

Big Data Anwendungen

Deskriptive Methoden zur Datenexploration

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
 - Betrachtung einer Stichprobe
 - Deskriptive Statistik
 - Visualisierung
 - Risiken der Datenvisualisierung
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Nutzen Betrachtung einer Stichprobe

- Idee
 - Betrachten der Rohdaten „zeilenweise“
- Vorteile
 - Diverse Informationen über die Daten sind „direkt“ sichtbar
 - Direktes Betrachten der Daten erfordert keine Tools
- Identifizierbare Informationen
 - Identifikation von Skalentypen
 - Identifikation von Wertebereichen
 - Erkennen von Formaten (z.B. Datum)
 - Entdecken von fehlenden Werten
 - Diskrepanz zu Datendokumentation
 - Entdecken von Fehlern in den Daten
 - Unterschiede zwischen Attributlänge und Datenlänge
 - Abgeschnittene Attribute
 - Leerzeichen in Attributen

Skalentypen

Skalentyp	Wertebereich	Mögliche Operationen	Beispiele
Nominale Größen	diskret, endlich	Gleichheit	Geschlecht Augenfarbe
Ordinale Größen	diskret, endlich, Ordnung	Gleichheit, größer / kleiner als	Prüfungsnoten Schulabschluß
Intervallgrößen	kontinuierlich bzw. ganzzahlig, unendlich	Gleichheit, größer / kleiner als Differenz	Datum Temperatur
Ratiogrößen	kontinuierlich bzw. ganzzahlig, unendlich	Gleichheit größer / kleiner als Differenz Verhältnis	Abstand Alter

- Anwendbarkeit der Statistiken abhängig vom Skalentyp
 - Mittelwert des Geschlechts
 - Modalwert der Prüfungsnoten

Beispiel

Guthaben	Telefon	CC seit	CC histor	Stand	Besitz	Alter	seit	
0	ja	6	kritisch_o_offene_Zahlung	0	Wohnung	67	4	
1	nein	48	alles_bezahlt	59.51	0W_verheiratet	Wohnung	22	2
nan	nein	12	kritisch_o_offene_Zahlung	20.96	0M_single	Wohnung	49	3
0	nein	42	alles_bezahlt	78.82	0M_single	Versicherung	45	3
0	nein	3	alles_bezahlt	3.7	0M_single	unbekannt	53	2
nan	ja	1	alles_bezahlt	55	M_single	unbekannt	35	2
1	ja	1	alles_bezahlt	48	0M_single	Auto	35	2
nan	nein	12	alles_bezahlt	30.59	3M_geschieden	Wohnung	61	3
1	nein	30	kritisch_o_offene_Zahlung	52.34	0M_verheiratet	Auto	28	0
1	nein	12	alles_bezahlt	12.95	0W_verheiratet	Auto	25	1
0	nein	48	alles_bezahlt	43.08	0W_verheiratet	Versicherung	24	1
1	ja	12	alles_bezahlt	15.67	0W_verheiratet	Auto	22	2
0	nein	24	kritisch_o_offene_Zahlung	11.99	0M_single	Auto	60	4
1	nein	15	alles_bezahlt	14.03	0W_verheiratet	Auto	3	2
1	nein	24	alles_bezahlt	12.82	1W_verheiratet	Auto	3	2

Ordinale Attribute in kategorischen versteckt

Fehlende Werte unterschiedlich kodiert (nan oder „ „)

Binäre Werte unterschiedlich kodiert ([0, 1] oder [ja, nein])

Zwei Attribute in einem Feld

Dauer in Monaten oder Jahren

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
 - Betrachtung einer Stichprobe
 - Deskriptive Statistik
 - Visualisierung
 - Risiken der Datenvisualisierung
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Zentralitätsmaße

- Mittelwert

- Durchschnitt über alle Werte

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

- Im Beispiel: $\bar{x} = \frac{11.69 + 59.51 + 20.96 + 78.82 + 78.82}{5} = 49.96$

Wert
11.69
59.51
20.96
78.82
78.82

- Median

- „Mittlere Wert“ aller sortierten Werte

- Allgemein: $\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ ungerade} \\ \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & n \text{ gerade} \end{cases}$

- Im Beispiel: $\tilde{x} = 59.51$

- Modalwert

- Häufigster Wert

- Im Beispiel: $\bar{x}_M = 78.82$

Quantile

- p -Quantil
 - p -Quantil x_p teilt Stichprobe, so dass p kleiner und $1 - p$ größer x_p
 - Allgemein: $x_p = \begin{cases} \frac{1}{2}(x_{n \cdot p} + x_{(n \cdot p)+1}), & \text{wenn } n \cdot p \text{ ganzzahlig} \\ x_{\lfloor n \cdot p + 1 \rfloor}, & \text{wenn } n \cdot p \text{ nicht ganzzahlig} \end{cases}$
 - mit $x_1 \leq x_2 \leq \dots \leq x_n$ und $\lfloor x \rfloor$ ist Abrundungsfunktion
 - Hinweis: Median \tilde{x} ist 0.5-Quantil $x_{0.5}$
 - Im Beispiel: $x_{0.2} = \frac{1}{2}(11.69 + 20.96) = 16.33$ und $x_{0.25} = 20.96$
- Interquartilabstand IQR
 - Abstand zwischen drittem und erstem Quartil
 - Allgemein: $IQR = x_{0.75} - x_{0.25}$
 - Im Beispiel: $IQR = 78.82 - 20.96 = 57.86$

