Daten/graphische Darstellungen

Statistik I

Sommersemester 2009

Daten
- Rohdaten/Konventionen
- Tabellen
  - Häufigkeiten
  - Anteilswerte

Graphische Darstellungen
- Kategoriale Daten
  - Eine Dimension
  - Zwei und mehr Dimensionen
- Kontinuierliche Daten
  - Eine Dimension
  - Zwei und mehr Dimensionen
  - Sonderfall: Kartogramme

Mißbrauch graphischer Darstellungen

Zusammenfassung

Zum Nachlesen

- Gehring/Weins: Kapitel 5
- Agresti/Finlay: Kapitel 3.1

Beispieldatensätze

- European Social Survey (drei Länder)
  - Insgesamt 5 463 Befragte
  - Aus Österreich, Italien, Schweden
  - Alter, Geschlecht, Links-Rechts etc.
  - Individual-/Mikrodaten
- Französische Regionalwahl 2004
  - 96 Départements auf dem französischen Festland
  - Stimmenanteil Front National, Arbeitslosenquote, Zuwanderer etc.
  - Aggregat-/Makrodaten

Was ist eine „Rohdatenmatrix“?

- „Rohe“ (nicht-bearbeitete) Meßwerte → Tabelle
- „Fälle“ (Untersuchungsobjekt = Personen, Länder, Départements etc.) → Zeilen
- „Variablen“ (Eigenschaft = Nationalität, Links-Rechts-Wert etc.) → Spalten

?

	cntry	idno	lrscale	trstplt	trstplc	vote
1.	IT	3183200	6	2	7	yes
2.	SE	202429	6	3	6	yes
3.	AT	602	5	no trust	8	yes
4.	AT	1934	5	6	9	not elig
5.	IT	3583300	left	6	8	yes
6.	IT	3457000	4	7	9	yes
7.	AT	1216	4	4	7	yes
8.	IT	3120400	6	5	7	yes
9.	IT	3029500	left	5	complete	yes

!

Welche Konventionen gibt es?

- In Formeln werden Variablen durch lateinische Kleinbuchstaben (meist vom Ende des Alphabets) abgekürzt: $u, v, w, x, y, \cdots$
- Fälle erhalten eine laufende Nummer („Index“)
- Die Zahl der Fälle wird mit dem Buchstaben n abgekürzt
- Deshalb nimmt der Index ganze Werte zwischen 1 und n an

Welche Konventionen gibt es? II

- Ohne Informationsverlust . . .

	Country	Age	LRscale
1	SE	55	6
2	AT	43	7
3	SE	32	9
4	IT	34	8
5	AT	26	5

Welche Konventionen gibt es? II

- Ohne Informationsverlust ...
- Können *ganze Zeilen* beliebig vertauscht werden

	Country	Age	LRscale
4	IT	34	8
1	SE	55	6
3	SE	32	9
2	AT	43	7
5	AT	26	5

Welche Konventionen gibt es? II

- Ohne Informationsverlust ...
- Können *ganze Zeilen* beliebig vertauscht werden
- Können *ganze Spalten* beliebig vertauscht werden

	Country	LRscale	Age
1	SE	6	55
2	AT	7	43
3	SE	9	32
4	IT	8	34
5	AT	5	26

Welche Konventionen gibt es? II

- Ohne Informationsverlust ...
- Können *ganze Zeilen* beliebig vertauscht werden
- Können *ganze Spalten* beliebig vertauscht werden
- Oder beides

	Country	LRscale	Age
4	IT	8	34
1	SE	6	55
3	SE	9	32
2	AT	7	43
5	AT	5	26

Welche Konventionen gibt es? II

- Ohne Informationsverlust ...
- Können *ganze Zeilen* beliebig vertauscht werden
- Können *ganze Spalten* beliebig vertauscht werden
- Oder beides
- Aber nicht Teile von Spalten/Zeilen

	Country	LRscale	Age
4	IT	8	34
1	SE	9	55
3	SE	6	32
2	7	AT	43
5	AT	5	26

Was sind Tabellen?

- Allgegenwärtiges Hilfsmittel zur Auswertung und Präsentation von Daten
- Besteht aus Zeilen und Spalten
- Meist zweidimensional, aber
  - Eindimensionale Tabellen: Liste
  - Mehrdimensionale Tabellen (Aufteilung in Untertabellen)
- Begrenzte Anzahl von Spalten/Zeilen $\rightarrow$ Spalten/Zeilen entsprechen kategorialen (oder kategorisierten) Variablen
- Grundlage für viele (aber nicht alle) graphischen Darstellungen

Häufigkeitsauszählung

- Einfachste Form der Datenauswertung: Wie häufig kommen die Ausprägungen einer einzigen *kategorialen* Variablen vor?
- Vorgehensweise: Rohdatenmatrix nach den Kategorien der betreffenden Variablen sortieren
- Anschließend Häufigkeiten auszählen

	Country	Age	LRscale
1	SE (3)	55	6
2	AT (1)	43	7
3	SE (3)	32	9
4	IT (2)	34	8
5	AT (1)	26	5

Häufigkeitsauszählung

- Einfachste Form der Datenauswertung: Wie häufig kommen die Ausprägungen einer einzigen *kategorialen* Variablen vor?
- Vorgehensweise: Rohdatenmatrix nach den Kategorien der betreffenden Variablen sortieren
- Anschließend Häufigkeiten auszählen

	Country	Age	LRscale
2	AT (1)	43	7
5	AT (1)	26	5
4	IT (2)	34	8
3	SE (3)	32	9
1	SE (3)	55	6

Häufigkeitsauszählung II

Country	Häufigkeit
AT (1)	2
IT (2)	1
SE (3)	2

Was sind „relative Häufigkeiten“?

- Absolute Häufigkeiten meistens relativ uninteressant
- Relative Häufigkeiten: Absolute Häufigkeit durch Zahl der Fälle

Country	absolute Häufigkeit	relative Häufigkeit
AT (1)	2	
IT (2)	1	
SE (3)	2	
Σ	5	

Der griechische Großbuchstabe Σ bedeutet „Summe“ (mehr dazu nächste Woche)

Was sind „relative Häufigkeiten“?

- Absolute Häufigkeiten meistens relativ uninteressant
- Relative Häufigkeiten: Absolute Häufigkeit durch Zahl der Fälle

Country	absolute Häufigkeit	relative Häufigkeit
AT (1)	2	$\frac{2}{5} = 0,4$
IT (2)	1	$\frac{1}{5} = 0,2$
SE (3)	2	$\frac{2}{5} = 0,4$
Σ	5	1

Der griechische Großbuchstabe Σ bedeutet „Summe“ (mehr dazu nächste Woche)

Was sind Prozente?

- Prozent = relative Häufigkeit $\times$ 100

Country	absolute Häufigkeit	relative Häufigkeit	Prozent
AT (1)	2	0,4	
IT (2)	1	0,2	
SE (3)	2	0,4	
Σ	5	1	

Was sind Prozente?

- Prozent = relative Häufigkeit $\times$ 100

Country	absolute Häufigkeit	relative Häufigkeit	Prozent
AT (1)	2	0,4	=40%
IT (2)	1	0,2	=20%
SE (3)	2	0,4	=40%
Σ	5	1	100%

Wie unterscheiden sich Prozente von Prozent*punkten*?

- Relative Häufigkeit = absolute Häufigkeit / n
- Prozente = Relative Häufigkeit $\times$ 100
- Prozentpunkte = Differenz zwischen Prozentsätzen
  - Beispiel SPD-Anteil
  - BTW 1994 36,4%, BTW 1998 40,9% der gültigen Zweitstimmen
  - Verbesserung um 4,5 Prozentpunkte
  - Entspricht einer Zunahme um (4,5/36,4)*100 $\approx 12,4$ Prozent
- Das heißt: Veränderungen zwischen Prozentsätzen können als absolute Veränderungen (in Prozentpunkten) oder wiederum als prozentuale Veränderungen ausgedrückt werden

Exkurs: Typische Prozentuierungen in der Wahlforschung

- Prozentuierung auf *Wahlberechtigte*
  - abgegebene Stimmen/Wahlberechtigte = Wahlbeteiligung → Anteil der Nichtwähler
  - Stimmen für eine Partei/Wahlberechtigte erlaubt Vergleich der Mobilisierungsleistung unabhängig von Wahlbeteiligung
- Prozentuierung auf *abgegebene Stimmen*
  - Nur interessant, um den Anteil der ungültigen Stimmen zu berechnen
- Prozentuierung auf *gültige Stimmen*
  - Wichtig für Sitzverteilung im Parlament
  - Unterscheidet sich in Deutschland kaum von Prozentuierung auf abgegebene Stimmen
  - (Frankreich, Belgien etc.)
- Sonderfall: Prozentuierung auf *gültige und berücksichtigungsfähige Stimmen*

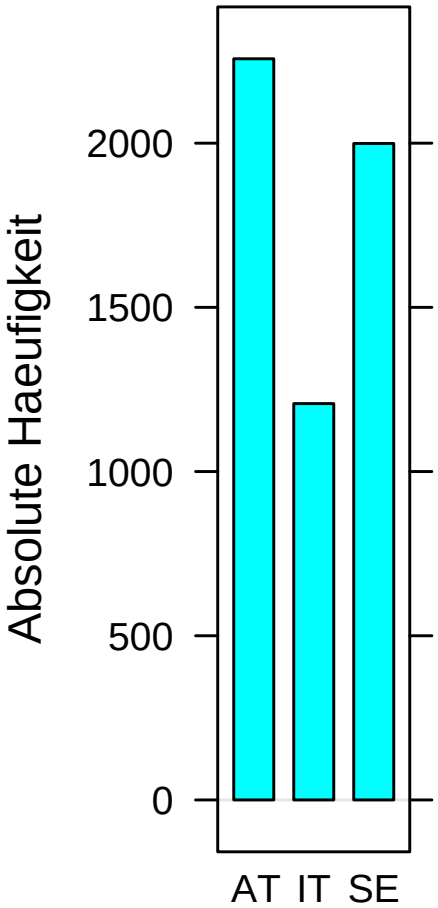
Was sind graphische Darstellungen?

