Methoden empirischer Sozialforschung

Metrische Merkmale

1 metrisches Merkmal

- In welchem Bereich kann man einen Mittelwert in der Grundgesamtheit erwarten?
  Konfidenzintervalle für den Mittelwert

1 metrisches Merkmal

- Numerische Beschreibung
  (Verteilungs-)Maßzahlen
- Grafische Beschreibung
  Histogramm, Boxplot
- Haben die Daten eine bestimmte Verteilungsform?
  QQ-Plot
- Ist ein Mittelwert in der Grundgesamtheit anders als in einer bestimmten Vorgabe?
  t-Test, Wilcoxon-Test

2 metrische Merkmale

- Grafische Beschreibung
  Streudiagramm
- Numerische Beschreibung
  Korrelationskoeffizient
- Wie stark ist der Zusammenhang zwischen zwei metrischen Merkmalen?
  Test des Korrelationskoeffizienten
- Unterscheiden sich die Mittelwerte zweier Merkmale, die an derselben Beobachtungseinheit erhoben wurden?
  Gepaarter t- und Wilcoxon-Test

Metrische Merkmale in R

- Metrische Merkmale werden in R einfach in numerischen Vektoren gespeichert.

- Eine Zusammenfassung mit einigen wichtigen Verteilungsmaßzahlen wird mit `summary()` erstellt.

Metrische Merkmale in R

```
R> x <- c(3.2, 4.7, 12, 0.3, 9.876, 3.9, 10.04, 11.8, 7)
R> class(x)
```

```
[1] "numeric"
```

```
R> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.30    3.90    7.00    6.98   10.04   12.00
```

```
R> remove(x)
```

Beispiel: Nitratbelastung

Nitratbelastung des Trinkwassers in NÖ.
(Quelle: NöSiWAG/WWF, 1998)

Information (in mg/l – Grenzwert ist 50mg/l) von 526 Messstellen

```
R> NOEWASSER <- read.table("noewasser.tab", header = TRUE)
R> dim(NOEWASSER)
```

```
[1] 526  12
```

```
R> attach(NOEWASSER)
```

Beispiel: Nitratbelastung

```
R> NITRAT
```

```
 [1] 12 12 23 23 32 32 32 12 12 12 12 12 12 12 23
[16] 12 12 12 12 23 23 23 20 20 20 20 20 20 20  4
[31] 23 12 12 20 20 12 12  4  1  1  1  1  1  1  9
...
```

Information aus den Rohdaten unübersichtlich.

- Wie kann man die Information aus den Daten zusammenfassen?

- Wie kann man die Daten grafisch beschreiben?

Beispiel: Gewichtsabnahme

Gewichtsreduktion nach 2-wöchiger Diät bei 20 Personen

```
R> DIET <- read.table("diet.tab", header = TRUE)
R> dim(DIET)

[1] 20  1

R> DIET[, "GEWICHT"]

 [1] 4.373033 5.748030 6.888957 3.820782 7.131452 3.809716 6.093950 2.004049
 [9] 2.567579 5.729797 3.176135 4.903371 6.811930 3.772525 3.428133 5.580575
[17] 2.596678 3.616996 6.459704 5.994147
```

Beispiel: Testergebnisse

Testergebnisse von 20 Studenten

```
R> LAGE <- read.table("lagemasse.tab", header = TRUE)
R> dim(LAGE)

[1] 20  4

R> names(LAGE)

[1] "TEST"    "GEMUESE" "KRANK"   "MAGAZIN"
```

Beispiel: Testergebnisse

```
R> attach(LAGE)
R> TEST

 [1] 2 3 3 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 8 9
```

Diese Daten sind *diskret*, können also nur bestimmte Werte annehmen.

1 metrisches Merkmal

Numerische Beschreibung

Wesentliche Begriffe

- Häufigkeitsverteilung (kurz: Verteilung)
  Welche Zahlen kommen wie oft vor?
- Numerische Zusammenfassungen, statistische Maßzahlen
  Wie können wir eine Verteilung durch einige wenige Zahlen beschreiben, die die wesentlichen Charakteristika enthalten?
- Grafische Präsentation
  Wie kann eine Verteilung grafisch dargestellt werden?

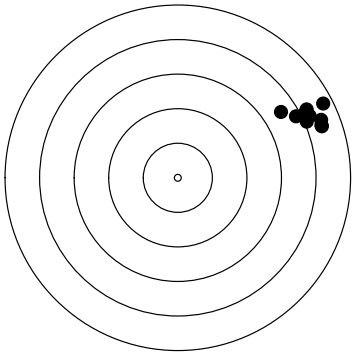
Numerische Beschreibung

Gestalt von Verteilungen wird beschrieben durch:

- Lage (Mittelwert, Median)
- Streuung (Varianz, Standardabweichung)
- Schiefe
- Wölbung

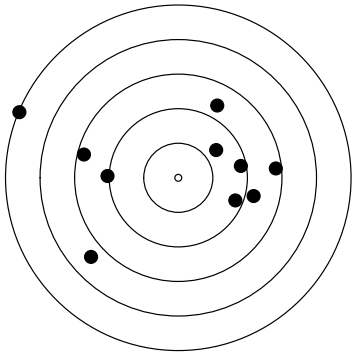
Numerische Beschreibung

Lage: schlecht − Streuung: gut

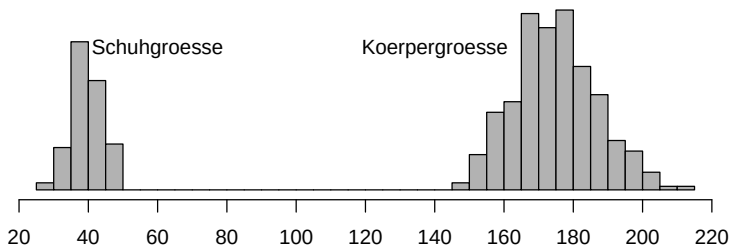


Numerische Beschreibung

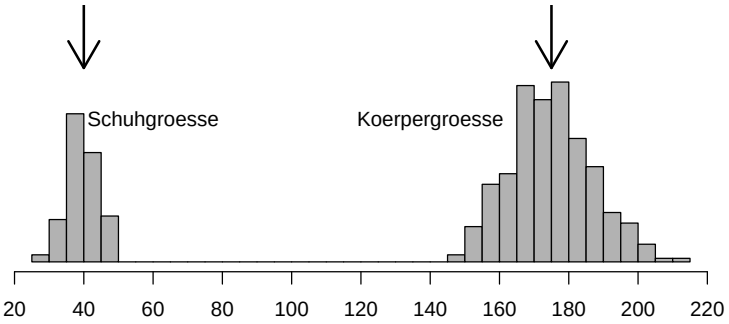
Lage: gut − Streuung: schlecht



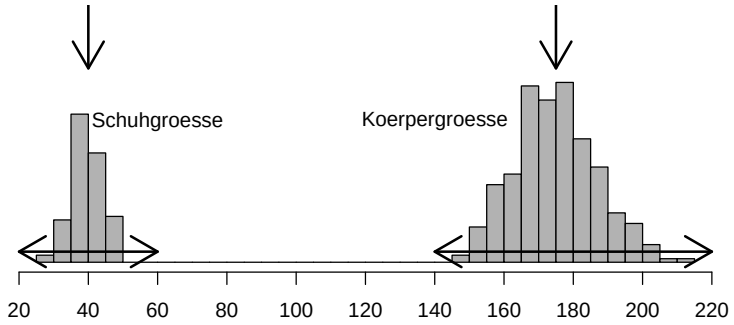
Numerische Beschreibung



Numerische Beschreibung



Numerische Beschreibung



Numerische Beschreibung

Lagemaße:

- **Mittelwert** – Durchschnittswert

$$\bar{x} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Median** – mittlerer Wert (50% Quantil $Q_{0.5}$, teilt die sortierte Datenliste in obere und untere Hälfte)

- **Modus** – häufigster Wert

	Mittelwert	Median	Modus
kategoriale Daten	–	–	ok
ordinale Daten	–	ok	ok
metrische Daten	ok	ok	ok

Numerische Beschreibung

Qual der Wahl bei metrischen Daten:

Bei eingipfligen symmetrischen Verteilungen mit nur einem Gipfel sind alle drei Lagemaß typischerweise sehr ähnlich.

```
R> mean(TEST)

[1] 5.4

R> median(TEST)

[1] 5.5

R> table(TEST)

TEST
2 3 4 5 6 7 8 9
1 2 3 4 5 3 1 1
```

Numerische Beschreibung

Streuungsmaße

- **Varianz** – mittlere quadratische Abweichung vom Mittelwert

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

- **Standardabweichung** – Wurzel der Varianz (s_x)
- **Interquartilsabstand** – Differenz zwischen dem 75%-Quantil (oberes Quartil) und dem 25%-Quantil (unteres Quartil)
- **Spannweite** – Differenz von größtem und kleinstem Wert

Numerische Beschreibung

	Standardabweichung	IQR	Spannweite
kategoriale Daten	–	–	?
ordinale Daten	–	?	?
metrische Daten	ok	ok	ok

Varianz, Standardabweichung und IQR sind nur sinnvoll interpretierbar, wenn eine ungefähr symmetrische, eingipflige Verteilung vorliegt.