Wert
11.69
59.51
20.96
78.82
78.82

Empirische Varianz

- Empirische Varianz

Wert
11.69
59.51
20.96
78.82
78.82

11.69

59.51

20.96

78.82

78.82

- Mittlere quadratische Abweichung der Beobachtungen vom Mittelwert

- Formal: $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$

- Hinweis: oft alternative Definition $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

- Im Beispiel: $s^2 = \frac{1}{n} ((11.69 - 49.96)^2 + (59.51 - 49.96)^2$

$$+ (20.96 - 49.96)^2 + (78.82 - 49.96)^2$$

$$+ (78.82 - 49.96)^2)$$

$$= \frac{1}{5} (38.27^2 + 9.55^2 + 29.00^2 + 28.86^2 + 28.86^2)$$

$$= 812.52$$

- Empirische Standardabweichung

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

- Im Beispiel: $s = \sqrt{812.52} = 28.50$

Schiefe

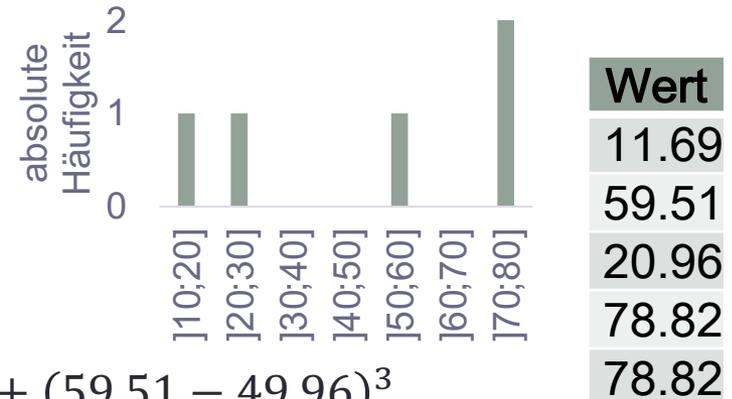
- Schiefe (engl.: Skewness)

- Maß für die Asymmetrie der Verteilung

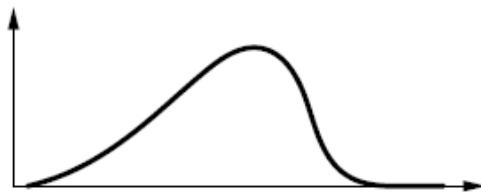
- Formal: $\alpha_3 = \frac{1}{n \cdot s^3} \sum_{i=1}^n (x_i - \bar{x})^3 = \frac{1}{n} \sum_{i=1}^n z_i^3$

mit $z_i = \frac{x_i - \bar{x}}{s}$

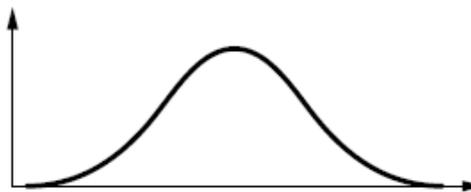
- Im Beispiel: $\alpha_3 = \frac{1}{5 \cdot 28.50^3} ((11.69 - 49.96)^3 + (59.51 - 49.96)^3 + (20.96 - 49.96)^3 + (78.82 - 49.96)^3 + (78.82 - 49.96)^3)$
 $= \frac{1}{5 \cdot 28.50^3} \cdot -31493.02 = -0.27$



- Schiefe (graphisch)



$\alpha_3 < 0$: rechtssteil



$\alpha_3 = 0$: symmetrisch



$\alpha_3 > 0$: linkssteil

Empirische Wölbung

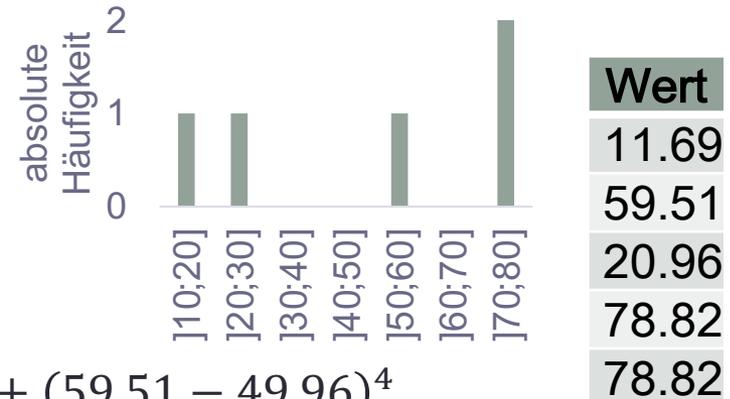
- Empirische Wölbung (Kurtosis)

- Maß für die Steilheit der Verteilung

- Formal: $\alpha_4 = \frac{1}{n \cdot s^4} \sum_{i=1}^n (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^n z_i^4$

mit $z_i = \frac{x_i - \bar{x}}{s}$

- Im Beispiel: $\alpha_4 = \frac{1}{5 \cdot 28.50^4} ((11.69 - 49.96)^4 + (59.51 - 49.96)^4 + (20.96 - 49.96)^4 + (78.82 - 49.96)^4 + (78.82 - 49.96)^4)$
 $= \frac{1}{5 \cdot 28.50^4} \cdot 4248074.75 = 1.29$

