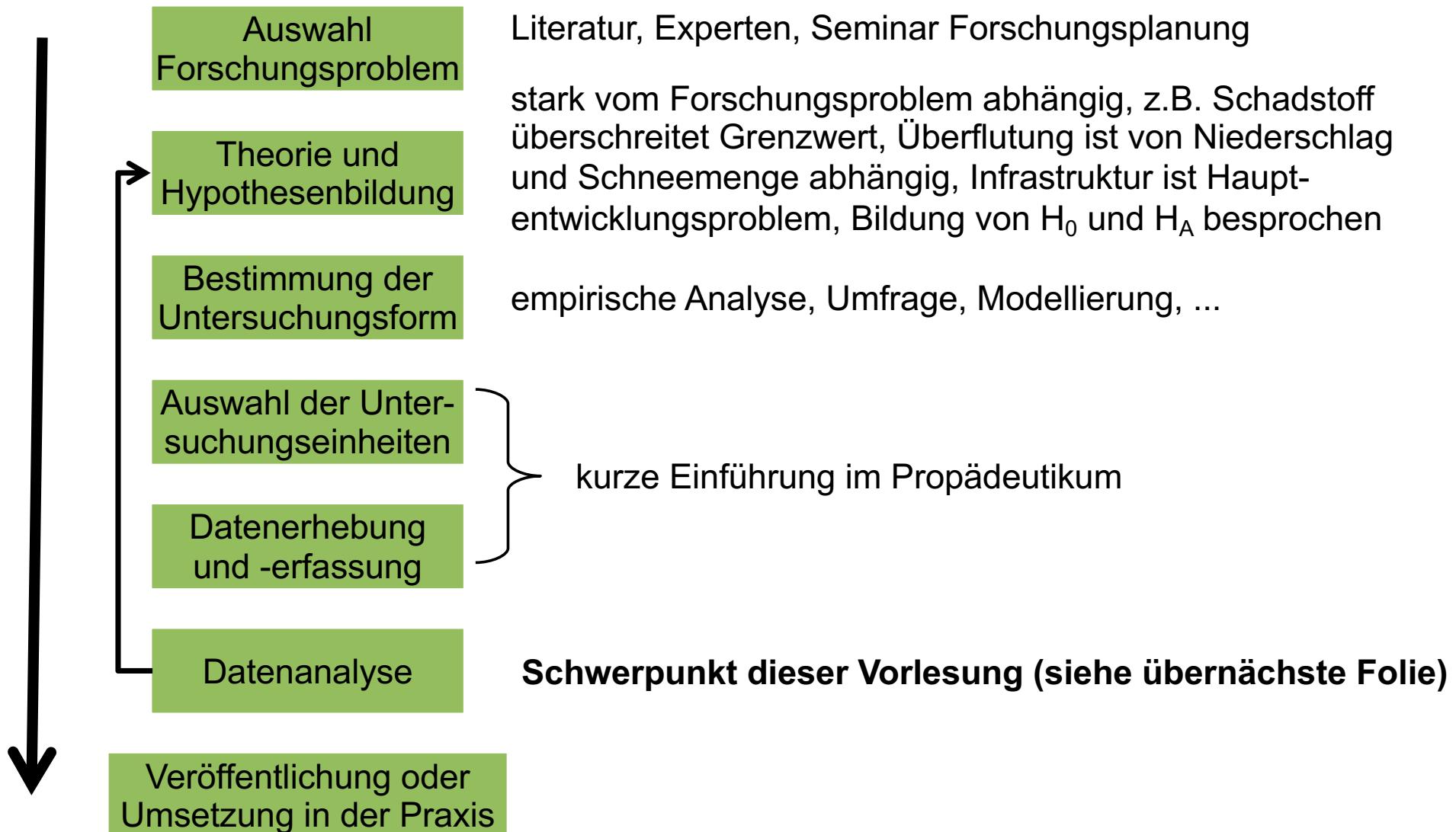


Quantitative Methoden in der Geographie

Herbstsemester 2024

Jörg Franke

Übersicht Forschungsprozess



Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik	Rohdaten visualisieren	Datenqualität prüfen	statistische Masszahlen
Schliessende Statistik	Unterschiede identifizieren	Zusammenhänge identifizieren	Abhängigkeiten modellieren
Fallen der Statistik	Wie wahrscheinlich sind die Daten der Stichprobe, wenn die Nullhypothese zutrifft?	Korrelation	Regression
weiterführende Methoden	Daten zusammenfassen	Extremwertstatistik	Hauptkomponenten-analyse
	Clusteranalyse	Zeitreihenanal. etc.	

Deskriptive Verfahren

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Ziel: Daten strukturieren, beschreiben

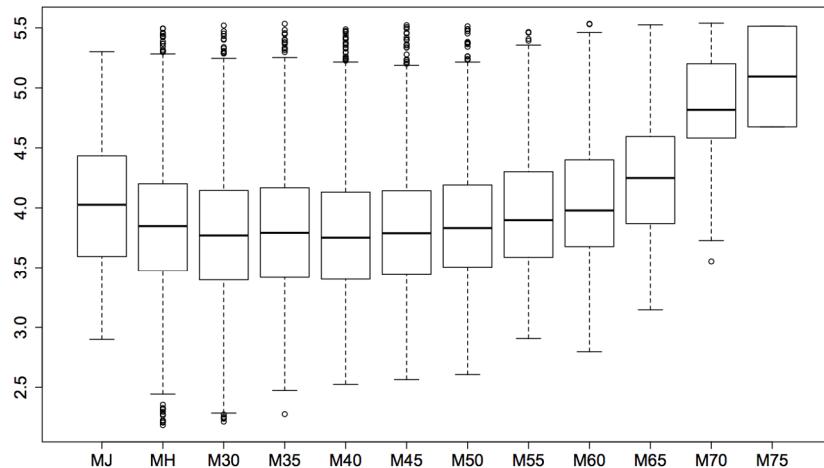


Abbildung 2.11: Box-Whisker-Plot von Laufzeit versus Altersklasse (Männer, Beispiel 2.4).

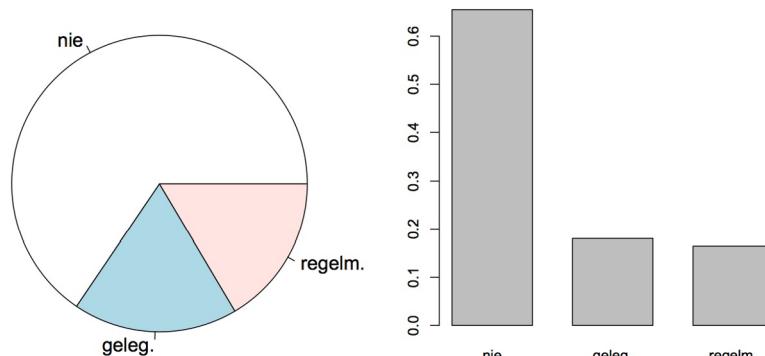
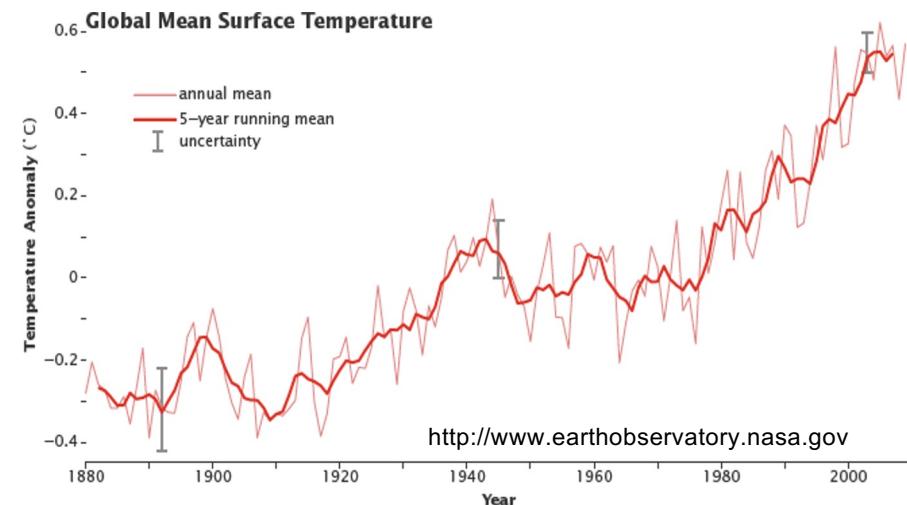


Abbildung 2.1: Kuchen- und Stabdiagramm der Variable Rauchen für Beispiel 2.1.

Das Auge erkennt sehr gut Strukturen
in den Daten!



<http://www.earthobservatory.nasa.gov>

Statistische Datenanalyse (Aufbau dieser Vorlesung)

Deskriptive Statistik

Rohdaten
visualisieren

Datenqualität
prüfen

statistische
Masszahlen

Schliessende Statistik

Unterschiede
identifizieren

Zusammenhänge
identifizieren

Abhängigkeiten
modellieren

Statistische Tests
Konfidenzintervalle

Korrelation

Regression

Wie wahrscheinlich
sind die Daten der
Stichprobe, wenn
die Nullhypothese
zutrifft?

Gibt es gemein-
same gleich- oder
entgegengerichtete
Variationen

Kausalzusammen-
hänge für Vorher-
sagen oder Inter-
polationen nutzen,
Modellvalidierung

Fallen der Statistik

weiterführende
Methoden

Daten
zusammenfassen

Extremwertstatistik

Hauptkomponenten-
analyse

Clusteranalyse

Zeitreihenanal. etc.

Grundgesamtheit vs. Stichprobe

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Grundgesamtheit = Alle Elemente für die eine Aussage gemacht werden soll

- > Untersuchung der Grundgesamtheit meist nicht durchführbar (unendlich gross oder zu hohe Kosten oder ...)
- > Griechische Buchstaben für **Grundgesamtheit** (μ für Mittelwert)

Stichprobe = repräsentative Teilmenge der Grundgesamtheit

- > Keine Regel für Mindestgrösse (meist >30 angestrebt)
- > Standard Römisches Alphabet für **Stichprobe** (m für Mittelwert)

Deskriptive/Beschreibende Verfahren lassen sich auf Grundgesamtheit und Stichprobe anwenden

Schliessende/Analytische Verfahren leiten aus Daten einer Stichprobe Eigenschaften einer Grundgesamtheit ab

Schliessende Verfahren

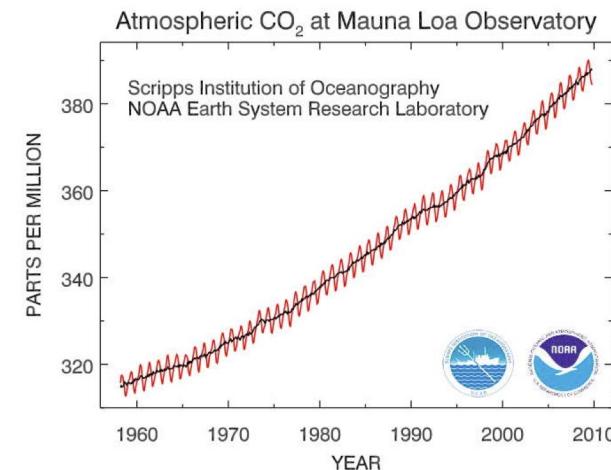
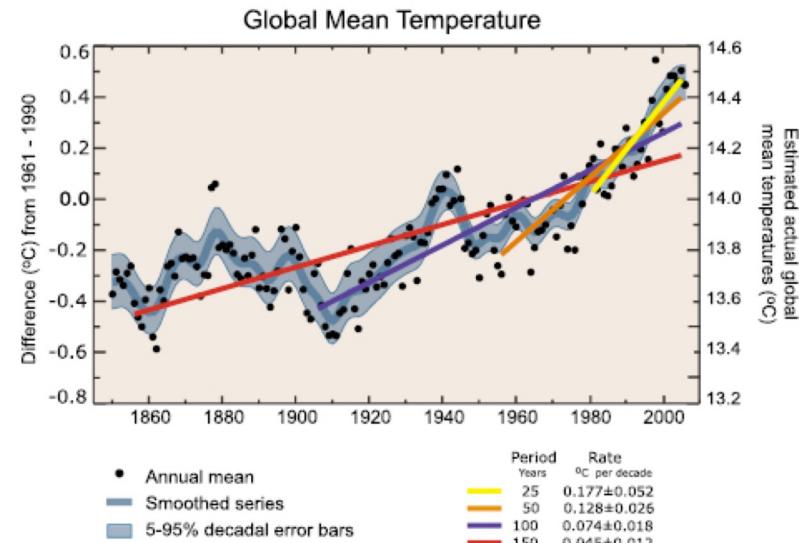
u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Ziele:

