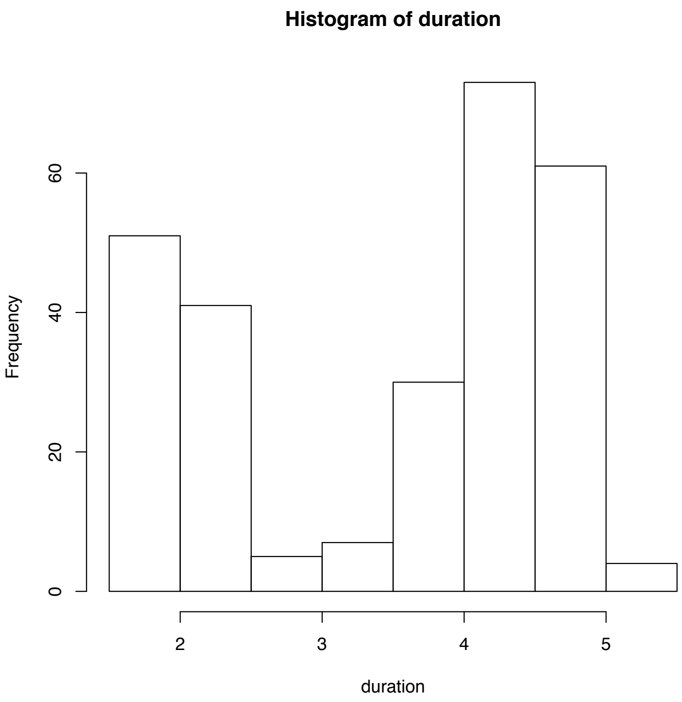
Ein Histogramm stellt die Häufigkeitsverteilung einer quantitativen Variablen graphisch dar.

Problem: Zeichnen Sie das Histogramm der Eruptionszeiten aus **faithful**.

Histogramm

Lösung:

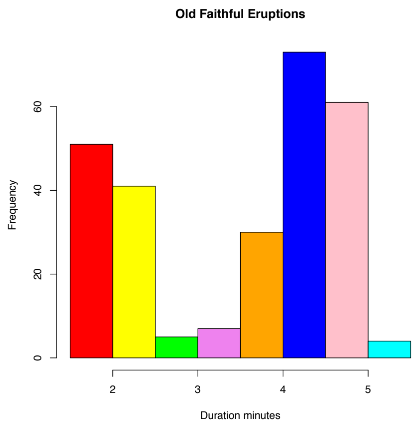
```
duration = faithful$eruptions
hist(duration, right=FALSE) # Intervalle sind rechts offen
```



Histogramm

Erweiterte Antwort: Wir verwenden Farben und fügen Titel sowie Achenbeschriftungen ein.

```
colors = c("red", "yellow", "green", "violet", "orange", "blue",
"pink", "cyan")
hist(duration, right=FALSE, col=colors,
main="Old Faithful Eruptions", xlab="Duration minutes")
```



Relative Häufigkeitsverteilung quantitativer Daten

Definition

Die relative Häufigkeitsverteilung einer quantitativen Variablen gibt an, wie sich die Anteile der Merkmalswerte über nicht-überlappende Intervalle verteilen.

Problem: Bestimmen Sie die relative Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Relative Häufigkeitsverteilung quantitativer Daten

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.relfreq = duration.freq/nrow(faithful)
duration.relfreq

## duration.cut
##    [1.5,2)    [2,2.5)    [2.5,3)    [3,3.5)    [3.5,4)
## 0.18750000 0.15073529 0.01838235 0.02573529 0.11029412
##    [4,4.5)    [4.5,5)    [5,5.5)
## 0.26838235 0.22426471 0.01470588
```

Relative Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Wir zeigen weniger Stellen an.

```
old = options(digits=1)
duration.relfreq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##    0.19    0.15    0.02    0.03    0.11    0.27    0.22
## [5,5.5)
##    0.01

options(old)  # alter Status
```

Relative Häufigkeitsverteilung quantitativer Daten

Erweiterte Antwort: Wir zeigen weniger Stellen an.