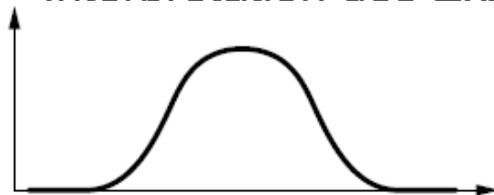


- Exzess

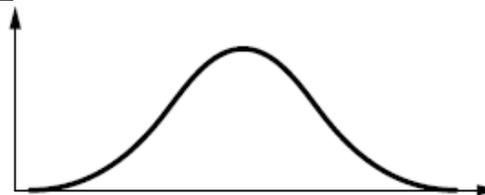
- Normierung der Wölbung

- Formal: $\gamma = \alpha_4 - 3$, Im Beispiel: $\gamma = 1.29 - 3 = -1.71$

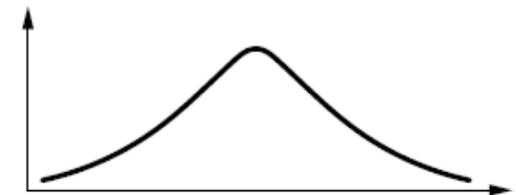
- Interpretation des Exzess



$\gamma < 0$: platykurtisch,
flachgipflig



$\gamma = 0$: mesokurtisch,
normalgipflig



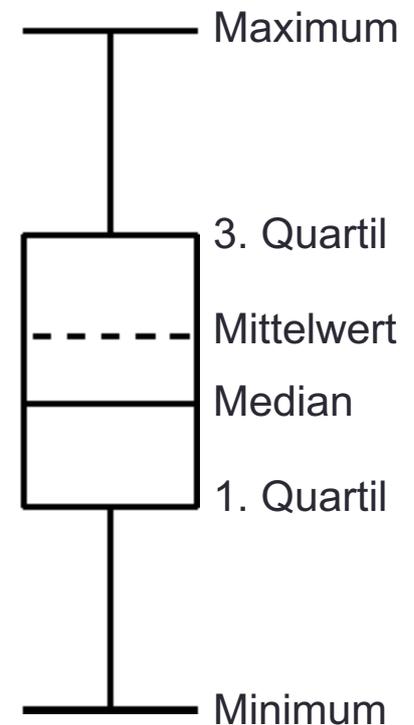
$\gamma > 0$: leptokurtisch,
steilgipflig

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
 - Betrachtung einer Stichprobe
 - Deskriptive Statistik
 - Visualisierung
 - Risiken der Datenvisualisierung
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Boxplot

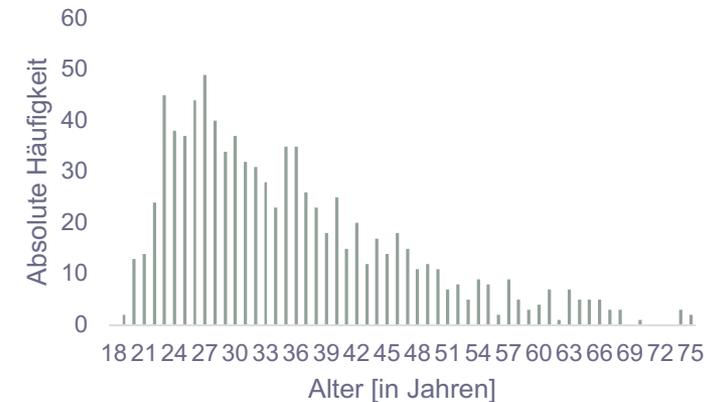
- Faßt mehrere statistische Masse zusammen
- Illustration von
 - Mittelwert: \bar{x}
 - Median: \tilde{x}
 - Interquartilsabstand über $x_{0.25}$ und $x_{0.75}$
 - Minimum und Maximum
(stattdessen oft $x_{0.05}$ und $x_{0.95}$)
- Vorteile der Darstellung
 - Fast alle deskriptiven Kennzahlen in einer Abbildung
 - Schiefe ablesbar
 - Empirische Wölbung ablesbar



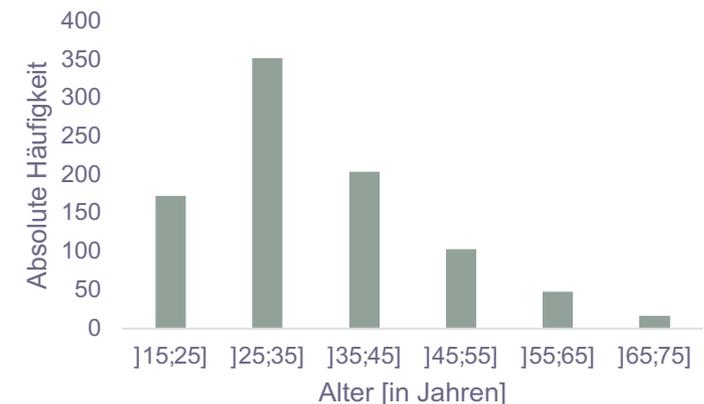
Histogramm

- Darstellung von Verteilungen
- Illustration von
 - Anzahl (bzw. Anteil) der Ausprägungen
 - Oft Gruppierung (binning) der Daten
- Vorteile Darstellung
 - Ausreißer und Verteilung identifizierbar
 - Peaks in Verteilung erkennbar
 - Schiefe ablesbar
 - Empirische Wölbung ablesbar

