Statistik

Vorlesung Statistik 1

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

Organisation

- Webseite (Folien, Skript, Probeklausuren, Organisation)
- Übungen
- Klausur (Anmeldung 1+2, Hilfsmittel)
- Vorlesung
- Wie bestehe ich? (Vorlesung, Lernen, Übungen)

Inhalt heute (Grundlagen)

- Was ist Statistik?
- Grundmodelle der Statistik
- Datenmatrix
- Skala
- Datentafel

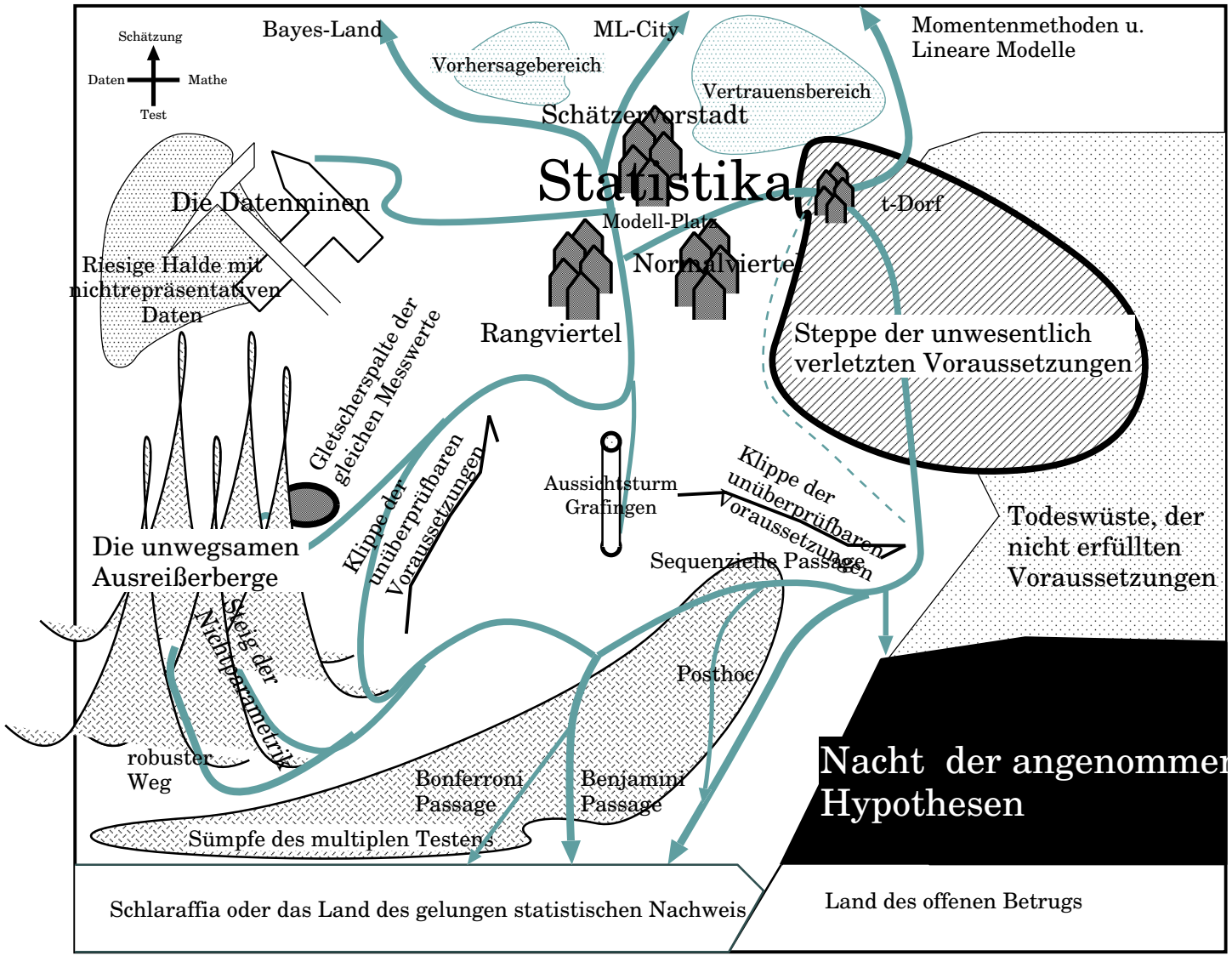
Was ist Statistik?