Numerische Beschreibung

Lagemaße

- Mittelwert (arithmetisches Mittel): `mean(x)`
- Median: `median(x)`
- Modus (via `table(x)`)

Numerische Beschreibung

Streungsmaße

- Varianz: `var(x)`
- Standardabweichung: `sd(x)`
- Mittlerer absoluter Abstand: `mad(x)`
- Interquartilsabstand: `IQR(x)`
- Spannweite: `range(x)`, `diff(range(x))`

Numerische Beschreibung

Weitere Kennzahlen/Zusammenfassungen

- Häufigkeitstabellen:
  `table(x)`
- Quantile:
  `quantile(x)`
- Five-Point Summary:
  `fivenum(x)`

Numerische Beschreibung

```
R> NOEWASSER <- read.table("noewasser.tab", header = TRUE)
R> dim(NOEWASSER)

[1] 526  12

R> names(NOEWASSER)

 [1] "GEMEINDE" "ENTNAHME" "DATUM"    "PHWERT"   "GESAMTH." "NITRAT"
 [7] "EISEN"    "MANGAN"   "CHLORID"  "SULFAT"   "ORT"      "GEWICHT"

R> attach(NOEWASSER)
```

Numerische Beschreibung

```
R> mean(NITRAT)

[1] 17.39163

R> median(NITRAT)

[1] 15
```

Numerische Beschreibung

```
R> var(NITRAT)

[1] 134.0254

R> sd(NITRAT)

[1] 11.57693

R> sqrt(var(NITRAT))

[1] 11.57693

R> mad(NITRAT)

[1] 11.8608

R> IQR(NITRAT)

[1] 17
```

Numerische Beschreibung

```
R> min(NITRAT)

[1] 1

R> max(NITRAT)

[1] 47

R> range(NITRAT)

[1]  1 47

R> diff(range(NITRAT))

[1] 46
```

Numerische Beschreibung

```
R> fivenum(NITRAT)

[1]  1  7 15 24 47

R> summary(NITRAT)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    7.00   15.00   17.39   24.00   47.00

R> quantile(NITRAT)

  0%  25%  50%  75% 100%
   1    7   15   24   47

R> quantile(NITRAT, 0.9)

90%
 32
```

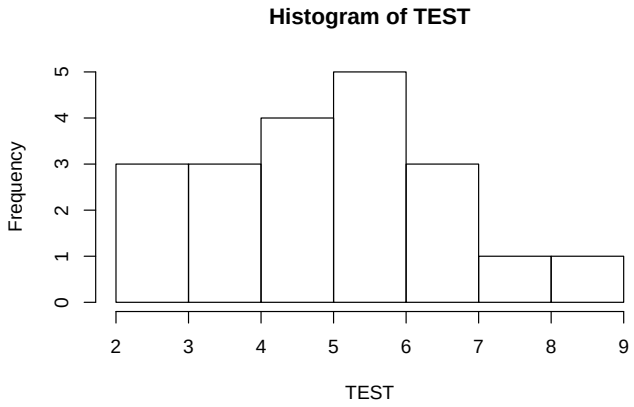
Grafische Beschreibung

Das **Histogramm** ist eine flächenproportionale Darstellung für die relativen Häufigkeiten einer numerischen Variablen.

```
R> table(TEST)

TEST
2 3 4 5 6 7 8 9
1 2 3 4 5 3 1 1

R> table(TEST)/length(TEST)

TEST
   2    3    4    5    6    7    8    9
0.05 0.10 0.15 0.20 0.25 0.15 0.05 0.05
```

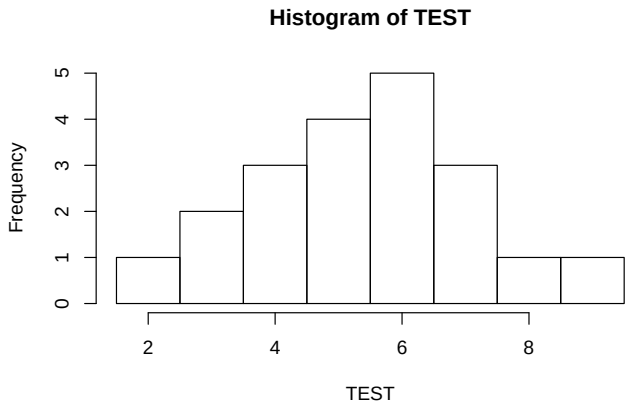
Grafische Beschreibung

```
R> hist(TEST)
```



Grafische Beschreibung

```
R> hist(TEST, breaks = 1:9 + 0.5)
```



Grafische Beschreibung

```
R> hist(TEST, breaks = 1:9 + 0.5, prob = TRUE)
```

Grafische Beschreibung

```
R> hist(TEST, breaks = 1:9 + 0.5, prob = TRUE)
R> abline(v = mean(TEST), col = 2)
```

Grafische Beschreibung

Aber wenn die Daten nicht diskret (genug) sind, sondern kontinuierlich:

```
R> table(GEWICHT)

GEWICHT
2.004049 2.567579 2.596678 3.176135 3.428133 3.616996 3.772525 3.809716
       1        1        1        1        1        1        1        1
3.820782 4.373033 4.903371 5.580575 5.729797  5.74803 5.994147  6.09395
       1        1        1        1        1        1        1        1
6.459704  6.81193 6.888957 7.131452
       1        1        1        1
```

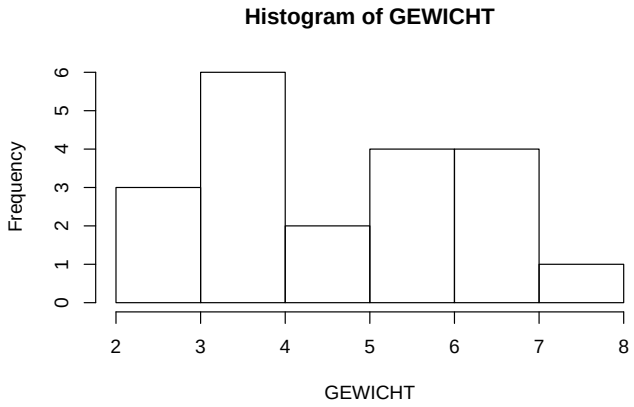
Bilde Intervalle und dann eine Häufigkeitstabelle für diese kategorisierten Daten.

Grafische Beschreibung

```
R> cut(GEWICHT, 2:8)

 [1] (4,5] (5,6] (6,7] (3,4] (7,8] (3,4] (6,7] (2,3] (2,3] (5,6] (3,4] (4,5]
[13] (6,7] (3,4] (3,4] (5,6] (2,3] (3,4] (6,7] (5,6]
Levels: (2,3] (3,4] (4,5] (5,6] (6,7] (7,8]

R> table(cut(GEWICHT, 2:8))

(2,3] (3,4] (4,5] (5,6] (6,7] (7,8]
    3     6     2     4     4     1

R> table(cut(GEWICHT, 2:8))/length(GEWICHT)

(2,3] (3,4] (4,5] (5,6] (6,7] (7,8]
 0.15  0.30  0.10  0.20  0.20  0.05
```

Grafische Beschreibung

```
R> hist(GEWICHT)
```



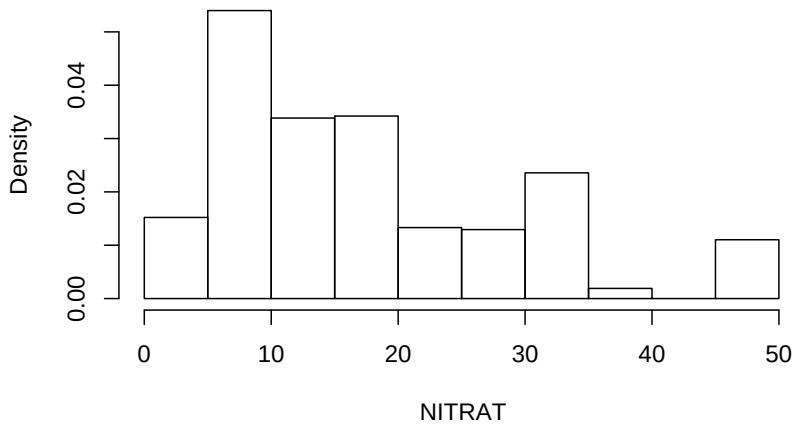
Grafische Beschreibung

Wie sollte diese Intervalleinteilung vorgenommen werden?

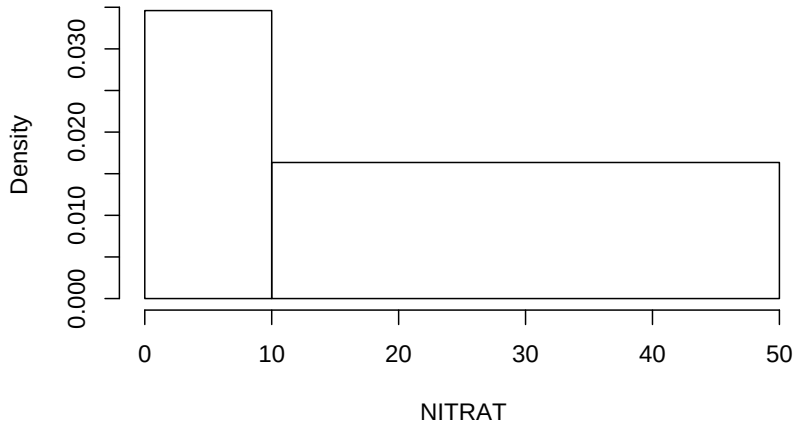
Die Intervalleinteilung beeinflusst den visuellen Eindruck substanziell.

