Statistik

Vorlesung Statistik 1

K.Gerald van den Boogaart
http://www.stat.boogaart.de/

Organisation

- Webseite (Folien, Skript, Probeklausuren, Organisation)

Organisation

- Webseite (Folien, Skript, Probeklausuren, Organisation)
- Übungen

Organisation

- Webseite (Folien, Skript, Probeklausuren, Organisation)
- Übungen
- Klausur (Anmeldung 1+2, Hilfsmittel)

Inhalt heute (Grundlagen)

- **Was ist Statistik?**
- **Grundmodelle der Statistik**
- **Datenmatrix**
  Wie werden Daten gespeichert?

Was ist Statistik?

Wortwurzel: Aufstellungen (lat. stare)
Bedeutungen:

- **Datensammlung**
  des Staats (ursprüngliche Bedeutung)

Was ist Statistik?

Wortwurzel: Aufstellungen (lat. stare)
Bedeutungen:

- **Datensammlung**
  des Staats (ursprüngliche Bedeutung)
- **Wissenschaft**
  von der Auswertung von Daten/vom Schließen aus Daten

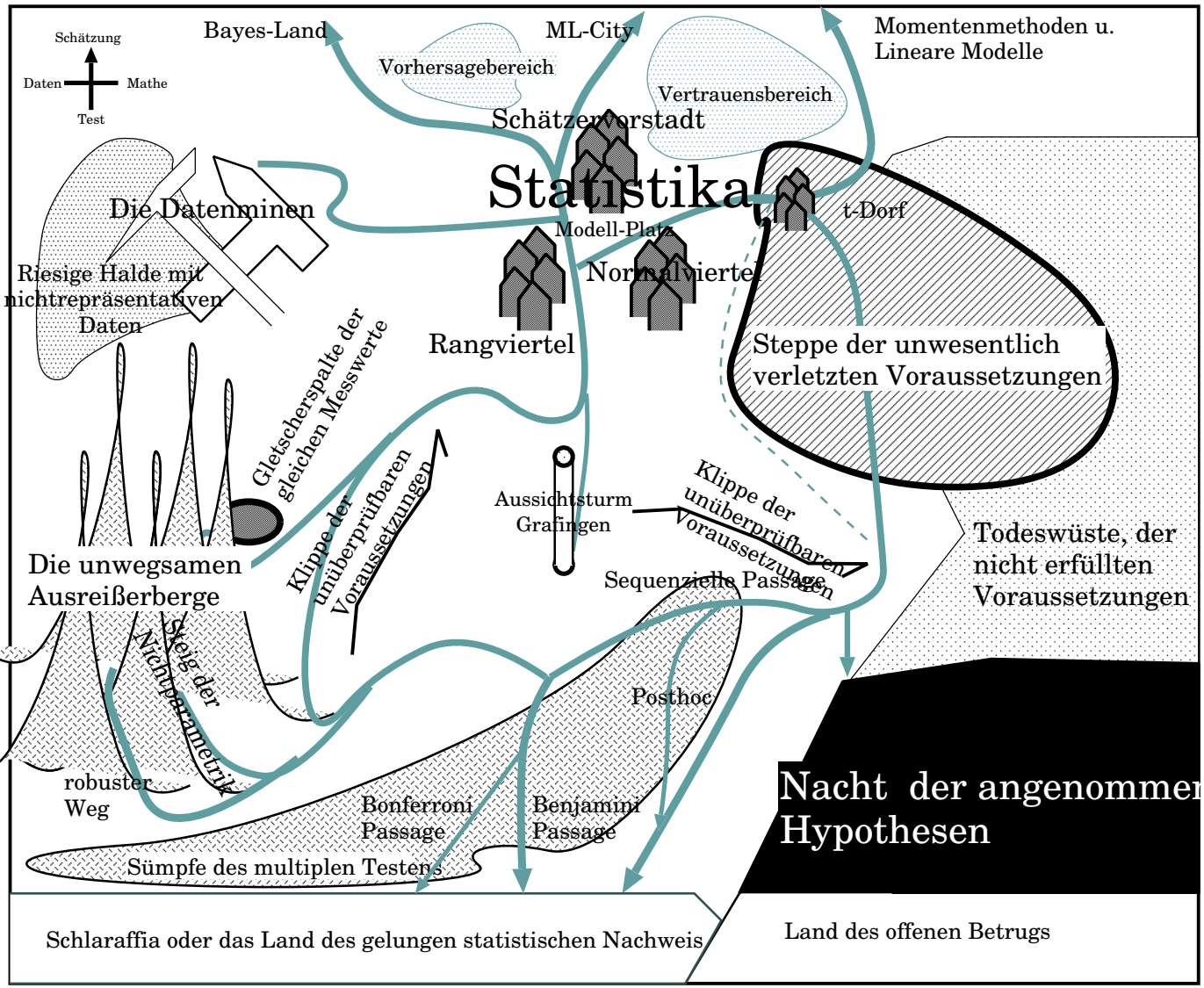
Was ist Statistik?

