Methoden empirischer Sozialforschung

Kategoriale Merkmale

1 kategoriales Merkmal

- Numerische und grafische Beschreibung
- Kommen alle Kategorien gleich häufig vor?
  χ^2-*Test auf Gleichverteilung*
- Entsprechen die Prozentsaätze in verschiedenen Kategorien aus einer Stichprobe den tatsächlichen Prozentsätzen in der Population?
  χ^2-*Test auf vorgegebene Verteilung*
- In welchen Grenzen kann den unbekannten Anteil einer Population erwarten?
  Konfidenzintervall für relative Häufigkeiten

2 kategoriale Merkmale

- Numerische und grafische Beschreibung
- Ist die Verteilung von Häufigkeiten in verschiedenen Gruppen gleich?
  χ^2-*Test auf Homogenität*
- Sind zwei kategoriale Merkmale voneinander unabhängig?
  χ^2-*Test auf Unabhängigkeit*
- Unterscheiden sich Chancen (Odds)?
  Fisher-Test und Konfidenzintervalle für Odds Ratios

Kategoriale Merkmale in R

- Kategoriale Variablen werden in R als Objekte der Klasse `"factor"` gespeichert und können durch `factor()` oder `as.factor()` kreiert werden.
- Häufigkeitstabellen werden von `summary()` und (besser) `table()` erstellt.

Kategoriale Merkmale in R

```
R> SEX <- c(1, 1, 1, 0, 0, 1, 0, 1, 0, 0)
R> SEX <- factor(SEX, levels = c(0, 1), labels = c("m", "f"))
R> class(SEX)

[1] "factor"

R> SEX

 [1] f f f m m f m f m m
Levels: m f

R> summary(SEX)

m f
5 5

R> table(SEX)

SEX
m f
5 5
```

Kategoriale Merkmale in R

```
R> GRADES <- ordered(c(1, 5, 3, 3, 3, 4, 4, 5), levels = 1:5, labels = c("A",
+     "B", "C", "D", "F"))
R> class(GRADES)

[1] "ordered" "factor"

R> GRADES

[1] A F C C C D D F
Levels: A < B < C < D < F

R> table(GRADES)

GRADES
A B C D F
1 0 3 2 2
```

Beispiel: Werbung

Problem: Vergleich von Host Selling Commercials mit Announcer Commercials

- Announcer Commercials: übliche Werbung,
- Host Selling Commercials: bekannter Showmaster bzw. TV-Persönlichkeit macht Werbung für ein Produkt,
- Studie an Kindern sollte untersuchen, ob Showmaster Werbung effektiver ist.

Fragestellung: Ist Werbung mit Showmaster effektiver als normale Werbung?

Beispiel: Werbung

Studie: 2 Gruppen von jeweils 121 Kindern (6-10 Jahre sehen eine Sendung mit Werbepausen.

Beworbenes Produkt: Canaray Crunch Frühstücksflocken

Drei kategoriale Variablen: `WERBUNG`, `RECALL`, `CEREAL`

`WERBUNG` gibt die Werbegruppe an:

- `SW` – Werbung durch Showmaster
- `NW` – normale Werbung

Beispiel: Werbung

RECALL gibt an, wieviel von der Werbung sie sich gemerkt haben (maximal 10 Punkte).

CEREAL gibt an, welches von vier Produkten sich die Kinder als Gratispackung aussuchen:

- BB – Boo Berries
- CC – Canary Crunch (beworben)
- FL – Fruit Loops
- KH – Kangaroo Hoops

Beispiel: Werbung

Daten (Ausschnitt):

```
R> CMMRCIAL <- read.csv2("cmmrcial.csv")
R> CMMRCIAL
```

```
  RECALL CEREAL WERBUNG
1      6     FL      SW
2      9     CC      SW
3      7     KH      SW
4      7     CC      SW
5      9     BB      SW
...
```

Beispiel: Werbung

```
R> names(CMMRCIAL)
```

```
[1] "RECALL"  "CEREAL"  "WERBUNG"
```

```
R> summary(CMMRCIAL)
```

```
     RECALL       CEREAL  WERBUNG
 Min.   : 0.000   BB:64   NW:121
 1st Qu.: 6.000   CC:81   SW:121
 Median : 8.000   FL:45
 Mean   : 7.364   KH:52
 3rd Qu.:10.000
 Max.   :10.000
```

```
R> attach(CMMRCIAL)
```

Beispiel: Werbung

Fragen:

- Wählen die Kinder das beworbene Produkt häufiger als die anderen Produkte? War die Werbung erfolgreich?
- Merken sich die Kinder mehr Details eines Produkts, wenn es von einem Showmaster beworben wird? *(→ später)*
- Würden die Kinder das beworbene Produkt häufiger wählen, wenn es von einem Showmaster präsentiert wird verglichen mit normaler Werbung?