(R wählt eine Anzahl gleichgroßer Intervalle nach einer Faustregel)

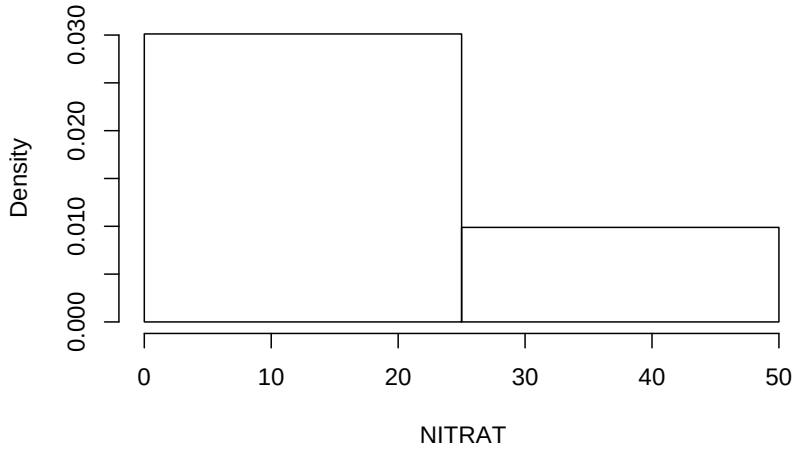
Grafische Beschreibung



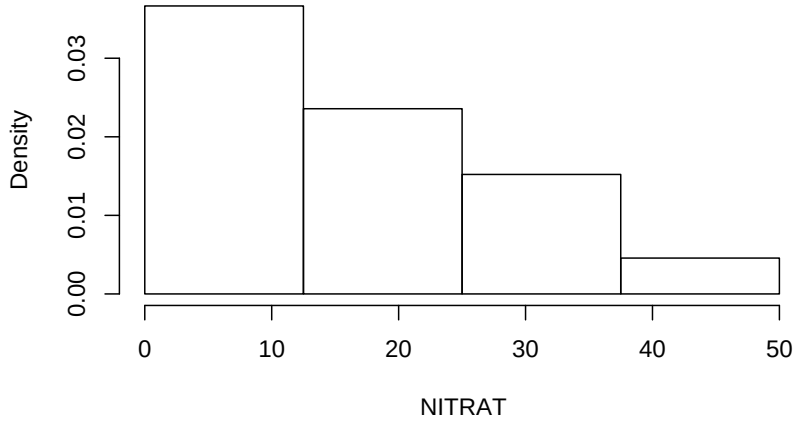
Grafische Beschreibung



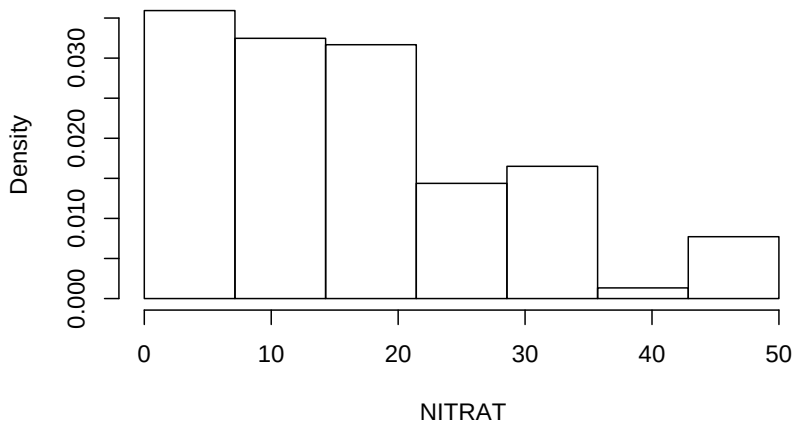
Grafische Beschreibung



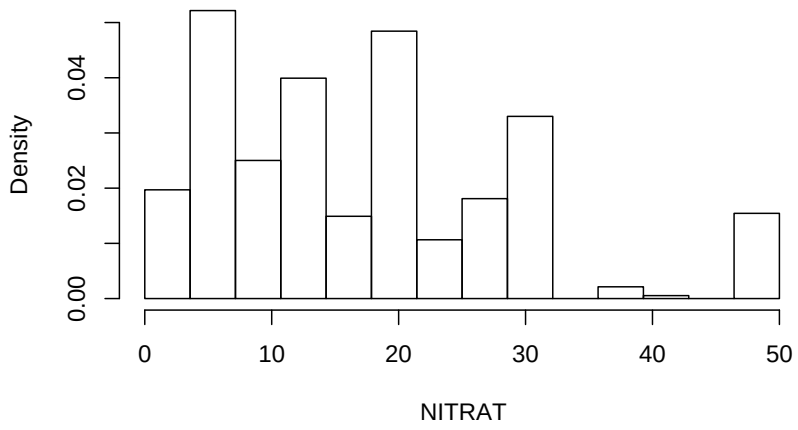
Grafische Beschreibung



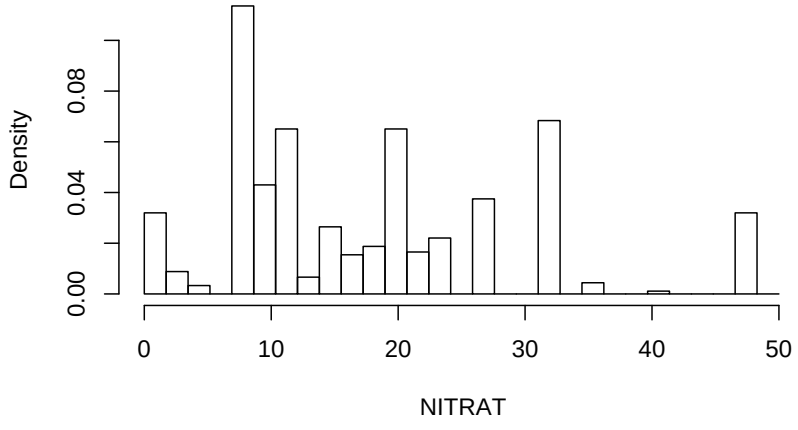
Grafische Beschreibung



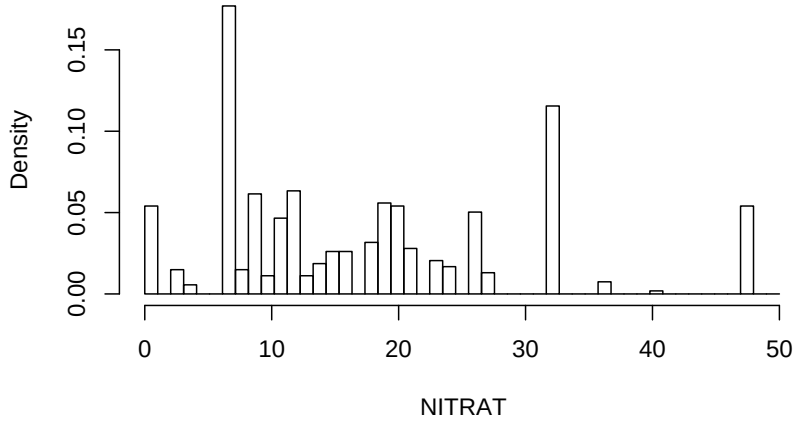
Grafische Beschreibung



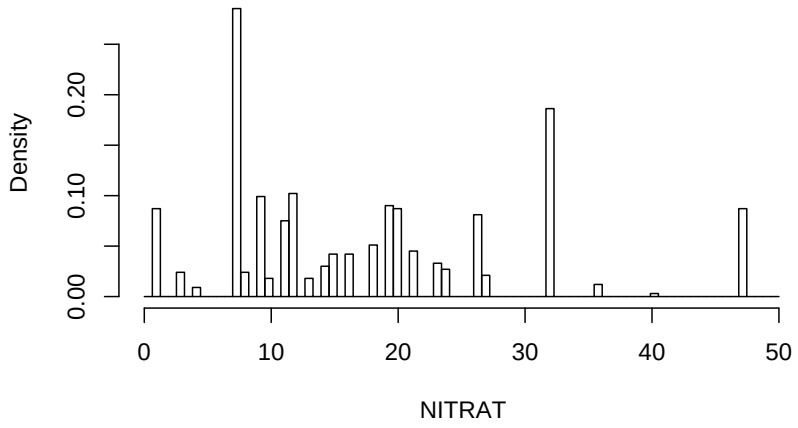
Grafische Beschreibung



Grafische Beschreibung



Grafische Beschreibung



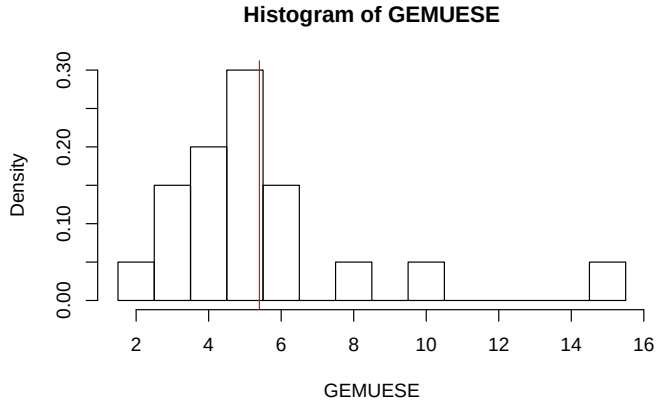
Grafische Beschreibung

Einige Lagemaße waren nur für symmetrische eingipflige Verteilungen geeignet. Dies kann nun anhand von Histogrammen auch grafisch verdeutlicht werden, bspw. für schiefe und U-förmige Verteilungen.