Histogramm



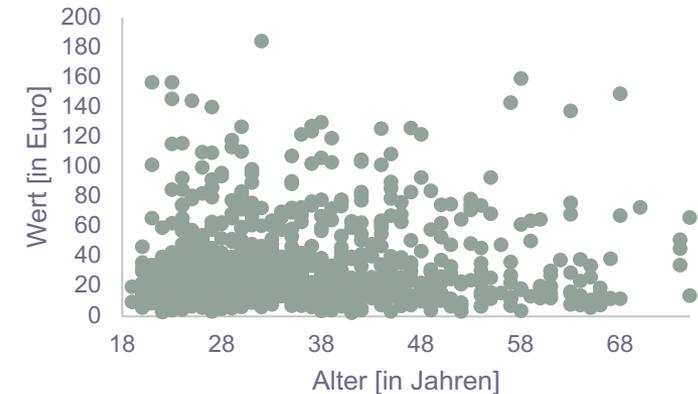
Histogramm (gebinnt)



Scatterplot

- Darstellung von Beobachtungen
- Illustration von
 - Unterschiedlichen Attributen...
... pro Achse (maximal 3)
- Vorteile Darstellung
 - Ausreißer identifizierbar
 - Korrelationen identifizierbar
 - Häufungen von Datenpunkten identifizierbar

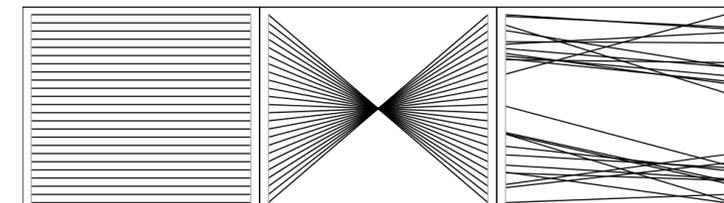
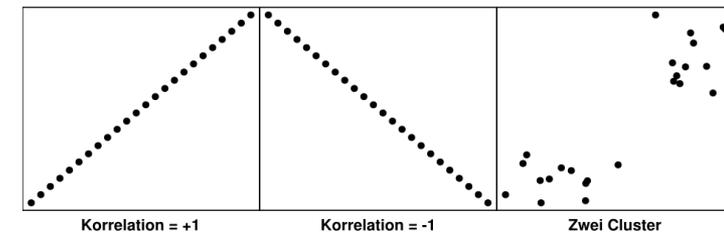
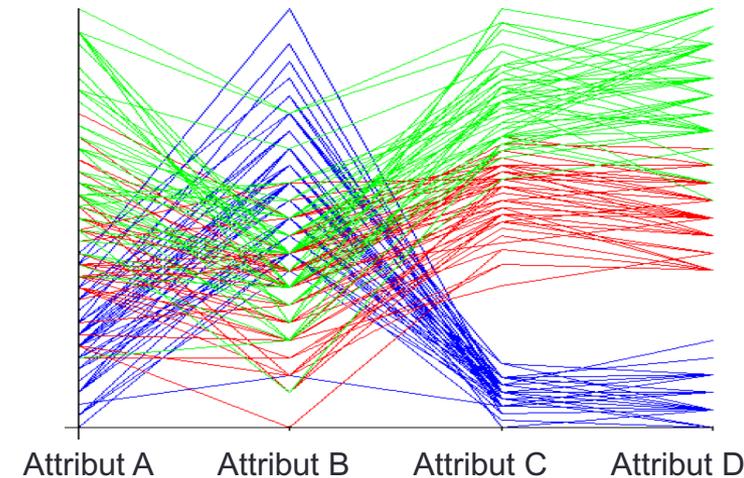
Scatterplot



Parallelkoordinaten

- Darstellung von Beobachtungen
- Illustration von
 - Unterschiedlichen Attributen...
... pro Achse (pro Achse eine Spalte)
 - Einfärbung nach Klassen möglich
- Vorteile Darstellung
 - Korrelationen identifizierbar
 - Häufungen von Datenpunkten identifizierbar
- Vergleich mit Scatterplot
 - Ähnliche Informationen visualisierbar
 - Erlaubt Darstellung von mehr als 3 Attributen

Parallelkoordinaten



Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
 - Betrachtung einer Stichprobe
 - Deskriptive Statistik
 - Visualisierung
 - Risiken der Datenvisualisierung
- Datenqualität
- Klassifikation
- Recommender Systems
- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Median vs. Mittelwert