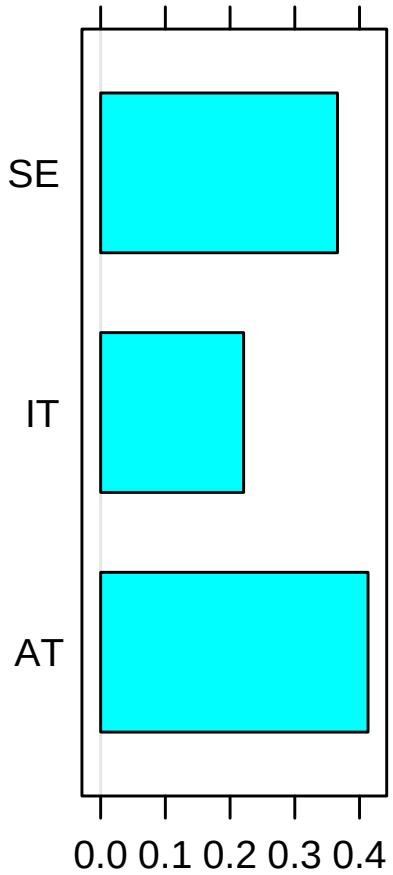
- Vor allem für kleinere Datensätze eigenständige Form der induktiven Analyse
- (Sehr) große Datensätze $\rightarrow$ Probleme
- Häufiger: Veranschaulichung tabellarischer und anderer Analysen
  - Ein Bild sagt mehr ...
  - Mißbrauch, Irreführung, überflüssige Darstellungen (eye candy)
- Vermeiden Sie nach Möglichkeit (überflüssige) dreidimensionale Darstellungen
- Möglichst klare Darstellung („viel Information pro Linie“)
- Konventionen
  - Waagerechte Achse = x-Achse
  - Senkrechte Achse = y-Achse

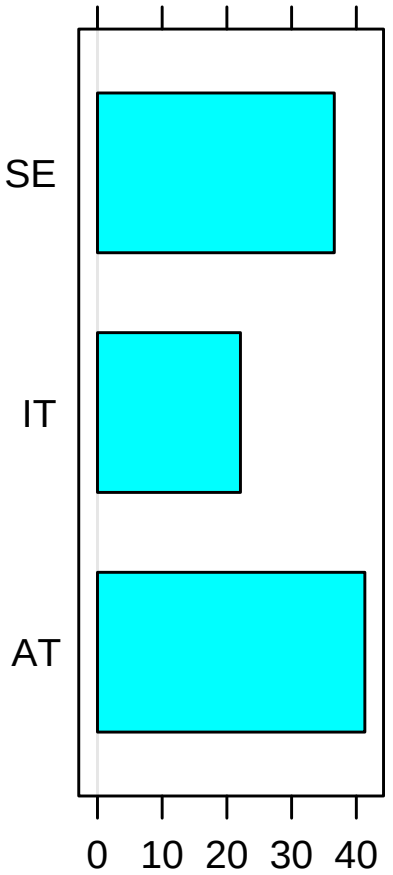
Balken- und Säulendiagramme

- Sind äquivalent (Unterschied nur in der Leserichtung des Diagramms)
- Können für nominal- und ordinalskalierte Daten verwendet werden
- Bei ordinalen Variablen muß die Reihenfolge der Kategorien in der Grafik erhalten bleiben
- Absolute Häufigkeiten, relative Häufigkeiten oder Prozente

Balken/Säulen: Nationalität





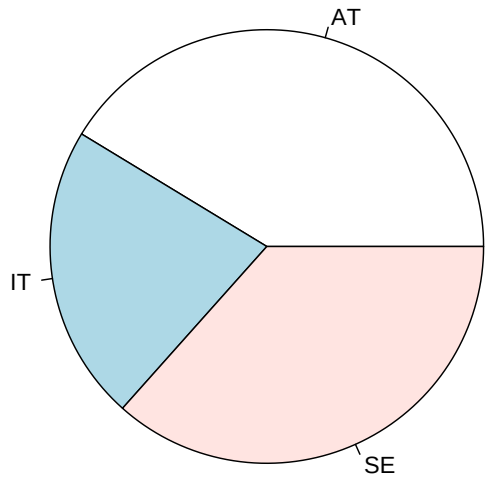


Kreis-/Tortendiagramme

- Darstellung von relativen Häufigkeiten/Prozenten
- Haben (unverdient?) schlechten Ruf
- Für Laien scheinbar anschaulich

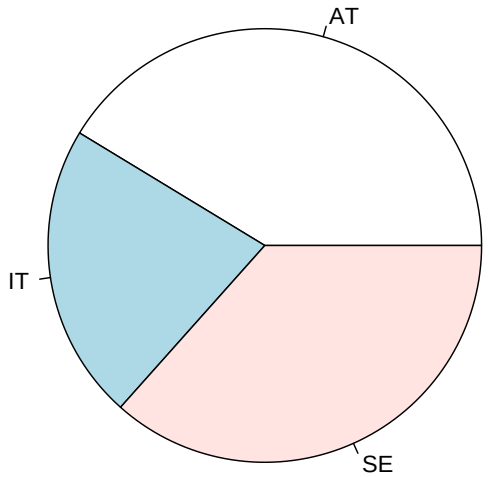
Kreis-/Tortendiagramme

- Darstellung von relativen Häufigkeiten/Prozenten
- Haben (unverdient?) schlechten Ruf
- Für Laien scheinbar anschaulich



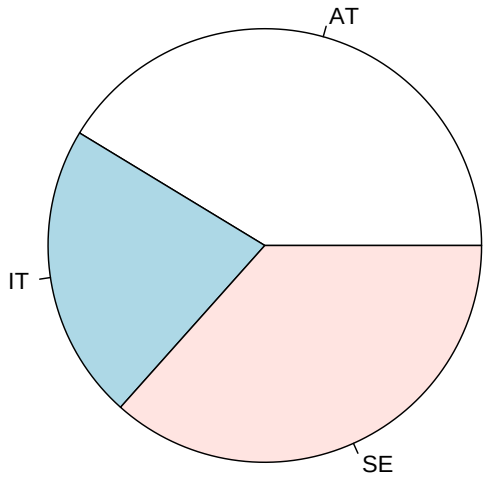
Kreis-/Tortendiagramme

- Darstellung von relativen Häufigkeiten/Prozenten
- Haben (unverdient?) schlechten Ruf
- Für Laien scheinbar anschaulich
- Fläche/Winkel entspricht Anteil $\longrightarrow$ Wahrnehmung häufig verzerrt



Kreis-/Tortendiagramme

- Darstellung von relativen Häufigkeiten/Prozenten
- Haben (unverdient?) schlechten Ruf
- Für Laien scheinbar anschaulich
- Fläche/Winkel entspricht Anteil $\longrightarrow$ Wahrnehmung häufig verzerrt
  - AT $\approx 2 \times$ IT?
  - AT $>$ IT?



Peinliche Pie-Charts

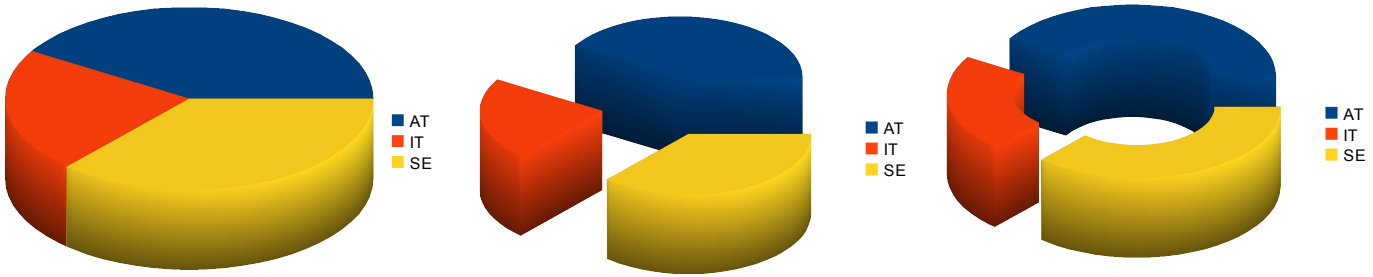
- Dreidimensionale Darstellungen verschärfen die Probleme

Peinliche Pie-Charts

- Dreidimensionale Darstellungen verschärfen die Probleme

Tabelle1

AT	2257
IT	1207
SE	1999

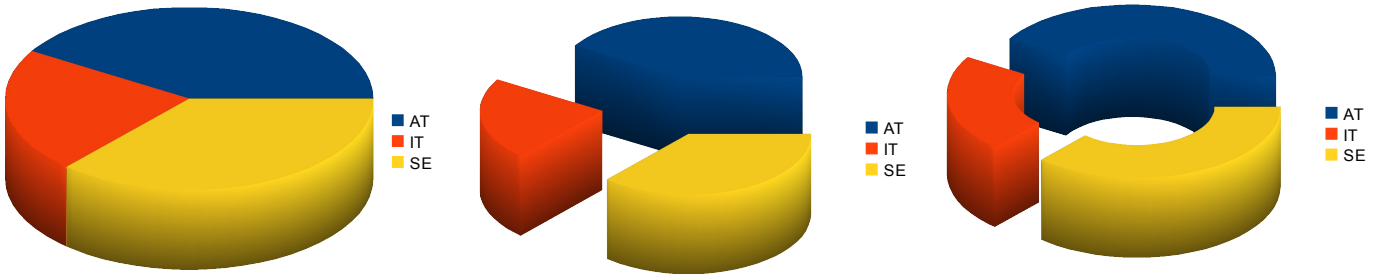


Peinliche Pie-Charts

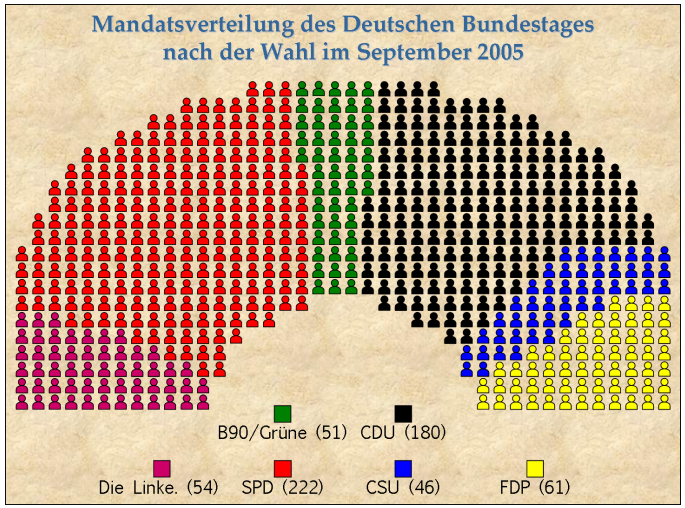
- Dreidimensionale Darstellungen verschärfen die Probleme
- Finger weg davon!

Tabelle1

AT	2257
IT	1207
SE	1999

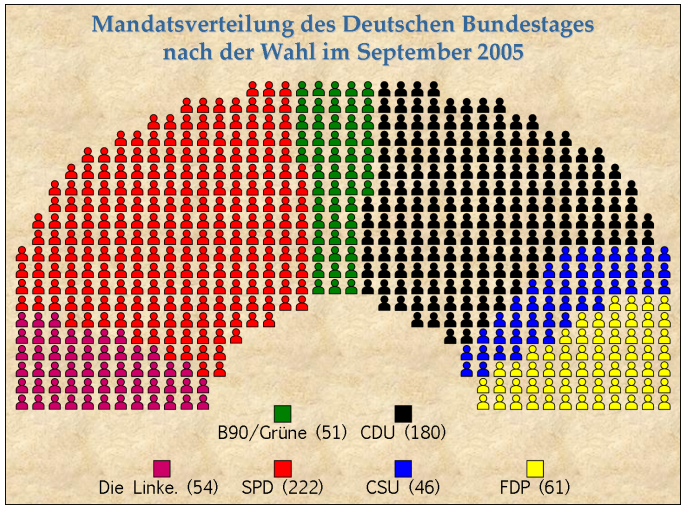


Halbkreise: Gut gemeint, schlecht gemacht



- Stimmenanteile von Parteien werden oft als Halb- oder Drei-Viertelkreis dargestellt
- Spiegelt Sitzordnung im Parlament wieder
- Völlige Verwirrung: $\frac{1}{4}$ Kreis = 50%
- Zusätzliche Komplikation: Halbkreis nicht völlig rund
- Wie groß ist die Regierungsmehrheit?

Halbkreise: Gut gemeint, schlecht gemacht



- Stimmenanteile von Parteien werden oft als Halb- oder Drei-Viertelkreis dargestellt
- Spiegelt Sitzordnung im Parlament wieder
- Völlige Verwirrung: $\frac{1}{4}$ Kreis $=$ 50%
- Zusätzliche Komplikation: Halbkreis nicht völlig rund
- Wie groß ist die Regierungsmehrheit?
- 73% $\times$ 180 Grad $\approx$ 130 Grad, verteilt auf drei Blöcke

Kategoriale Variablen: Zwei und mehr Dimensionen

- Vertrauen in das nationale Parlament . . .