```
duration.percentage = duration.relfreq*100
old = options(digits=3)
head(cbind(duration.freq, duration.percentage),5)

##          duration.freq duration.percentage
## [1.5,2)             51               18.75
## [2,2.5)             41               15.07
## [2.5,3)              5                1.84
## [3,3.5)              7                2.57
## [3.5,4)             30               11.03

options(old)
```

Kumulierte Häufigkeitsverteilung

Definition

Die kumulierte Häufigkeitsverteilung einer quantitativen Variablen summiert die Anteile der Merkmalswerte über nicht-überlappende Intervalle.

Problem: Bestimmen Sie die kumulierte Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Kumulierte Häufigkeitsverteilung

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumfreq

## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##      51      92      97     104     134     207     268
## [5,5.5)
##     272
```

Kumulierte Häufigkeitsverteilung

Erweiterte Antwort: Wir präsentieren das Ergebnis als Spalte.

```
cbind(duration.cumfreq)

##           duration.cumfreq
## [1.5,2)                 51
## [2,2.5)                 92
## [2.5,3)                 97
## [3,3.5)                104
## [3.5,4)                134
## [4,4.5)                207
## [4.5,5)                268
## [5,5.5)                272
```

Kumulierte Häufigkeitsverteilungskurve

Definition

Die kumulierte Häufigkeitsverteilungskurve einer quantitativen Variablen stellt die summierten Häufigkeiten der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte Häufigkeitsverteilungskurve der Eruptionszeiten aus **faithful**.

Kumulierte Häufigkeitsverteilungskurve

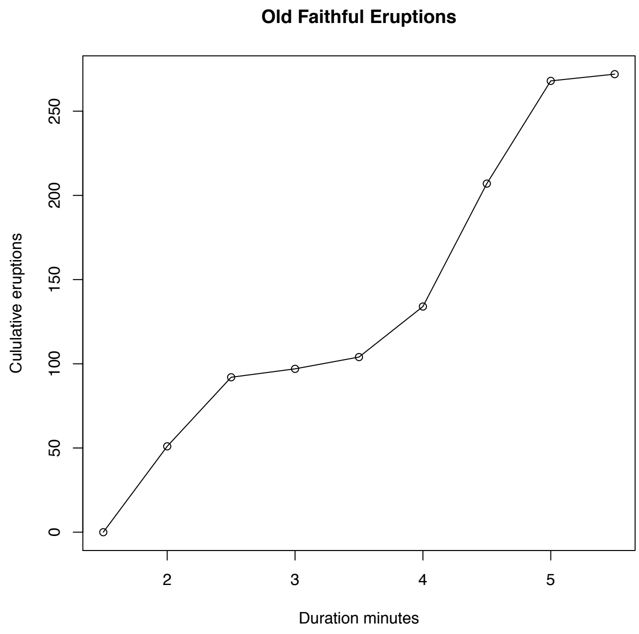
Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq0 = c(0,cumsum(duration.freq))
```

Kumulierte Häufigkeitsverteilungskurve

Lösung:

```
plot(breaks, duration.cumfreq0, main="Old Faithful Eruptions",
xlab="Duration minutes", ylab="Culaltive eruptions")
lines(breaks, duration.cumfreq0)
```



Kumulierte relative Häufigkeitsverteilung

Definition

Die kumulierte Häufigkeitsverteilung einer quantitativen Variablen stellt die summierten Anteile der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte relative Häufigkeitsverteilung der Eruptionszeiten aus **faithful**.

Kumulierte relative Häufigkeitsverteilung

Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumrelfreq = duration.freq/nrow(faithful)
duration.cumrelfreq

## duration.cut
##    [1.5,2)    [2,2.5)    [2.5,3)    [3,3.5)    [3.5,4)
## 0.18750000 0.15073529 0.01838235 0.02573529 0.11029412
##    [4,4.5)    [4.5,5)    [5,5.5)
## 0.26838235 0.22426471 0.01470588
```

Kumulierte relative Häufigkeitsverteilung

Erweiterte Antwort: Wir drucken weniger Stellen.

```
old = options(digits=2)
duration.cumrelfreq

## duration.cut
## [1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
##   0.188   0.151   0.018   0.026   0.110   0.268   0.224
## [5,5.5)
##   0.015

options(old)
```

Kumulierte relative Häufigkeitsverteilungskurve

Definition

Die kumulierte relative Häufigkeitsverteilungskurve einer quantitativen Variablen stellt die summierten Anteile der Merkmalswerte über nicht-überlappenden Intervallen graphisch dar.

Problem: Bestimmen Sie die kumulierte relative Häufigkeitsverteilungskurve der Eruptionszeiten aus **faithful**.