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.
Repräsentativ heißt:

- *identisch verteilt*: alle Beobachtungen folgen dem gleichen (uns interessierenden) Zufallsgesetz.
- *stochastisch unabhängig*: jede Beobachtung ist neu nach diesem Zufallsgesetz zustandegekommen.

Es gibt zwei grundsätzlich verschiedene Wege zu repräsentativen Daten:

- **Zufällige und faire** Auswahl einer *Stichprobe* aus einer **Grundgesamtheit**.
- **Unabhängiger** Wiederholung **identischer** *Zufallsexperimente*.

Repräsentative Daten durch Stichprobe und Grundgesamtheit

Grundbegriffe

- Grundgesamtheit
- statistisches Individuen
- Stichprobe
- repräsentativ
- Zufallsvariable
- Realisierung der Zufallsvariable

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.

Zufallsexperimente

Repräsentative Daten

durch

Zufallsexperimente

Grundbegriffe

- Vorschrift für ein Zufallsexpriment
- Zufallsexperiment
- identisch verteilt
- unabhängig
- repräsentativ
- Zufallsvariable
- Realisierung der Zufallsvariable

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
12	29	A	M

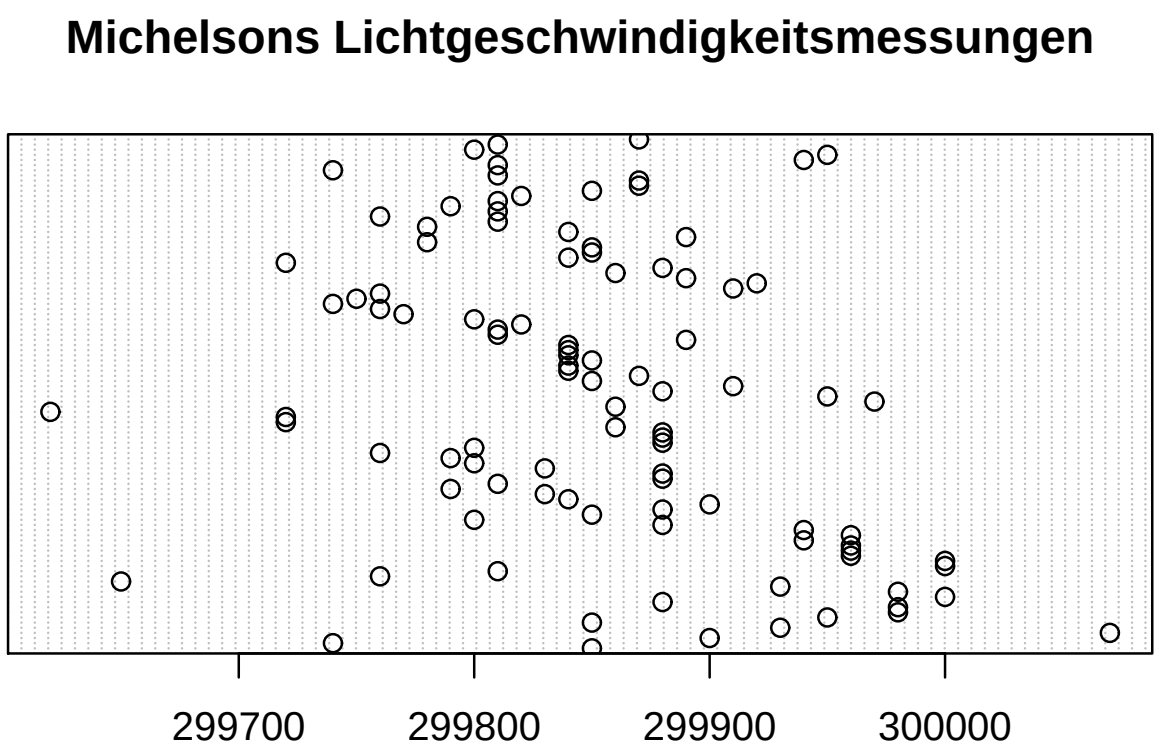
Beispiel: Lichtgeschwindigkeitsmessungen

```
> 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 Lichtgeschwind
```



Zusammenfassung zur 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
```

Repräsentation statistischer Daten

- Datenliste
- Datenmatrix
  - Fälle
  - Variablen
  - Skala (bestimmt die Auswertung!!!)
- Datentafel

Beispiel einer Datenliste