Wortwurzel: Aufstellungen (lat. stare)
Bedeutungen:

- **Datensammlung**
  des Staats (ursprüngliche Bedeutung)
- **Wissenschaft**
  von der Auswertung von Daten/vom Schließen aus Daten
- Aus beobachteten Zufallsvariablen berechnete weitere **Zufallsvariablen** (z.B. der Mittelwert)

$$\bar{X} = \frac{1}{n}\left(X_1 + \ldots + X_n\right)$$

Die Landkarte der Vorlesung



Repräsentative Daten

Nur mit repräsentativen Daten kann man statistische Schlüsse ziehen.

Grundbegriffe

- **Grundgesamtheit**
- **statistisches Individuen**
- **Stichprobe**
  Die Menge der statistische Individuen zu den man tatsächlich Daten erhebt.
  z.B. die Menge der Befragten Wahlberechtigten

Grundbegriffe

- **Grundgesamtheit**
- **statistisches Individuen**
- **Stichprobe**
- **repräsentativ**
  Eine Stichprobe heißt repräsentativ, wenn jedes statistische Individuum aus der Grundgesamtheit unabhängig von allen anderen mit der gleichen Wahrscheinlichkeit in die Stichprobe gelangen kann.

Grundbegriffe

- **Grundgesamtheit**
- **statistisches Individuen**
- **Stichprobe**
- **repräsentativ**
- **Zufallsvariable**
  das beim i-ten statistischen Individum beobachtete Merkmal (z.B. bevorzugte Partei) wird durch eine Zufallsvariable X_i modeliert.

Beispiel: Bodenqualität

- **Grundgesamtheit:** Alle Punkte des Bodens im Untersuchungsgebiet.
- **Stichprobe:** Zufällig ausgewählte Untersuchungspunkte.
- **Zufallsvariablen:** Nährstoffgehalt in an diesen Stellen genommenen Bodenproben.
- **Realisierungen:** $5.34\%, 7, 45\%, \ldots$

Beispiel: Werkstückprüfung

- **Grundgesamtheit:** Alle gefertigten Zahnräder der Teilenummer 45632N.
- **Stichprobe:** Zufällig zu Testzwecken entnommen Zahnräder.
- **Zufallsvariablen:** Betriebstunden im Testbetrieb bis Defekt.
- **Realisierungen:** $5343h, 7342h, \ldots$

Vollerhebung

- Die **Vollerhebung** ist eine spezielle Art der Stichprobennahme.
- Bei Vollerhebung ist die Stichprobe gleich der Grundgesamtheit.
- Unabhängigkeit: alle kommen unabhängig von allen anderen sicher in die Stichprobe.
- gleiche Wahrscheinlichkeit: Wahrscheinlichkeit in die Stichprobe zu kommen ist 1.

Grundbegriffe

- **Vorschrift für ein Zufallsexpriment**
- **Zufallsexperiment**
- **identisch verteilt**
- **unabhängig**
- **repräsentativ**
- **Zufallsvariable**
  die beim i-ten Zufallsexpriment zu machende Beobachtung X_i.
  z.B. die Bruchspannung.