- `GEMUESE` – Anzahl Beschäftigter in 20 Gemüsegeschäften
- `KRANK` – Krankenstandstage von 20 Arbeitern
- `MAGAZIN` – Anzahl gelesener Ausgaben einer Monatszeitschrift (erhoben bei 20 Personen)

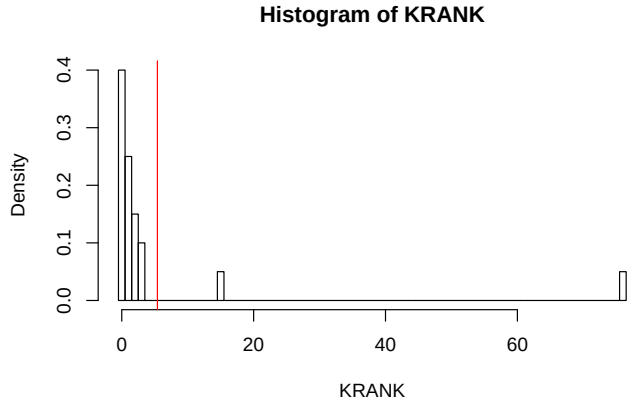
Grafische Beschreibung

```
R> hist(GEMUESE, prob = TRUE, breaks = 1:15 + 0.5)
R> abline(v = mean(GEMUESE), col = 2)
```



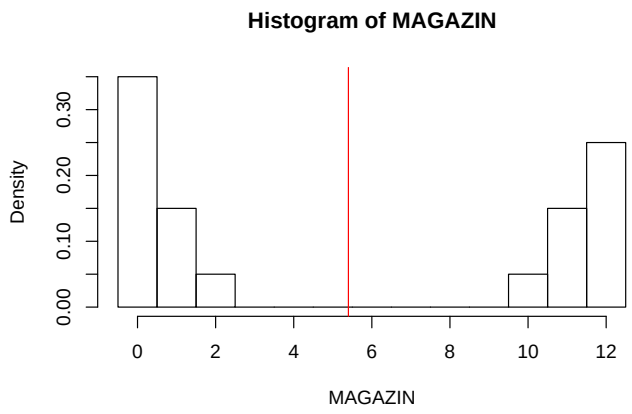
Grafische Beschreibung

```
R> hist(KRANK, prob = TRUE, breaks = -1:76 + 0.5)
R> abline(v = mean(KRANK), col = 2)
```



Grafische Beschreibung

```
R> hist(MAGAZIN, prob = TRUE, breaks = -1:12 + 0.5)
R> abline(v = mean(MAGAZIN), col = 2)
```



Grafische Beschreibung

Zwei Arten von Daten:

1. Etwa symmetrische Verteilung mit einem Gipfel.
   Mittelwert ist ein typischer Wert.

2. Schiefe, U-förmige, mehrgipflige Verteilungen oder Gleichverteilungen.
   Mittelwert ist *kein* typischer Wert.

Ausserdem: was tun bei Ausreißern?

Grafische Beschreibung

Manchmal beeinflussen sehr wenige Werte die statistischen Kennzahlen sehr stark, diese nennt man *Ausreißer*.

Mögliche Ursachen für Ausreißer:

- Schreib-/Tipp-Fehler
  Falls möglich korrigieren, sonst weglassen.

- Tatsächliche extreme Beobachtung:
  Robuste Methoden verwenden.
  Falls sinnvoll: Wert(e) weglassen und getrennt behandeln.

Ein Maßzahl heißt robust, falls sie sich durch einen einzelnen Ausreißer nur beschränkt ändern kann.

Grafische Beschreibung

```
R> KRANK
 [1]  0  0  0  0  0  0  0  0  1  1  1  1  1  2  2  2  3  3 15 76
R> KRANK[20]
[1] 76
R> mean(KRANK)
[1] 5.4
R> mean(KRANK[-20])
[1] 1.684211
```

Grafische Beschreibung

```
R> median(KRANK)

[1] 1

R> median(KRANK[-20])

[1] 1
```

Lagemaße: Median und Modus sind robust, der Mittelwert nicht.

Streuungsmaße: Der Interquartilsabstand ist robust, die Standardabweichung (bzw. Varianz) nicht.

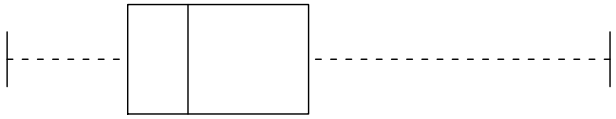
Grafische Beschreibung

Eine Darstellung, die die Daten noch mehr komprimiert ist der *Boxplot*.

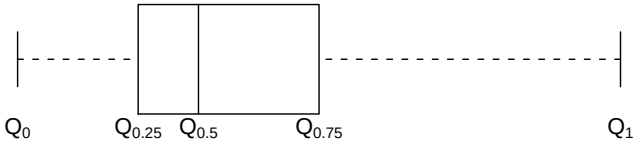
Dieser ist eine Darstellung für die Five-Point-Summary.

In R: `boxplot(x)`

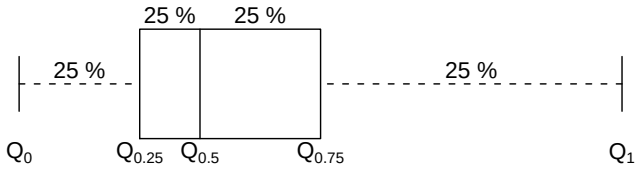
Grafische Beschreibung



Grafische Beschreibung

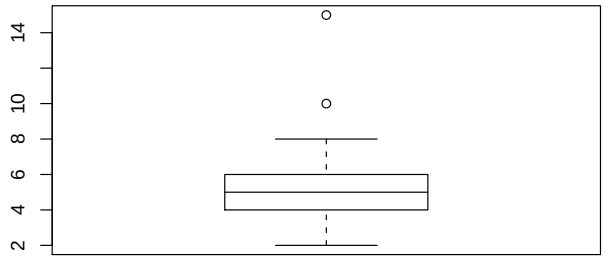


Grafische Beschreibung



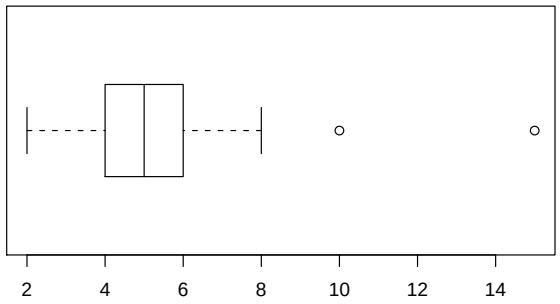
Grafische Beschreibung

```
R> boxplot(GEMUESE)
```



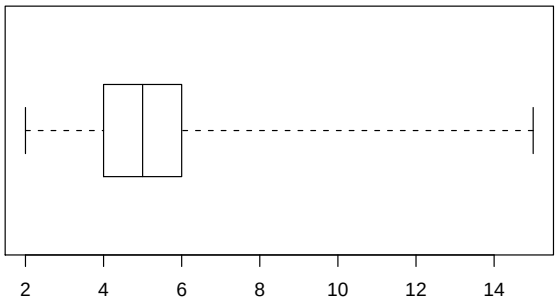
Grafische Beschreibung

```
R> boxplot(GEMUESE, horizontal = TRUE)
```



Grafische Beschreibung

```
R> boxplot(GEMUESE, horizontal = TRUE, range = 0)
```



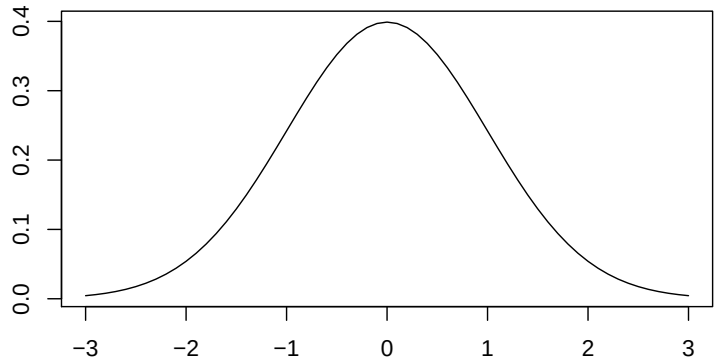
Verteilungsform

Als typische symmetrische eingipflige Verteilung wird häufig die *(Standard-)Normalverteilung* herangezogen.

Sie approximiert viele Verteilungen recht gut und wird oft als Referenzverteilung herangezogen.

Viele Verfahren funktionieren nur gut, wenn die untersuchten Daten annähernd normalverteilt sind.

Verteilungsform



Verteilungsform

Um zu beurteilen, ob ein Datensatz etwa normalverteilt ist, trägt man dessen empirische Quantile gegen die Quantile einer Standardnormalverteilung ab.