- > Hypothesen testen, z.B. Es gibt einen Erwärmungstrend über die Zeit
- > Parameter schätzen (Vorhersagen machen), z.B. Trend: die Temperatur nimmt um 1K pro 100 Jahre zu
- > Abhängigkeiten untersuchen, z.B. Temperatur und CO₂
- > Konfidenz-/Vertrauensintervalle angeben, z.B. mit 95%iger Sicherheit nimmt die Temperatur um 0.9 bis 1.1K pro 100 Jahre zu



Skalen

Nominalskala (keine Rangordnung)

- > z.B. Farben (grün, rot, blau), Wohnort
- > $=/\neq$

Ordinalskala (Rangordnung)

- > z.B. A, B, C; Rangliste
- > $=/\neq, >/<$

Kardinalskala

- > **Intervallskala** (x....y)
 - z.B. Temperatur [$^{\circ}\text{C}$]
 - $=/\neq, >/<, +/-$
- > **Verhältnisskala** (0....z)
 - z.B. Niederschlag in mm
 - $=/\neq, >/<, +/-, \times/\div$



kategoriale Variablen



metrische Variablen

diskret stetig

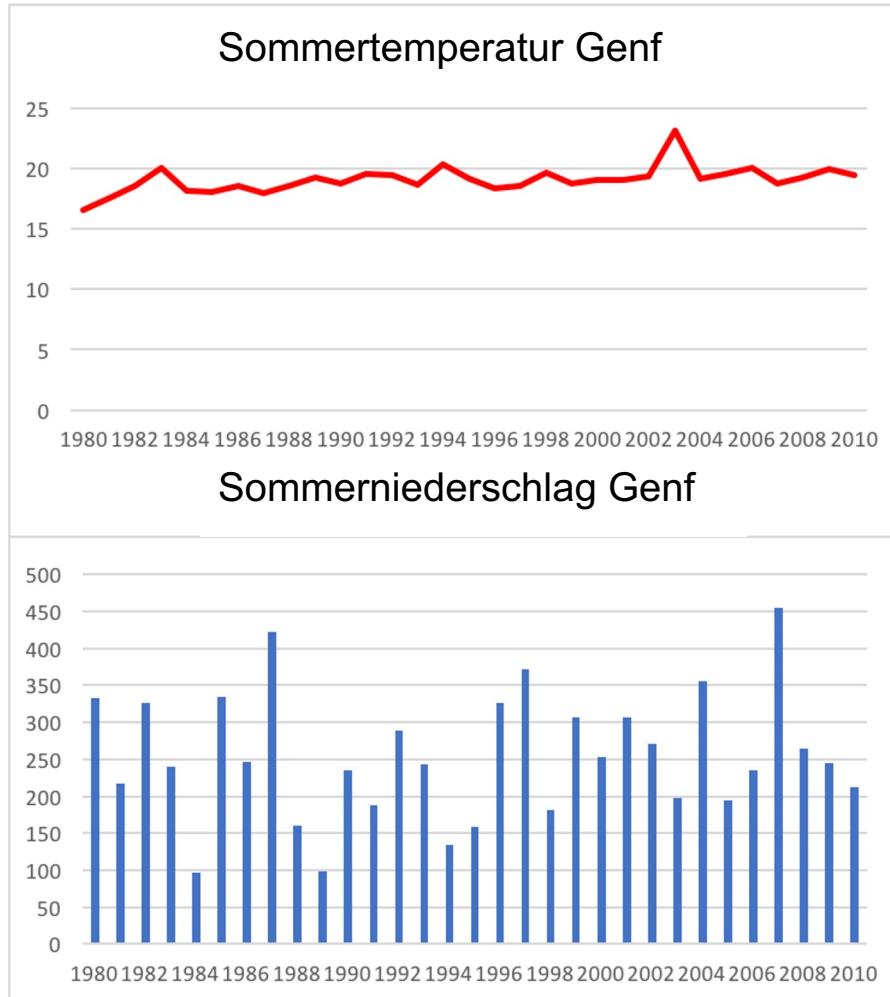
z.B. Anzahl z.B. Temperatur
Nebeltage

Deskriptive Verfahren

Was erkennt ihr in den Daten?

1980	16.53333	332.2
1981	17.53333	218
1982	18.56667	326.3
1983	20.03333	240.2
1984	18.13333	96.6
1985	18.06667	335.2
1986	18.56667	246.1
1987	17.93333	421.5
1988	18.53333	160.2
1989	19.2	97.9
1990	18.76667	234.3
1991	19.5	188.6
1992	19.4	289.2
1993	18.66667	243.5
1994	20.36667	133.7
1995	19.16667	157.8
1996	18.3	326.1
1997	18.5	372.5
1998	19.63333	181.2
1999	18.73333	307.3
2000	19.06667	252.9
2001	19.03333	307.2
2002	19.36667	271.5
2003	23.13333	197
2004	19.13333	355.2
2005	19.5	193.9
2006	20	234.9
2007	18.73333	455.6
2008	19.26667	264.1
2009	19.96667	244.1
2010	19.46667	212.8

Deskriptive Verfahren



1980	16.53333	332.2
1981	17.53333	218
1982	18.56667	326.3
1983	20.03333	240.2
1984	18.13333	96.6
1985	18.06667	335.2
1986	18.56667	246.1
1987	17.93333	421.5
1988	18.53333	160.2
1989	19.2	97.9
1990	18.76667	234.3
1991	19.5	188.6
1992	19.4	289.2
1993	18.66667	243.5
1994	20.36667	133.7
1995	19.16667	157.8
1996	18.3	326.1
1997	18.5	372.5
1998	19.63333	181.2
1999	18.73333	307.3
2000	19.06667	252.9
2001	19.03333	307.2
2002	19.36667	271.5
2003	23.13333	197
2004	19.13333	355.2
2005	19.5	193.9
2006	20	234.9
2007	18.73333	455.6
2008	19.26667	264.1
2009	19.96667	244.1
2010	19.46667	212.8

Lageparameter / Masse der Zentraltendenz

- > Mittelwert, Median, Modus, ...

Streuungsparameter

- > Spannweite, Varianz, Standardabweichung, Quantile, ...

Häufigkeitstabellen

Sonstiges

- > absolute Werte vs. Anomalien
- > Standardisierung/Transformationen
- > Freiheitsgrade

Lageparameter

Mittelwert, Median, Modus

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Modus

- > Wert der am häufigsten Auftritt (mehrere Werte, wenn multimodal)
- > Ordinal- und Nominaldaten

Median

- > Wert an mittlerer Stelle, wenn man Werte nach Grösse sortiert
- > Metrische und ordinale Daten
- > **verteilungsunabhängig**

arithmetischer Mittelwert

- > Metrische Daten, symmetrische Verteilung
- > Empfindlich gegenüber Ausreisern

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Kenntnis wird in der Prüfung vorausgesetzt

Streuungsparameter Quantile, Quartile, Median

u^b

b
UNIVERSITÄT
BERN

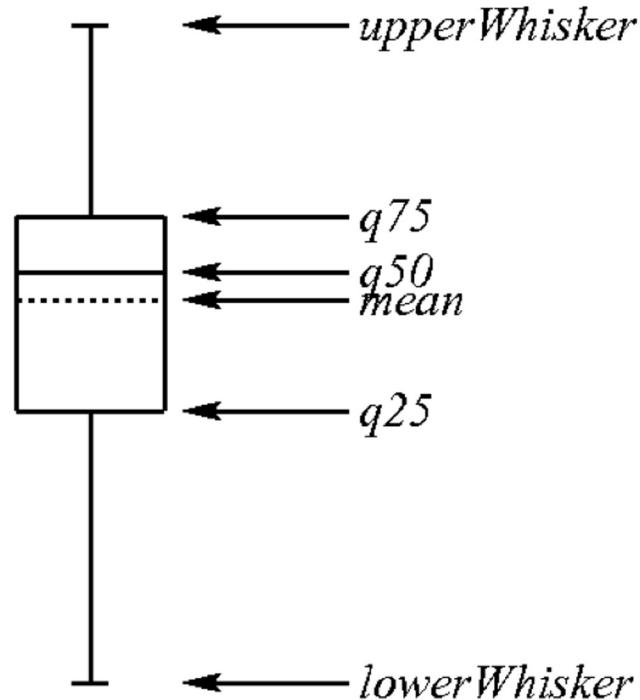
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Quantile

- > q% Quantil ist der Wert in einer geordneten Datenreihe, unterhalb dessen q% der Variablenwerte liegen
- > Metrische, ordinale Daten
- > Unempfindlich gegenüber Ausreisern
- > Auch für NICHT-symetrisch verteilte Daten

Spezielle Quantile

- > Median ist 50%-Quantil ($Q_{0.5}$)
- > Quartile ($Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$)
- > Whisker im Boxplot sind uneinheitlich definiert



Streuungsparameter Spannweite, Varianz, Standardabweichung

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Spannweite

- > Metrische Daten
- > Stark extremwertabhängig

$$x_{\max} - x_{\min}$$

Kenntnis wird in der Prüfung vorausgesetzt

Varianz

- > Mittlere quadratische Abweichung von arithmetischen Mittelwert
- > Sinnvoll, wenn arithmetischer Mittelwert sinnvoll (metrische Daten, symmetrische Verteilung)
- > Starker Einfluss von Extremwerten durch das Quadratieren

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

Kenntnis wird in der Prüfung vorausgesetzt

Standardabweichung (Quadratwurzel der Varianz)

- > Gleiche Einheiten wie Ausgangsdaten

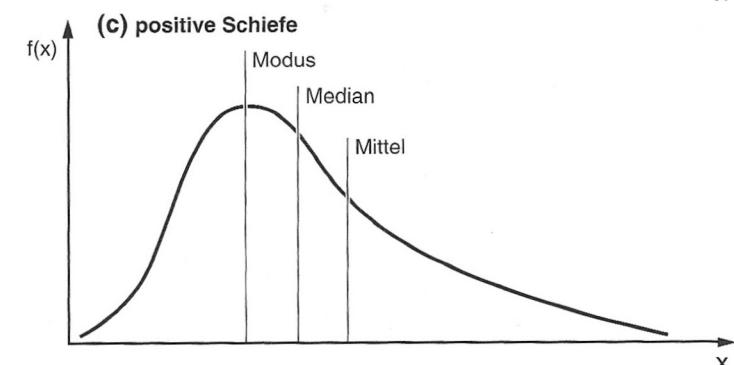
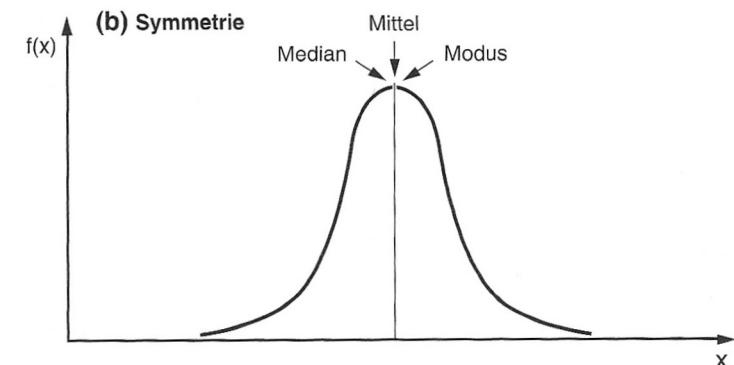
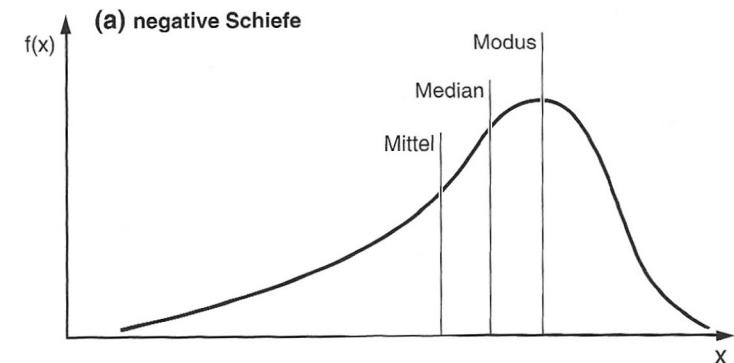
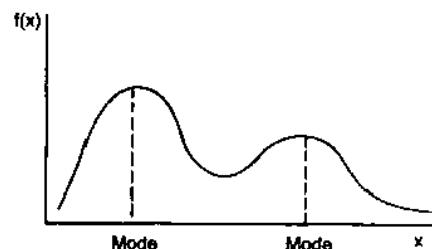
$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Kenntnis wird in der Prüfung vorausgesetzt

Schiefe

Einfaches Mass für die Schiefe:

- > $\text{Schiefe} = \frac{\text{arithm. Mittel} - \text{Median}}{\text{Standardabweichung}}$
- > Negative Schiefe = linksschief, rechtssteil
- > Positive Schiefe = rechtsschief, linkssteil
- > VORSICHT bei bi-, multi-modalen Daten!



Kreuztabelle / Kontingenztafel

Häufigkeitstabelle

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Für nominale Daten
- > Ordinale und metrische Daten können in nominale Daten transformiert werden (z.B. Grenzwertüberschreitung ja/nein)
- > Beispiel: Es werden 2000 Personen darüber befragt, ob sie Produkt A oder B bevorzugen. Das Ergebnis wird nach Geschlecht des Befragten ausgewertet.