Fadenbrüche

Anzahl Fadenbrüche bei verschiedenen Rahmenbedingungen:

```
> warpbreaks
```

	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L
7	51	A	L
8	26	A	L
9	67	A	L
10	18	A	M
11	21	A	M

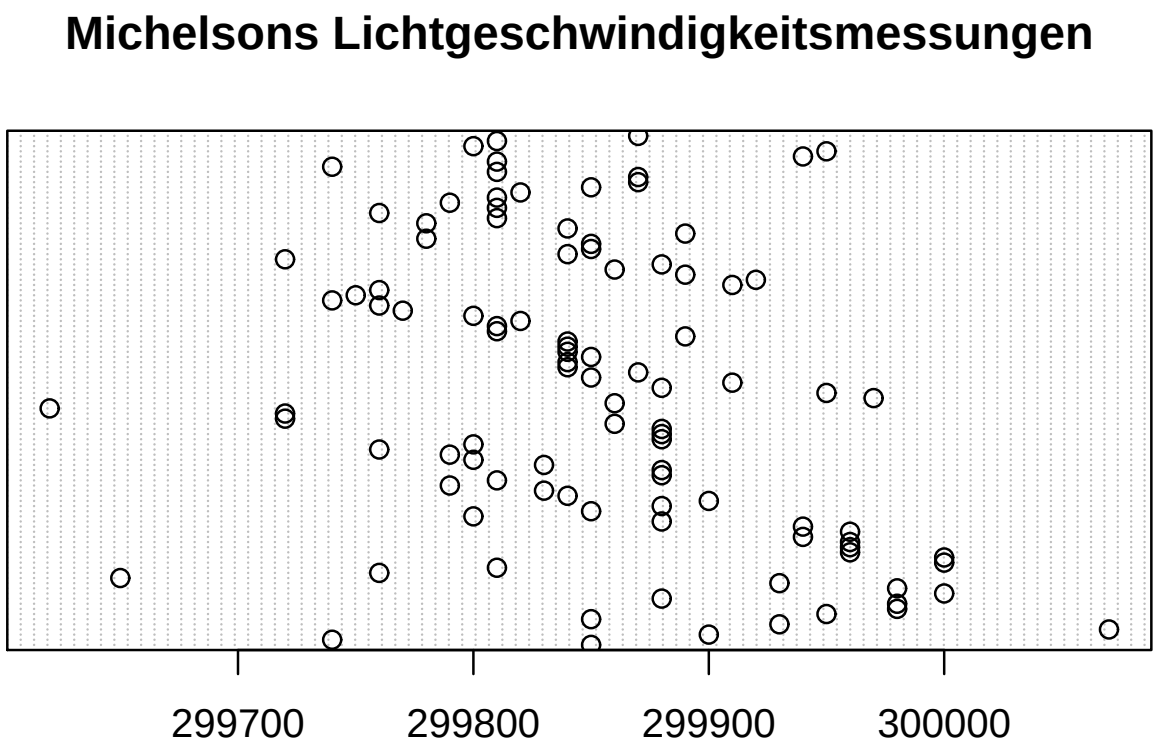
Beispiel: Lichtgeschwindigkeitsmessungen

```
> lightspeeds

  [1] 299850 299740 299900 300070 299930 299850 299950 299980
  [9] 299980 299880 300000 299980 299930 299650 299760 299810
 [17] 300000 300000 299960 299960 299960 299940 299960 299940
 [25] 299880 299800 299850 299880 299900 299840 299830 299790
 [33] 299810 299880 299880 299830 299800 299790 299760 299800
 [41] 299880 299880 299880 299860 299720 299720 299620 299860
 [49] 299970 299950 299880 299910 299850 299870 299840 299840
 [57] 299850 299840 299840 299840 299890 299810 299810 299820
 [65] 299800 299770 299760 299740 299750 299760 299910 299920
 [73] 299890 299860 299880 299720 299840 299850 299850 299780
 [81] 299890 299840 299780 299810 299760 299810 299790 299810
 [89] 299820 299850 299870 299870 299810 299740 299810 299940
 [97] 299950 299800 299810 299870
```