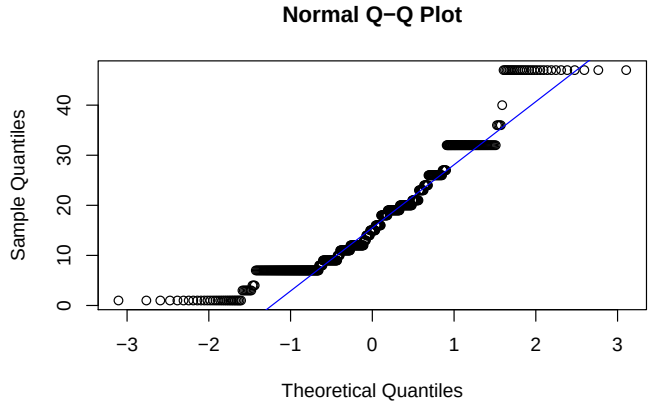
Bei einer Normalverteilung bilden die abgetragenen Punkte eine exakte Grade. Sind die Abweichungen von einer Gerade nicht allzu groß (insbesonders nicht an den Enden), so ist die Approximation nicht schlecht.

In R:

```
qqnorm(x)
qqline(x)
```

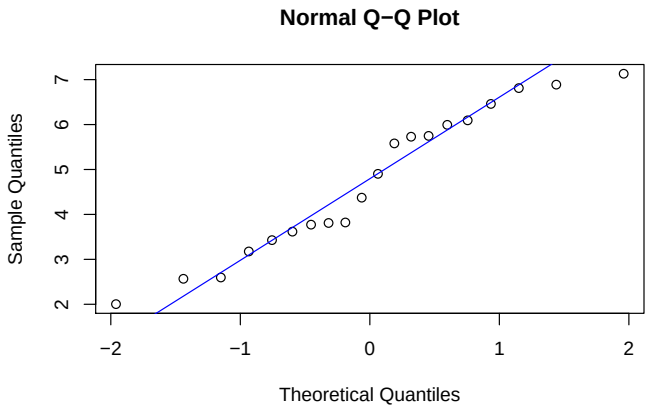
Verteilungsform

```
R> qqnorm(NITRAT)
R> qqline(NITRAT, col = 4)
```



Verteilungsform

```
R> qqnorm(GEWICHT)
R> qqline(GEWICHT, col = 4)
```



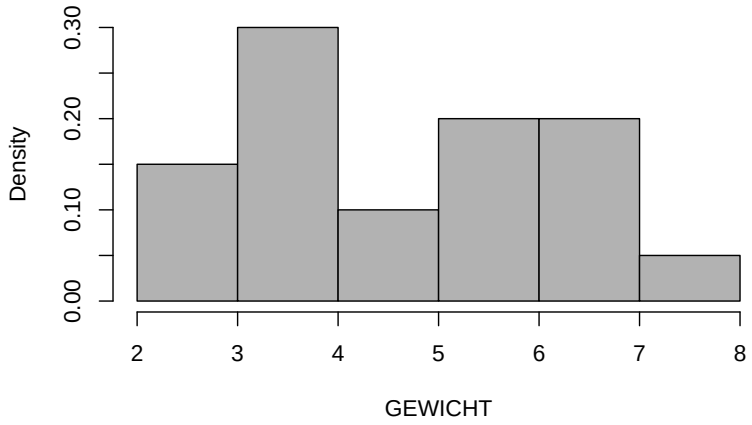
Test des Mittelwerts

Ist ein Mittelwert in der Grundgesamtheit anders als in einer bestimmten Vorgabe?

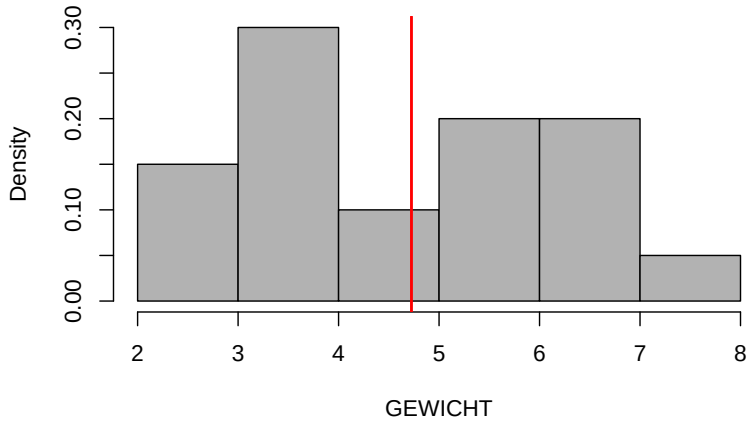
Beispiel: Von der Diät aus dem `DIET` Datensatz wird behauptet, daß sie zu einer Gewichtsreduktion von mindestens 5kg führt.

Frage: Ist die mittlere Gewichtsabnahme von 4.725kg aus einer Stichprobe vom Umfang 20 ausreichend, um nachzuweisen, daß in der Grundgesamtheit die Gewichtsabnahme unter 5kg liegt?

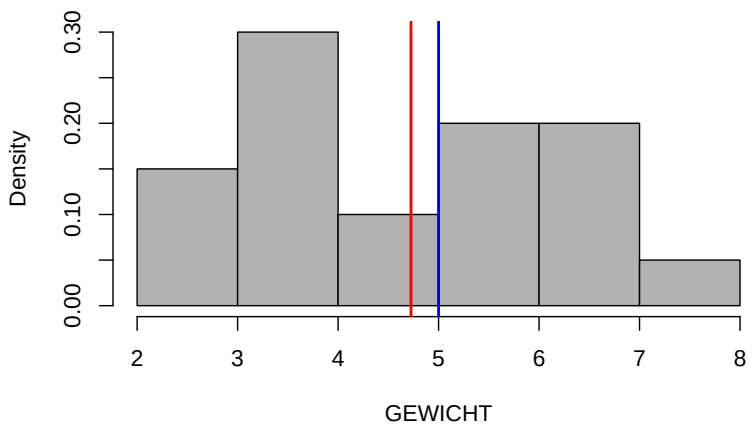
Test des Mittelwerts



Test des Mittelwerts



Test des Mittelwerts



Test des Mittelwerts

Frage: Weicht der *Mittelwert der Stichprobe* stark von dem hypothetisch unterstellten *Erwartungswert der Grundgesamtheit* ab?

Anders gefragt: Ist ein empirischer Mittelwert von $\bar{x} = 4.725$ mit einem unterstellten Erwartungswert von $\mu = 5$ vereinbar?

Zur Beantwortung brauchen wir eine Methode, um das Ergebnis der Stichprobe beurteilen zu können: *statistischer Test*

Test des Mittelwerts

Es kommen zwei Antworten in Frage:

H der Erwartungswert ist $\mu = \mu_0 = 5$ (oder sogar größer)
A der Erwartungswert ist $\mu < \mu_0$

H nennt man auch *Nullhypothese*, die hier besagt: die Gewichtsabnahme durch die Diät beträgt (mindestens) 5kg.

A nennt man auch *Alternative*, die hier besagt: die Gewichtsabnahme beträgt weniger als 5kg.

Test des Mittelwerts

Durchführung von statistischen Tests:

1. Nimm an, daß die Nullhypothese wahr ist.

2. Berechne eine Teststatistik, die die Diskrepanz zwischen den Beobachtungen und den Erwartungen (unter der Nullhypothese) einfängt.

3. Berechne die Wahrscheinlichkeit, eine solche Teststatistik (oder eine extremere) zu beobachten. Diese Wahrscheinlichkeit heißt *p-Wert*.

Test des Mittelwerts

Berechnung der Teststatistik: betrachte Differenz

$$\bar{x} - \mu_0$$

Der Mittelwert einer Stichprobe unterliegt Zufallsschwankungen, deshalb standardisiere mit der geschätzten Standardabweichung des Mittelwerts

$$\frac{s_x}{\sqrt{n}}$$

Test des Mittelwerts

Die Teststatistik

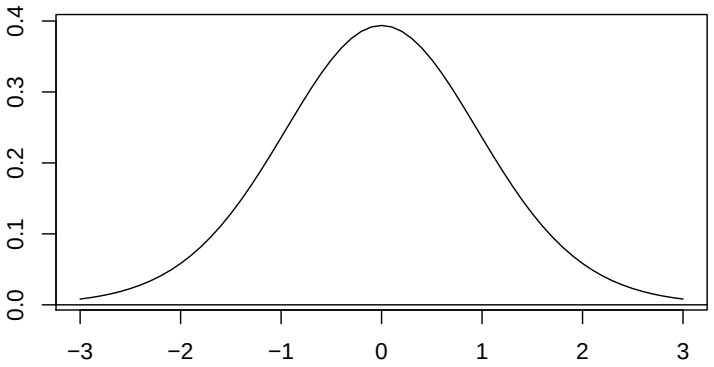
$$T \quad = \quad \sqrt{n}\,\frac{\bar{x} - \mu_0}{s_x}$$

ist approximativ t-verteilt mit $n - 1$ Freiheitsgraden, falls die Daten tatsächlich aus einer Grundgesamtheit mit Erwartungswert μ_0 stammen.

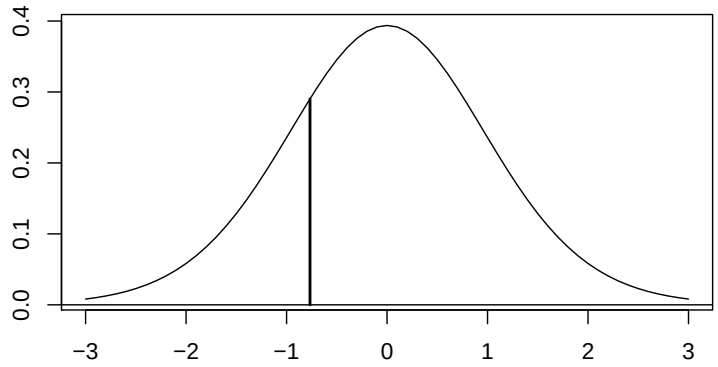
Paßt nun die empirische Teststatistik von $T = -0.767$ zu einer t_{19} Verteilung oder ist sie unwahrscheinlich klein?