Kumulierte relative Häufigkeitsverteilungskurve

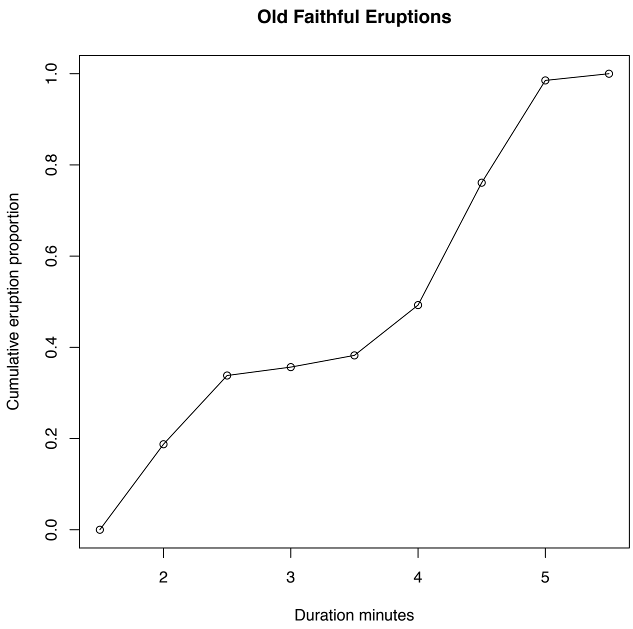
Lösung:

```
duration = faithful$eruptions
breaks = seq(1.5, 5.5, by=0.5)
duration.cut = cut(duration, breaks, right=FALSE)
duration.freq = table(duration.cut)
duration.cumfreq = cumsum(duration.freq)
duration.cumrelfreq = duration.cumfreq/nrow(faithful)
duration.cumrelfreq0 = c(0, duration.cumrelfreq)
```

Kumulierte relative Häufigkeitsverteilungskurve

Lösung:

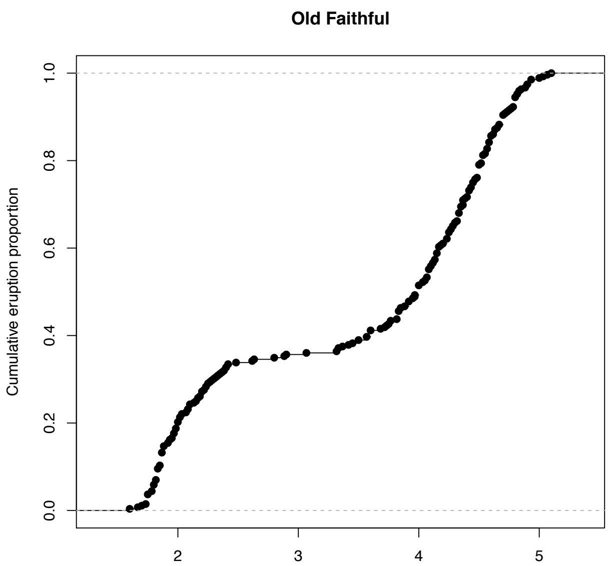
```
plot(breaks, duration.cumrelfreq0, main="Old Faithful Eruptions",
xlab="Duration minutes", ylab="Cumulative eruption proportion")
lines (breaks, duration.cumrelfreq0)
```



Kumulierte relative Häufigkeitsverteilungskurve

Erweiterte Antwort: Wir interpolieren die relative Häufigkeitsverteilung mit dem Befehl `ecdf`.

```
Fn = ecdf(duration)
plot(Fn, main="Old Faithful", xlab="Duration minutes",
ylab="Cumulative eruption proportion")
```



Streudiagramm

Definition

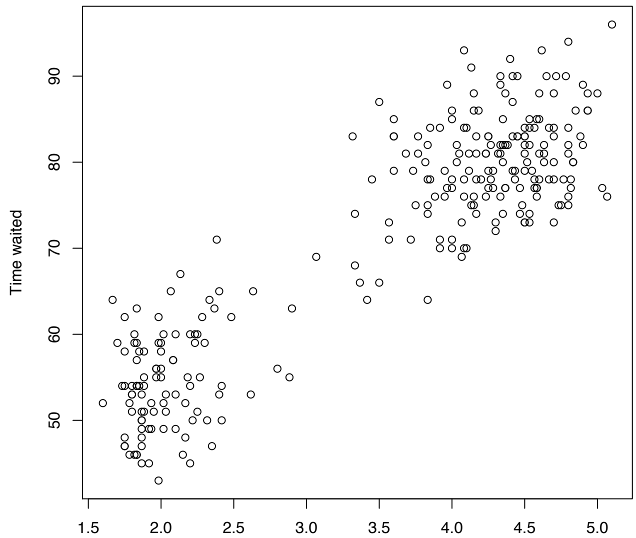
Ein Streudiagramm ist die graphische Darstellung von beobachteten Wertepaaren zweier statistischer Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt.