1 kategoriales Merkmal

Numerische Beschreibung

- absoluten Häufigkeiten werden durch `table` berechnet
- relative Häufigkeiten erhält man durch Division durch die Anzahl der Beobachtungen (also die *Länge* des Datenvektors oder die Summer der absoluten Häufigkeiten)
- Prozentwerte durch Multiplikation mit 100

Numerische Beschreibung

```
R> table(CEREAL)

CEREAL
BB CC FL KH
64 81 45 52

R> table(CEREAL)/length(CEREAL)

CEREAL
       BB        CC        FL        KH
0.2644628 0.3347107 0.1859504 0.2148760
```

Numerische Beschreibung

```
R> absH <- table(CEREAL)
R> relH <- absH/length(CEREAL)
R> cbind(absH = absH, relH = relH)

   absH      relH
BB   64 0.2644628
CC   81 0.3347107
FL   45 0.1859504
KH   52 0.2148760
```

Numerische Beschreibung

Manchmal hat man nicht die Rohdaten, sondern schon vortabellierte Daten:

```
R> absH <- as.table(c(keine = 2310, gering = 3783, hoch = 4397,
+     "sehr hoch" = 844))
R> absH
```

```
    keine    gering      hoch sehr hoch
     2310      3783      4397       844
```

Numerische Beschreibung

Relativen Häufigkeiten:

```
R> relH <- absH/sum(absH)
R> relH
```

```
     keine     gering       hoch  sehr hoch
0.20381154 0.33377448 0.38794777 0.07446621
```

Kumulierte relative Häufigkeiten durch Teilsummen:

```
R> cumsum(relH)
```

```
[1] 0.2038115 0.5375860 0.9255338 1.0000000
```

Numerische Beschreibung

Alles übersichtlich in einer Tabelle:

```
R> cbind("abs H" = absH, "rel H" = relH, "kum rel H" = cumsum(relH))
```

```
          abs H      rel H kum rel H
keine      2310 0.20381154 0.2038115
gering     3783 0.33377448 0.5375860
hoch       4397 0.38794777 0.9255338
sehr hoch   844 0.07446621 1.0000000
```

Grafische Beschreibung

- Balkendiagramm

  ```
  plot(factorobj, ...)
  barplot(tableobj, ...)
  ```

  nur mit vertikalen Linien (Stabdiagramm)

  ```
  plot(tableobj, ...)
  ```

- Sektorendiagramm (Tortendiagramm)

  ```
  pie(tableobj, ...)
  ```

Grafische Beschreibung

Faktor:

```
R> CEREAL

 [1] FL CC KH CC BB BB FL FL BB KH
[11] KH CC BB CC FL ...
```

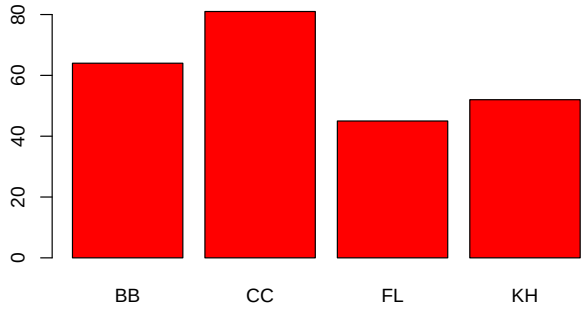
Tabelle:

```
R> tab <- table(CEREAL)
R> tab

CEREAL
BB CC FL KH
64 81 45 52
```

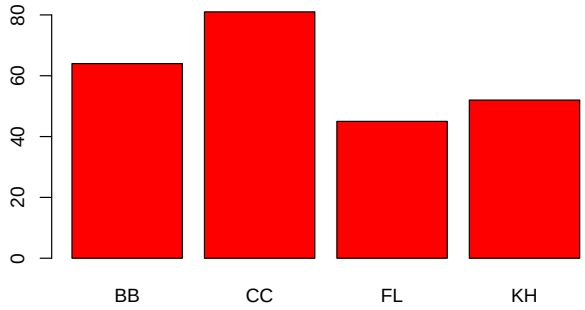
Grafische Beschreibung

```
R> plot(CEREAL)
```