Produkt/Geschlecht	weiblich	männlich	Summe
Produkt A	660	440	1100
Produkt B	340	560	900
Summe	1000	1000	2000

**Freiheitsgrade = Anzahl der Werte, die man frei ändern kann,
ohne das Ergebnis zu ändern.**

Beispiel 1 Mittelwert:

- > Ein Produkt kostet in einem über 30 Geschäfte gemittelten Durchschnitt 19.63 CHF.
- > Die Preise von 29 Geschäften können nun frei variiieren und nur beim 30. Geschäft muss der Preis so angepasst werden, dass der Mittelwert 19.63 ergibt. Das arithmetische Mittel hat also n-1 Freiheitsgrade.

- > Anzahl Beobachtungen abzüglich Anzahl geschätzter Parameter.
- > Beispiel: Standardabweichung aus Stichprobe mit n Beobachtungen.

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

- > Der Mittelwert wurde bereits aus den Beobachtungen geschätzt. Wenn man den Mittelwert und alle Beobachtungen ausser der letzten kennt ($n-1$) dann kann man diese berechnen, es besteht also keine "Freiheit" mehr (oder: die letzte Beobachtung ist "überflüssig").

Freiheitsgerade

R Beispiel

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

```
> # Grundgesamtheit
> gg=rnorm(100000,mean=0,sd=10)
> sqrt(sum((gg-mean(gg))^2)/100000)

> # Stichprobe ziehen
> st=sample(gg,30)
> sqrt(sum((st-mean(st))^2)/30)
> sqrt(sum((st-mean(st))^2)/(30-1))

> # Ergebnis von Stichprobenziehung abhängig,
> # deshalb viele Stichproben ziehen
> s=vector()
> sm1=vector()
> for (i in 1:1000) {
>   st=sample(gg,30)
>   s[i] <- sqrt(sum((st-mean(st))^2)/30)
>   sm1[i] <- sqrt(sum((st-mean(st))^2)/(30-1))
> }
> mean(s)
> mean(sm1)
```

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

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

Take-home messages

Metrische und symetrisch verteilte Daten

- > Mittelwert und Standardabweichung sind aussagekräftig

NICHT-symetrisch verteilte Daten

- > Median und Quantile sind robustere Lage- und Verteilungsmasse

Nomiale Daten

- > Können mit Modus und Kontingenztabellen beschrieben werden

Absolute vs. relative Zahlen?

Wichtige Begriffe

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Objekt	Untersuchungseinheit	Beispiel: Person
Merkmal	Theoretische Dimension, die mehreren Objekten gemeinsam ist	Körpergrösse, Haarfarbe
Konstante	Merkmal mit nur einer beobachteten Ausprägung	schwarz
Variable	Merkmal mit mehreren beobachteten Ausprägungen	schwarz, braun, blond, ...
Wert	Ausprägung einer Variablen	180cm Körpergrösse
Skala	Systematische Zuordnung von Zahlen oder Symbolen zu den Ausprägungen einer Variablen	Nominal, ordinal, kardinal/metrisch
Messen	Einordnen von Objekten auf einer Skala (umfasst auch Zählen)	32

Häufige Abkürzungen

u^b

^b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

- > Skalare gekennzeichnet durch nicht-fett gedruckte Kleinbuchstaben (x),
- > Vektoren durch fett gedruckte Kleinbuchstaben (**x**) und
- > Matrizen durch fett gedruckte Grossbuchstaben (**X**)
- > n: Stichprobengrösse
- > H_0 : Nullhypothese (beinhaltet normalerweise das „=“ Zeichen)
- > H_1 oder H_A : Alternativhypothese (beinhaltet normalerweise das „≠“, „>“ oder „<“ Zeichen)
- > α : Signifikanznivau, maximal zulässige Irrtumswahrscheinlichkeit für Fehler 1. Art (α -Fehler)
- > p-Wert: Probability, also Wahrscheinlichkeit, diese Stichprobenwerte zu erhalten, wenn H_0 zutrifft
- > x/μ : Mittelwert der Stichprobe/Grundgesamtheit
- > s/σ : Standardabweichung der Stichprobe/Grundgesamtheit
- > s^2/σ^2 : Varianz der Stichprobe/Grundgesamtheit
- > s_x : Standardfehler zeigt die theoretische Streubreite des Stichprobenmittelwerts $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$
- > $s_{\widehat{\beta}_1}$: Standardfehler des Regressionskoeffizienten
- > s_{xy} : Kovarianz zwischen x und y
- > r/ρ : Korrelationskoeffizient
- > y: Beobachtungen von y
- > \hat{y} : Das „Dach“ steht für eine Schätzung hier von y, z.B. aus Regressionmodell
- > β_0 : Regressionskonstante (Schnittpunkt mit der y-Achse bei $x=0$)
- > β_1 : Regressionskoeffizient (Steigung der Regressionsgeraden)
- > R^2 : Bestimmtheitsmass bei der Regression ($r^2=R^2$)
- > ε : error, also Fehler, bei der Regression die Residuen ($y - \hat{y}$)

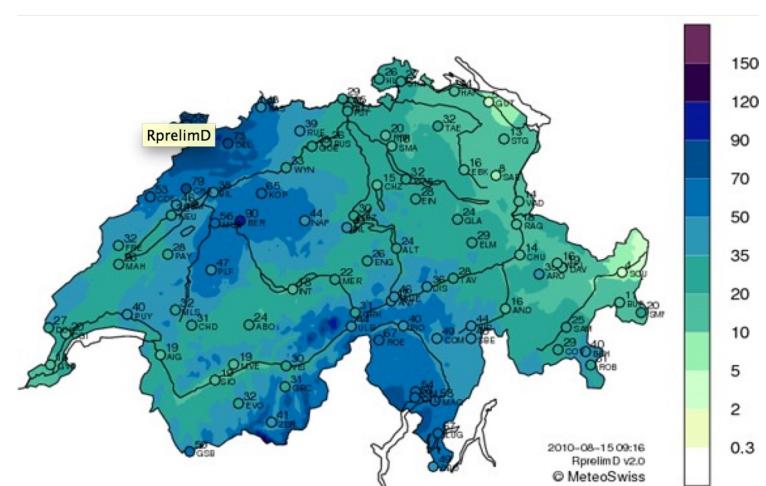
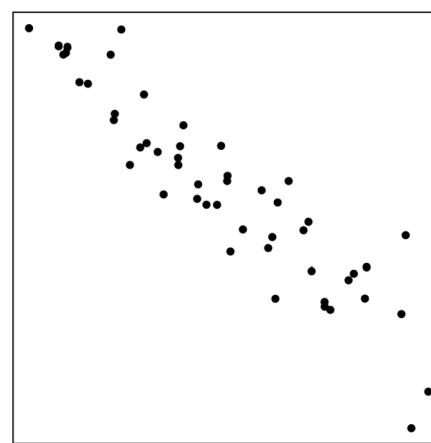
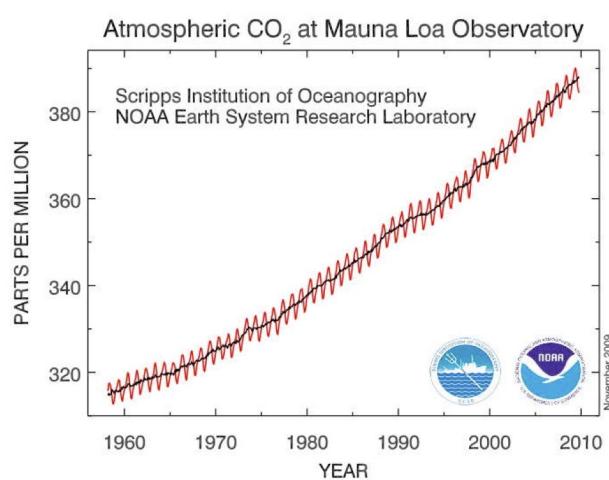
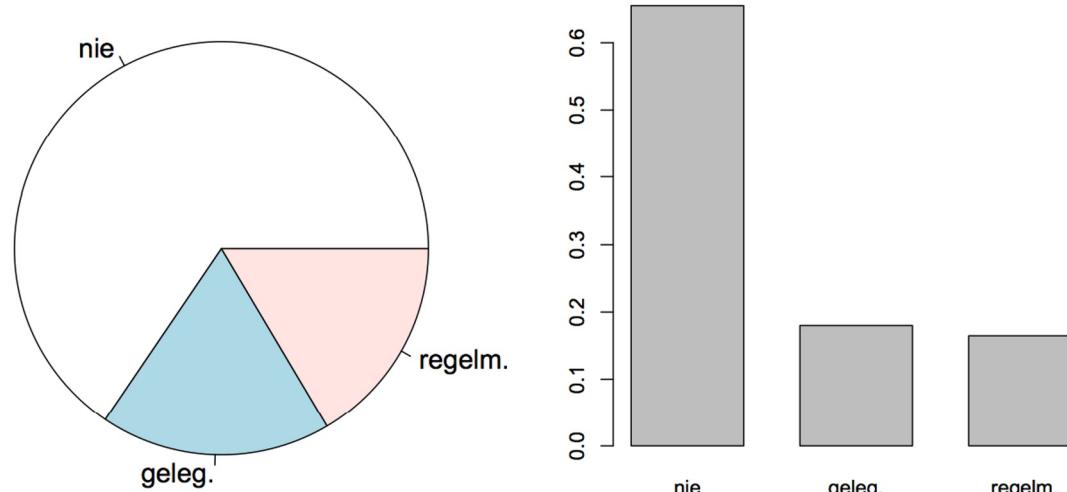
Visualisierung

Das Auge erkennt sehr gut Strukturen in den Daten!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Graphiken sollten benutzt werden, wenn:

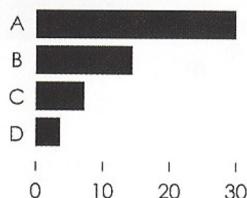
- > für schnelle Übersichten unübersichtlicher Zahlen
- > genaue Zahlen nicht interessieren
- > als Argumente in Vorträgen, da Graphiken mehr Information in kurzer Zeit vermitteln und gut im Gedächtnis bleiben

Balken- und Kreisdiagramme

Categories

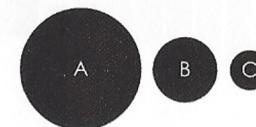
When your data is straightforward, with a value for each category, these are easy to read and create.

Bar graph



With length as visual cue, useful for straightforward comparisons

Symbol plot

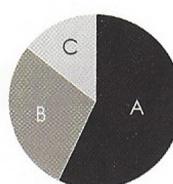


Can be used in place of bars, but can be hard to see small differences

Parts of a whole

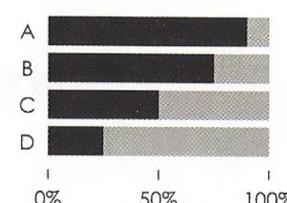
The categorical breakdown within a population can be interesting, and you might want to keep the groups together, although often not essential.

Pie chart



Parts add to 100 percent, typically sorted clockwise for readability

Stacked bar chart

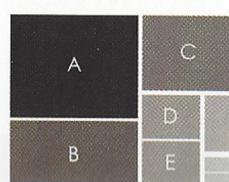


Often used to show poll results and can also be used for raw counts

Subcategories

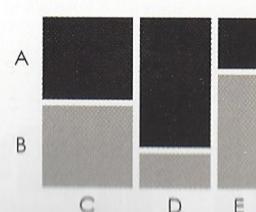
Data can have a hierarchical structure, which can be important in data interpretation and it often allows for different points of view.

Treemap



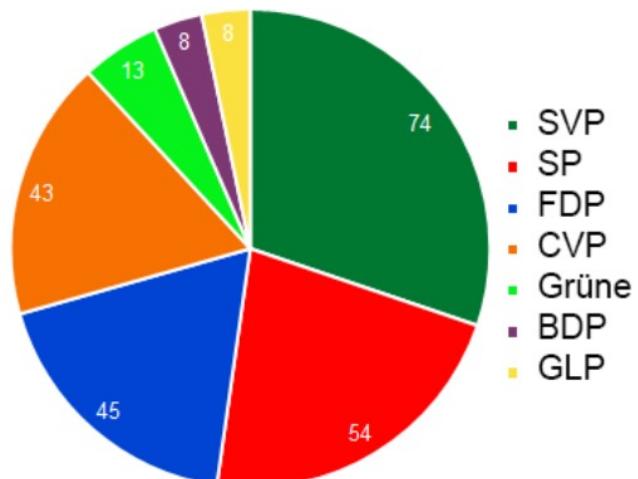
Shows hierarchical structure in a compact space, area often combined with color

Mosaic plot



Allows comparison across multiple categories in one view

Kuchendiagramme



Anzahl Ratsmitglieder
pro Fraktion (21.6.19)

18.7

Man liest im Uhrzeigersinn,
beginnend auf 12 Uhr.
Wichtige Tortenstücke
auf die 12-h-Position setzen.