Wie wahrscheinlich ist es unter der Nullhypothese eine solche Statistik T oder eine noch extremere zu beobachten?

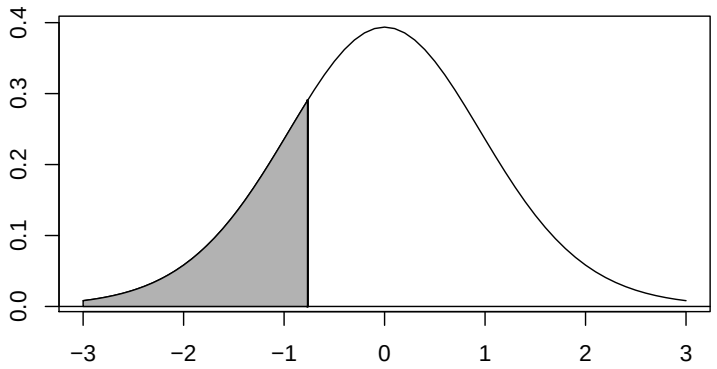
Test des Mittelwerts



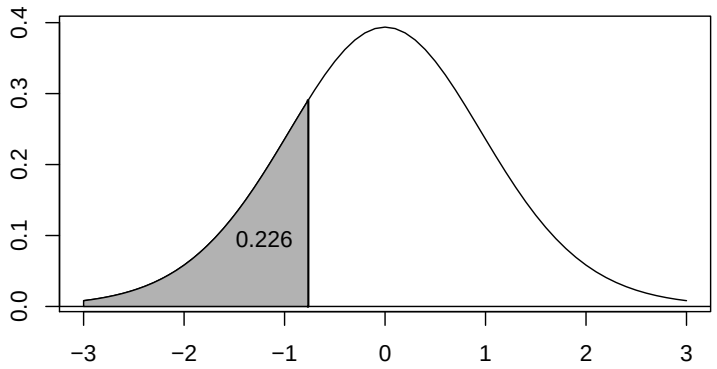
Test des Mittelwerts



Test des Mittelwerts



Test des Mittelwerts



Test des Mittelwerts

Zwei mögliche Ausgänge:

(a) Der p-Wert ist sehr klein (typischerweise: kleiner 5%). Die Daten widersprechen der Nullhypothese und diese wird verworfen. Das Ergebnis ist *signifikant*.

(b) Der p-Wert ist nicht sehr klein. Die Daten liefern keinen statistischen Beweis dafür, dass die Nullhypothese falsch ist, diese wird beibehalten. Das Ergebnis ist *nicht signifikant*.

Hier widersprechen die Daten offensichtlich nicht der Annahme, daß der Erwartungswert in Wirklichkeit größer als 5 ist.

Test des Mittelwerts

Bei solchen Entscheidung können zwei Arten von Fehlern gemacht werden:

	H wird nicht verworfen	H wird verworfen
H trifft zu	Entscheidung richtig	Fehler 1. Art
H trifft nicht zu	Fehler 2. Art	Entscheidung richtig

Der Fehler 1. Art wird durch das Signifkanzniveau des Tests (typischerweise 5%) kontrolliert. Deshalb ist das Verwerfen der Nullhypothese ein statistischer Beweis dafür, daß sie falsch ist.

Test des Mittelwerts

In R:

```
t.test(x, mu = 0, alternative = "two.sided", ...)
```

Dabei spezifiziert

- `x` – die Beobachtungen
- `mu` – den Erwartungswert unter der Hypothese
- `alternative` – die Alternative, die entweder `"two.sided"` (ungleich), `"less"` (kleiner) oder `"greater"` (größer) sein kann.

Test des Mittelwerts

```
R> t.test(GEWICHT, mu = 5, alternative = "less")

        One Sample t-test

data:  GEWICHT
t = -0.7669, df = 19, p-value = 0.2263
alternative hypothesis: true mean is less than 5
95 percent confidence interval:
     -Inf 5.344555
sample estimates:
mean of x
 4.725377
```

Test des Mittelwerts

R gibt außerdem ein Konfidenzintervall für den wahren Erwartungswert der Grundgesamtheit an. Dieses überdeckt den Erwartungswert μ mit einer vorgegebenen Wahrscheinlichkeit (standardmäßig 95%).

Dies ist äquivalent zur Durchführung des Tests mit Hilfe des p-Werts. Wenn das Konfidenzintervall zum 95% Niveau den unterstellten Wert μ_0 miteinschließt, so widerspricht es offenbar nicht der Annahme. Hier

```
R> t.test(GEWICHT, alternative = "less")$conf.int

[1]     -Inf 5.344555
attr(,"conf.level")
[1] 0.95
```

Test des Mittelwerts

Der t-Test funktioniert dann besonders gut, wenn die Daten etwa symmetrisch und eingipflig und etwa normalverteilt sind.

Sind die Daten zwar symmetrisch und eingipflig, aber nicht besonders normalverteilt – etwa weil es Ausreißer gibt – dann gibt es einen anderen Test, den Wilcoxon-Test.

Idee: Betrachte die Ränge der Beobachtungen links und rechts vom unterstellten Median.

Test des Mittelwerts

```
R> wilcox.test(GEWICHT, mu = 5, alternative = "less")

        Wilcoxon signed rank test

data:  GEWICHT
V = 83, p-value = 0.2152
alternative hypothesis: true mu is less than 5
```

2 metrische Merkmale

Grafische Beschreibung

Zwei metrische Merkmale X und Y werden an denselben Merkmalsträgern gemessen. Stichprobe: (x_i, y_i), $i = 1, \ldots, n$.

Fragen:

- Besteht ein Zusammenhang zwischen den beiden Variablen?
- Wie kann ich den Zusammenhang grafisch und numerisch beschreiben?

Grafische Beschreibung

Beispiel: Ausgaben für Alkohol und Tabak

In 11 Regionen Großbrittanniens wurden 1981 die durchschnittlichen Ausgaben (in britischen Pfund, GBP) pro Haushalt und Woche für Alkohol und Tabakwaren erhoben.

Frage: Wie stark ist der Zusammenhang zwischen den Ausgaben für Alkohol und Tabak?

Grafische Beschreibung

```
R> ALCTOBAC <- read.table("alctobac.tab", header = TRUE)
R> dim(ALCTOBAC)

[1] 11  3

R> ALCTOBAC

            REGION ALCOHOL TOBACCO
1            North    6.47    4.03
2        Yorkshire    6.13    3.76
3        Northeast    6.19    3.77
4     EastMidlands    4.89    3.34
5     WestMidlands    5.63    3.47
6       EastAnglia    4.52    2.92
7        Southeast    5.89    3.20
8        Southwest    4.79    2.71
9            Wales    5.27    3.53
10        Scotland    6.08    4.51
11 NorthernIreland    4.02    4.56

R> attach(ALCTOBAC)
```

Grafische Beschreibung

Zur grafischen Beschreibung wird am einfachsten ein *Streudiagramm* verwendet.

In R:

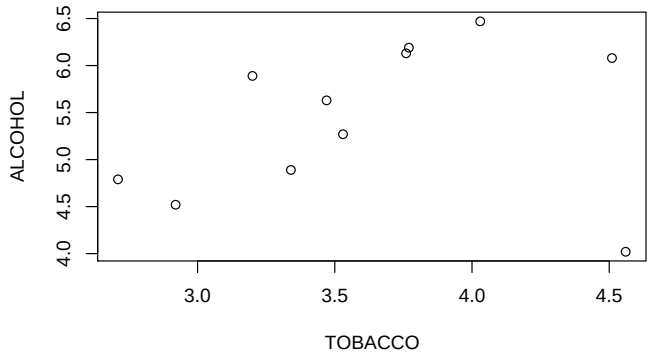
```
plot(x, y)
```

oder

```
plot(y ~ x)
```

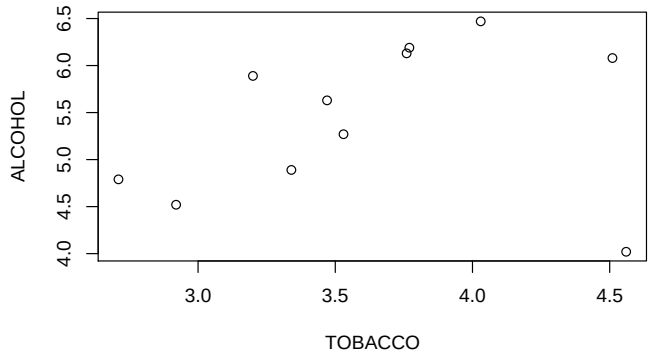
Grafische Beschreibung

```
R> plot(TOBACCO, ALCOHOL)
```



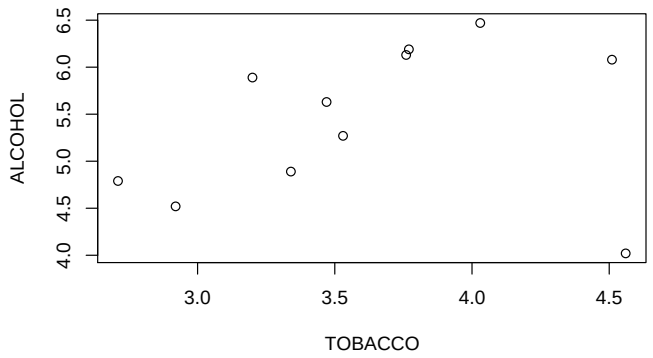
Grafische Beschreibung

```
R> plot(ALCOHOL ~ TOBACCO)
```



Grafische Beschreibung

```
R> plot(ALCOHOL ~ TOBACCO, data = ALCTOBAC)
```



Grafische Beschreibung

Eigenschaften des Streudiagramms

- $(\bar{x}, \bar{y})$ ist der Mittelpunkt der Punktewolke
- Projektion der Punktewolke auf die x-Achse ergibt das Punktediagramm der Datenliste $x_1, \ldots, x_n$.
- Projektion der Punktewolke auf die y-Achse ergibt das Punktediagramm der Datenliste $y_1, \ldots, y_n$.