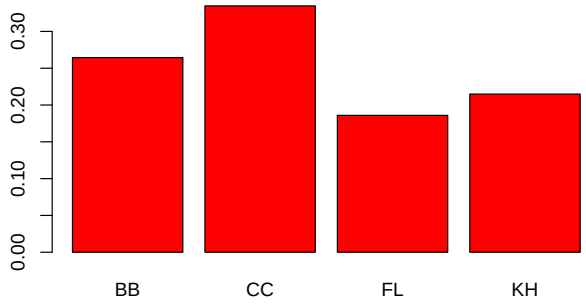
Grafische Beschreibung

```
R> barplot(tab)
```



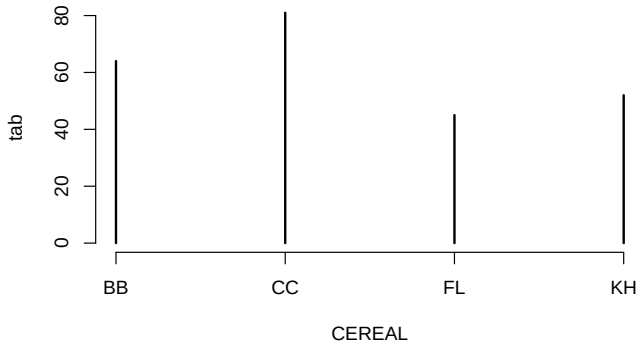
Grafische Beschreibung

```
R> barplot(tab/sum(tab))
```



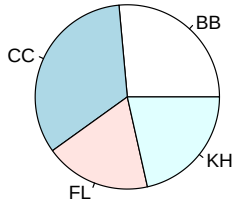
Grafische Beschreibung

```
R> plot(tab)
```



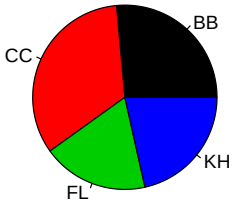
Grafische Beschreibung

```
R> pie(tab)
```



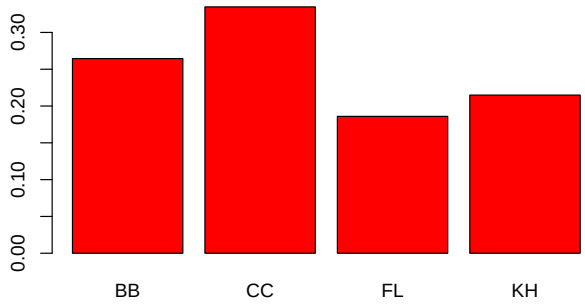
Grafische Beschreibung

```
R> pie(tab, col = 1:4)
```



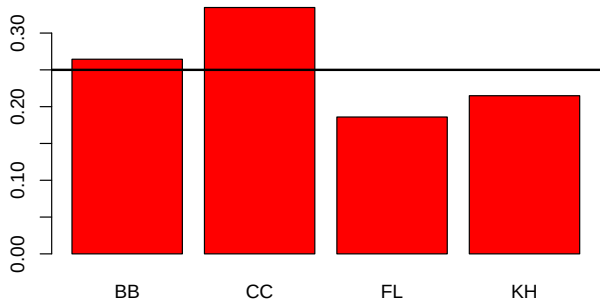
Test auf Gleichverteilung

Kommen alle Kategorien gleich häufig vor?



Test auf Gleichverteilung

Kommen alle Kategorien gleich häufig vor?



Test auf Gleichverteilung

Ab wann glauben wir nicht mehr an eine Gleichverteilung?

	BB	CC	FL	KH	gleichverteilt?
Experiment 1	61	60	60	61	ja
Experiment 2	57	62	65	58	ja?
Experiment 3	0	242	0	0	nein
Experiment 4	12	188	25	17	nein?
Experiment 5	64	81	45	52	nein?

Test auf Gleichverteilung

Ab wann glauben wir nicht mehr an eine Gleichverteilung?