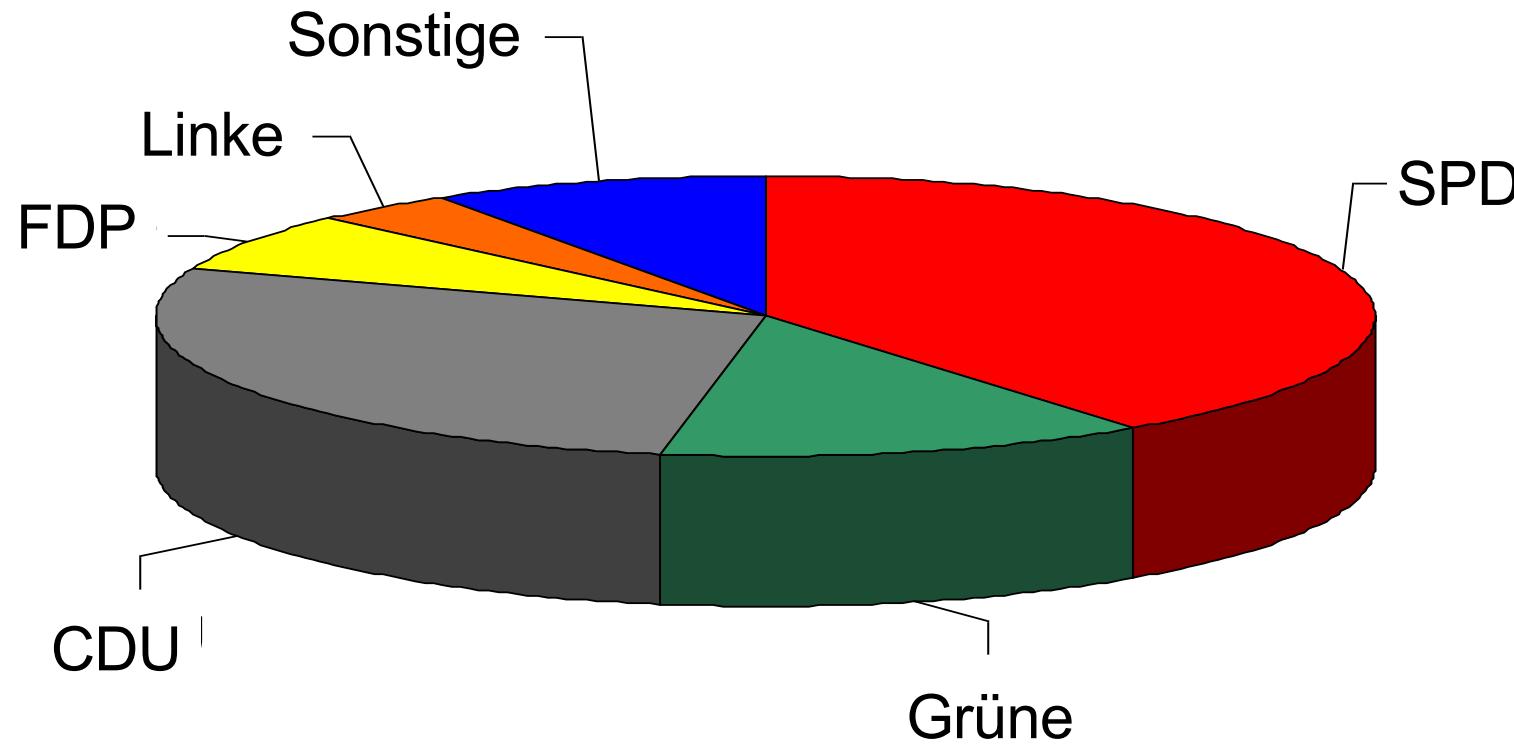
Kuchendiagramme

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

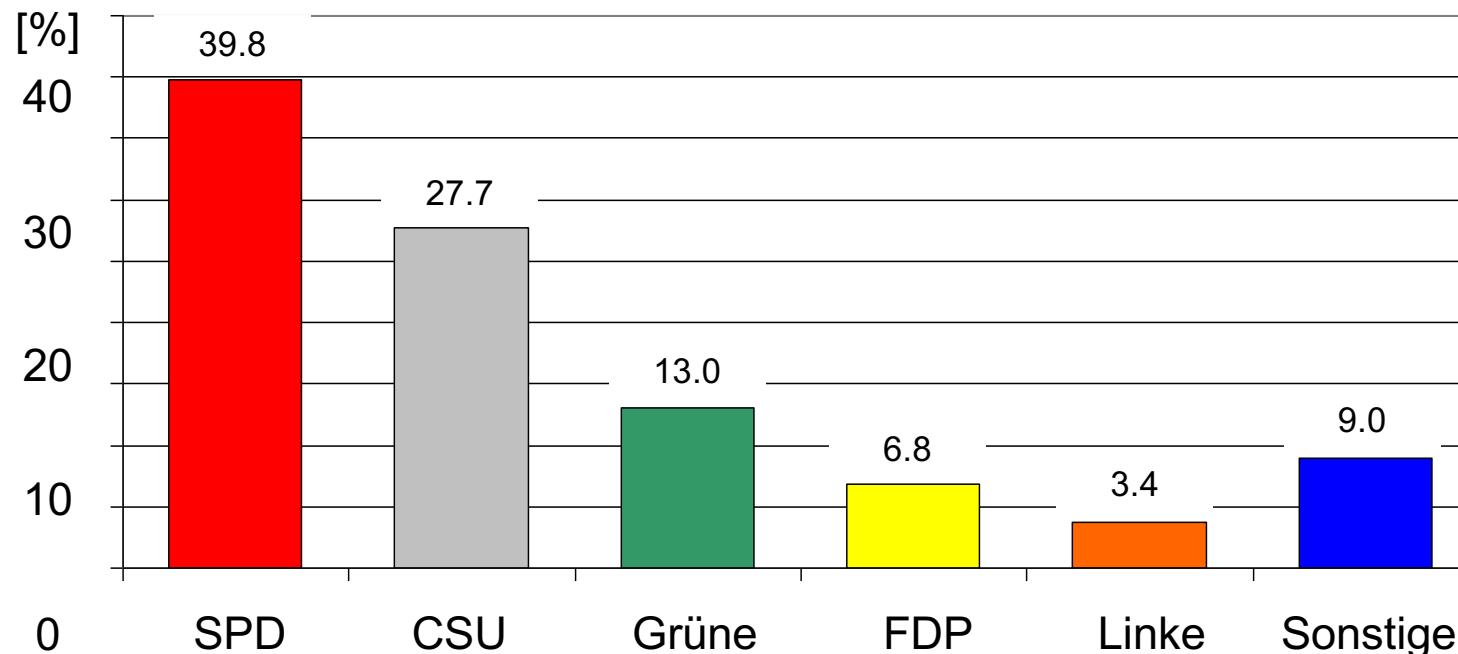
Was zusammengehört,
gehört nebeneinander!



Winkel bei 3D-Darstellung
Schwer zu interpretieren

Säulendiagramme

- > Wenn der Vergleich von Anteilen im Vordergrund steht
- > Für Torten mit zu vielen Anteilen



- > Säulen betonen die Anteilsunterschiede und Rangfolge
- > Torten die Aufteilung eines Ganzen auf Teile

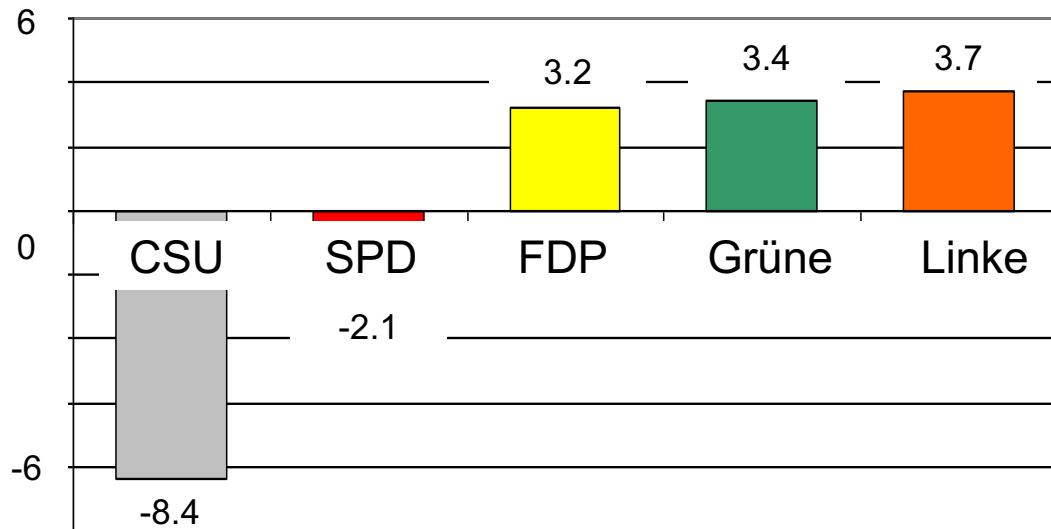
Säulenvariationen

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

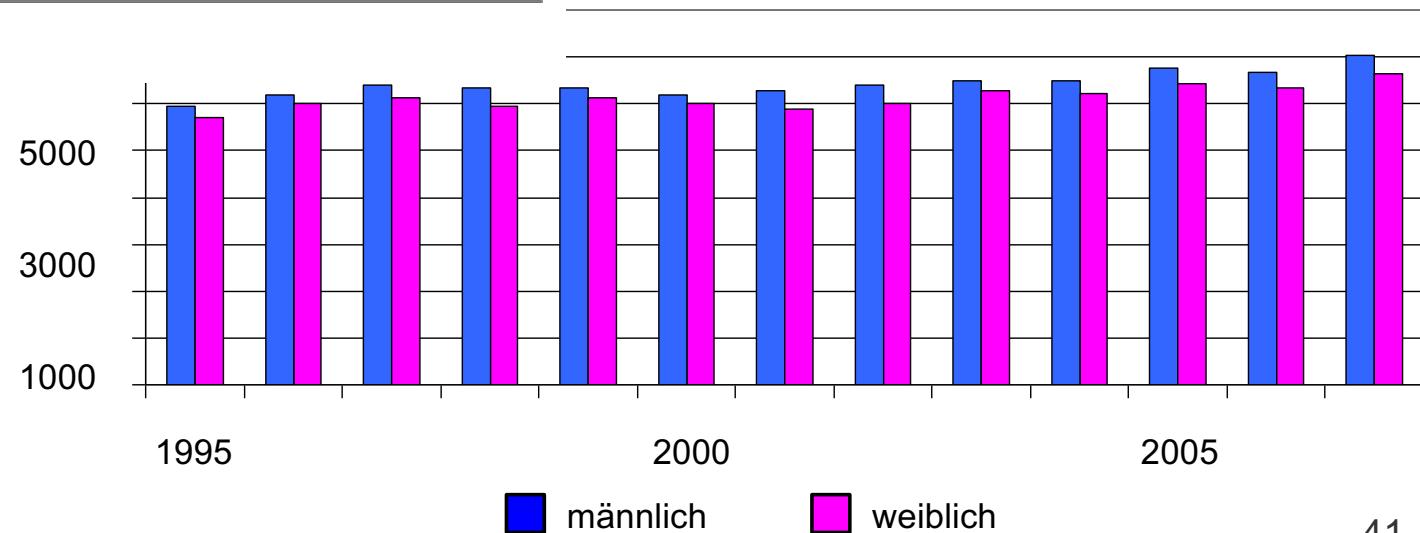
Gewinne und Verluste bei den Wahlen



Auch negative Werte
darstellbar (nicht bei
Torte)