```
> 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 Datenlisten

```
$setosa
 [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8
[16] 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7
[31] 4.8 5.4 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1
[46] 4.8 5.1 4.6 5.3 5.0

$versicolor
 [1] 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6
[16] 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7
[31] 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6
[46] 5.7 5.7 6.2 5.1 5.7

$virginica
 [1] 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8
[16] 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2
[31] 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7
[46] 6.7 6.3 6.5 6.2 5.9
```

Beispiel einer Datenmatrix

Ausschnitt eines Datensatzes:

```
> X

    Sepal.Length Sepal.Width    Species
1            5.1           3.5     setosa
2            4.9           3.0     setosa
3            4.7           3.2     setosa
53           6.9           3.1 versicolor
54           5.5           2.3 versicolor
56           5.7           2.8 versicolor
58           4.9           2.4 versicolor
```

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
- Jede Spalte $X_{\cdot j}$ 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.
- Die Einträge einer Datenmatrix sind nicht unbedingt reelle Zahlen!

Fälle

Die Zeile der Datenmatrix heißen Fälle. Sie entsprechen den statistischen Individuen.

```
> X
```

```
    Sepal.Length Sepal.Width    Species
1            5.1           3.5     setosa
2            4.9           3.0     setosa
3            4.7           3.2     setosa
53           6.9           3.1 versicolor
54           5.5           2.3 versicolor
56           5.7           2.8 versicolor
58           4.9           2.4 versicolor
```

Der Begriff der Skala

Zu jeder Variable gehört eine **Skala**, also ein Wertebereich mit gewissen sinnvollen mathematischen Operationen. Kriterien zur Bestimmung der Skala sind:

- Welche Werte sind möglich?
- Wieviele Werte sind möglich?
- Sind die möglichen Werte geordnet? (Fachabi<Abi?)
- Sind die Abstände der Werte vergleichbar?
- Ist die Differenz ein guter Unterschiedsbegriff?
- Ist das Verhältnis ein guter Unterschiedsbegriff?

Skalen

- **diskrete Skalen**
  haben voneinander getrennte Werte
  nominal (), dichotom $(= \mathrm{NOT})$, kategoriell(=),
  ordinal(=<), Intervallskala(=< −), Anzahlen(=< −∗)
- **stetige Skalen**
  Anteil< /, positiv< ∗/, reell< ∗−
- **spezielle Skalen**
  z.B. Richtungen, Zusammensetzungen,
  Orientierungen, Winkel, Zuordnungen, . . .

Diskrete Skalen

- nominal
- kategoriell
- ordinal
- dichotom
- intervallskaliert
- Anzahl

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
- ratio / positiv reell /Verhältnisskala
- Anteilskala / Wahrscheinlichkeitskala