Beispiel: Lichtgeschwindigkeitsmessungen

```
> dotchart(lightspeeds, main = "Michelsons Lichtgeschw
```



Repräsentativität

Allgemein (resultierende Zufallsvariablen)

- identisch verteilt
- stochastisch unabhängig

Stichproben (zufällige Auswahl)

- mit der gleichen Wahrscheinlichkeit
- unabhängig voneinander

Zufallsexperimente (Experiment mit zufälligem Ausgang)

- nach gleicher Vorschrift durchgeführt
- unabhängig voneinander

Mehrstichprobenmodell

Oft finden wir in einem Datensatz **zwei oder mehrer Gruppen** von Daten, die von unterschiedlichen

- Grundgesamtheit oder
- Zufallsexperimenten (Experimentiervorschriften)

herrühren.
Ein Datensatz kann also mehrer Stichproben enthalten.
Man spricht dann von einer **Zweistichproben- oder Mehrstichprobensituation**.

Zufälligkeit der Daten

Ein repräsentativer Datensatz ist grundsätzlich zufällig, da

- die Auswahl der Beobachtungen zufällig zustandegekommen ist, oder
- die Experimente zufällige Ergebnisse haben.

Wir interessieren uns aber nicht für die konkreten Daten, sondern für die dahinterstehenden Gesetzte: z.B. für die Zahlräder, die tatsächlich ausgeliefert werden, was alle Deutschen wählen, oder welche Maschieneneinstellung in Zukunft die besten Ergebnisse liefert.

Zufälligkeit der Kenngrößen

Das erste Ergebnis einer statistischen Analyse sind oft Kenngrößen, wie z.B. der Mittelwert.

- **Der Mittelwert als Zufallsvariable und Statistik**

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$

Der Mittelwert ist selbst **zufällig**!!!.

- **Der Mittelwert als abstrakte Realisierung**

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

- **Der realisierte Mittelwert**

```
[1] 299852.4
```

Die Datenmatrix

- $X_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, m$ sind die Einträge einer Datenmatrix.
- Jede Zeile $X_{i\cdot}$ gehört zu einem statistischen Individuum

Die Datenmatrix

- $X_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, m$ sind die Einträge einer Datenmatrix.
- Jede Zeile $X_{i.}$ gehört zu einem statistischen Individuum
- Jede Spalte $X_{.j}$ gehört zu einem Merkmal

Die Datenmatrix

- $X_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, m$ sind die Einträge einer Datenmatrix.
- Jede Zeile $X_{i.}$ gehört zu einem statistischen Individuum
- Jede Spalte $X_{.j}$ gehört zu einem Merkmal
- Der Eintrag X_{ij} entspricht der Ausprägung des j-ten Merkmals am i-ten Individuum.

Die Datenmatrix

- $X_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, m$ sind die Einträge einer Datenmatrix.
- Jede Zeile $X_{i.}$ gehört zu einem statistischen Individuum
- Jede Spalte $X_{.j}$ gehört zu einem Merkmal
- Der Eintrag X_{ij} entspricht der Ausprägung des j-ten Merkmals am i-ten Individuum.
- Die Einträge einer Datenmatrix sind Zufallsvariablen bzw. ihre Realisierungen.

Diskrete Skalen

- **nominal**
- **kategoriell**
  (gr. katagorein = unterordnen, einordnen)
  Kategorien, jedes statistische Individuum wird in eine von ein paar vor dem Expriment feststehende Kategorien eingeordent.

Diskrete Skalen

- **nominal**
- **kategoriell**
- **ordinal**
  (angeordnet)
  wie kategoriell, nur dass die Kategorien in eine natürliche Reihenfolge gebracht werden können.
  z.B. tot, krank, gesund

Die diskreten Skalen

	Name	Geschlecht	Fach	Stufe	Note	Kinder
1	Maier	m	Chemie	Abi	4	0
2	Huber	w	Biologie	Vordiplom	1	1
3	Mueller	m	Geographie	Hauptdiplom	2	4