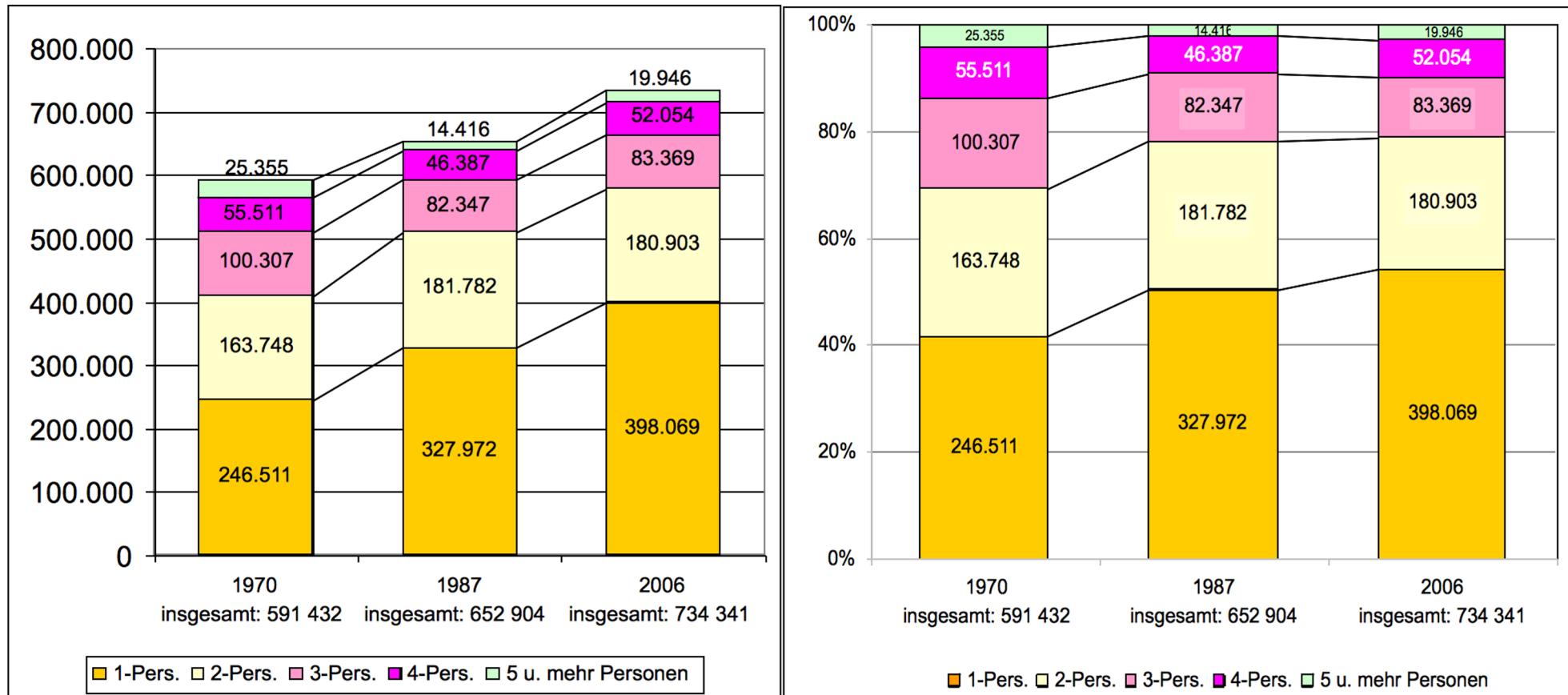
Geburten 1995-2007

Doppel-
Säulen-
Diagramm



Säulenvariationen

Privathaushalte in München 1970, 1987, 2006



Additives Säulendiagramm

100%-Säulendiagramme stellen
Verschiebung der Anteile heraus

Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Beispiel: Geschlechterverteilung 1974 bis 2004 in %

Jahr	Frauenanteil	Männeranteil
1974	6.8	93.2
1984	10.8	89.2
1994	23.6	76.7
2004	33.2	66.8

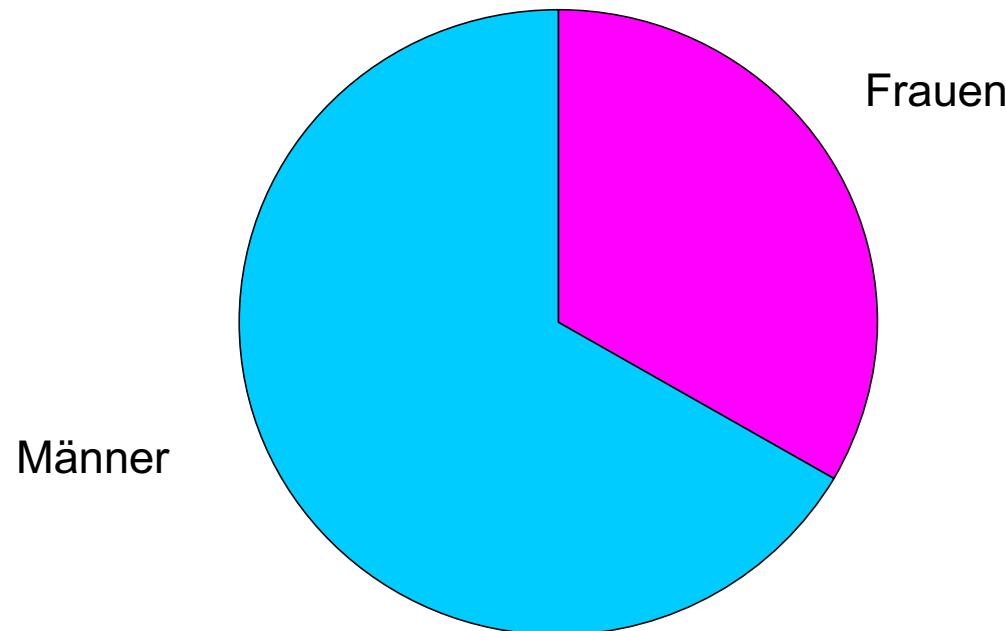
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Tortendiagramme betonen die Anteile



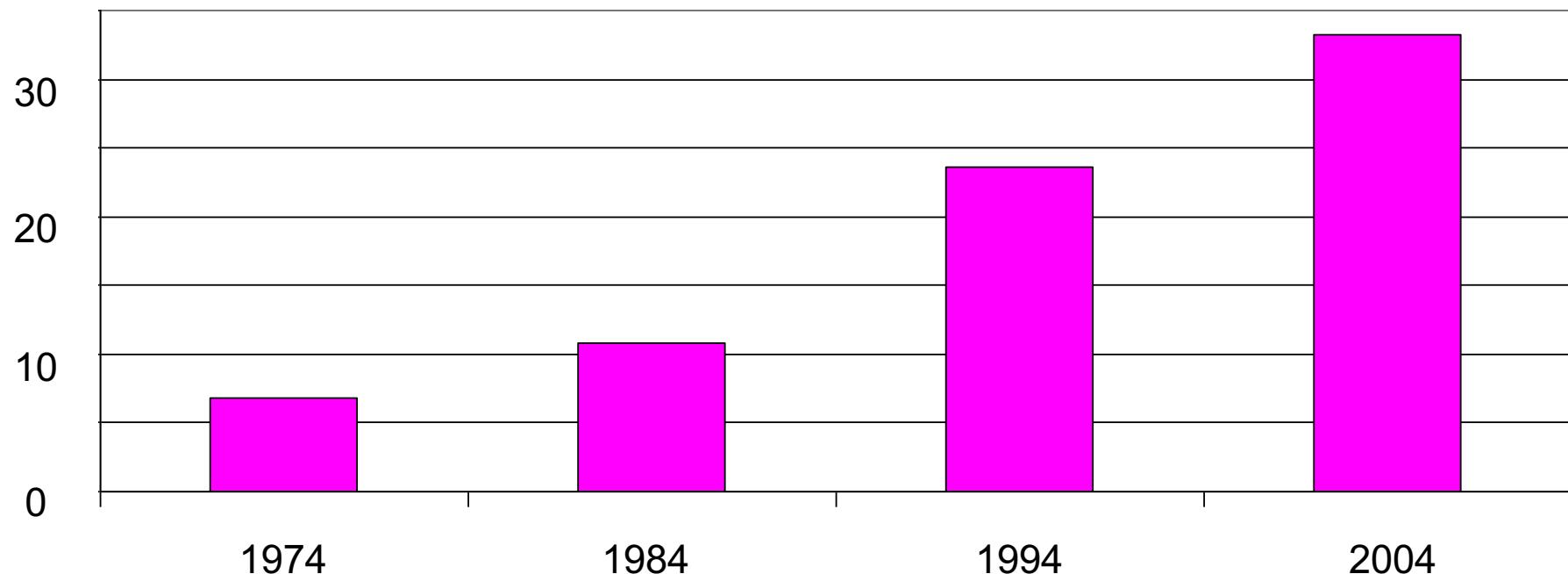
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Säulen betonen die Unterschiede von Anteilen im Verlauf



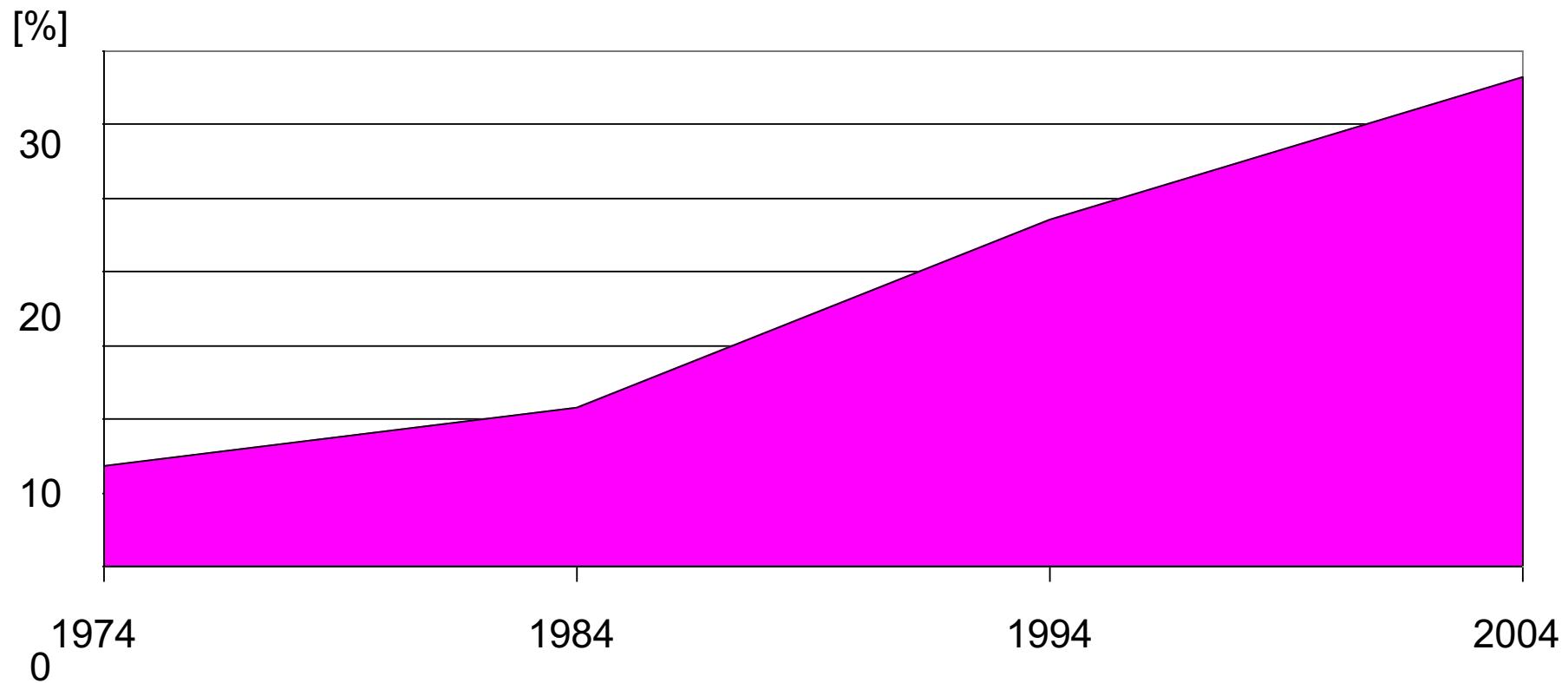
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Kurvendiagramme betonen den Trend



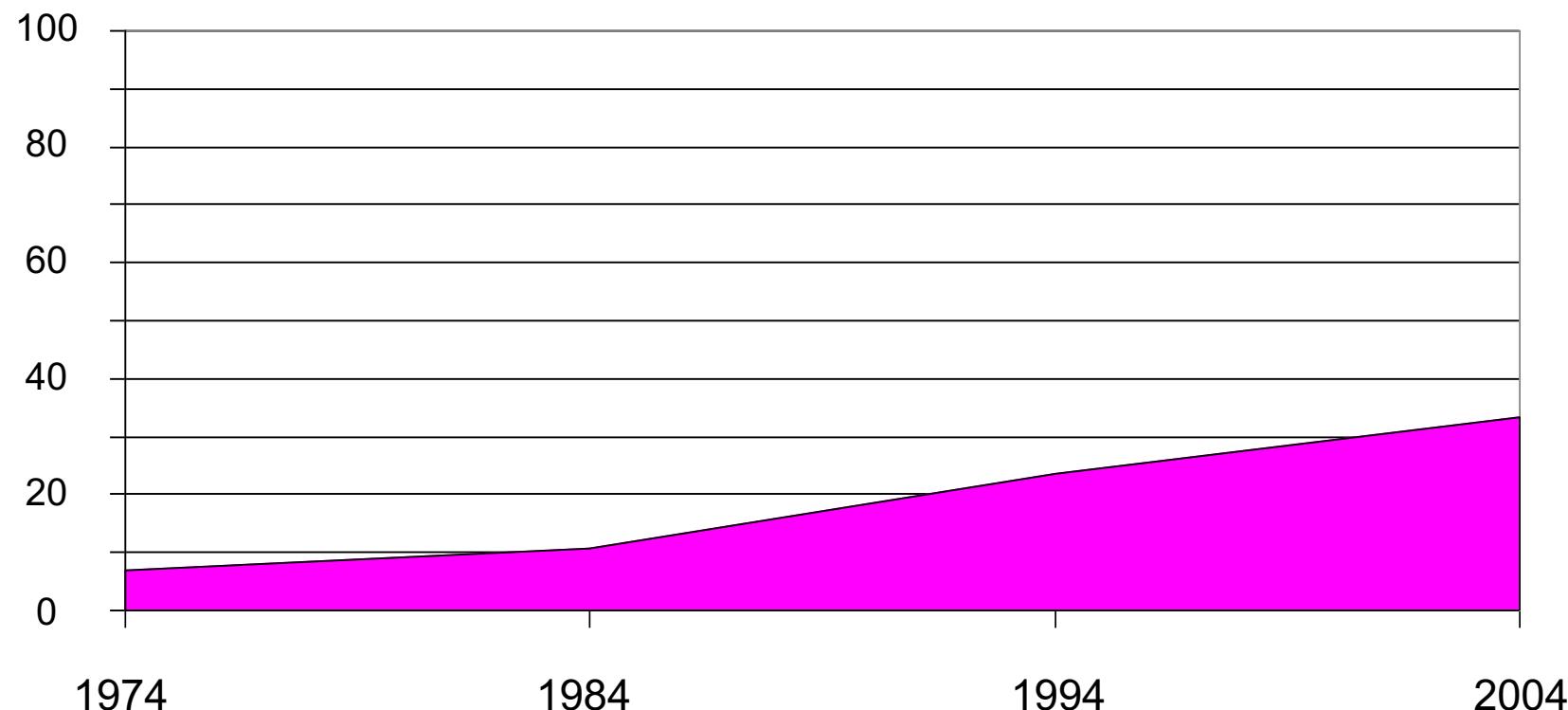
Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