Problem: Bestimmen Sie das Streudiagramm der Eruptions- und Wartezeiten aus **faithful**.

Streudiagramm

Lösung:

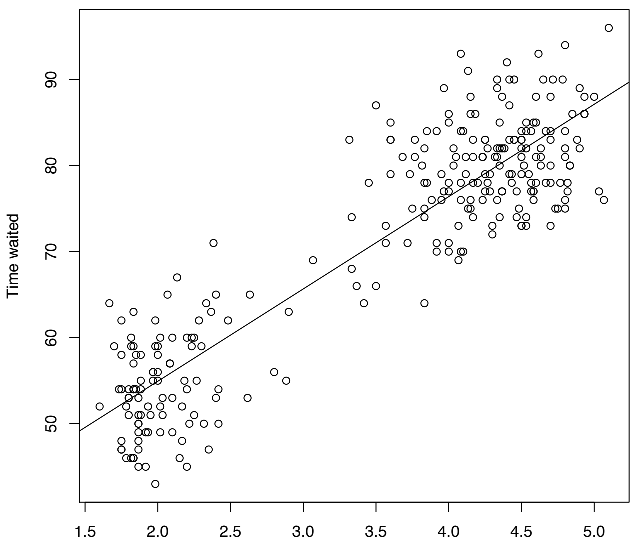
```
duration = faithful$eruptions
waiting = faithful$waiting
plot(duration, waiting, xlab="Erution duration",
ylab="Time waited")
```



Streudiagramm

Erweiterte Antwort: Wir berechnen mit `lm` ein lineares Model der beiden Variablen und fügen mit `ablilne` eine Trendline hinzu.

```
plot(duration, waiting, xlab="Eruption duration",
ylab="Time waited")
abline(lm(waiting ~ duration))
```



CAS Datenanalyse HS16 - DeskStat

Qualitative Daten

Qualitative Daten

- Als qualitative (nominale) Merkmale bezeichnet man Merkmale, bei denen sich die Merkmalsausprägungen zwar eindeutig in Kategorien unterscheiden lassen, diese Antworten jedoch keinen mathematischen Wert annehmen können.

Qualitative Daten

- Als qualitative (nominale) Merkmale bezeichnet man Merkmale, bei denen sich die Merkmalsausprägungen zwar eindeutig in Kategorien unterscheiden lassen, diese Antworten jedoch keinen mathematischen Wert annehmen können.
- Streng genommen zählen auch ordinale Merkmale zu den qualitativen Merkmalen. Bei ordinalen Merkmalen kann eine Hierarchie erstellt werden, eine genaue numerische Skalierung ist aber nicht möglich.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.
- **painters** enthält Informationen zu Malern des 18. Jahrhunderts.

Beispiel: painters

- Wir verwenden den von R mitgelieferten data frame **painters**.
- **painters** enthält Informationen zu Malern des 18. Jahrhunderts.

```
library(MASS)
head(painters,3)  # oder painters[1:3,]

##              Composition Drawing Colour Expression School
## Da Udine              10       8     16          3      A
## Da Vinci              15      16      4         14      A
## Del Piombo             8      13     16          7      A
```

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.
- Die Variable `school` ist damit qualitativ.

Beispiel: painters

- Die letzte Spalte klassifiziert die Maler bezüglich einer Schule.
- Die Schulen sind mit `A, B, ...` bezeichnet.
- Die Variable `school` ist damit qualitativ.

```
painters$School
```

```
##  [1] A A A A A A A A A A B B B B B B C C C C C C D D D D D D
## [29] D D D D E E E E E E E F F F F G G G G G G G H H H H
## Levels: A B C D E F G H
```

Häufigkeitsverteilung

Definition

Die Häufigkeitsverteilung gibt an, wie oft eine Merkmalsausprägung in einer Variable vorkommt.

Problem: Bestimmen Sie die Häufigkeitsverteilung der Variablen `School` aus `painters`.