Kategoriale Variablen: Zwei und mehr Dimensionen

- Vertrauen in das nationale Parlament . . .
- Nach Ländern

Kategoriale Variablen: Zwei und mehr Dimensionen

- Vertrauen in das nationale Parlament ...
- Nach Ländern
- Und nach Geschlecht

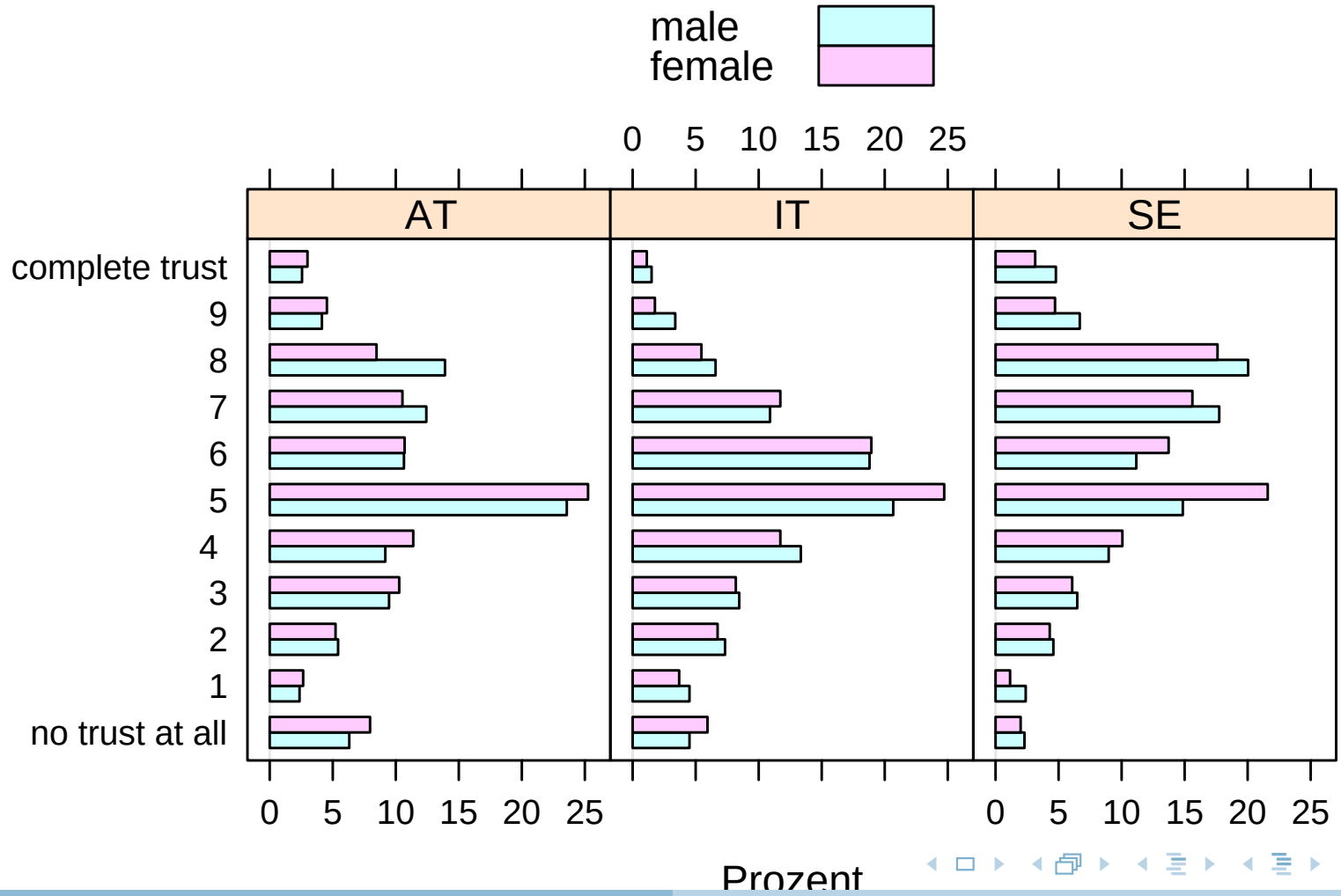
Kategoriale Variablen: Zwei und mehr Dimensionen

- Vertrauen in das nationale Parlament . . .
- Nach Ländern
- Und nach Geschlecht
- Unterschiedliche Verteilungen zwischen Ländern
- Unterschiede innerhalb von Ländern

Politisches Vertrauen nach Geschlecht und Land

	Male			Female		
	AT	IT	SE	AT	IT	SE
no trust at all	6	5	2	8	6	2
1	2	5	2	3	4	1
2	5	7	5	5	7	4
3	9	8	6	10	8	6
4	9	13	9	11	12	10
5	24	21	15	25	25	22
6	11	19	11	11	19	14
7	12	11	18	11	12	16
8	14	7	20	8	5	18
9	4	3	7	5	2	5
complete trust	3	2	5	3	1	3
Σ	100	100	100	100	100	100

Politisches Vertrauen nach Geschlecht und Land



Darstellung von kontinuierlichen Daten

- Sind politikwissenschaftliche Daten jemals wirklich kontinuierlich?
  - Beschränkter Meßbereich (Vertrauen von 0 10)
  - Beschränkte Zahl von Meßwerten (ganze Zahlen $0, 1, 2, \cdots 10$)
    $\rightarrow$ beschränkte Genauigkeit oder „Auflösung“ der Messung
  - Alter?
- „Kontinuierlich“ eine *konzeptuelle* Eigenschaft der Messung
- Plausible Annahme?
- Ab fünf/sieben verschiedenen Meßwerten meist zu rechtfertigen

Histogramme, Polygonzüge, Dichteschätzung I

- Für intervall- und ratioskalierte Daten geeignet
- Relevant sind sowohl die x-Achse (waagerechte Achse) als auch die y-Achse (senkrechte Achse) des Diagramms
  - X-Achse: Ausprägung der Variablen
  - Y-Achse: Häufigkeit/Wahrscheinlichkeit der Ausprägung
- Kontinuierliche Merkmale $\longrightarrow$ Rechtecke werden direkt nebeneinander gezeichnet
- Häufigkeit/Wahrscheinlichkeit von Meßwerten aus einem bestimmten *Intervall* wird durch die *Fläche* repräsentiert
- Intervalle sollten gleich breit sein, um Verwirrung zu reduzieren
- Intervallbreite sollte „ansprechend“ sein (wichtig bei kleinen Datensätzen)

Was ist ein Intervall?

- Kontinuierliche Variable wird für Darstellung im Histogramm in Bereiche eingeteilt
- Intervall = Wertebereich
  - In einem „geschlossenen“ Intervall sind die Grenzwerte des Bereichs mitenthalten → eckige Klammern
  - In einem „offenen“ Intervall sind die Grenzwerte *nicht* enthalten → runde Klammern
  - Ein „halboffenes Intervall“ enthält *einen* der beiden Grenzwerte

Beispiel: Alter von Erwerbstätigen (16-66)

Intervall	Wertebereich
$[16; 31)$	$16 \leqslant \text{Alter} < 31$
$[31; 46)$	$31 \leqslant \text{Alter} < 46$
$[46; 51)$	$46 \leqslant \text{Alter} < 51$
$[51; 66]$	$51 \leqslant \text{Alter} \leqslant 66$

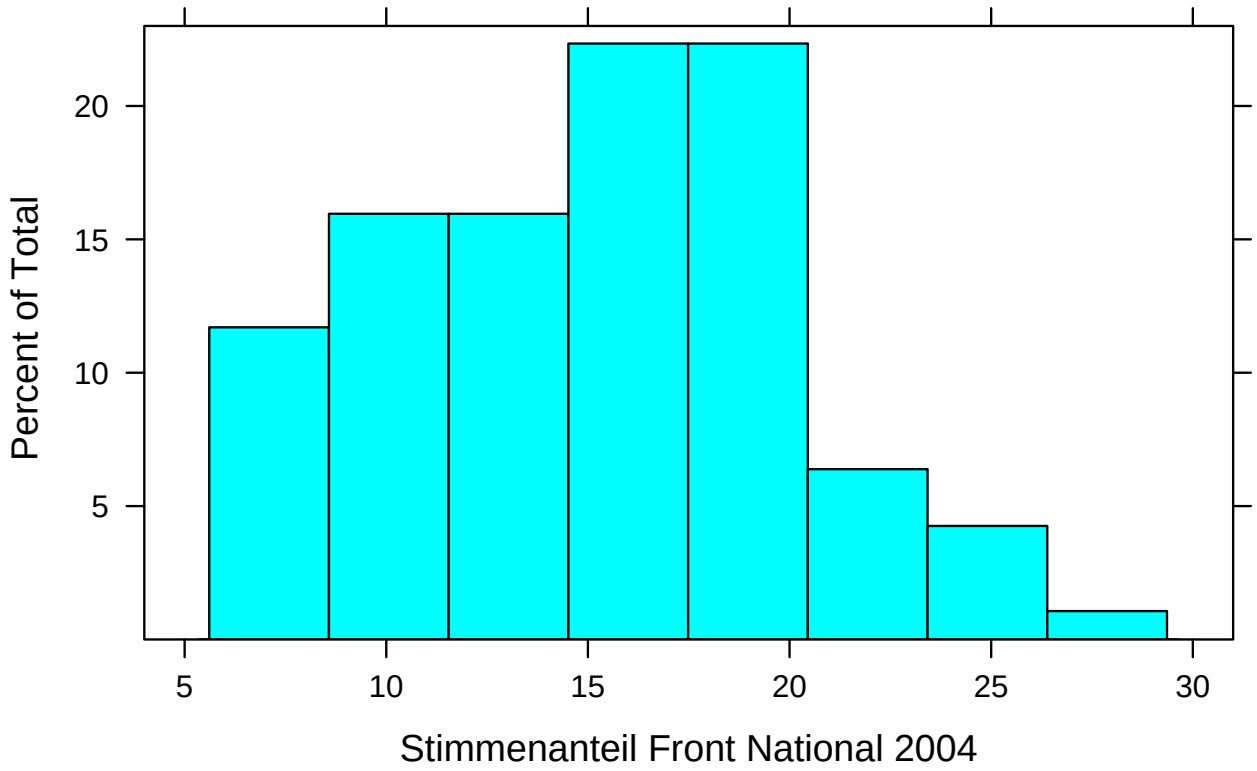
Unterstützung für den Front National 2004: Histogramm I

- Regionalwahl 2004
- Départements als Stimmbezirke
- Relativ komplexes Verhältniswahlsystem
- 94 Départements auf dem französischen Festland
- Stimmenanteil für den FN in der ersten Runde in Prozent
  - Kontinuierlich
  - Auf Wertebereich 6,5-28,5 beschränkt

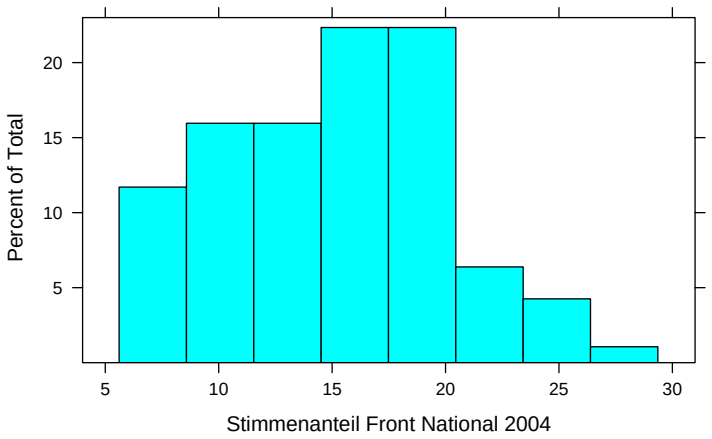
Fall	Département	FN 2004
1	Ain	20.5
2	Aisne	24.1
3	Allier	10.8
4	Alpes de Haute Provence	15.6
...	...	...

Unterstützung für den Front National 2004: Histogramm II

Intervallgrenzen: 5.6, 8.6, . . . 29.4; Breite: 2.97 Prozentpunkte



Was können wir hier sehen?