Zur Beantwortung brauchen wir eine Methode, um das Ergebnis eines solchen Experiments beurteilen zu können: *statistischer Test*

Es kommen zwei Antworten in Frage:

1. die Häufigkeiten sind gleichverteilt
2. die Häufigkeiten sind **nicht** gleichverteilt

1. nennt man auch *Nullhypothese*, die hier besagt: alle Kategorien kommen mit gleicher Wahrscheinlichkeit vor, nämlich mit 1/4.

2. nennt man auch *Alternative*, die hier besagt: mindestens eine Kategorie kommt mit höherer Wahrscheinlichkeit vor.

Test auf Gleichverteilung

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 auf Gleichverteilung

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*.

Test auf Gleichverteilung

Beobachtete Werte: observed

$$o = (64, 81, 45, 52)$$

Erwartete Werte: Welche Häufigkeiten würden wir bei einer Gleichverteilung erwarten? Jede Kategorie käme mit Wahrscheinlichkeit 0.25 vor, d.h. bei 242 Beobachtungen würden wir in jeder Kategorie $242 \cdot 0.25 = 60.5$ Beobachtungen erwarten.

$$e = (60.5, 60.5, 60.5, 60.5)$$

Test auf Gleichverteilung

Diskrepanz zwischen kann durch die folgende Teststatistik eingefangen werden (Pearson's X^2).

$$X^2 = \sum_{j=1}^{J} \frac{(o_j - e_j)^2}{e_j}$$

Unter der Nullhypothese ist diese Statistik approximativ χ^2-verteilt mit $J - 1$ Freiheitsgraden. Durch Vergleich mit dieser Verteilung wird der p-Wert berechnet.

Test auf Gleichverteilung

Mittels der Funktion `chisq.test` kann man verschiedene χ^2-Tests für Häufigkeitstabellen ausführen.

Der *χ^2-Test auf Gleichverteilung* ist der einfachste Fall: hier muss als einziges Argument die Tabelle der absoluten Häufigkeiten (aber *nicht* die Rohdaten) spezifiziert werden:

```
chisq.test(tableobj, ...)
```

Test auf Gleichverteilung

```
R> chisq.test(tab)

        Chi-squared test for given probabilities

data:  tab
X-squared = 12.314, df = 3, p-value = 0.006381
```

Der p-Wert 0.006 ist kleiner als 0.05, daher verwerfen wir die Nullhypothese. Die Kategorien kommen nicht gleichhäufig vor, d.h. die Werbung hat einen Effekt gehabt.

Test auf Gleichverteilung

Mit dem Ergebnis von `chisq.test` kann man auch weiter rechnen:

```
R> testErg <- chisq.test(table(CEREAL))
R> typeof(testErg)

[1] "list"

R> class(testErg)

[1] "htest"

R> names(testErg)

[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals"
```

Test auf Gleichverteilung

```
R> rbind("beobachtete H" = testErg$observed, "erwartete H" = testErg$expected)

                BB   CC   FL   KH
beobachtete H 64.0 81.0 45.0 52.0
erwartete H   60.5 60.5 60.5 60.5
```

Test auf Gleichverteilung

```
R> beob <- table(CEREAL)
R> erw <- sum(beob) * rep(1/4, 4)
R> stat <- (beob - erw)^2/erw
R> stat

CEREAL
       BB        CC        FL        KH
0.2024793 6.9462810 3.9710744 1.1942149

R> stat <- sum(stat)
R> stat

[1] 12.31405
```

Test auf Gleichverteilung

Für die Berechnung von kritischen Werten ersetzt die Funktion `qchisq` das Nachschlagen in der Tabelle der χ^2-Verteilung:

```
R> qchisq(0.95, df = 3)
```

```
[1] 7.814728
```

Dieser ist kleiner als der Wert der Teststatistik 12.314.

Test auf Verteilung

Entsprechen die Häufigkeiten bestimmten Vorgaben?

Repräsentativität einer Meinungsumfrage: Häufigkeiten von Familienarten in der Stichprobe und in Österreich (Population)

	Stichprobe		Österreich
Familienart	absolut	relativ	relativ
Lebensgem. kinderlos	23	0.115	0.358
Lebensgem. mit Kindern	133	0.665	0.517
alleinerz. Mütter	39	0.195	0.108
alleinerz. Väter	5	0.025	0.017

Test auf Verteilung

Benutze dieselbe Idee wie vorher: berechne die erwarteten Häufigkeiten, hier

$$e = (71.6, 103.4, 21.6, 3.4)$$

und damit dann die X^2-Statistik.