Lösung:

```
library(MASS)                    # das MASS-Paket laden
school = painters$School         # die Schulen der Maler
school.freq = table(school)      # Anwenden der table-Funktion
```

Beispiel: Häufigkeitsverteilung

Antwort: Die Häufigkeitsverteilung der Variablen `School` ist:

```
school.freq

## school
##  A  B  C  D  E  F  G  H
## 10  6  6 10  7  4  7  4
```

Beispiel: Häufigkeitsverteilung

Erweiterte Antwort: Mit `cbind` stellen wir das Ergebnis in Spalten dar.

```
cbind(school.freq)

##   school.freq
## A          10
## B           6
## C           6
## D          10
## E           7
## F           4
## G           7
## H           4
```

Relative Häufigkeitsverteilung

Definition

Die relative Häufigkeitsverteilung gibt an, welchen Anteil die Merkmalsausprägungen einer Variable einnehmen.

Problem: Bestimmen Sie die relative Häufigkeitsverteilung der Variablen `School` aus `painters`.

Lösung:

```
library(MASS)                     # das MASS-Paket laden
school = painters$School          # die Schulen der Maler
school.freq = table(school)       # Anwenden der table-Funktion
school.relfreq = school.freq / nrow(painters)
```

Beispiel: Relative Häufigkeitsverteilung

Antwort: Die Häufigkeitsverteilung der Variablen `School` ist

```
school.relfreq

## school
##          A          B          C          D          E
## 0.18518519 0.11111111 0.11111111 0.18518519 0.12962963
##          F          G          H
## 0.07407407 0.12962963 0.07407407
```

Beispiel: Relative Häufigkeitsverteilung

Erweiterte Antwort: Wir drucken Spalten und weniger Stellen.

```
old=options(digits=3)
head(cbind(school.relfreq*100))
```

```
##    [,1]
## A 18.52
## B 11.11
## C 11.11
## D 18.52
## E 12.96
## F  7.41
```

Balkendiagramm

Definition

Ein Balkendiagramm stellt die Häufigkeitsverteilung von qualitativen Daten durch vertikale Balken graphisch dar.

Problem: Ein Balkendiagramm der Variable `school` von **painters** gibt mit vertikalen Balken die Anzahl der Maler pro Schule an.

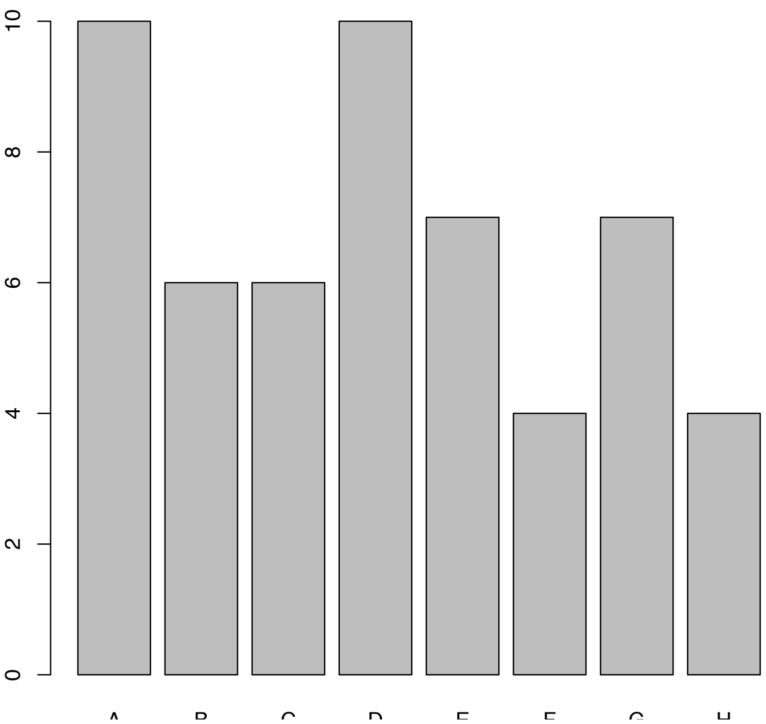
Lösung:

```
options(old)
school=painters$School
school.freq=table(school)
```

Beispiel: Balkendiagramm

Lösung:

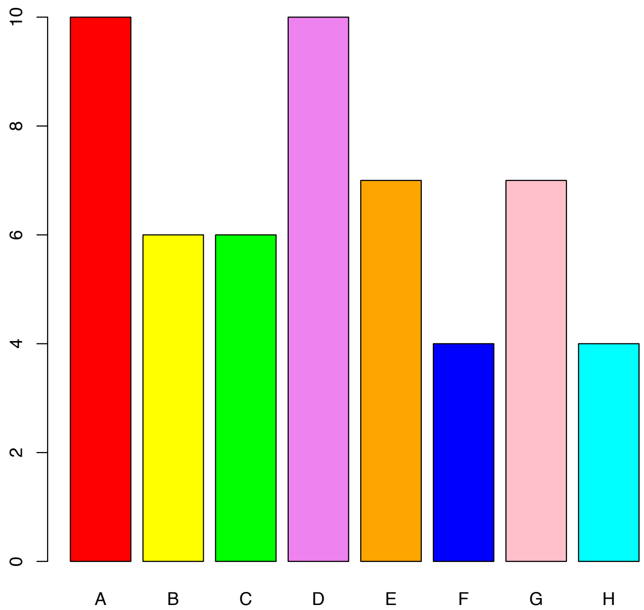
```
barplot(school.freq)
```



Beispiel: Balkendiagramm

Erweiterte Antwort:

```
farben=c("red","yellow","green","violet","orange","blue","pink","c
barplot(school.freq,col=farben)
```



Kuchendiagramm

Definition

Ein Kuchendiagramm stellt die Häufigkeitsverteilung von qualitativen Daten durch Pizzastücke graphisch dar.

Problem: Ein Kuchendiagramm der Variable `school` von **painters** gibt mit Pizzastücken die Anzahl der Maler pro Schule an.

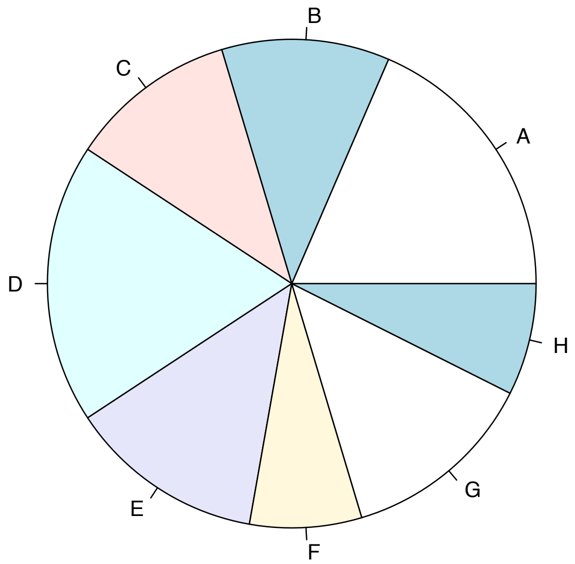
Lösung:

```
school=painters$School
school.freq=table(school)
```

Beispiel: Kuchendiagramm

Lösung:

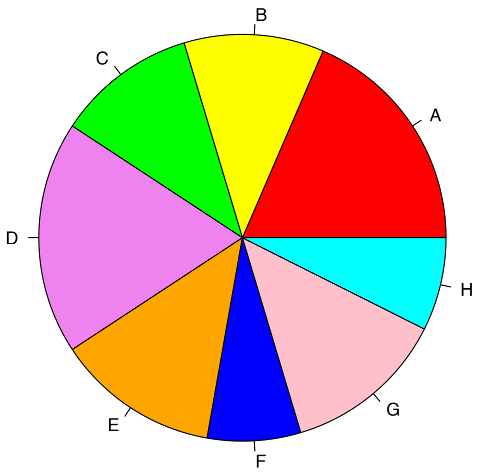
```
pie(school.freq)
```



Beispiel: Kuchendiagramm

Erweiterte Antwort:

```
farben=c("red","yellow","green","violet","orange","blue","pink","c
pie(school.freq,col=farben)
```



Gruppenstatistik

Problem: Bestimmen Sie den durchschnittlichen Wert von `Composition` in der Schule C.

Lösung:

```
school=painters$School
c_school= school=="C"
c_painters = painters[c_school, ]
mean(c_painters$Composition)

## [1] 13.16667
```

Gruppenstatistik

Erweiterte Antwort: Anstatt den Durchschnittswert von `Composition` jeder Schule manuell zu bestimmen, verwenden wir die Funktion `tapply`:

```
tapply(painters$Composition, painters$School, mean)
```

```
##        A        B        C        D        E        F
## 10.40000 12.16667 13.16667  9.10000 13.57143  7.25000
##        G        H
## 13.85714 14.00000
```