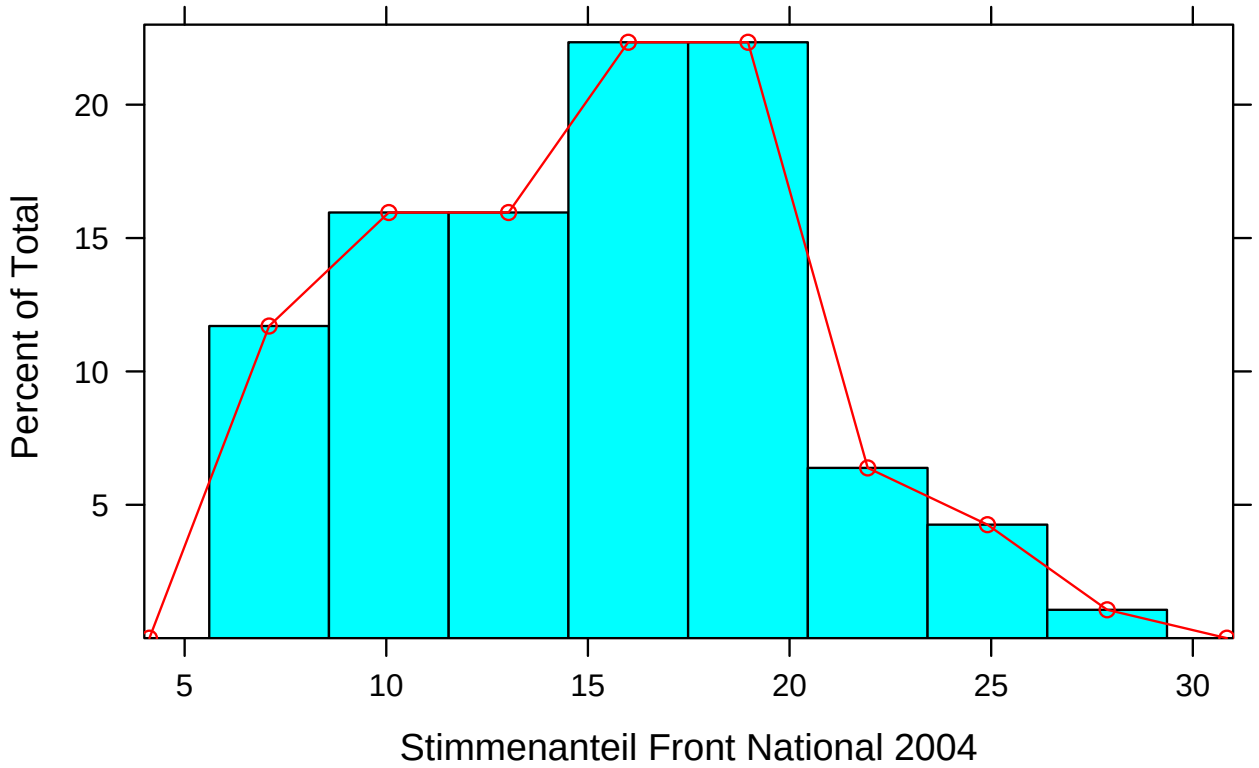
- In knapp der Hälfte aller Départements (44%) erhält der FN zwischen 14,5 und 20,5% der Stimmen
- In rund 11% der Départements sind es mehr als 20,5%

- Nur in 12% der Départements sind es weniger als 8,9% der Stimmen
- Nirgends sind es weniger als 5,6%
- Die Verteilung ist eingipflig, aber schief (mehr Fälle auf der linken Seite – weniger Stimmen. Mehr dazu nächste Woche)

Polygonzüge

1. Warum?
   - Variable auf x-Achse kontinuierlich $\longrightarrow$ „Dichteschätzung“
   - (Boxen unendlich schmal machen, Linie zwischen realen Datenpunkten extrapolieren)
   - Polygonzug: Gleichzeitige Darstellung von zwei Histogrammen, z.B. Stimmenanteil 2004/1998
2. Wie?
   - Mittelpunkt der „Deckel“ der Boxen miteinander verbinden
   - Linie zu den Mittelpunkten der (gedachten) Boxen links und rechts vom Histogramm herunterziehen
   - Fläche unter Polygonzug entspricht Fläche des Histogramms entspricht 100%

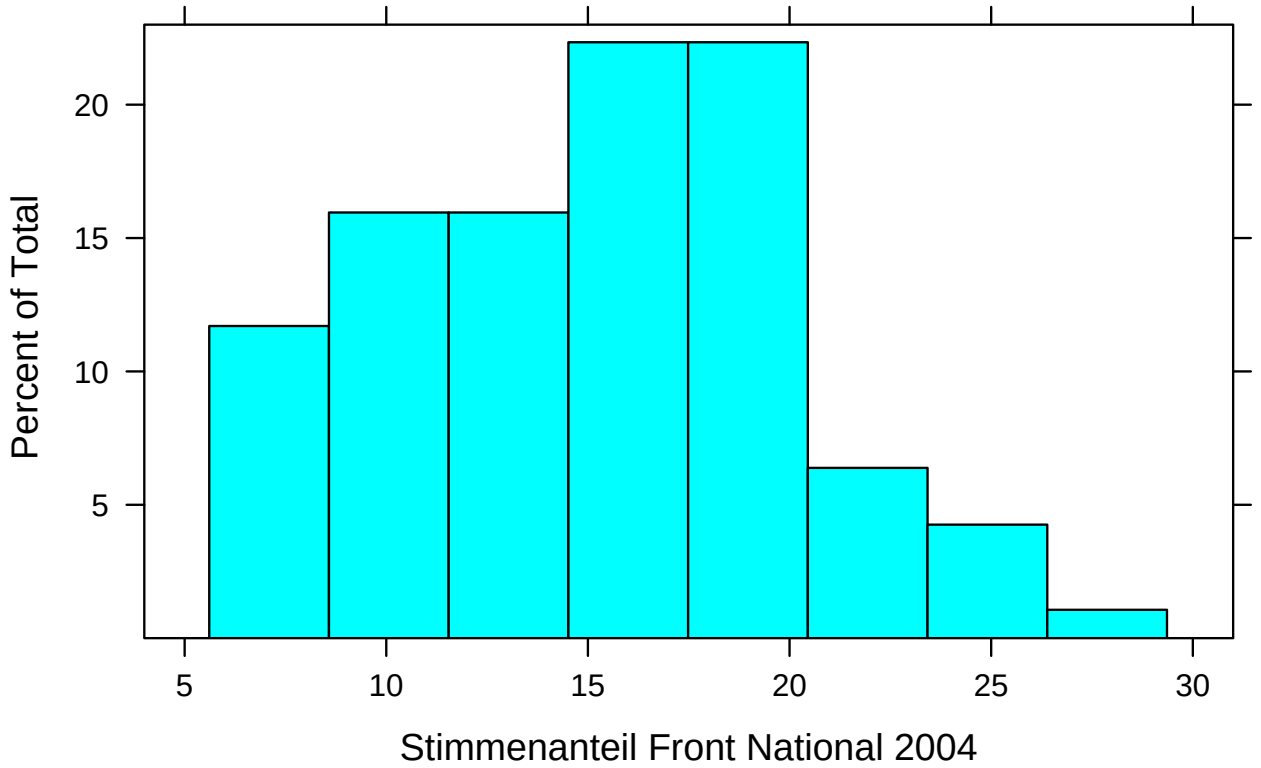
Polygonzüge



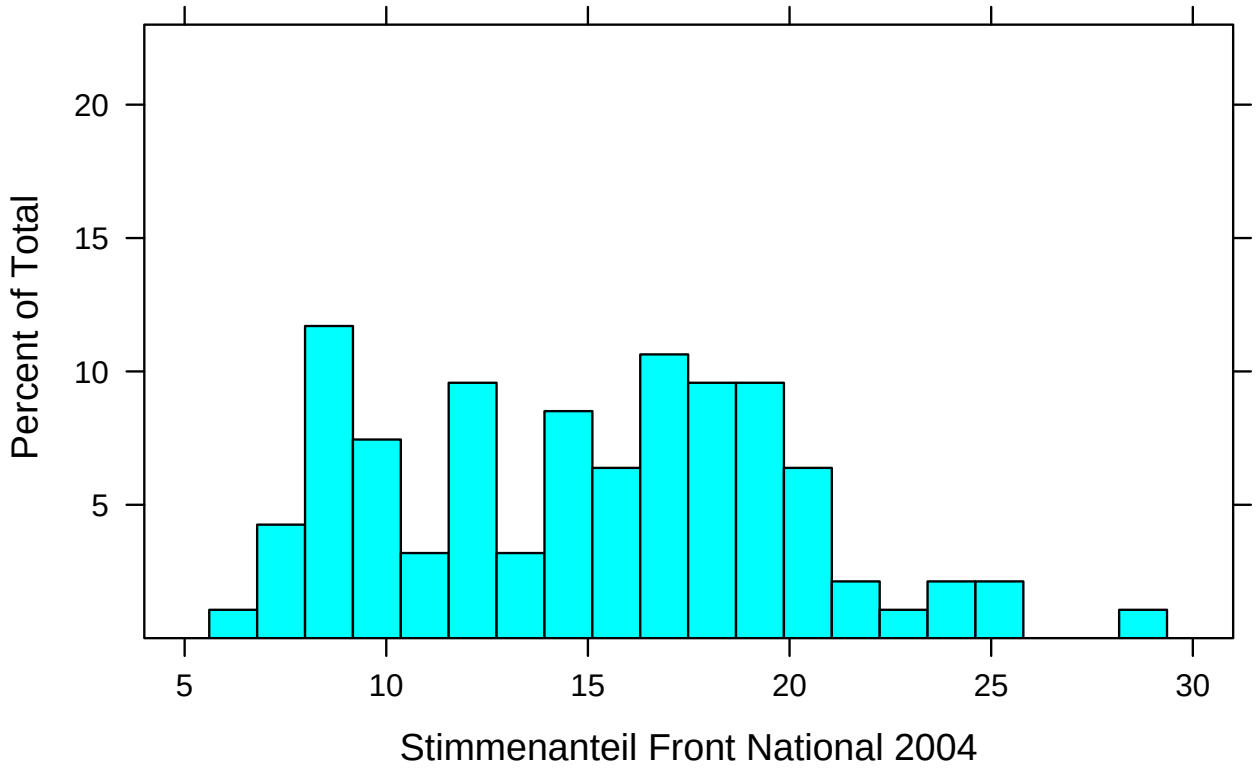
Probleme mit Histogrammen?