Stetige Skalen

- **reell**
  (reelle Zahlen)
  Jede beliebige reelle Zahl kann vorkommen. $+, -, *$ sind sinnvolle Operationen. Der Abstand von 10 zu 5 ist genauso groß wie der Abstand von 5 zu 0.
  z.B. Temperaturänderung

Stetige Skalen

- **reell**
- **ratio / positiv reell /Verhältnisskala**
  (ratio =Verhältnis)
  Nur positive Zahlen können beobachtet werden. $*, /$ sind sinnvolle Operationen. Der Abstand von 10 zu 1 ist genauso groß, wie der Abstand von $1 zu 0.1$.
  z.B. Gewicht, Länge

Stetige Skalen

- **reell**
- **ratio / positiv reell /Verhältnisskala**
- **Anteilskala / Wahrscheinlichkeitskala**
  (Anteil vom Ganzen)
  Nur Werte zwischen 0 und 1 können beobachtet werden. Die Werte sind als Anteile interpretierbar.

Stetige Skalen

- reell
- ratio / positiv reell /Verhältnisskala
- Anteilskala / Wahrscheinlichkeitskala

Die stetigen Skalen

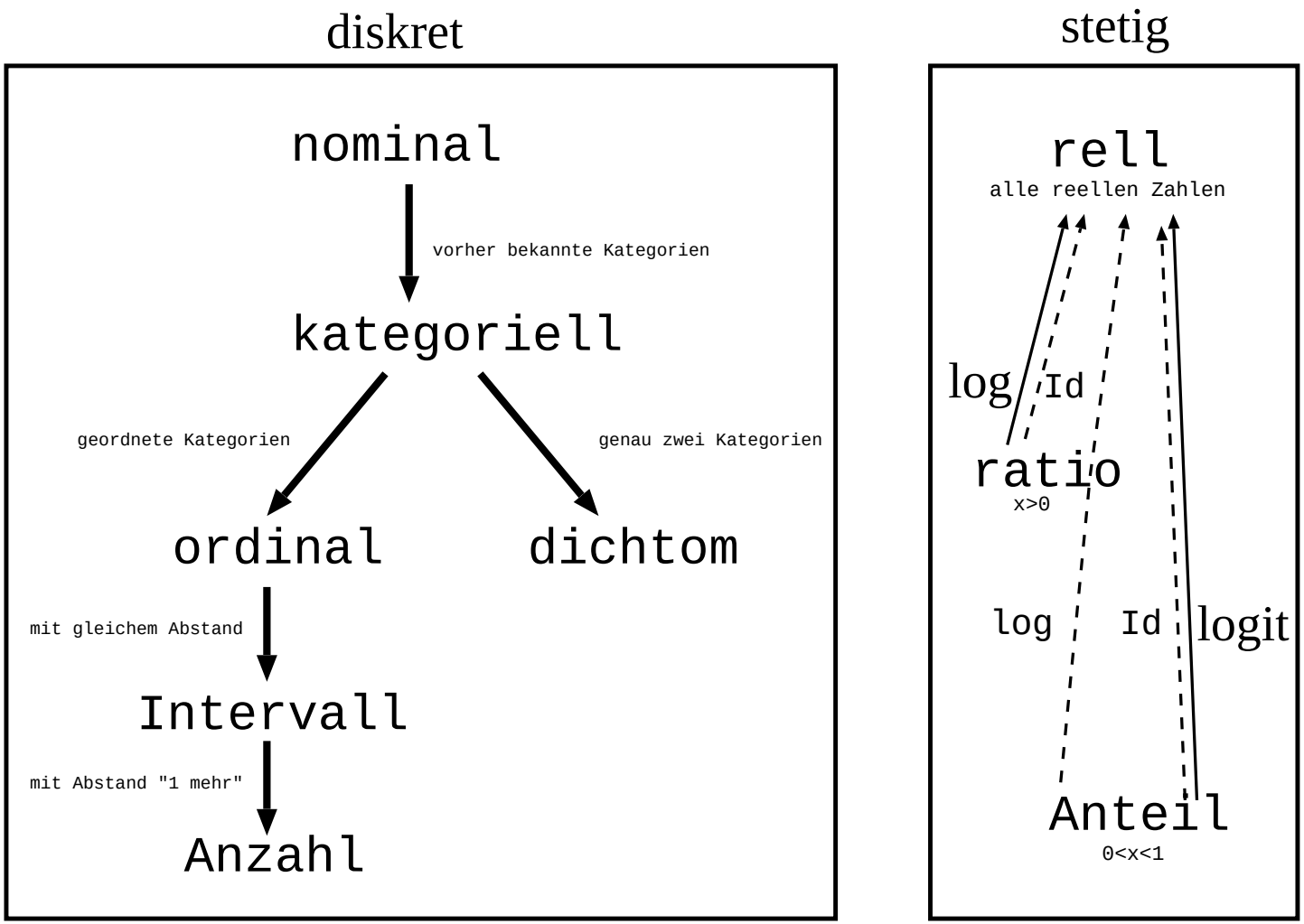
	AlkoholAnteil	Menge	Temperatur
1	0.1	0.125	16
2	0.3	0.500	5
3	0.7	1.000	-20