\$45,000



\$15,000



\$10,000



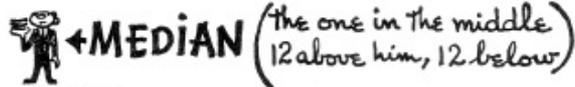
\$5,700



\$5,000



\$3,700



\$3,000



\$2,000

← **ARITHMETICAL AVERAGE**

(the one in the middle)
12 above him, 12 below

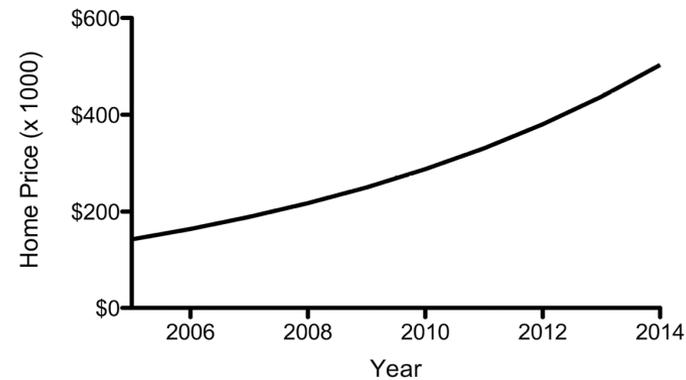
← **MODE**
(occurs most frequently)



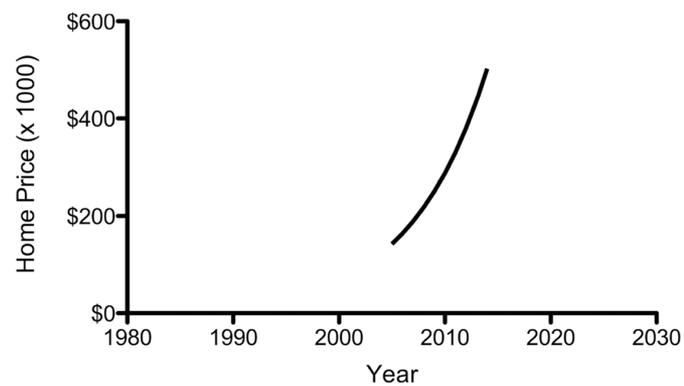
»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