1. „Sprünge“ im Histogramm
2. Aussehen des Histogramms hängt von oberer/unterer Schranke ab
3. Aussehen des Histogramms hängt von (willkürlicher) Klassenbreite ab

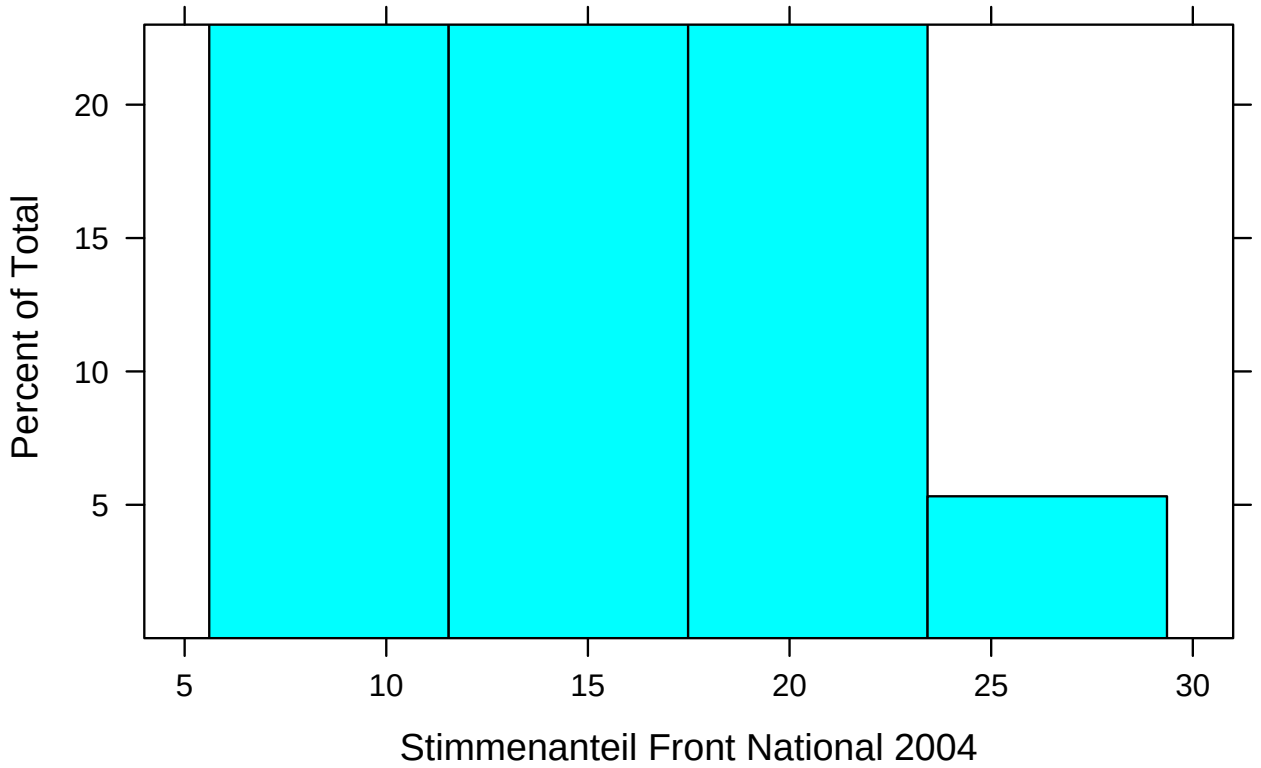
Probleme mit Histogrammen?



Probleme mit Histogrammen?



Probleme mit Histogrammen?



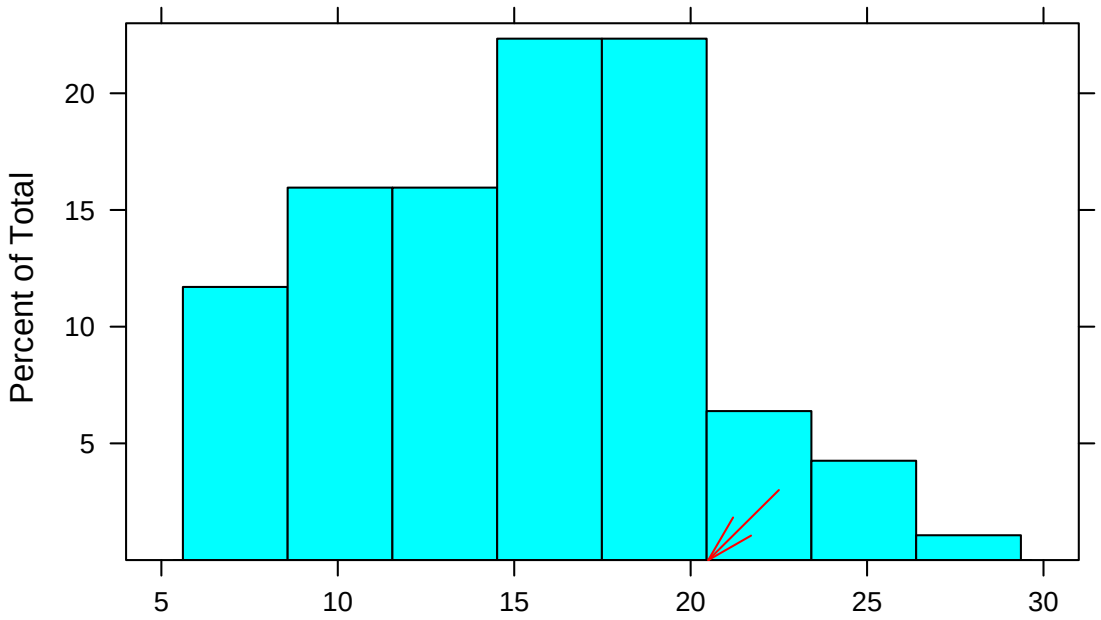
Dichteschätzung als Alternative

	Département	FN 2004
1	Ain	20,51
2	Haut-Rhin	20,50
3	Territoire de Belfort	20,41

- Kontinuierliche Variable können unendlich viele Werte annehmen $\rightarrow$ „Sprünge“ im Histogramm irreführend.
- In welche von mehreren Boxen ein realer kontinuierlicher Wert fällt, ist in gewisser Weise zufällig
- Wahrscheinlichkeit eines Front National Ergebnisses im Bereich 20,41-20,51 fällt nicht schlagartig ab

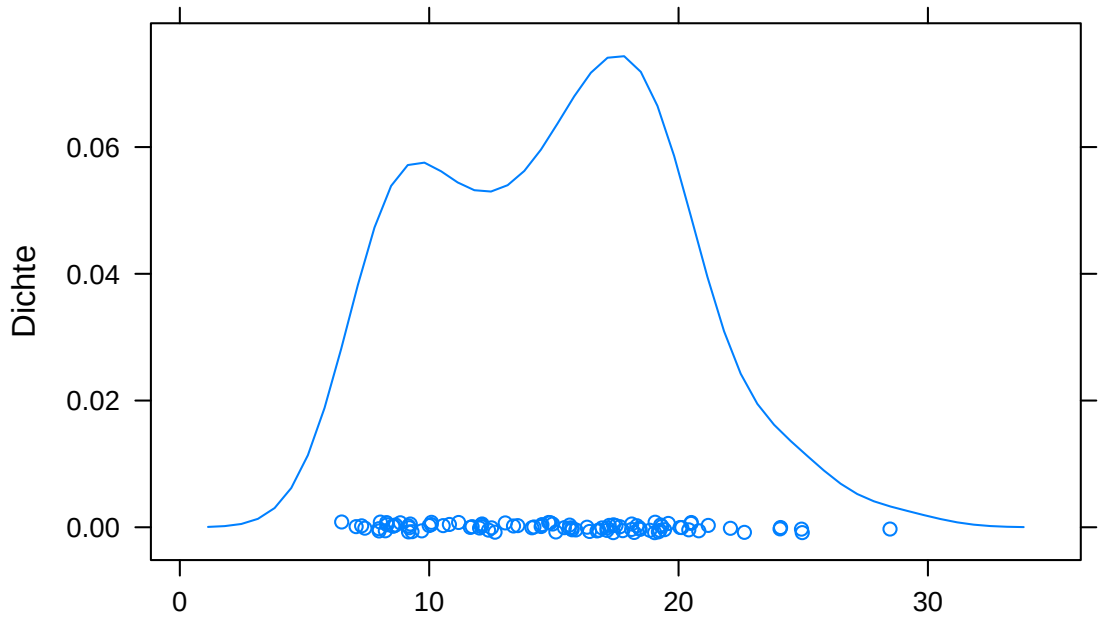
Dichteschätzung als Alternative

	Département	FN 2004
1	Ain	20,51
2	Haut-Rhin	20,50
3	Territoire de Belfort	20,41

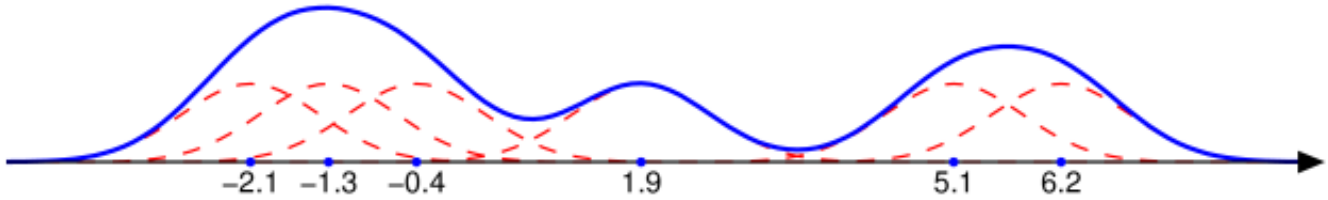


Dichteschätzung als Alternative

	Département	FN 2004
1	Ain	20,51
2	Haut-Rhin	20,50
3	Territoire de Belfort	20,41



Dichteschätzung als Alternative II

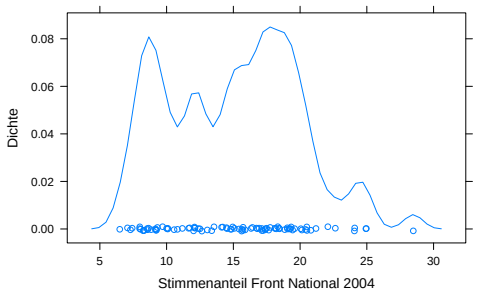


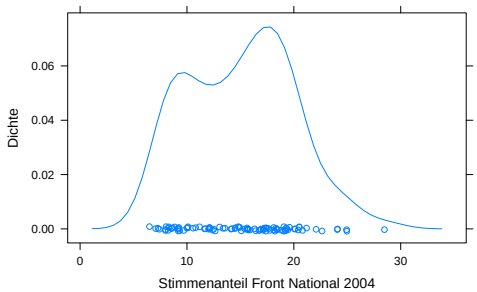
http://en.wikipedia.org/wiki/Kernel_density_estimation

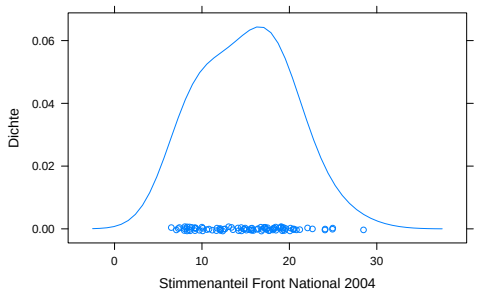
- Jeder Fall wird durch eine kleine Glockenkurve repräsentiert
- Gipfel bei tatsächlichem Wert
- Abfallende Verteilung, weil Wert mit zusehends geringerer Wahrscheinlichkeit auch in der Nachbarschaft liegen könnte
- Individuelle Kurven werden überlagert $\longrightarrow$ Schätzung für die Gesamtverteilung

Dichteschätzung als Alternative III

- Löst Problem der „Sprünge“
- Problem der oberen/unteren Schranke besteht in ähnlicher Weise
- Problem der Klassenbreite $\longrightarrow$ Problem der Bandbreite