Zur Durchführung in R wendet man wieder `chisq.test` auf die beobachteten Häufigkeiten an und spezifiziert die erwarteten relativen Häufigkeiten mittels des Argumentes `p`.

Test auf Verteilung

```
R> absH <- as.table(c(LGo = 23, LGm = 133, Mu = 39, Va = 5))
R> absH
```

```
LGo LGm  Mu  Va
 23 133  39   5
```

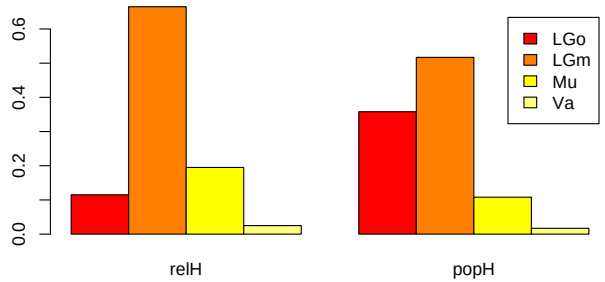
```
R> popH <- c(0.358, 0.517, 0.108, 0.017)
R> chisq.test(absH, p = popH)
```

```
        Chi-squared test for given probabilities

data:  absH
X-squared = 56.2314, df = 3, p-value = 3.749e-12
```

Test auf Verteilung

```
R> relH <- absH/sum(absH)
R> barplot(cbind(relH, popH), beside = TRUE, legend = TRUE)
```



Test von Anteilen

Mit der Funktion `binom.test` kann man Inferenz über Anteile treiben: sie liefert sowohl Konfidenzintervalle als auch Tests von Hypothesen.

Beispiel:

- Weicht der Anteil von Canary Crunch signifikant von 0.25 ab?
- In welchem Bereich ist der Anteil von Canary Crunch zu erwarten?

Test von Anteilen

```
R> cc <- CEREAL
R> levels(cc)

[1] "BB" "CC" "FL" "KH"

R> levels(cc)[-2] <- "nicht CC"
R> tab <- table(cc)
R> tab

cc
nicht CC       CC
     161       81
```

Test von Anteilen

Approximativer Test:

```
R> chisq.test(tab, p = c(0.75, 0.25))

        Chi-squared test for given probabilities

data:  tab
X-squared = 9.2617, df = 1, p-value = 0.00234
```

Test von Anteilen

Exakter Test:

```
R> binom.test(81, 242, p = 0.25)

        Exact binomial test

data:  81 and 242
number of successes = 81, number of trials = 242, p-value = 0.002952
alternative hypothesis: true probability of success is not equal to 0.25
95 percent confidence interval:
 0.2755460 0.3979829
sample estimates:
probability of success
             0.3347107
```

Test von Anteilen

99% Konfidenzintervall für den Anteil von Canary Crunch:

```
R> binom.test(81, 242, p = 0.25, conf.level = 0.99)$conf.int

[1] 0.2584569 0.4176306
attr(,"conf.level")
[1] 0.99
```

2 kategoriale Merkmale

Numerische Beschreibung

Kategoriale Information wird typischerweise in Kontingenztafeln zusammengefaßt, der Kreuzklassifikation der Variablen.

In R mit

```
table(a, b)
```

oder mit

```
xtabs(~ a + b)
```

Numerische Beschreibung

```
R> tab <- table(CEREAL, WERBUNG)
R> tab
```

```
      WERBUNG
CEREAL NW SW
    BB 31 33
    CC 37 44
    FL 26 19
    KH 27 25
```

Numerische Beschreibung

```
R> xtabs(~CEREAL + WERBUNG, data = CMMRCIAL)
```

```
      WERBUNG
CEREAL NW SW
    BB 31 33
    CC 37 44
    FL 26 19
    KH 27 25
```

Numerische Beschreibung

Die Randinformation erzeugt man aus Tabellen durch Bildung von Zeilen- beziehungsweise Spaltensummen.