Skalierungstricks - Hauspreise

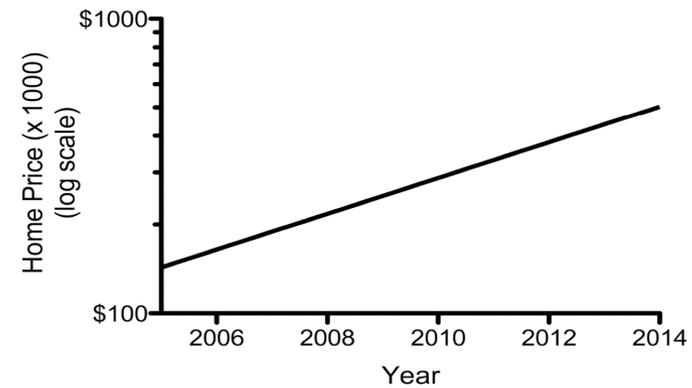
Tatsächlicher Zusammenhang



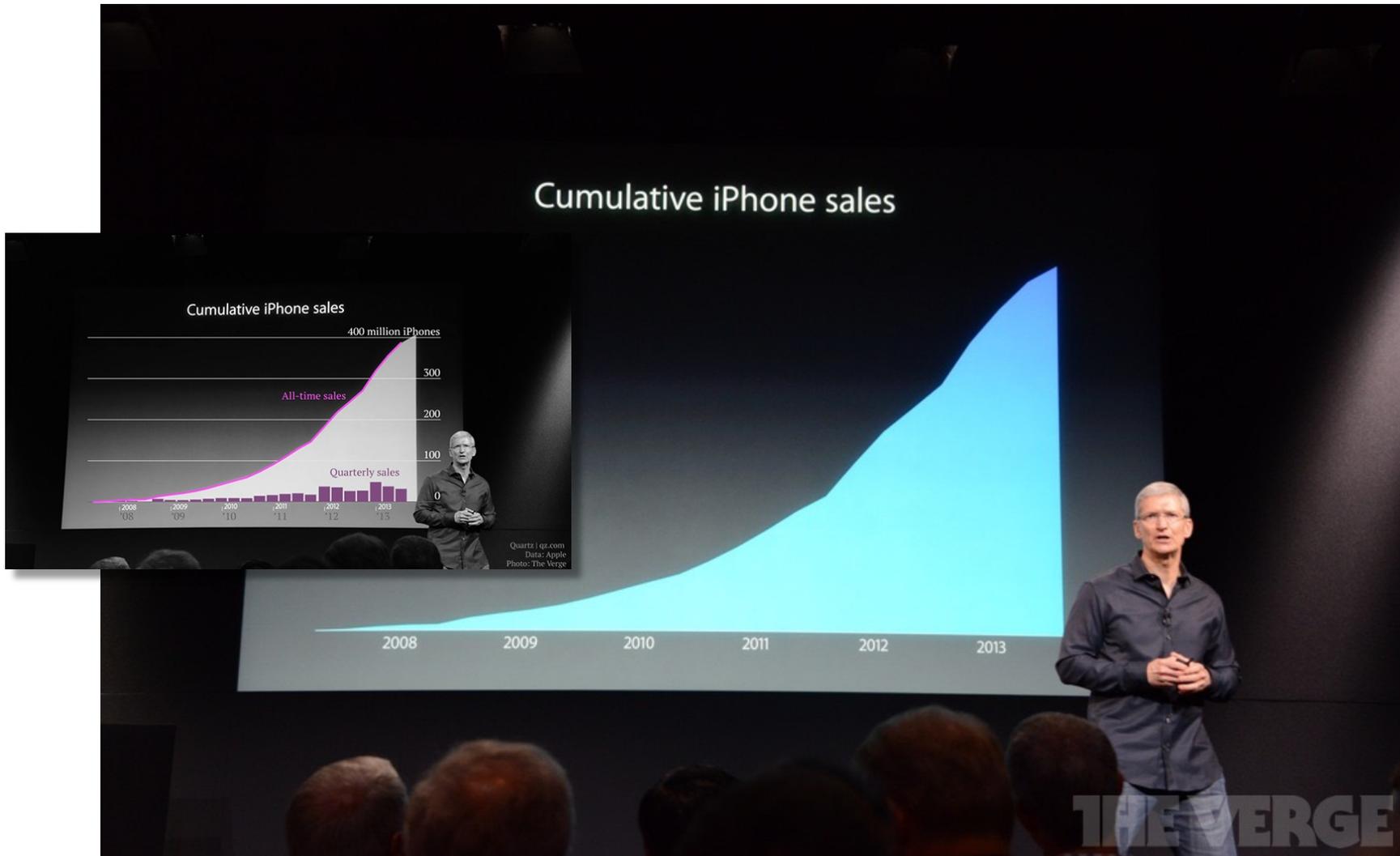
Skalierung der x-Achse



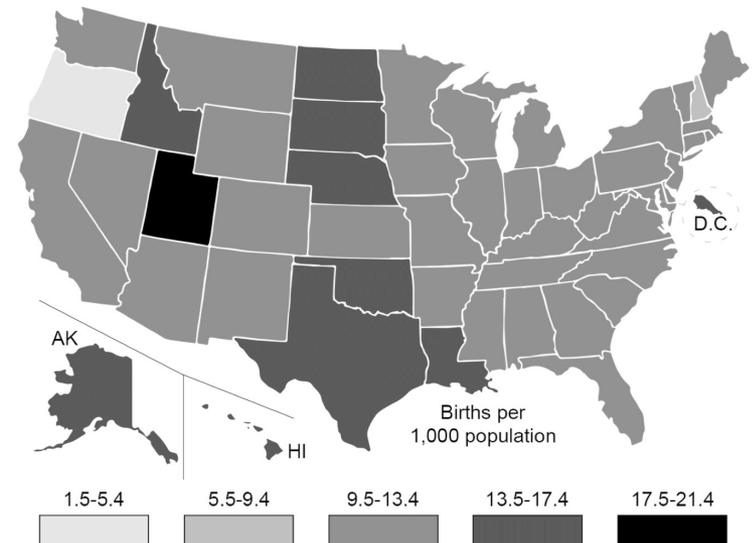
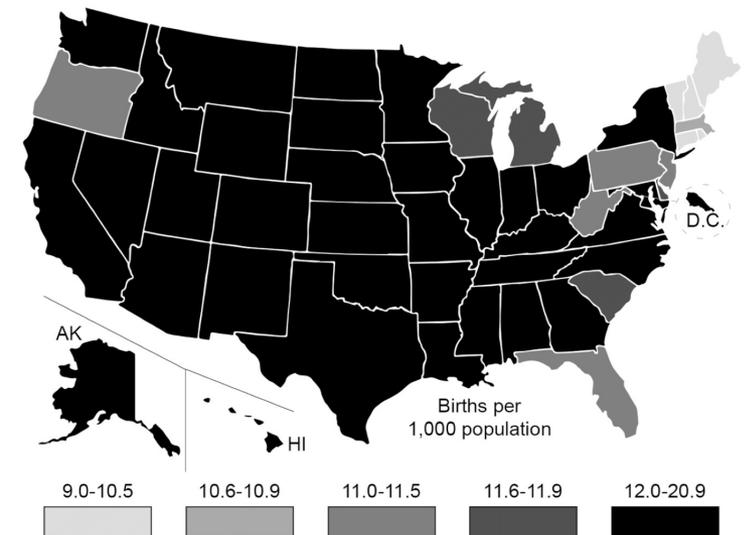
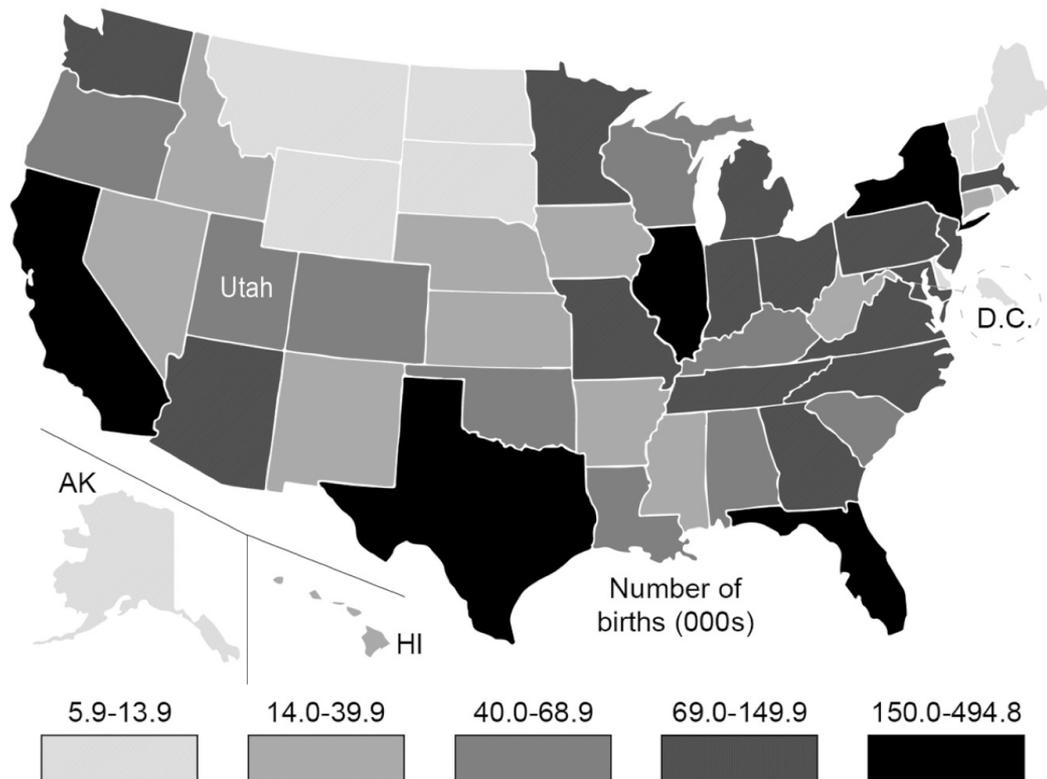
Skalierung der y-Achse