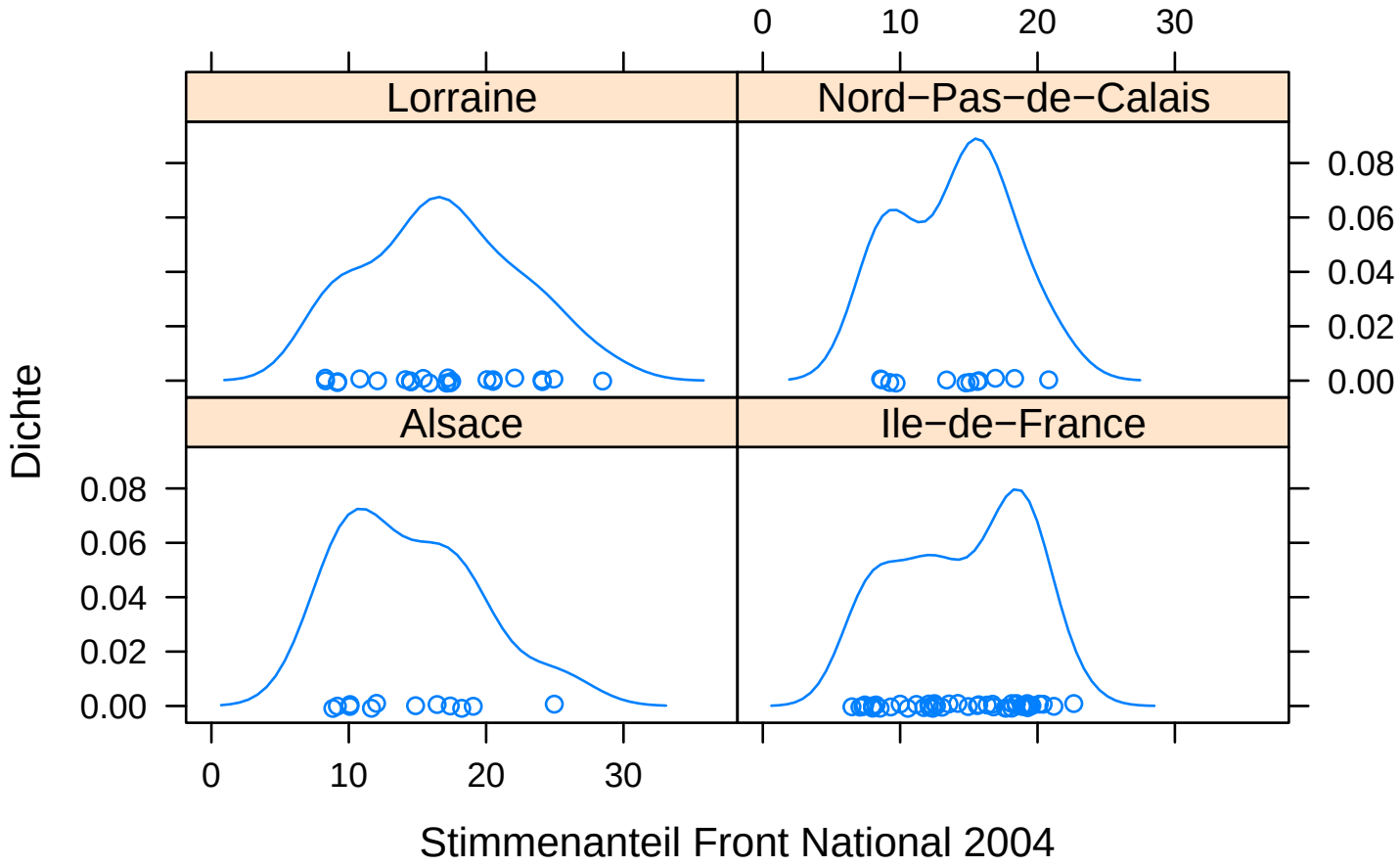




Zwei und mehr Dimensionen

1. Eine kontinuierliche, eine kategoriale Variable (z.B. Region)
   - Gleiches Prinzip wie bei zwei kategorialen Variablen
   - Panels

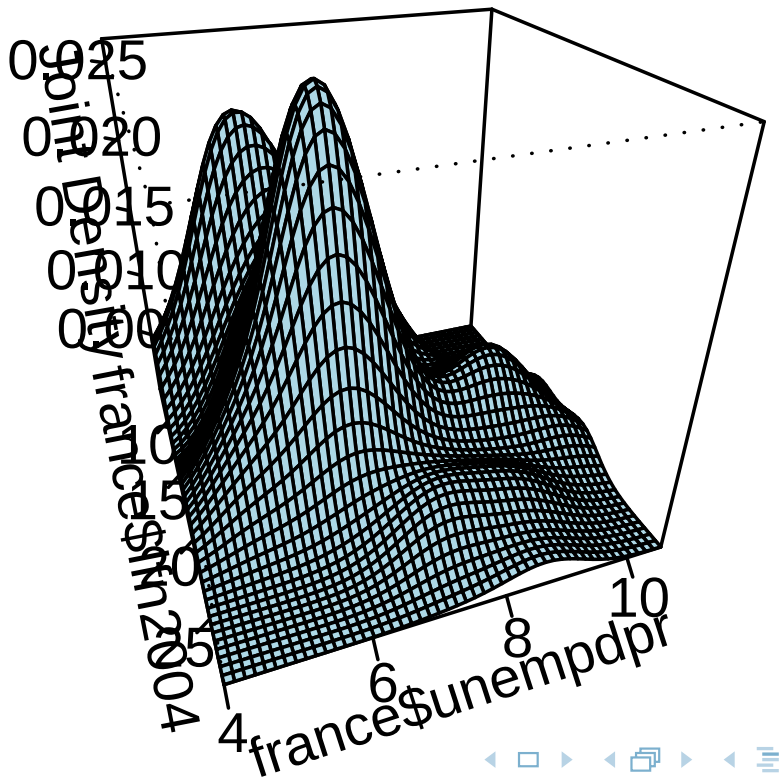
FN-Erfolge in vier französischen Regionen



Zwei und mehr Dimensionen

1. Eine kontinuierliche, eine kategoriale Variable (z.B. Region)
   - Gleiches Prinzip wie bei zwei kategorialen Variablen
   - Panels
2. Zwei kontinuierliche Variablen (z.B. Stärke FN und Arbeitslosenquote im Département)
   - Dreidimensionale Darstellung
   - Oder topographische Darstellung (wie eine Wanderkarte)
   - Oder tomographischer Plot ...
   - *Am besten und einfachsten: Streudiagramm (scatterplot)*

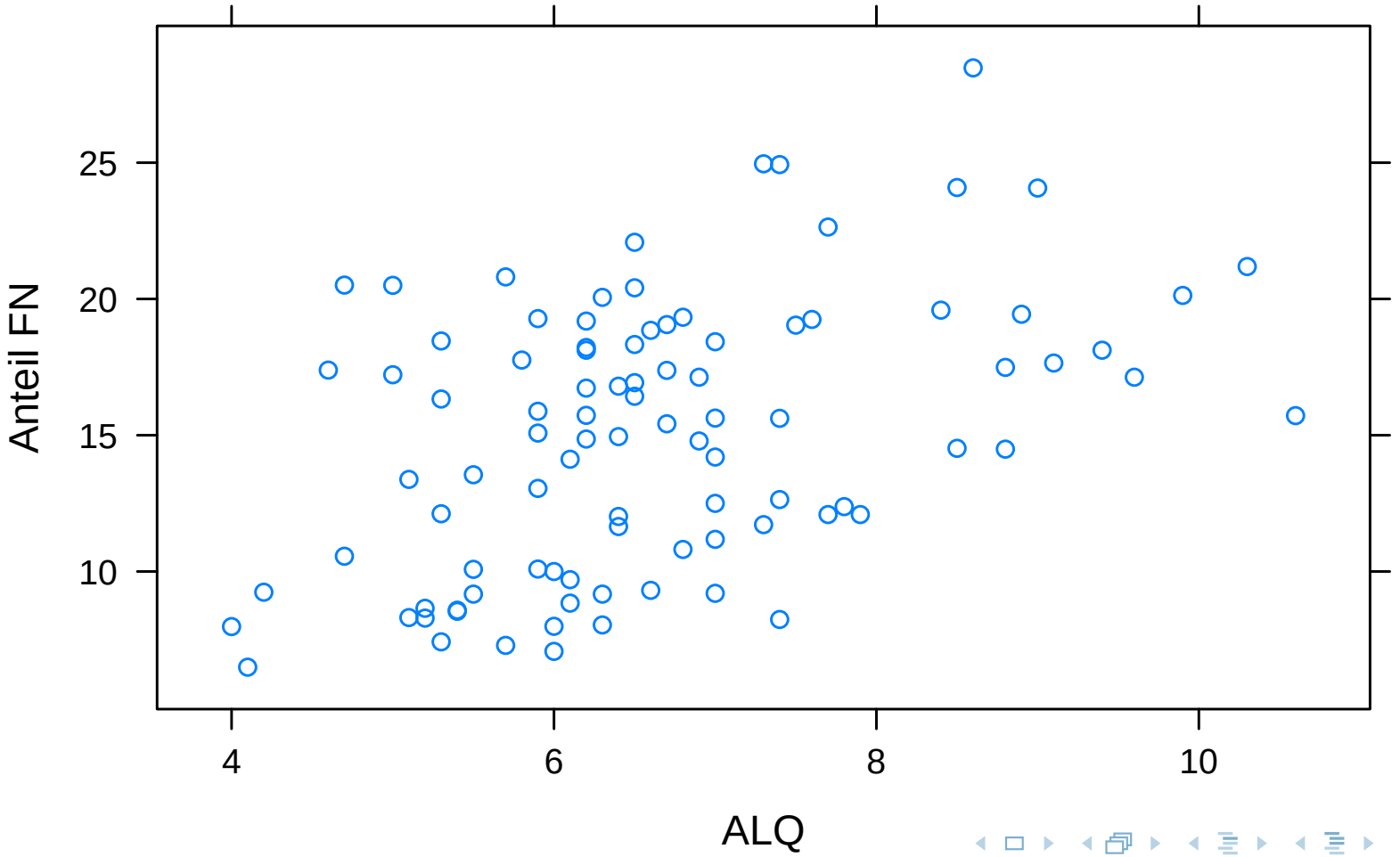
FN-Erfolge und Arbeitslosigkeit



Streudiagramme/Scatterplots

- Standarddiagramm für zwei kontinuierliche Variablen
- Jede Beobachtung wird durch Punkt oder anderes Symbol in zwei Dimensionen repräsentiert
- Zeigt (oder suggeriert) bivariaten Zusammenhang
- Funktioniert nicht für sehr große Datensätze
- Problematisch, wenn Variablen nur wenige Ausprägungen haben (pseudo-kontinuierliche Daten $\longrightarrow$ jitter)