In R am flexibelsten mit Hilfe der Funktion `apply` (mit der man beliebigdimensionale Ränder von beliebigdimensionalen Feldern berechnen kann):

```
R> apply(tab, 1, sum)
```

```
BB CC FL KH
64 81 45 52
```

```
R> apply(tab, 2, sum)
```

```
 NW  SW
121 121
```

Numerische Beschreibung

Für Zeilen- und Spaltensummen von Matrizen (d.h. 2-dimensionalen Feldern) gibt es auch die Bequemlichkeitsfunktionen `rowSums` und `colSums`:

```
R> rowSums(tab)
```

```
BB CC FL KH
64 81 45 52
```

```
R> colSums(tab)
```

```
 NW  SW
121 121
```

Numerische Beschreibung

Zur Berechnung der bedingten Information muss man die Elemente in den jeweiligen Zeilen (Spalten) durch die entsprechenden Zeilensummen (Spaltensummen) teilen. Auch hier gibt es eine allgemeine Funktion, `sweep`. Zeilenhäufigkeiten:

```
R> rowProp <- sweep(tab, 1, apply(tab, 1, sum), "/")
R> rowProp
```

```
      WERBUNG
CEREAL NW        SW
    BB 0.4843750 0.5156250
    CC 0.4567901 0.5432099
    FL 0.5777778 0.4222222
    KH 0.5192308 0.4807692
```

Numerische Beschreibung

Spaltenhäufigkeiten:

```
R> colProp <- sweep(tab, 2, apply(tab, 2, sum), "/")
R> colProp
```

```
      WERBUNG
CEREAL NW        SW
    BB 0.2561983 0.2727273
    CC 0.3057851 0.3636364
    FL 0.2148760 0.1570248
    KH 0.2231405 0.2066116
```

Probe:

```
R> apply(colProp, 2, sum)
```

```
NW SW
 1  1
```

Numerische Beschreibung

Sowohl für die Randinformation als auch für bedingte Information gibt es Bequemlichkeitsfunktionen: `margin.table` und `prop.table`.

```
R> margin.table(tab, 1)
```

```
CEREAL
BB CC FL KH
64 81 45 52
```

```
R> prop.table(tab, 2)
```

```
      WERBUNG
CEREAL NW        SW
    BB 0.2561983 0.2727273
    CC 0.3057851 0.3636364
    FL 0.2148760 0.1570248
    KH 0.2231405 0.2066116
```

Grafische Beschreibung

- Gruppierte Balkendiagramme

  ```
  barplot(tableobj, beside = TRUE, ...)
  ```

- Gestapelte Balkendiagramme

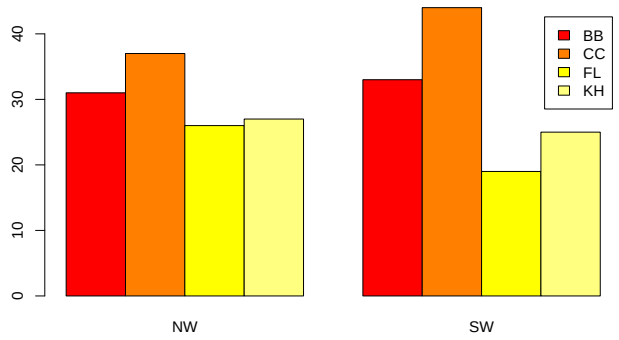
  ```
  barplot(tableobj, ...)
  ```

- Mosaicplots

  ```
  mosaicplot(tableobj, ...)
  ```

Grafische Beschreibung

```
R> barplot(tab, beside = TRUE, legend = TRUE)
```



Grafische Beschreibung

Umsortierung der Levels, so daß die betrachtete Kategorie CC zuerst kommt:

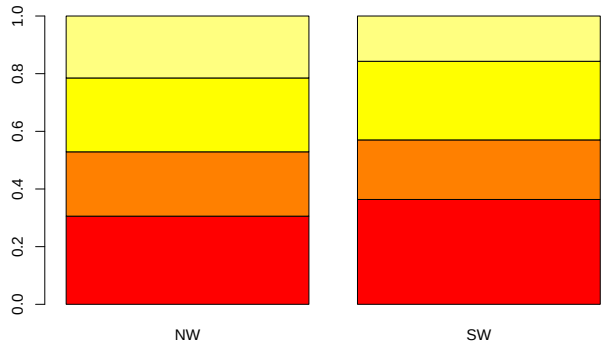
```
R> CEREAL2 <- factor(CEREAL, levels = c("CC", "KH", "BB", "FL"))
R> CEREAL2[1:10]

 [1] FL CC KH CC BB BB FL FL BB KH
Levels: CC KH BB FL
```

```
R> tab <- table(CEREAL2, WERBUNG)
```