Die stetigen Skalen

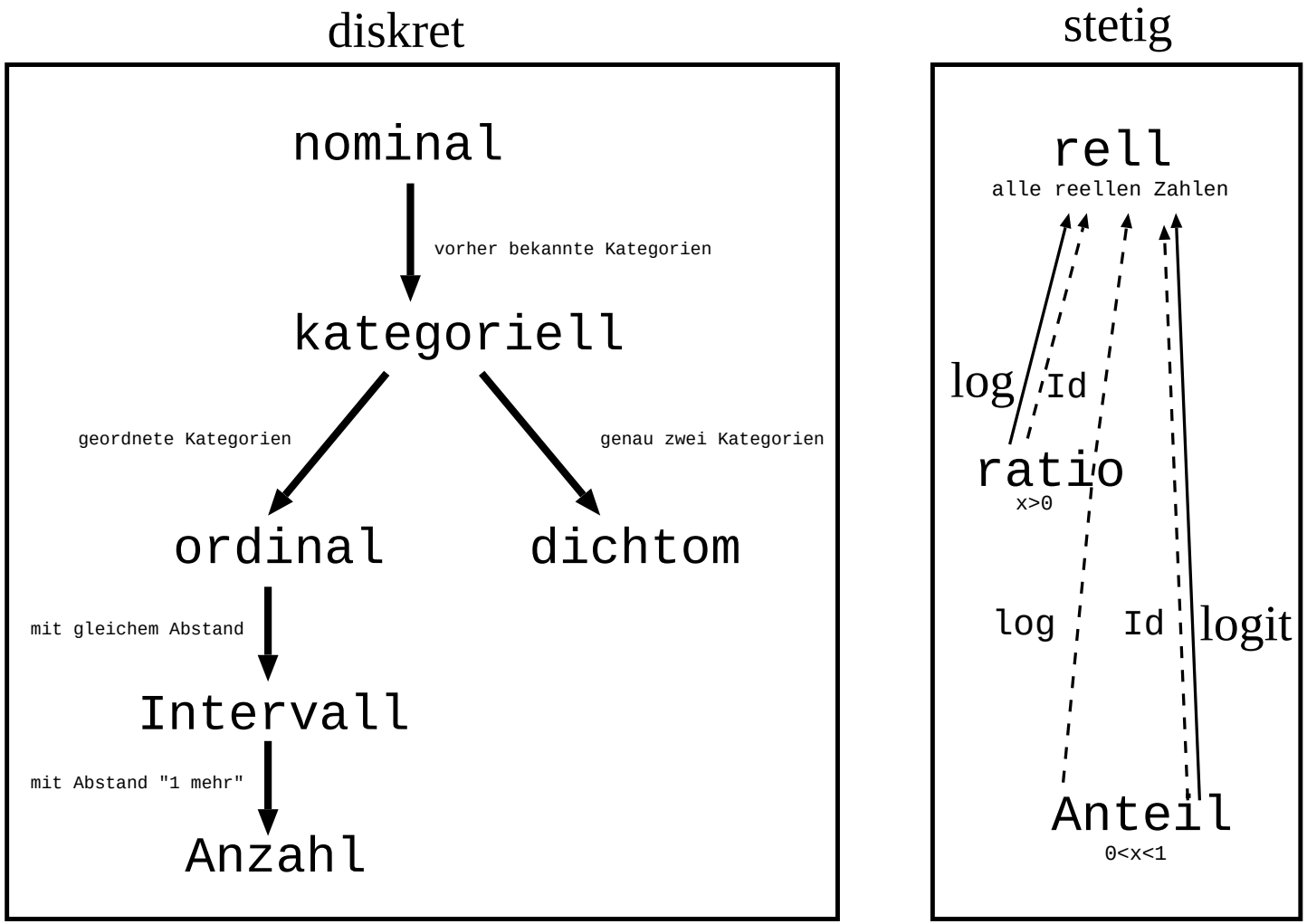
	AlkoholAnteil	Menge	Temperatur
1	0.1	0.125	16
2	0.3	0.500	5
3	0.7	1.000	-20

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?

Wozu Skala?

- Die Skala bestimmt welche weiteren Verfahren angewendet werden sollten.
- Die Skala gibt Hinweise was in der weiteren Analyse beachtet werden sollte.
- Die Skala bestimmt, wie die Daten zusammengefaßt und beschrieben werden können.
- Die Bestimmung der Skala der Variablen ist daher der erste Schritt jeder Datenanalyse.

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

Erklaerung Datentafel

Die Datentafel

- Jede Zelle der Datenmatrix enthält die Anzahl statistischer Individuen in der Stichprobe mit der gegeben Faktorkombination.

Erste 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?**
- **Sind die Daten für die Grundgesamtheit repräsentativ?**

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?**
- **Sind die Daten für die Grundgesamtheit repräsentativ?**

Repräsentation statistischer Daten

- Datenliste
  - Nur ein Merkmal!!!
  - alle Skalen
- Datenmatrix
  - mehrere Variablen
  - alle Skalen
- Datentafel
  - mehrere Variablen
  - nur kategorielle Skalen