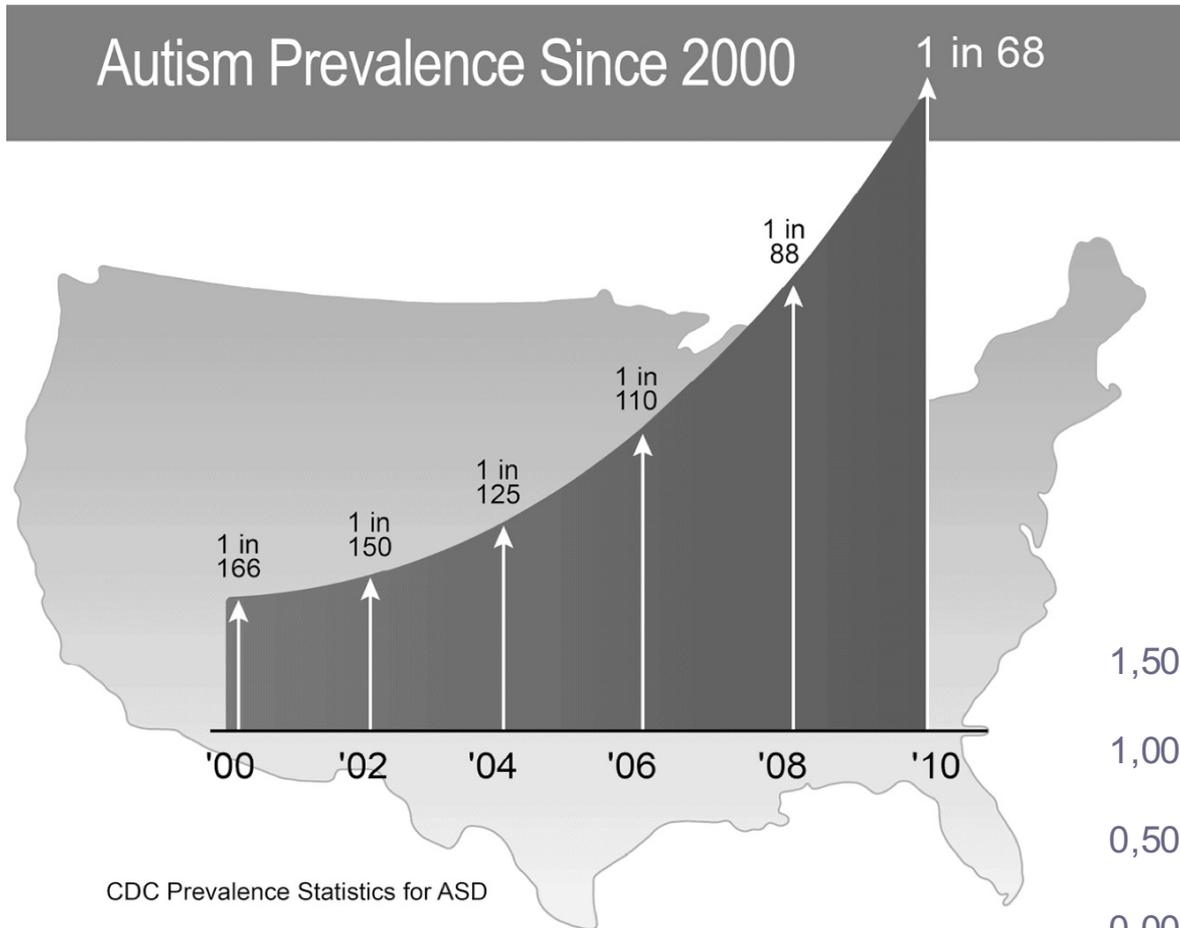
Kumulative Verteilung (Apple läuft)