Numerische Beschreibung

X und Y sind **positiv gekoppelt**, wenn *tendenziell* gilt

- je größer x, desto größer y
- je kleiner x, desto kleiner y

X und Y sind **negativ gekoppelt**, wenn *tendenziell* gilt

- je größer x, desto kleiner y
- je kleiner x, desto größer y

Numerische Beschreibung

Dabei heißt *groß*: größer als der Mittelwert.
Und *klein*: kleiner als der Mittelwert.
Betrachte also:

$$x - \bar{x}$$
$$y - \bar{y}$$

Es gilt:

- $(x - \bar{x}) \cdot (y - \bar{y}) > 0$: Der Punkt (x, y) liegt im 1. oder 3. Quadranten
- $(x - \bar{x}) \cdot (y - \bar{y}) < 0$: Der Punkt (x, y) liegt im 2. oder 4. Quadranten

Numerische Beschreibung

Berücksichtige zusätzlich zum Vorzeichen die Größe von $(x - \bar{x}) \cdot (y - \bar{y})$.

Die *Kovarianz* der Daten ist

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

Problem: Die Kovarianz hängt ab von der verwendeten Maßeinheit.
Ausweg: Normiere durch Standardabweichungen.

Der *Korrelationskoeffizient* r_{xy} (nach K. Pearson) ist

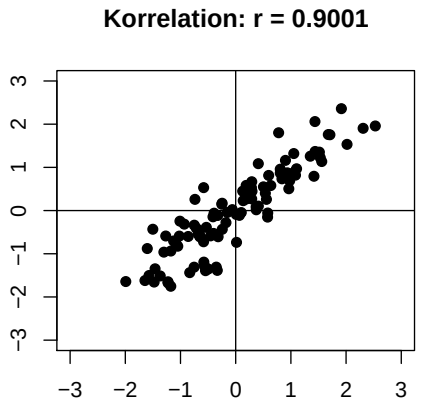
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Numerische Beschreibung

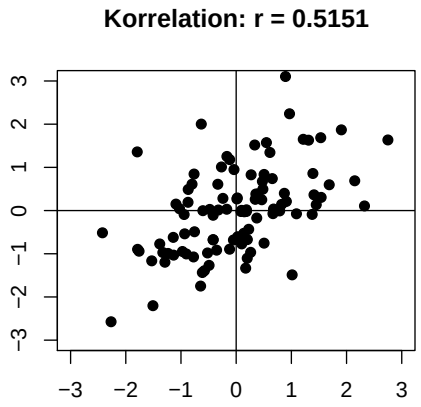
Es gilt:

- symmetrisch: $r_{xy} = r_{yx}$
- identische Datenlisten haben die Korrelation 1: $r_{xx} = 1$
- $-1 \leq r_{xy} \leq 1$

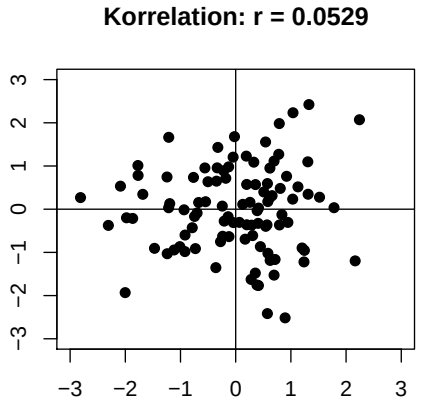
Numerische Beschreibung



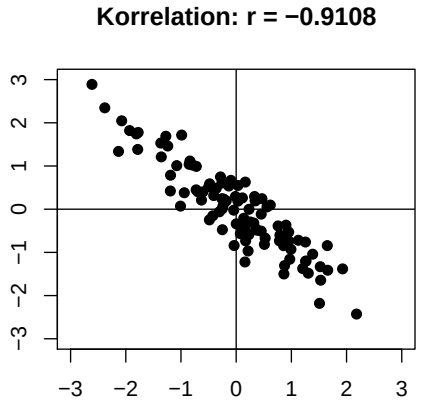
Numerische Beschreibung



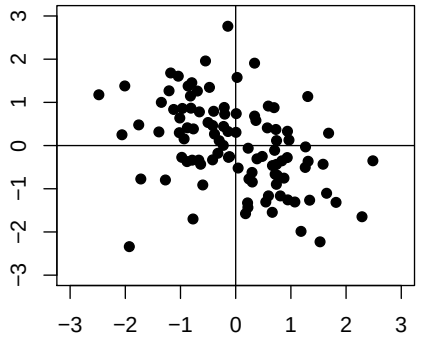
Numerische Beschreibung



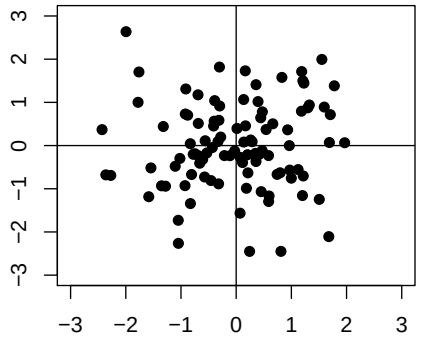
Numerische Beschreibung



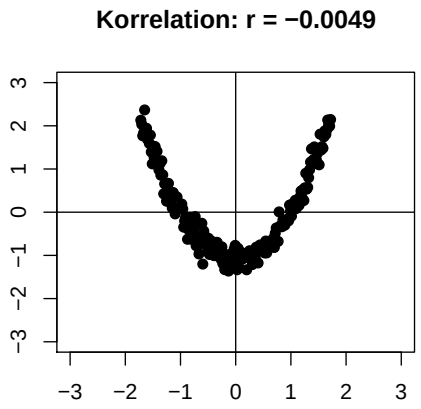
Numerische Beschreibung



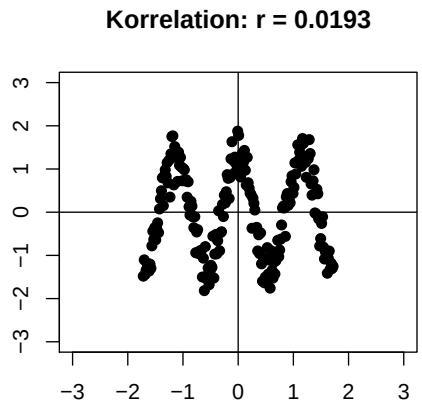
Numerische Beschreibung



Numerische Beschreibung



Numerische Beschreibung



Numerische Beschreibung

In R: `cor(x, y)`

Für den Zusammenhang von Tabak- und Alkoholausgaben:

```
R> cor(TOBACCO, ALCOHOL)
```

```
[1] 0.2235721
```

Pearson-Korrelation ist besonders geeignet für ungefähr normalverteilte Variablen.

Numerische Beschreibung

Wenn es Ausreißer gibt bzw. die Daten nicht sehr normalverteilt sind, dann kann auch Spearmans Rangkorrelation verwendet werden.

Idee: Bilde Ränge von X und Y und berechne die Pearson-Korrelation der Ränge.

```
R> ALCOHOL
```

```
 [1] 6.47 6.13 6.19 4.89 5.63 4.52 5.89 4.79 5.27 6.08 4.02
```

```
R> rank(ALCOHOL)
```

```
 [1] 11  9 10  4  6  2  7  3  5  8  1
```

Numerische Beschreibung

```
R> sort(ALCOHOL)

 [1] 4.02 4.52 4.79 4.89 5.27 5.63 5.89 6.08 6.13 6.19 6.47

R> rank(sort(ALCOHOL))

 [1]  1  2  3  4  5  6  7  8  9 10 11

R> cor(ALCOHOL, TOBACCO)

[1] 0.2235721

R> cor(rank(ALCOHOL), rank(TOBACCO))

[1] 0.3727273

R> cor(ALCOHOL, TOBACCO, method = "spearman")

[1] 0.3727273
```

Numerische Beschreibung

Weitere Alternative: Kendalls τ.

```
R> cor(ALCOHOL, TOBACCO, method = "kendall")

[1] 0.3454545
```

Test der Korrelation

Frage: Besteht ein Zusammenhang zwischen zwei Merkmalen?