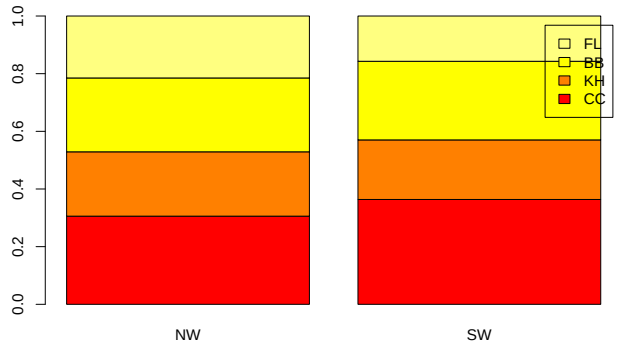
Grafische Beschreibung

```
R> barplot(prop.table(tab, 2))
```



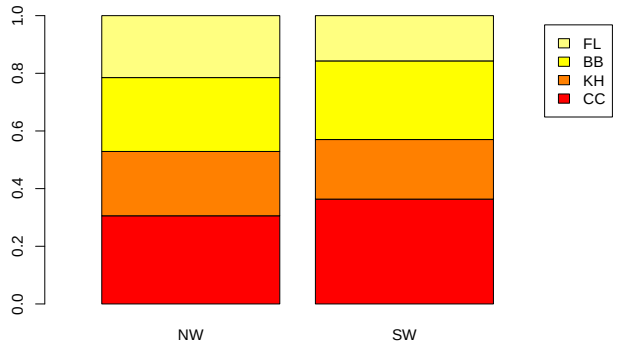
Grafische Beschreibung

```
R> barplot(prop.table(tab, 2), legend = TRUE)
```



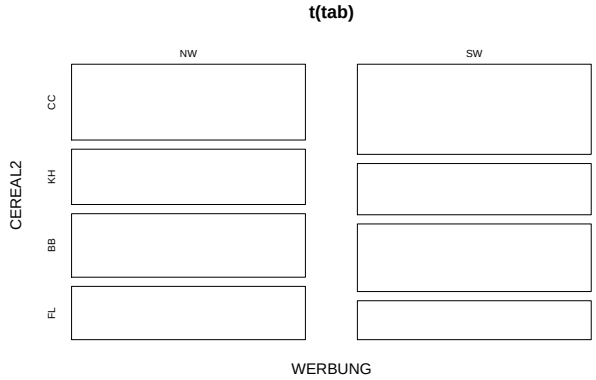
Grafische Beschreibung

```
R> barplot(prop.table(tab, 2), legend = TRUE, xlim = c(0, 3))
```



Grafische Beschreibung

```
R> mosaicplot(t(tab))
```



Test auf Homogenität

Ist die Verteilung von Häufigkeiten in verschiedenen Gruppen gleich?

Beispiel: Wählen Kinder, die die normale Werbung gesehen haben, die Frühstocksflocken mit der gleichen Verteilung, wie die Kinder, die die Showmaster-Werbung gesehen haben?

Wenn dies der Fall ist, dann ist die Verteilung in beiden Gruppen gleich der Randverteilung.

Test auf Homogenität

```
R> rowSums(tab)/sum(tab)

       CC        KH        BB        FL
0.3347107 0.2148760 0.2644628 0.1859504
```

Hieraus kann man wieder erwartete Häufigkeiten e berechnen und diese in einer χ^2-Statistik mit den beobachteten Werten o vergleichen.

Hier haben beide Gruppen (`NW` und `SW`) einen Stichprobenumfang von 121, damit sind die erwarteten Häufigkeiten jeweils:

```
R> rowSums(tab)/sum(tab) * 121

  CC   KH   BB   FL
40.5 26.0 32.0 22.5
```

Test auf Homogenität

Für den χ^2-Homogenitätstest verwendet man die Funktion `chisq.test`, wobei man als Argumente entweder die beiden Vektoren mit den Werten der kategorialen Variablen oder deren Kreuztabelle nimmt. (Ein Bilden von bedingten Häufigkeiten ist nicht erforderlich.)

```
R> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 1.8333, df = 3, p-value = 0.6077

R> chisq.test(tab)$expected

       WERBUNG
CEREAL2   NW   SW
     CC 40.5 40.5
     KH 26.0 26.0
     BB 32.0 32.0
     FL 22.5 22.5
```

Test auf Homogenität

```
R> chisq.test(CEREAL2, WERBUNG)

        Pearson's Chi-squared test

data:  CEREAL2 and WERBUNG
X-squared = 1.8333, df = 3, p-value = 0.6077
```

Test auf Unabhängigkeit

Sind zwei kategoriale Merkmale voneinander unabhängig?