Grobeinteilung der Skalen

Die Skala bestimmt welche statistischen Verfahren angewendet werden können. Oft genügt im ersten Schritt schon eine Grobeinteilung:

- **diskret**
  Variablen mit diskreten Skalen heißen oft auch Faktor. Die Möglichen Werte heißen dann Stufen des Faktors.
- **stetig**
  Variablen mit stetigen Skalen können ein unendlich viele verschiedene Zahlenwerte annehmen. Treten dabei der gleiche Wert mehrfach auf, so spricht man von **Bindungen**.
- **spezielle**
  Variablen, die nicht ins Schema passen haben eine spezielle Skala.

Das feinste Skalenniveau



Versuchen wir es selbst

Ausschnitt des Iris Blueten Datensatzes:

```
> X
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
56	5.7	2.8	4.5	1.3	versicolor
58	4.9	2.4	3.3	1.0	versicolor

Welche Spalte hat welche Skala?

Datentafel

Die Datentafel ist eine alternative Darstellung zur Datenmatrix, wenn nur diskrete Skalen auftreten.

Datentafel (Beispiel)

```
> data(Titanic)
> ftable(Titanic, col.vars = c("Class", "Survived"))
```

		Class	1st		2nd		3rd		Crew	
		Survived	No	Yes	No	Yes	No	Yes	No	Yes
Sex	Age									
Male	Child		0	5	0	11	35	13	0	0
	Adult		118	57	154	14	387	75	670	192
Female	Child		0	1	0	13	17	14	0	0
	Adult		4	140	13	80	89	76	3	20

Erste Analyseschritte

Eine Datenauswertung beginnt grundsätzlich mit den folgenden Analyseschritten:

- **Wie liegen die Daten vor?**
- **Welche Variablen gibt es und was bedeuten Sie?** Wir lesen die Beschreibung!!!

Wozu die ersten Analyseschritte?

Eine Datenauswertung beginnt grundsätzlich mit den folgenden Analyseschritten:

- **Wie liegen die Daten vor?**
- **Welche Variablen gibt es und was bedeuten Sie?** Wir müssen den Datensatz verstehen, um Ihn auswerten zu können.

Wozu die ersten Analyseschritte?

Eine Datenauswertung beginnt grundsätzlich mit den folgenden Analyseschritten:

- **Wie liegen die Daten vor?**
- **Welche Variablen gibt es und was bedeuten Sie?**
- **Welche Skala haben die einzelnen Variablen?**
  Die Skala bestimmt die Auswahl der Analyseverfahren.

Wozu die ersten Analyseschritte?

Eine Datenauswertung beginnt grundsätzlich mit den folgenden Analyseschritten:

- **Wie liegen die Daten vor?**
- **Welche Variablen gibt es und was bedeuten Sie?**
- **Welche Skala haben die einzelnen Variablen?**
- **Ein-, Zwei- oder Mehrstichprobensituation?**
- **Was sind die Grundgesamtheiten?**
  Alle Analyseergebnisse beziehen sich nur auf diese Grundgesamtheit.