Hier: Ist eine empirische Korrelation $r_{xy} = 0.224$ ausreichend um nachzuweisen, daß in der Grundgesamtheit die wahre Korrelation ρ ungleich 0 ist?

$$H : \rho = 0 \quad \text{vs.} \quad A : \rho \neq 0$$

Teststatistik:

$$T = \sqrt{n-2}\,\frac{r_{xy}}{\sqrt{1-r_{xy}^2}}$$

Test der Korrelation

In R:

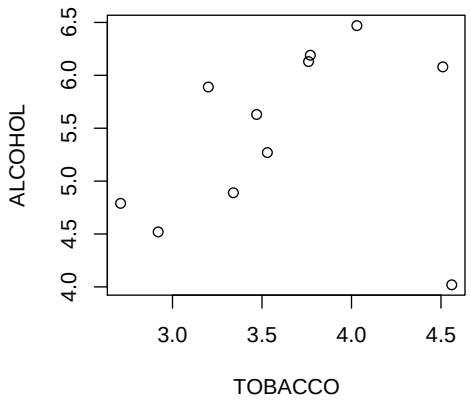
```
R> cor.test(ALCOHOL, TOBACCO)

        Pearson's product-moment correlation

data:  ALCOHOL and TOBACCO
t = 0.6881, df = 9, p-value = 0.5087
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4345878  0.7260700
sample estimates:
      cor
0.2235721
```

Dies liefert auch ein Konfidenzintervall für die wahre Korrelation ρ.

Test der Korrelation

```
R> cor.test(ALCOHOL, TOBACCO, method = "spearman")

        Spearman's rank correlation rho

data:  ALCOHOL and TOBACCO
S = 138, p-value = 0.2544
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3727273

R> cor.test(ALCOHOL, TOBACCO, method = "kendall")

        Kendall's rank correlation tau

data:  ALCOHOL and TOBACCO
T = 37, p-value = 0.1646
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3454545
```

Test der Korrelation



Test der Korrelation

Die vorliegenden Daten haben offensichtlich einen Ausreißer.

In R kann dieser folgendermaßen identifiziert werden:

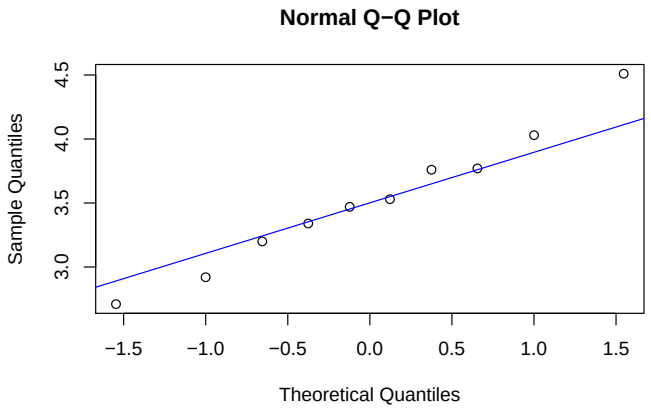
```
R> outlierIndex <- which((ALCOHOL < 5) & (TOBACCO > 4))
R> outlierIndex

[1] 11

R> ALCTOBAC[outlierIndex, ]

            REGION ALCOHOL TOBACCO
11 NorthernIreland    4.02    4.56
```

Test der Korrelation

```
R> ALCOHOL

 [1] 6.47 6.13 6.19 4.89 5.63 4.52 5.89 4.79 5.27 6.08 4.02

R> ALCOHOL < 5

 [1] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE

R> which(ALCOHOL < 5)

[1]  4  6  8 11

R> (ALCOHOL < 5) & (TOBACCO > 4)

 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE

R> which((ALCOHOL < 5) & (TOBACCO > 4))

[1] 11
```

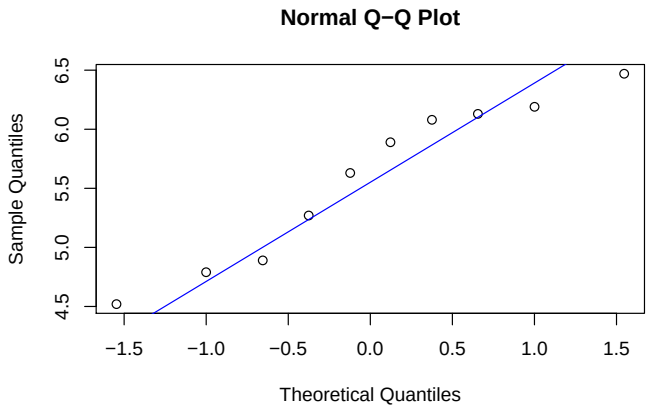
Test der Korrelation

```
R> qqnorm(TOBACCO[-outlierIndex])
R> qqline(TOBACCO[-outlierIndex], col = 4)
```



Test der Korrelation

```
R> qqnorm(ALCOHOL[-outlierIndex])
R> qqline(ALCOHOL[-outlierIndex], col = 4)
```



Test der Korrelation

Ohne Nordirland

```
R> cor.test(TOBACCO[-outlierIndex], ALCOHOL[-outlierIndex])

        Pearson's product-moment correlation

data:  TOBACCO[-outlierIndex] and ALCOHOL[-outlierIndex]
t = 3.5756, df = 8, p-value = 0.007234
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3055382 0.9465163
sample estimates:
      cor
0.7842873
```

Test der Mittelwertdifferenz

Frage: Unterscheiden sich die Mittelwerte zweier Merkmale, die an derselben Beobachtungseinheit erhoben wurden?

Beispiel: Einschätzung der Attraktivität im Laufe eines Abends bei Alkoholkonsum.
Wird die Attraktivität nach Konsum von Alkohol höher eingeschätzt?

Test der Mittelwertdifferenz

```
R> ALCATTR <- read.table("alcattrneu.tab", header = TRUE)
R> dim(ALCATTR)
```

```
[1] 180   2
```

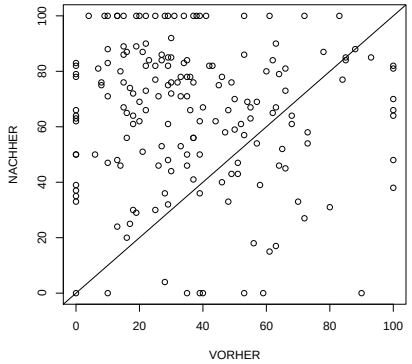
```
R> summary(ALCATTR)
```

```
     VORHER          NACHHER
 Min.   :  0.00   Min.   :  0.00
 1st Qu.: 18.00   1st Qu.: 48.00
 Median : 33.00   Median : 69.00
 Mean   : 37.29   Mean   : 64.54
 3rd Qu.: 53.25   3rd Qu.: 83.00
 Max.   :100.00   Max.   :100.00
```

```
R> attach(ALCATTR)
```

Test der Mittelwertdifferenz

```
R> plot(VORHER, NACHHER)
R> abline(0, 1)
```



Test der Mittelwertdifferenz

Bei diesen sogenannten *gepaarten* Beobachtungen geht man einfach zu Differenzen über und führe dann die Methoden wie im Einstichprobenfall durch.

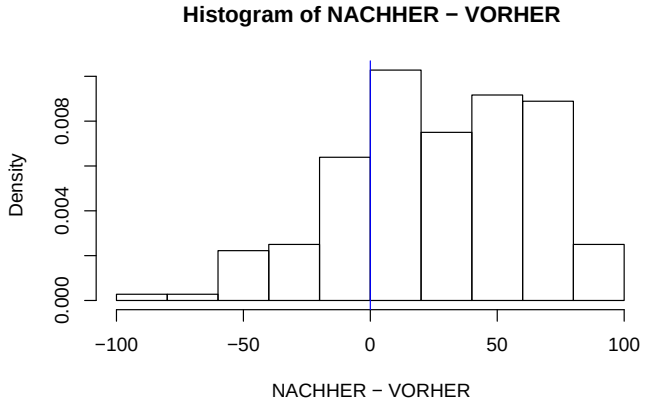
Betrachte hier also den Attraktivitätszugewinn.

```
R> summary(NACHHER - VORHER)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -90.00    4.00   32.00   27.25   57.00   96.00
```

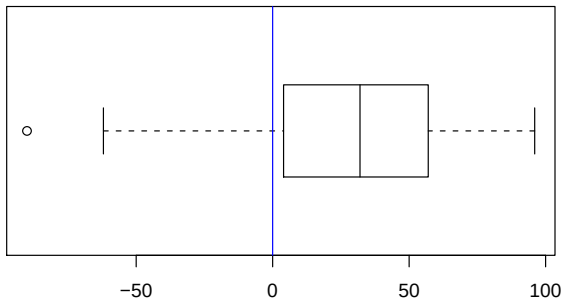
Test der Mittelwertdifferenz

```
R> hist(NACHHER - VORHER, prob = TRUE)
R> abline(v = 0, col = 4)
```



Test der Mittelwertdifferenz

```
R> boxplot(NACHHER - VORHER, horizontal = TRUE)
R> abline(v = 0, col = 4)
```



Test der Mittelwertdifferenz

```
R> t.test(NACHHER - VORHER, alternative = "greater")

        One Sample t-test

data:  NACHHER - VORHER
t = 9.7997, df = 179, p-value = < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 22.65239      Inf
sample estimates:
mean of x
    27.25
```

Test der Mittelwertdifferenz

```
R> t.test(NACHHER, VORHER, paired = TRUE, alternative = "greater")

        Paired t-test

data:  NACHHER and VORHER
t = 9.7997, df = 179, p-value = < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 22.65239      Inf
sample estimates:
mean of the differences
                  27.25
```

Test der Mittelwertdifferenz

```
R> wilcox.test(NACHHER, VORHER, paired = TRUE, alternative = "greater")

        Wilcoxon signed rank test with continuity correction

data:  NACHHER and VORHER
V = 13348, p-value = 5.551e-16
alternative hypothesis: true mu is greater than 0
```