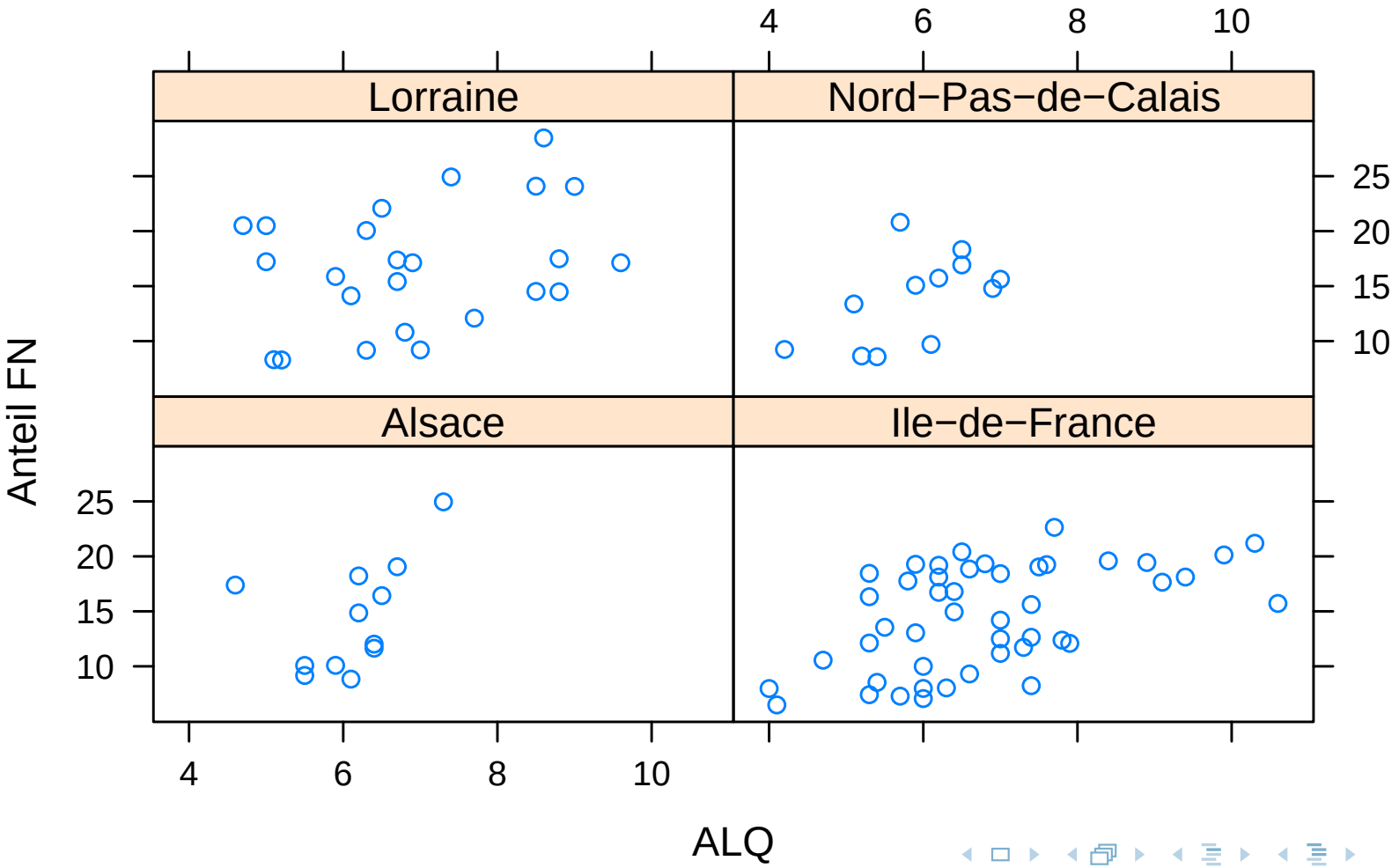
FN-Erfolge und Arbeitslosigkeit



Streudiagramme/Scatterplots

- Standarddiagramm für zwei kontinuierliche Variablen
- Jede Beobachtung wird durch Punkt oder anderes Symbol in zwei Dimensionen repräsentiert
- Zeigt (oder suggeriert) bivariaten Zusammenhang
- Funktioniert nicht für sehr große Datensätze
- Problematisch, wenn Variablen nur wenige Ausprägungen haben (pseudo-kontinuierliche Daten $\longrightarrow$ jitter)
- Durch Panels Erweiterung auf drei oder mehr Dimensionen möglich

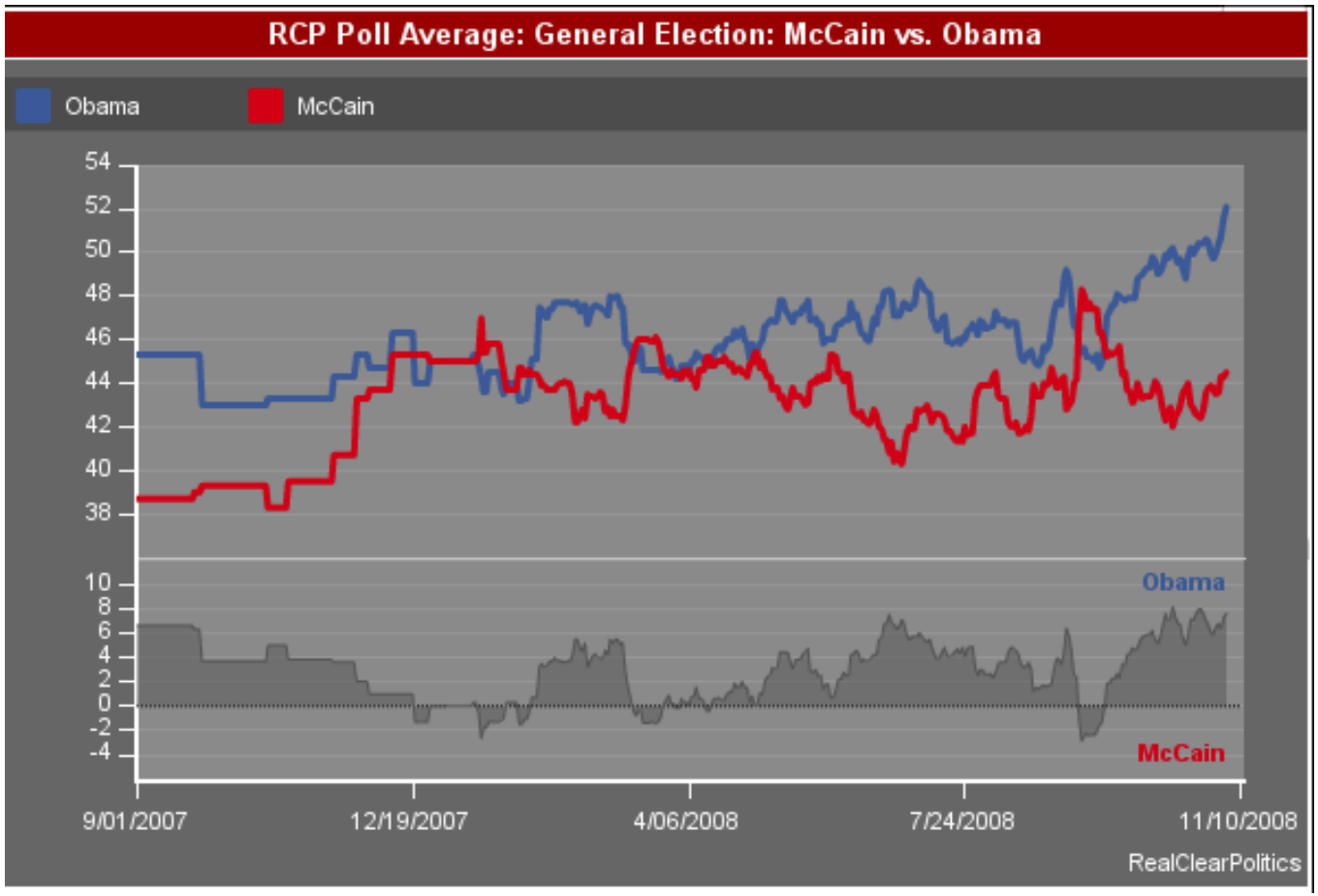
FN-Erfolge und Arbeitslosigkeit in vier Regionen



Sonderfall des Streudiagramms: Zeitreihe

- Messung *am selben Objekt* wird über Zeit wiederholt
- Ein Datenpunkt für jede Messung
- Zeit auf x-Achse
- Mehrere Zeitreihen können in einem Diagramm kombiniert werden, wenn Maßstab vergleichbar ist

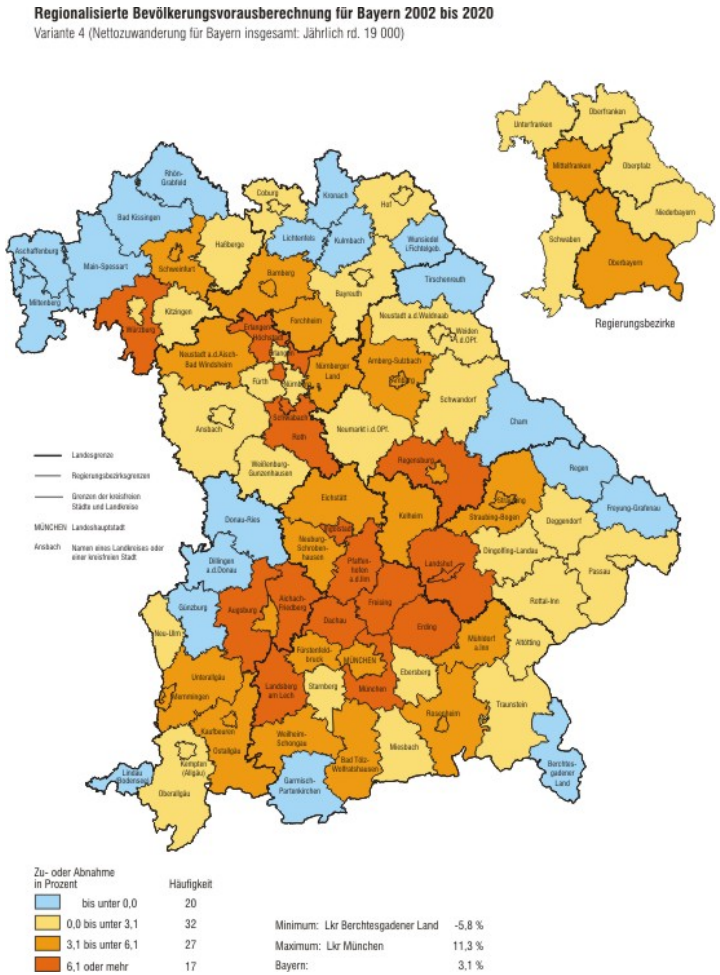
Wahlabsicht Obama vs. McCain



Kartogramme

- Sonderform der zwei- oder mehrdimensionalen Darstellung
- Eine kontinuierliche oder kategoriale Variable
- Zweite Variable: räumliche Position

Bevölkerungsprognose Bayern

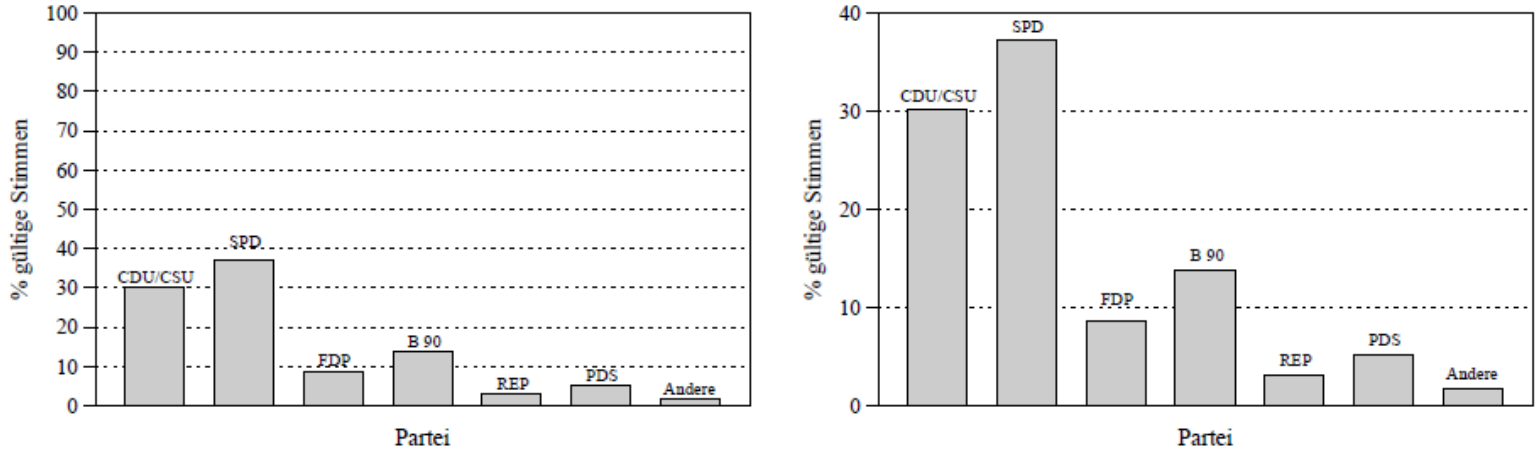


Manipulation graphischer Darstellungen

- Grafiken sollen der Veranschaulichung, nicht der Manipulation dienen
- Keine Histogramme, Polygonzüge, Dichteschätzungen für nominal- und ordinalskalierte Daten
- Maßstab so wählen, daß Unterschiede erkennbar sind, aber nicht übertrieben werden
- Unterbrechungen in Zeitreihen kennzeichnen
- „Künstlerische“ Darstellungen (Figuren etc. vermeiden)
- Bei vergleichbaren Grafiken identischen Maßstab wählen
- Zwei Zeitreihen innerhalb einer Grafik sollten immer denselben Maßstab haben
- y-Achse muß bei null beginnen. Ansonsten Unterbrechung der Achse markieren