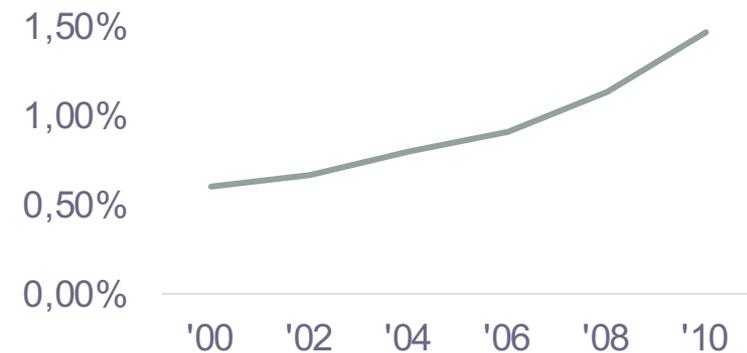
Geburtenstatistik der USA, 2013



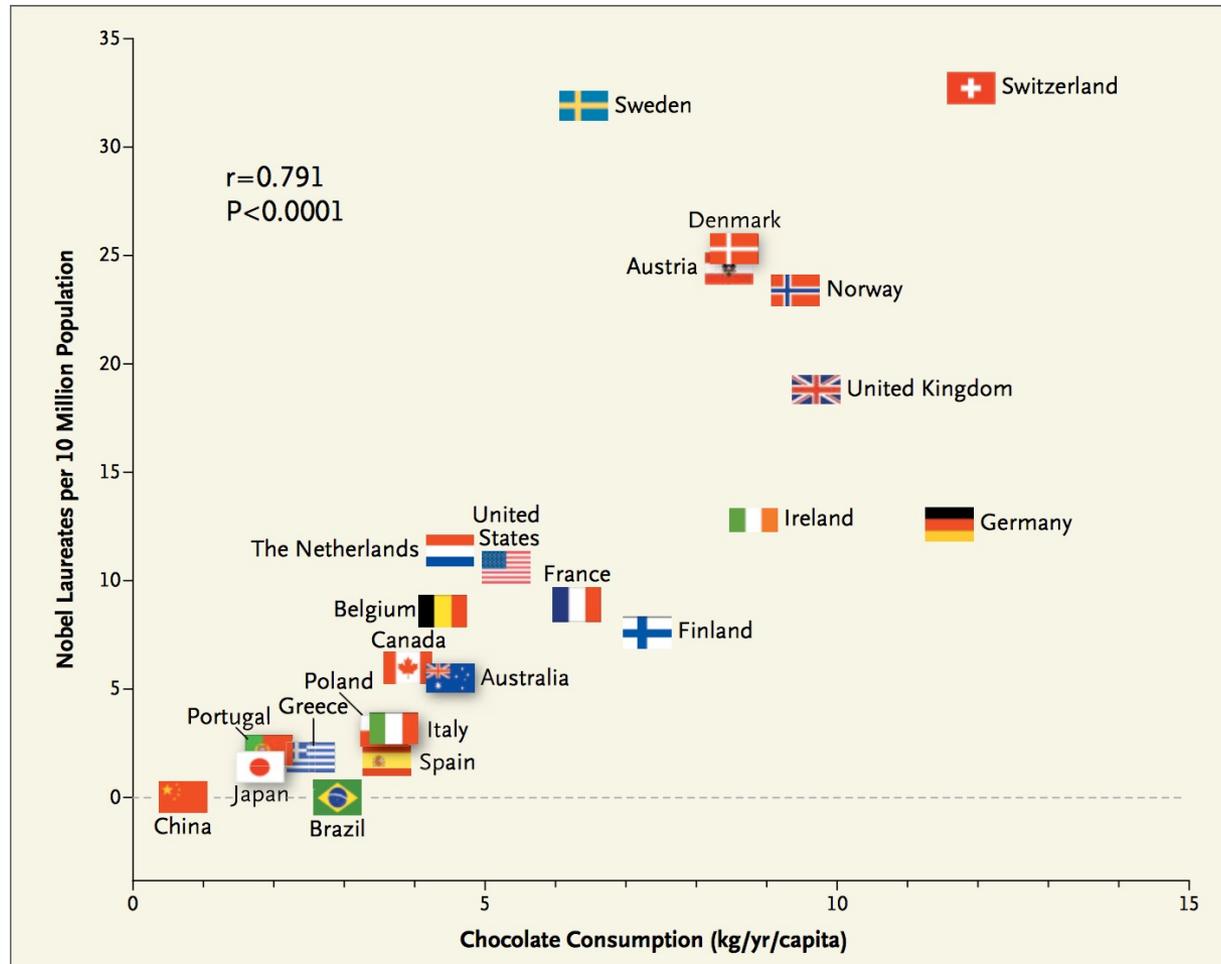
Entwicklung von Autismus in den USA



Umrechnung in Prozent



Schokolade macht klug



- Messerli, 2012: Chocolate Consumption, Cognitive Function, and Nobel Laureates, New England Journal of Medicine