Diese Kurve betont den Abstand zur 100%-Linie

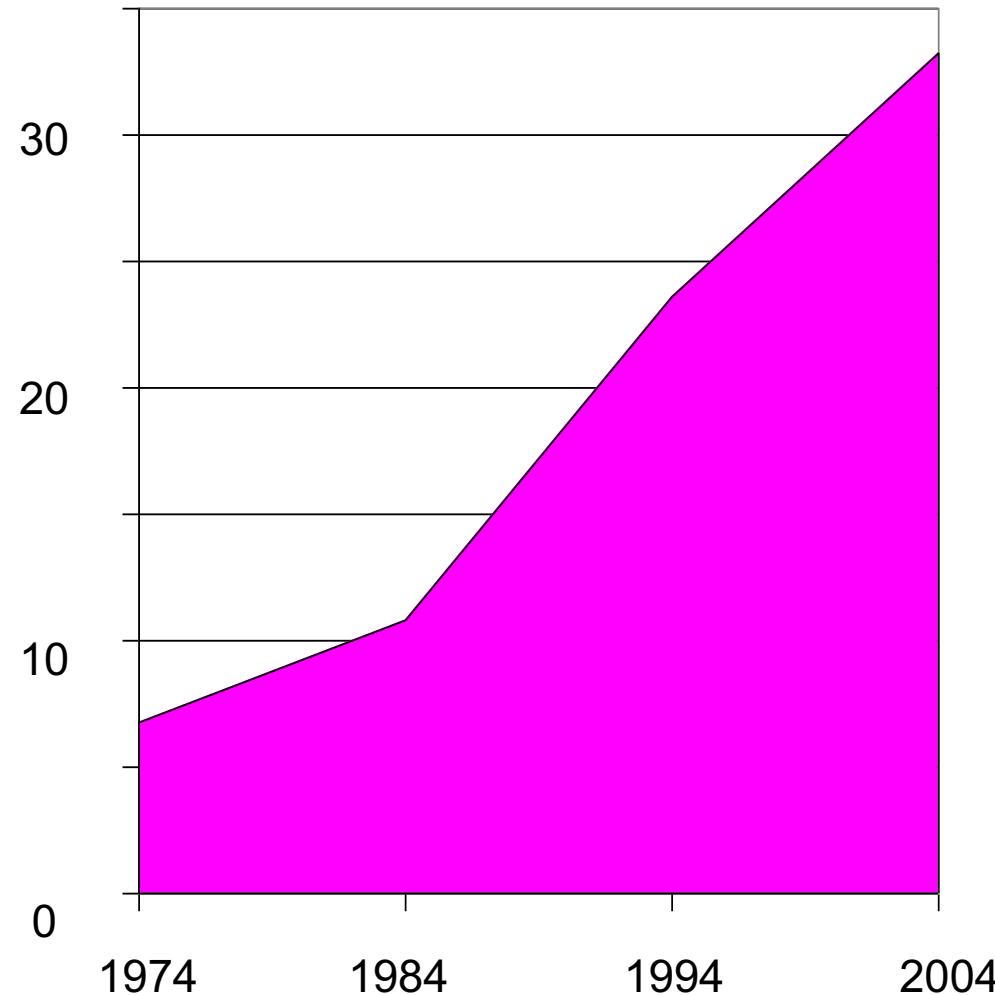


Welche Grafiktypen verwenden?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



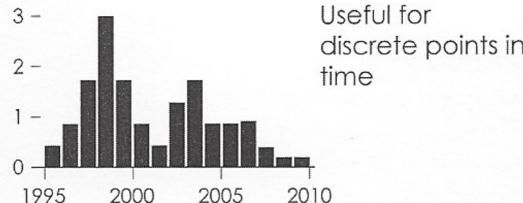
Bei dieser Kurve
wächst der
Frauenanteil
dramatisch!

Zeitreihen

Time series

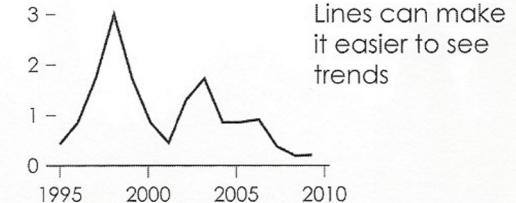
There are a variety of ways to see patterns over time, using cues such as length, direction, and position.

Bar graph



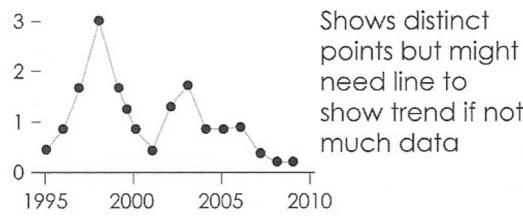
Useful for discrete points in time

Line chart



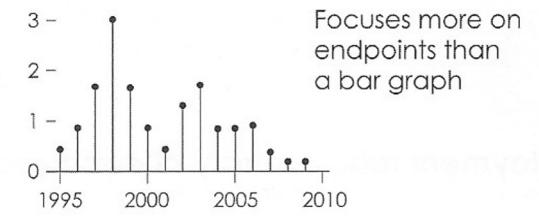
Lines can make it easier to see trends

Dot plot



Shows distinct points but might need line to show trend if not much data

Dot-bar graph

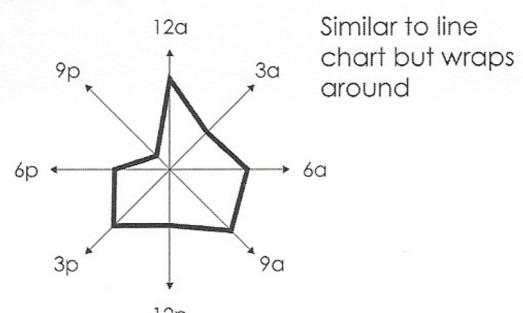


Focuses more on endpoints than a bar graph

Cycles

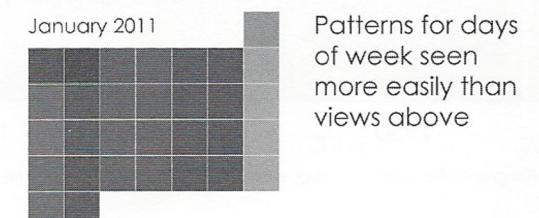
Time of day, day of the week, and month of the year repeat themselves, so it is often beneficial to align the segments in time.

Radial plot



Similar to line chart but wraps around

Calendar



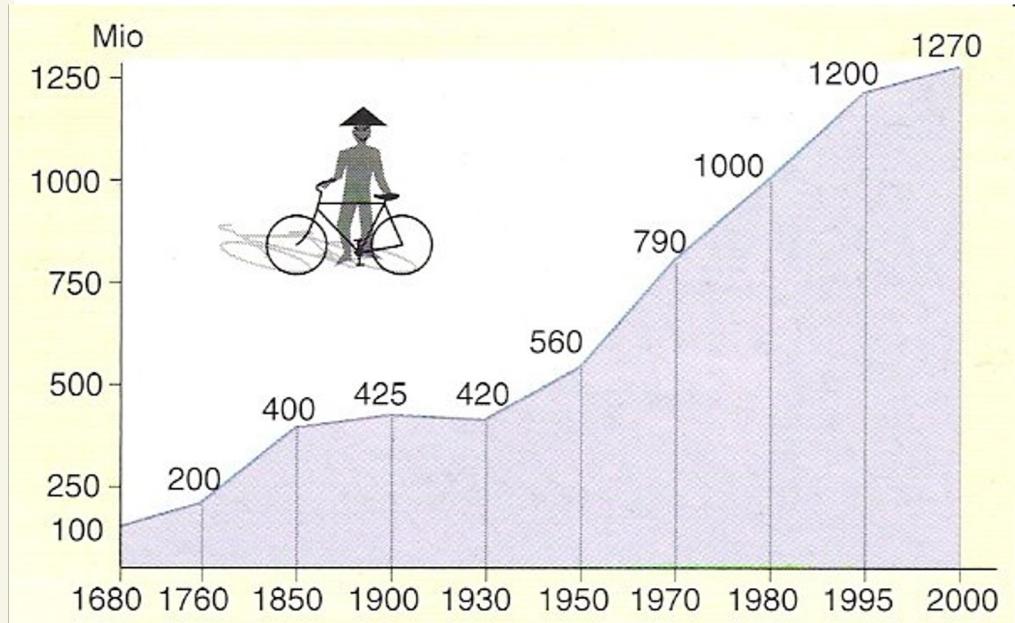
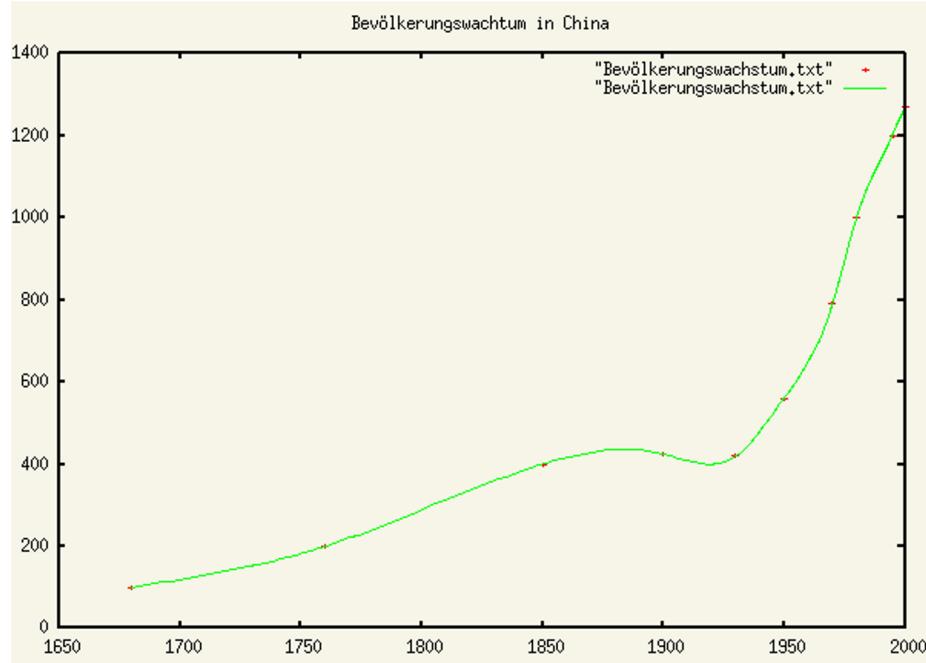
Patterns for days of week seen more easily than views above

Woher kommen die Unterschiede?