Manipulation durch Wahl des Maßstabs

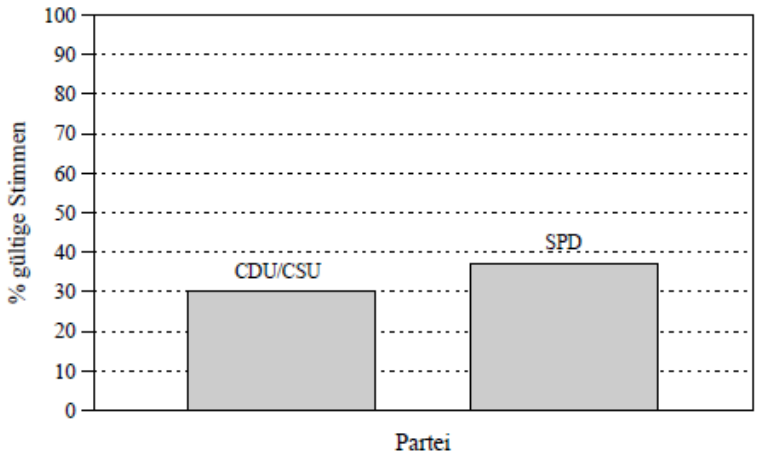
Abbildung 5.7: Wahlabsicht bei Veränderung des y-Achsen-Maßstabes

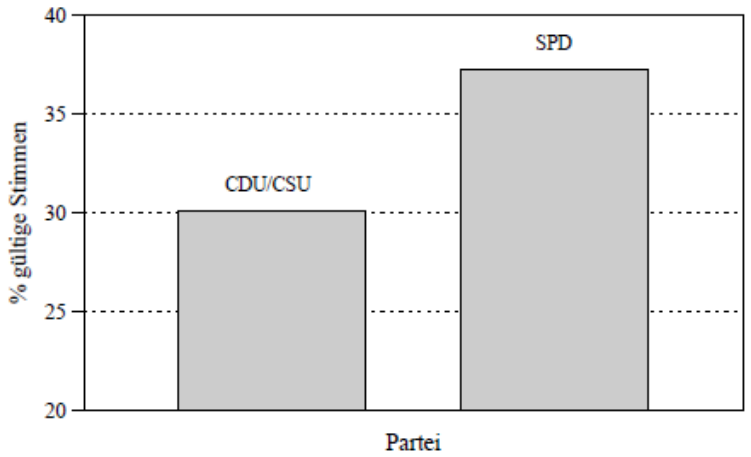


Quelle: ALLBUS 1994, n=2298

Manipulation durch Wahl der Grundlinie

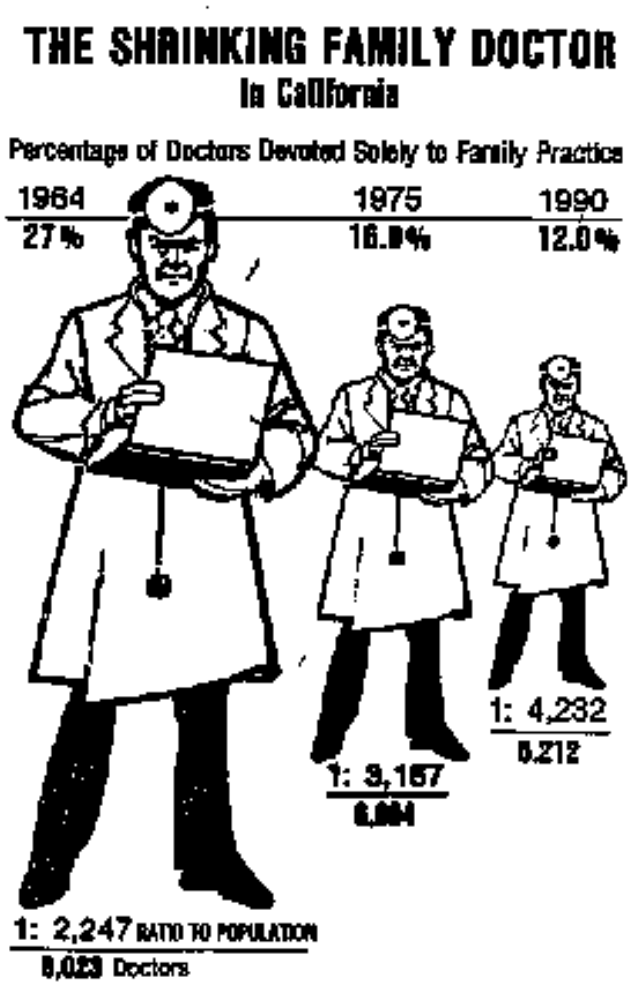
Abbildung 5.8: Wahlabsicht mit korrekter und falscher Grundlinie



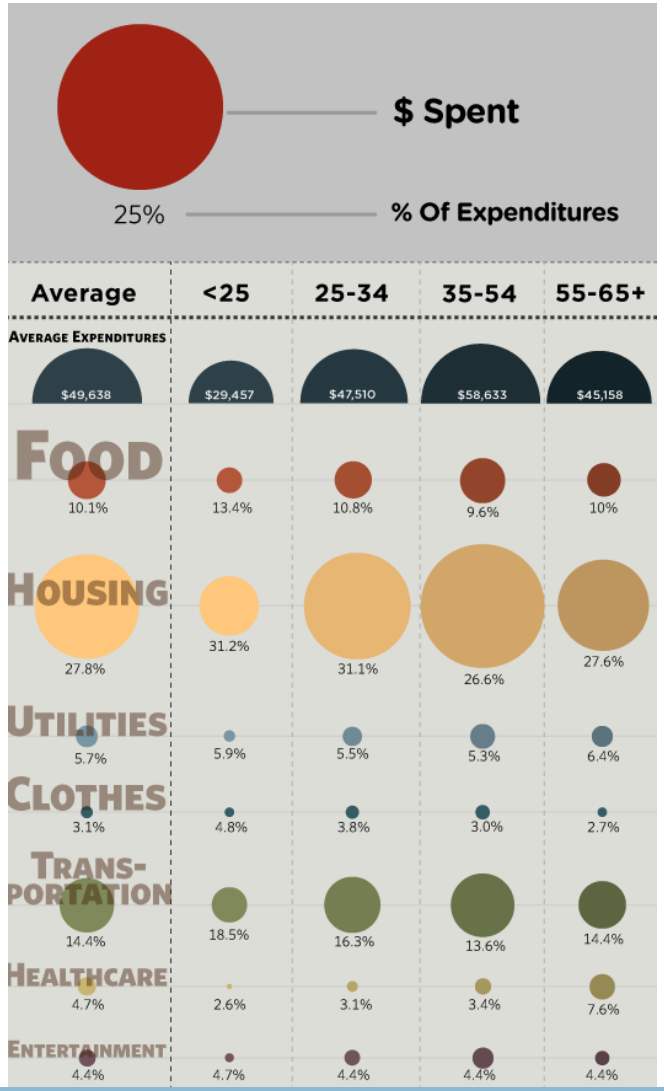


ALLBUS 1994, n=2298

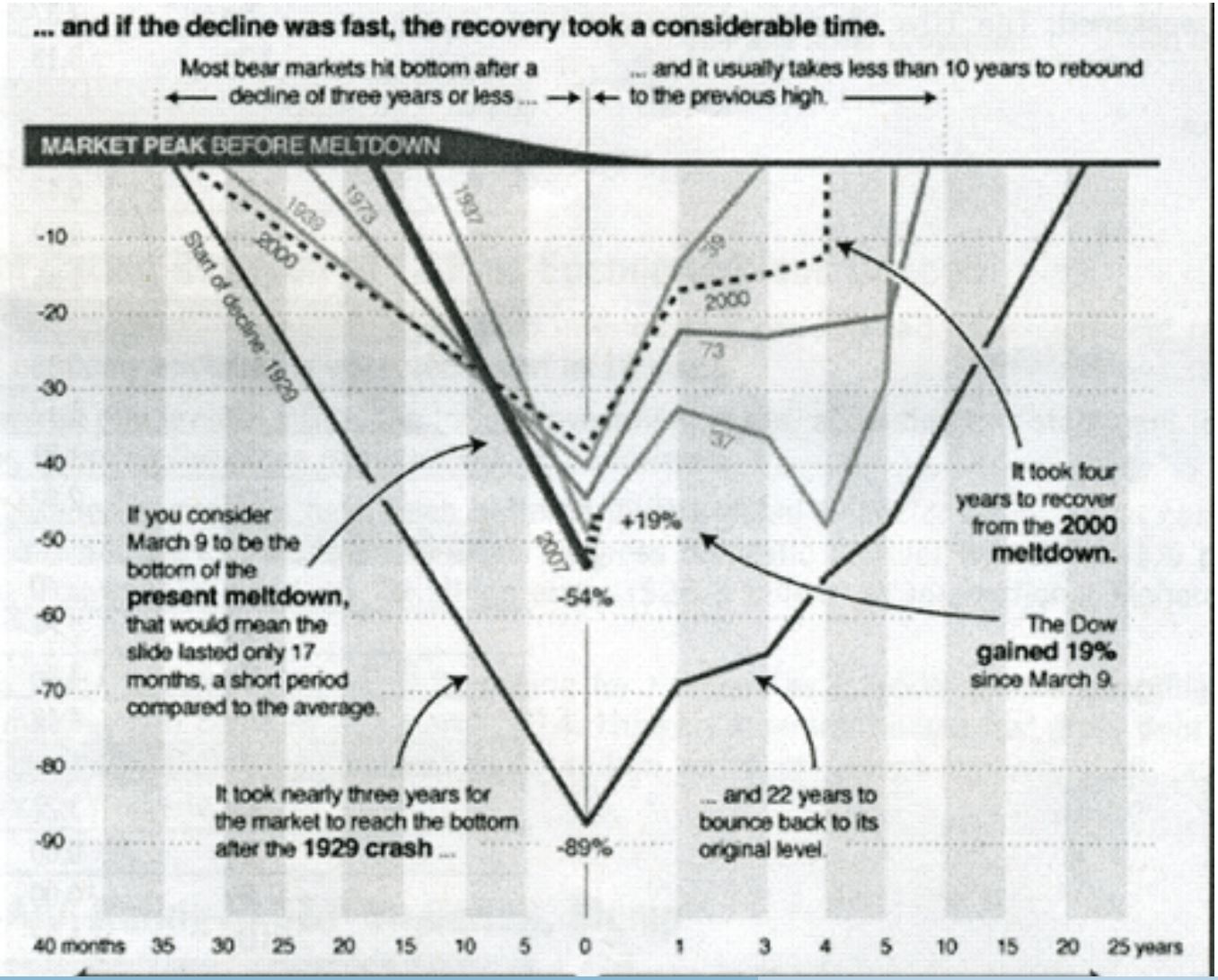
Figuren etc. führen in die Irre



Totale Konfusion



Maßstabsproblem



Zusammenfassung

- Berechnung von Anteilswerten: einfachstes aber nützliches Verfahren
- Grafiken vor allem für kleinere Datensätze sehr nützlich
- Darstellung von ein, zwei oder mehr Dimensionen möglich
- Viele veröffentlichte Grafiken führen in die Irre (oder sind zumindest nicht hilfreich)