Beispiel: Unabhängigkeit von erwarteten Verkäufen im In- und Ausland.

```
R> SALES <- read.csv2(file = "expectedsales.csv")
R> SALES

  COUNT DOMESTIC  FOREIGN
1    78 Increase Increase
2    88 Increase Decrease
3    49 Decrease Increase
4    37 Decrease Decrease
```

Test auf Unabhängigkeit

Hier sind die Werte vortabelliert, die Kontigenztabelle wird wieder mit `xtabs` erzeugt:

```
R> tab <- xtabs(COUNT ~ FOREIGN + DOMESTIC, data = SALES)
R> tab

          DOMESTIC
FOREIGN    Decrease Increase
  Decrease 37       88
  Increase 49       78
```

Auch hier kann man wieder erwartete Häufigkeiten berechnen: wenn `FOREIGN` voneinander unabhängig sind, dann sind die erwarteten relativen Häufigkeiten das Produkt der Randverteilungen.

Test auf Unabhängigkeit

Relative Häufigkeiten für einen Anstieg bei den Inlands- bzw. Auslandsverkäufen:

```
R> domest <- margin.table(tab, 2)/sum(tab)
R> domest

DOMESTIC
 Decrease  Increase
0.3412698 0.6587302

R> foreign <- margin.table(tab, 1)/sum(tab)
R> foreign

FOREIGN
 Decrease  Increase
0.4960317 0.5039683
```

Die gemeinsamen relativen Häufigkeiten unter Unabhängigkeit kann man dann durch Produktbildung berechnen.

Test auf Unabhängigkeit

Die erwartete absolute Häufigkeit für einen gleichzeitigen Anstieg im In- und Ausland ist unter Unabhängigkeit also:

```
R> domest[1] * foreign[1] * sum(tab)

Decrease
42.65873
```

Analog für die anderen Zellen. Dies wird wieder per `chisq.test` durchgeführt.

Test auf Unabhängigkeit

```
R> chisq.test(tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 1.8792, df = 1, p-value = 0.1704

R> chisq.test(tab)$expected

          DOMESTIC
FOREIGN    Decrease Increase
  Decrease 42.65873 82.34127
  Increase 43.34127 83.65873
```

Test auf Unabhängigkeit

```
R> chisq.test(tab, correct = FALSE)

        Pearson's Chi-squared test

data:  tab
X-squared = 2.2611, df = 1, p-value = 0.1327
```

In beiden Varianten weist der große p-Wert darauf hin, daß kein Zusammenhang zwischen dem erwarteten Zuwachs an Verkäufen im Ausland (`FOREIGN`) und im Inland (`DOMESTIC`) besteht.

Test auf Unabhängigkeit

Sowohl der Unabhängigkeitstest als auch der Homogenitätstest werden auf dieselbe Art durchgeführt, was ist also der Unterschied?

Bei der Homogenitätshypothese gibt es eine *wenn – dann* Beziehung. Die Wenn-Variable nennt man *erklärende* Variable, die Dann-Variable die *abhängige* Variable.

Bei der Unabhängigkeitshypothese gibt es eine solche Beziehung nicht, beide Variablen sind gleichberechtigt und miteinander vertauschbar.

Test auf Unabhängigkeit

Der χ^2-Test ist wieder ein approximativer Test. Ähnlich wie beim Test für Anteile gibt es auch hier einen exakten Test, den sogenannten Fisher-Test.

In R:

```
R> fisher.test(tab)

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.1453
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.382347 1.168422
sample estimates:
odds ratio
 0.6703796
```

Test auf Unabhängigkeit

Für 2×2 Tabellen gibt `fisher.test` ein Assoziationsmaß mit aus, das *Chancenverhältnis* oder *Odds Ratio*.

Hier: die Chance auf eine Abnahme der Inlandsverkäufe ist 0.67 mal höher bei einer Abnahme der Auslandsverkäufe. In diesem Fall weicht er aber nicht signifikant von 1 ab. Die Chancen sind also für beide Gruppen gleich.