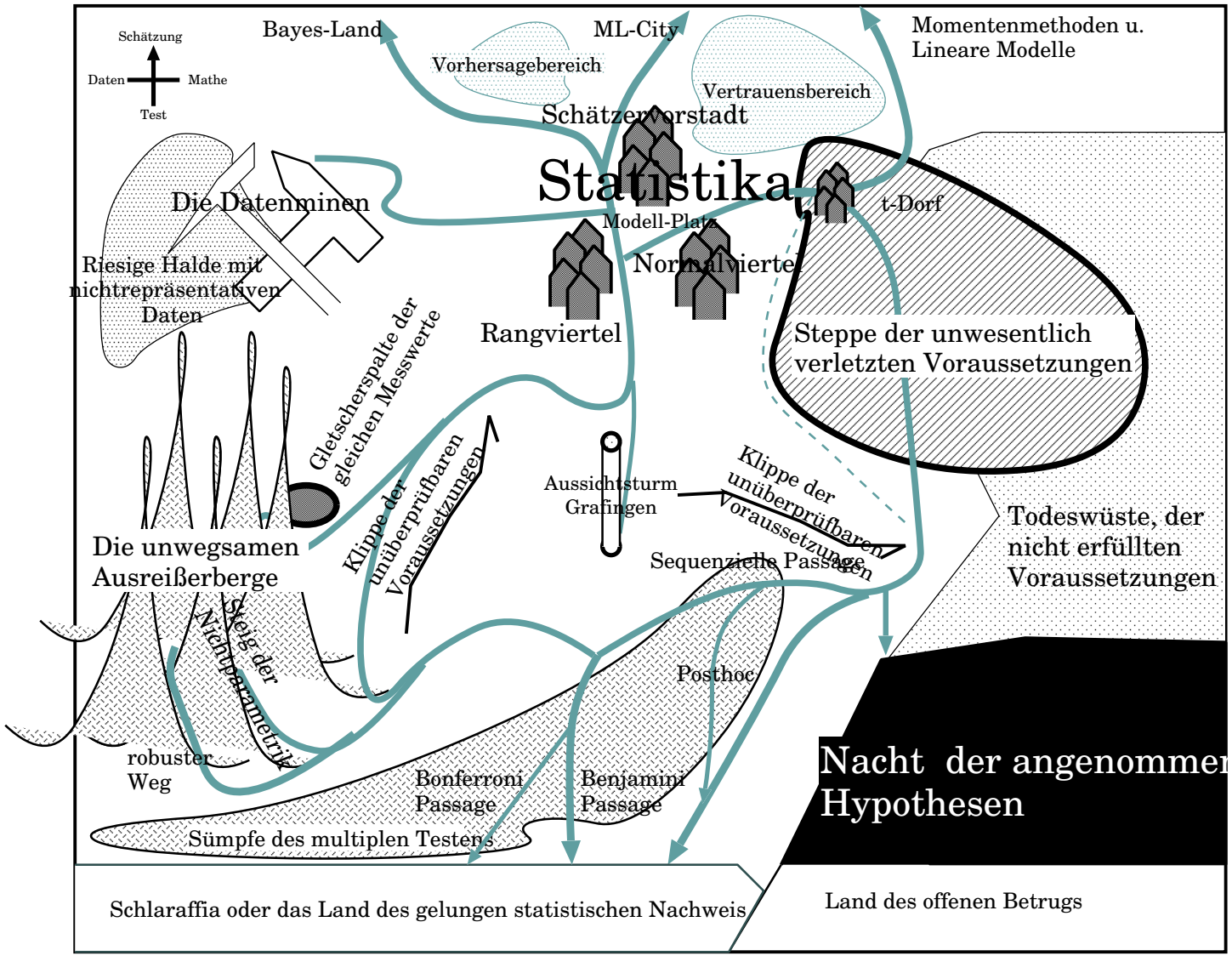
Zusammenfassung

- Repräsentativität statistischer Daten
- Repräsentation statistischer Daten
- Skalen statistischer Daten
- Zufälligkeit statistischer Daten

Zusammenfassung

- **Repräsentativität statistischer Daten**
  Nur diese Daten erlauben Rückschlüsse.
- **Repräsentation statistischer Daten**
  Nur diese Daten versteht jemand.
- **Skalen statistischer Daten**
  Das bestimmt das Auswertungsverfahren.
- **Zufälligkeit statistischer Daten**
  Das ist das Kernproblem bei der Auswertung.

Einordnung


Schätzung
Daten
Mathe
Test
Bayes-Land
ML-City
Momentenmethoden u.
Lineare Modelle
Vorhersagebereich
Vertrauensbereich
Schätzervorstadt
Statistika
Modell-Platz
t-Dorf
Die Datenminen
Riesige Halde mit
nichtrepräsentativen
Daten
Normalviertel
Rangviertel
Steppe der unwesentlich
verletzten Voraussetzungen
Gletscherspalte der
gleichen Messwerte
Klippe der
unüberprüfbaren
Voraussetzungen
Aussichtsturm
Grafingen
Klippe der
unüberprüfbaren
Voraussetzungen
Todeswüste, der
nicht erfüllten
Voraussetzungen
Die unwegsamen
Ausreißerberge
Sequenzielle Passagen
Steig der
Nichtparametrik
Posthoc
robuster
Weg
Bonferroni
Passage
Benjamini
Passage
Nacht der angenommer
Hypothesen
Sümpfe des multiplen Testens
Schlaraffia oder das Land des gelungen statistischen Nachweis
Land des offenen Betrugs