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

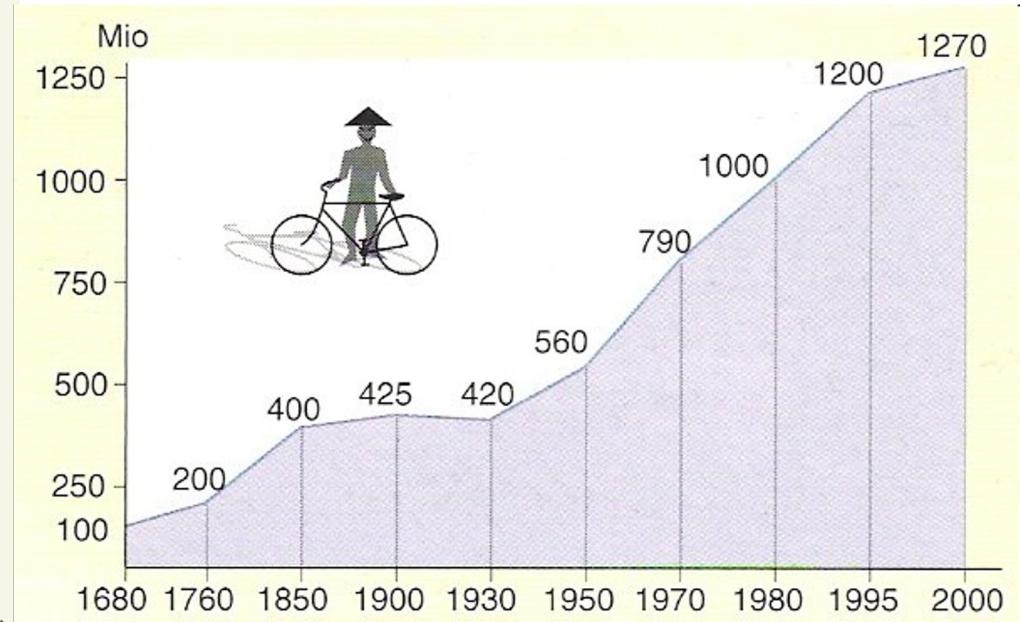
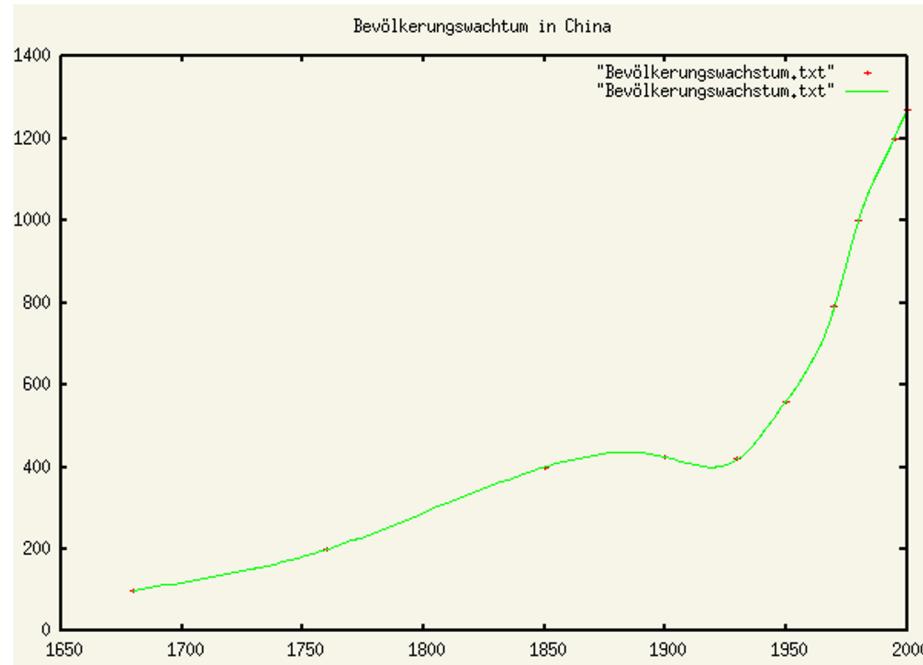


Achtung: Manipulationsgefahr!

u^b

b
UNIVERSITÄT
BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



- > Auf Achsenbeschriftung achten
- > Achsen von eigenen Abbildungen IMMER gut beschriften einschliesslich der Einheiten!

Warum nimmt das Flugzeug diese Route?

u^b

b
UNIVERSITÄT
BERN

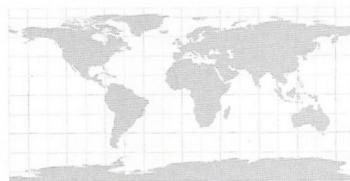
OESCHGER CENTRE
CLIMATE CHANGE RESEARCH



Map projections

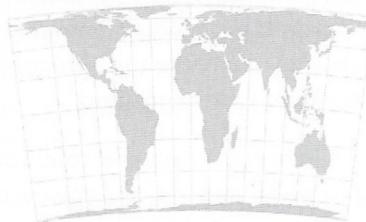
Equirectangular

Typically used for thematic mapping, but doesn't preserve area or angle



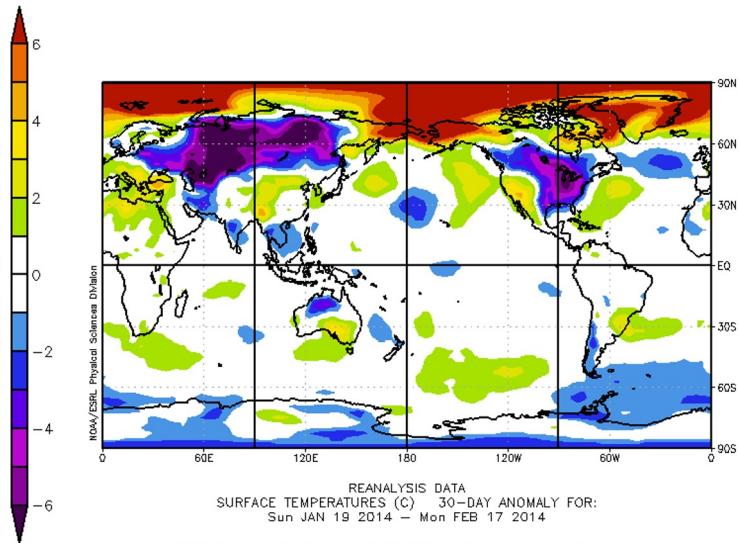
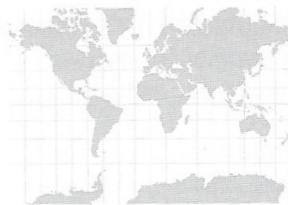
Albers

Scale and shape not preserved; angle distortion is minimal



Mercator

Preserves angles and shapes in small areas, making it good for directions



Lambert conformal conic

Better for showing smaller areas and often used for aeronautical maps.



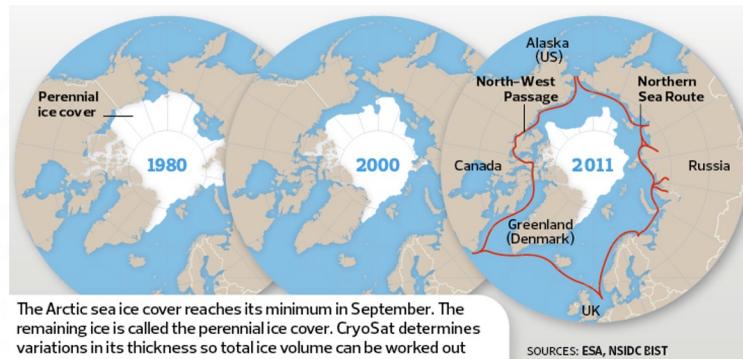
Sinusoidal

Preserves area; useful for areas near the prime meridian



Polyconic

Was often used to show US in the mid-1900s; little distortion in small areas near meridian



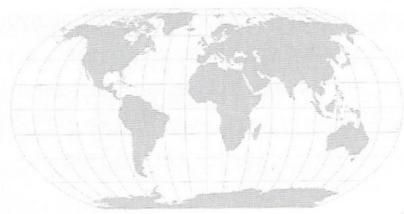
Winkel Tripel

Minimized area, angle, and distance distortion; good choice for world map



Robinson

A compromise between preserving areas and angles; good to show world map



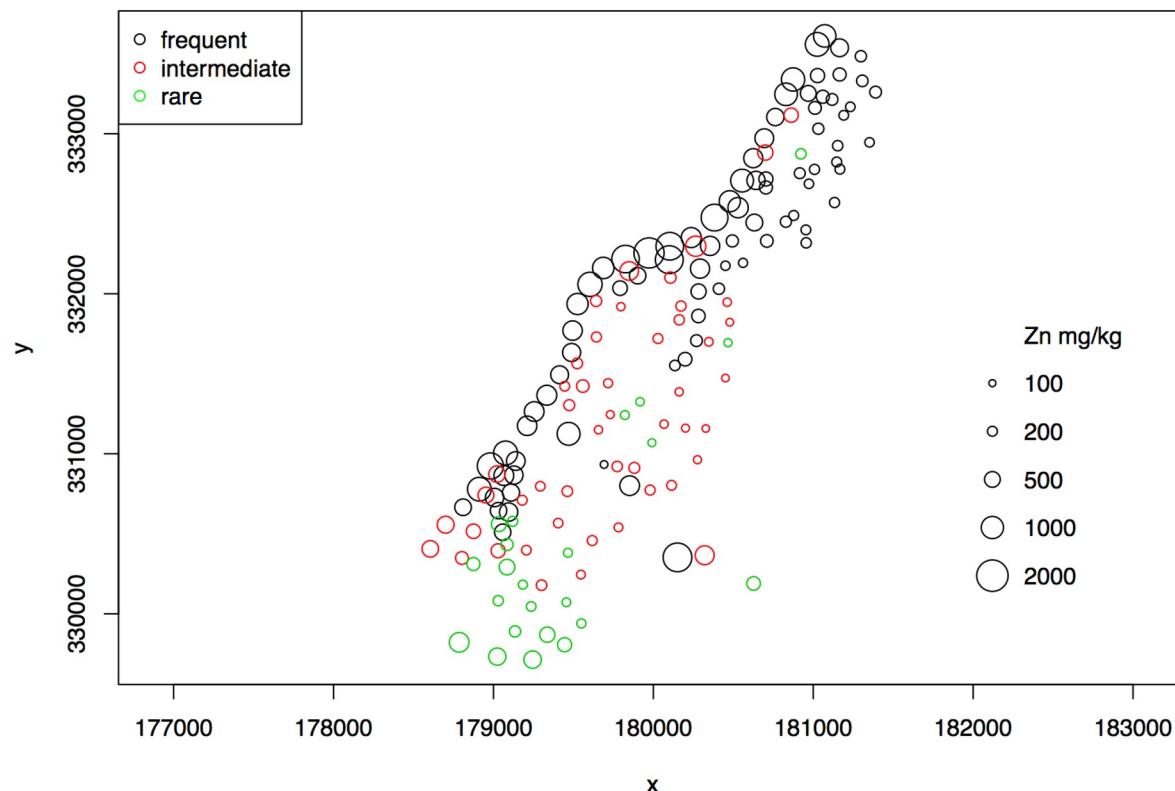
Orthographic

Representing a 3-D object in 2-D, need to rotate to area of interest



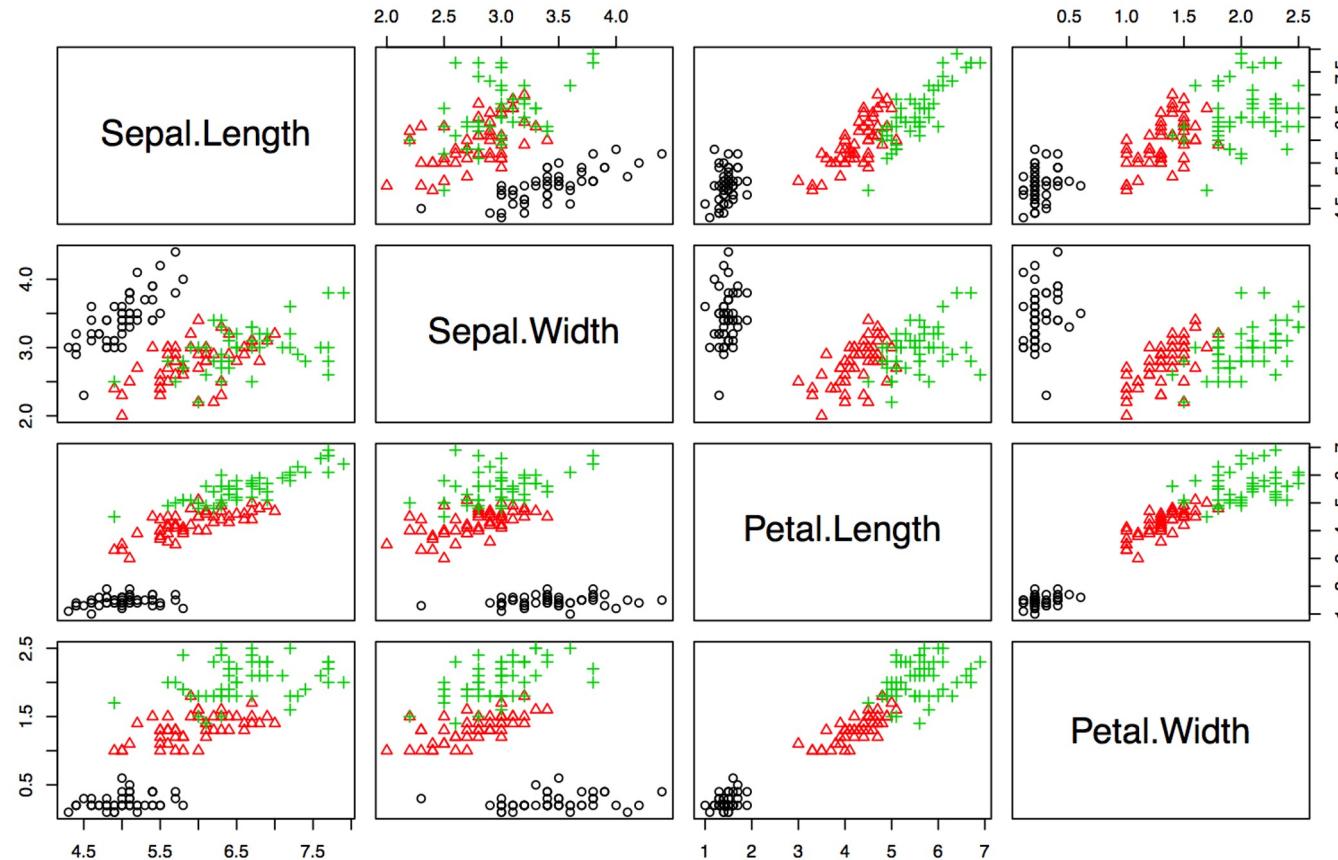
Multivariate Darstellungen

Beispiel Streudiagramm mit Farben und variabler Symbolgrösse in Relation zu Zinkgehalten im Boden

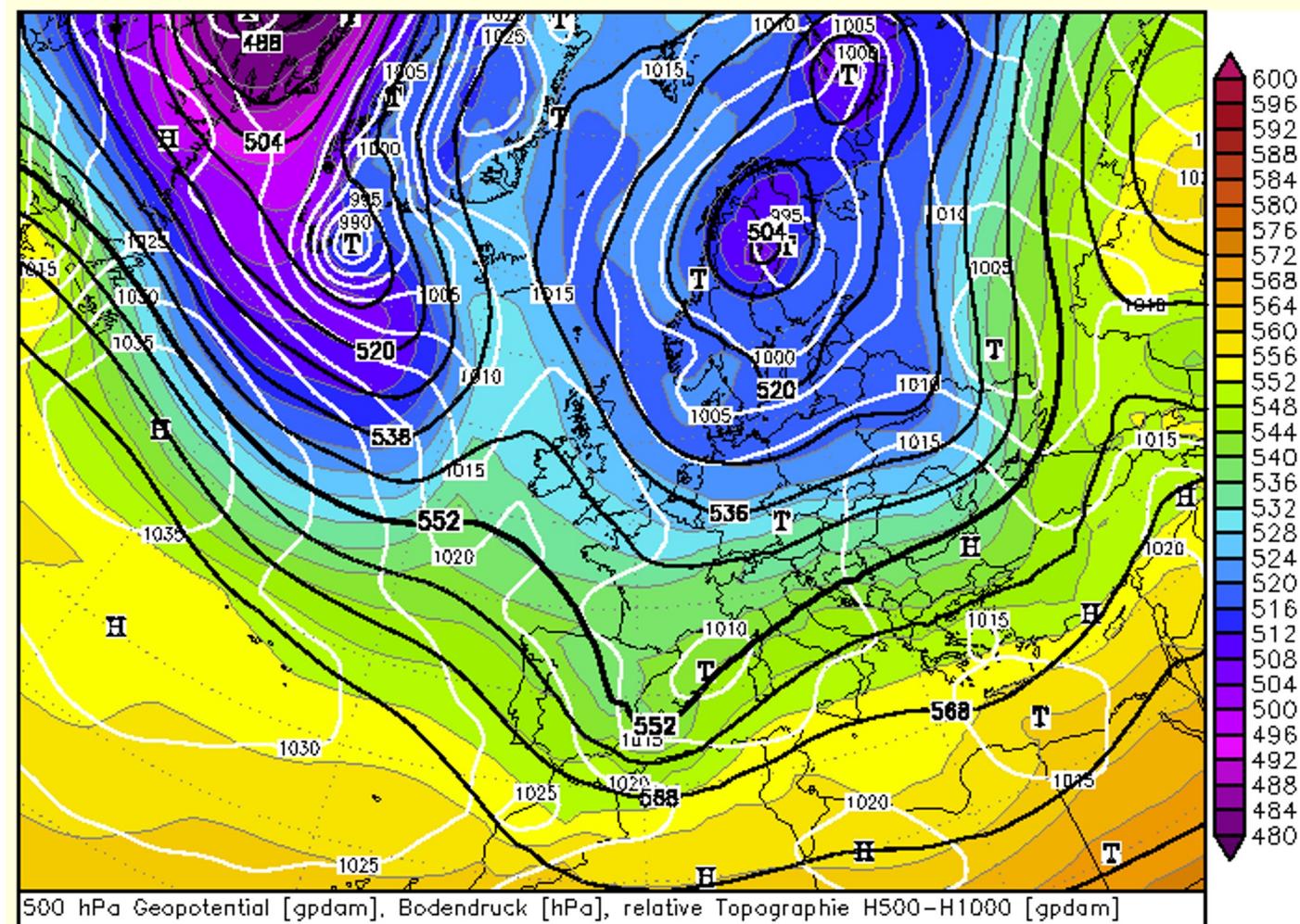


Multivariate Darstellungen

Streudiagrammmmatrix für 4 Variablen und 3 Pflanzenarten

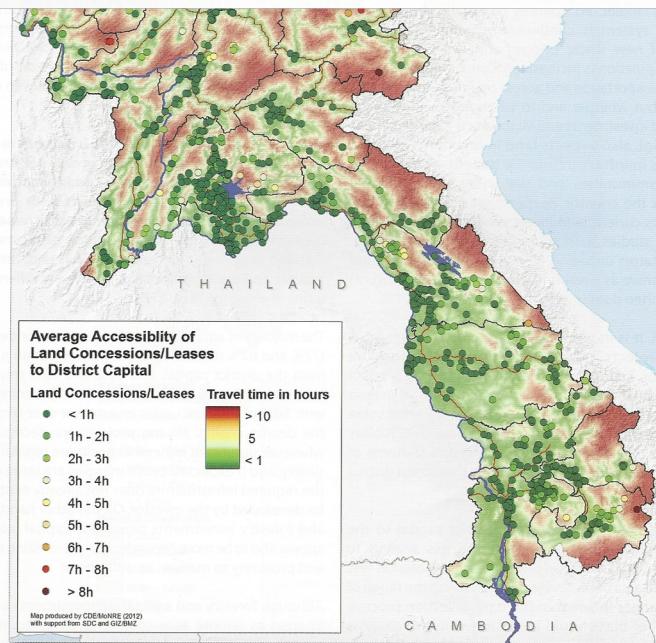
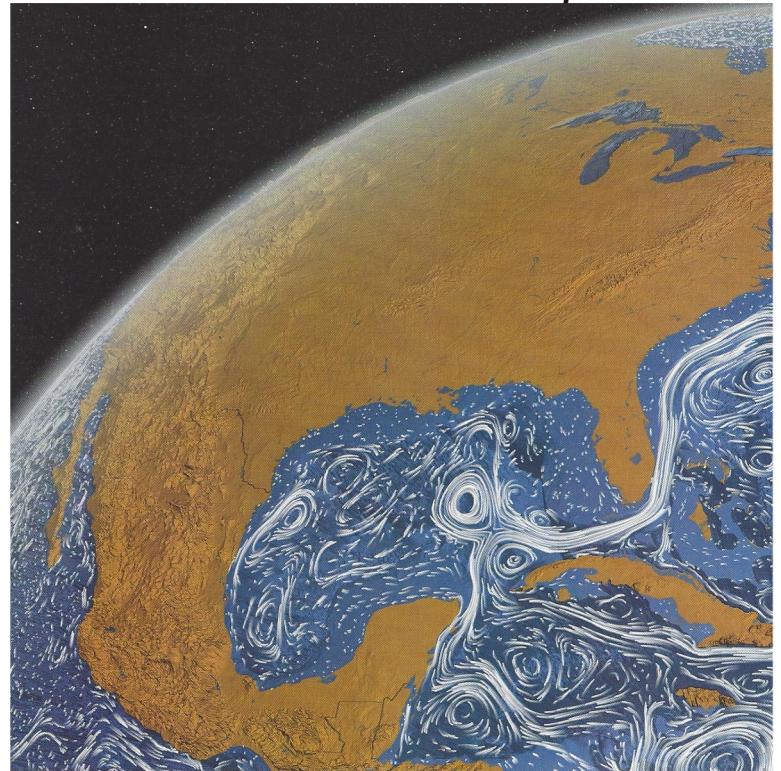
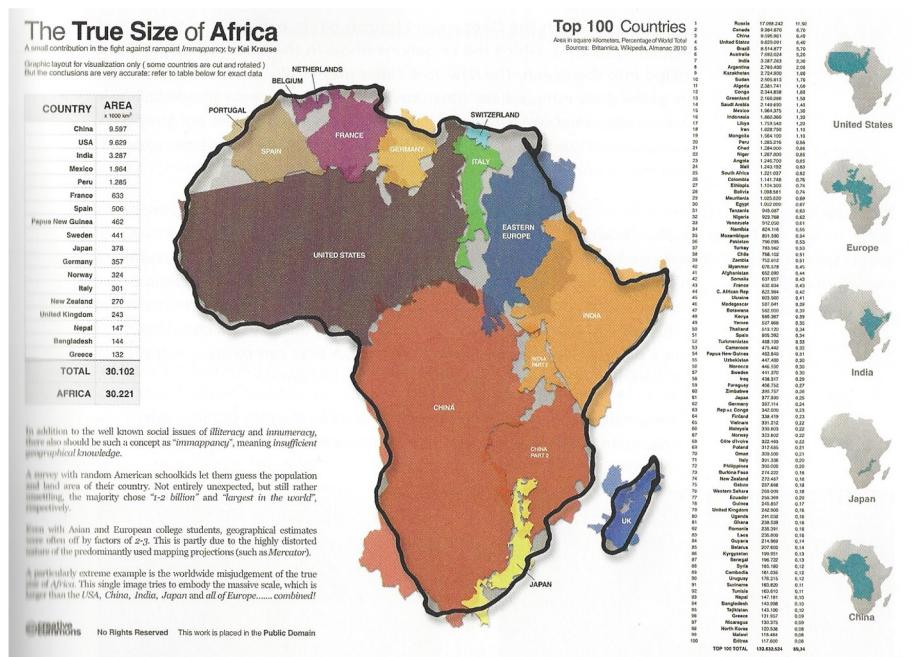


Multivariate Darstellungen



Faktenbezogen oder künstlerisch

thetruesize.com

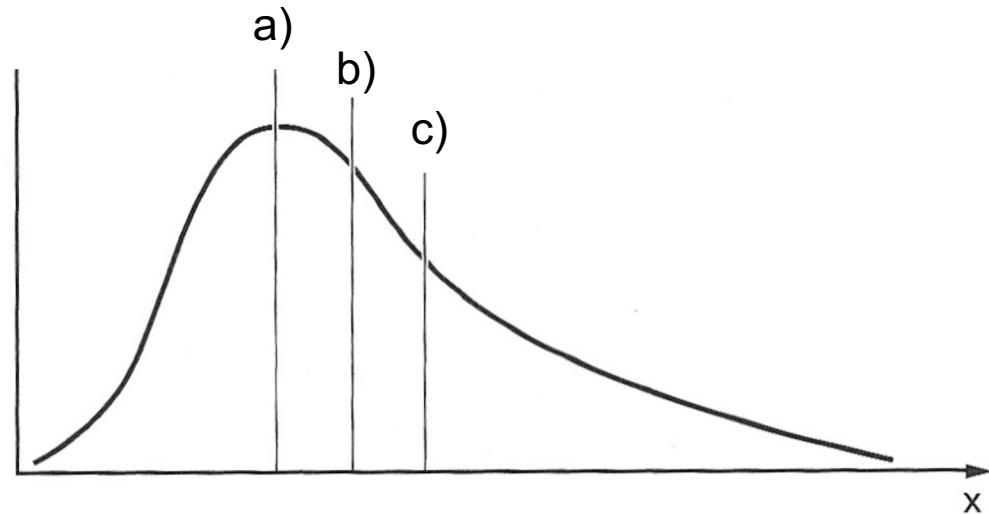


Take-home messages

- > Daten immer zuerst plotten, das Auge erkennt Strukturen, aber auch Fehlwerte sehr gut
- > Art der Abbildung gut auswählen und bei fremden Graphiken vor allem genau auf die Achsen achten

Beispiel Prüfungsfrage

- > Ordne a, b und c zu:
 - Mittelwert
 - Median
 - Modus



- > Welche Streuungsmasse eignen sich für die Daten mit der abgebildeten Verteilung
 - Varianz
 - Standardabweichung
 - Quantile
 - Minimum/Maximum