Universität Zürich[UZH]

Datenanalyse
(PHY231)
Herbstsemester 2017

Olaf Steinkamp

36-J-05 olafs@physik.uzh.ch 044 63 55763

Vorlesungsprogramm

- **Einführung, Messunsicherheiten, Darstellung von Messdaten**
- **Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik**
  - Mittelwert, Standardabweichung, Kovarianz und Korrelation
- **Fehlerfortpflanzungsgesetz**
- **Wahrscheinlichkeitsverteilungen**
  - diskrete Verteilungen, kontinuierliche Verteilungen
  - zentraler Grenzwertsatz
- **Monte-Carlo Methode**
- **Wahrscheinlichkeitsverteilungen II**
  - Faltung zweier Verteilungen
  - Verteilungen zweier Variablen
- **Stichproben und Schätzfunktionen**
  - Maximum-Likelihood Methode
  - Methode der kleinsten Quadrate
- **Interpretation von Messergebnissen**
  - Konfidenzintervalle, Testen von Hypothesen

Beispielprogramme im Verzeichnis

`/disk/puma/da/vorl/stat`

Mittelwert einer Verteilung

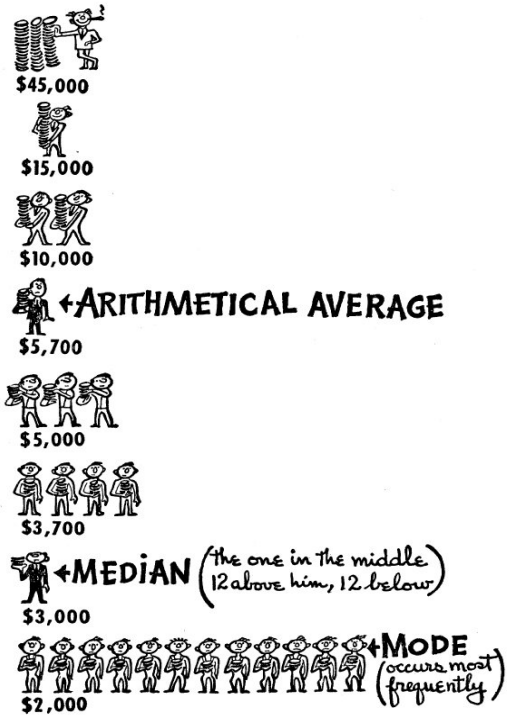
Für eine Verteilung aus N Werten $x_1, x_2, \ldots, x_N$

- **arithmetischer Mittelwert:**

$$\bar{x} \equiv \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

- **Median:**
  die Hälfte aller Werte ist grösser,
  die Hälfte aller Werte ist kleiner

- **Modus:**
  der am häufigsten vorkommende Wert



im folgenden: **"Mittelwert" ≡ arithmetischer Mittelwert**

Für eine Funktion $f(x_i)$

$$\bar{f} \equiv \frac{1}{N} \cdot \sum_{i=1}^{N} f(x_i)$$

$$f(x) = x^2 \quad \Rightarrow \quad \bar{f} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i^2$$

Gewichteter Mittelwert

Gewichteter Mittelwert von N Werten x_i mit Gewichten w_i

$$\overline{x} \equiv \frac{\sum_{i=1}^{N} w_i \cdot x_i}{\sum_{i=1}^{N} w_i}$$

- **wichtige Anwendung: gewichteter Mittelwert von N voneinander unabhängigen Messungen mit unterschiedlichen Messunsicherheiten σ_i**

$$w_i = 1/\sigma_i^2$$

Herleitung später

- **Mittelwert eines Histogramms mit N Intervallen:**

$$\overline{x} \equiv \frac{\sum_{i=1}^{N} n_i \cdot x_i}{\sum_{i=1}^{N} n_i}$$

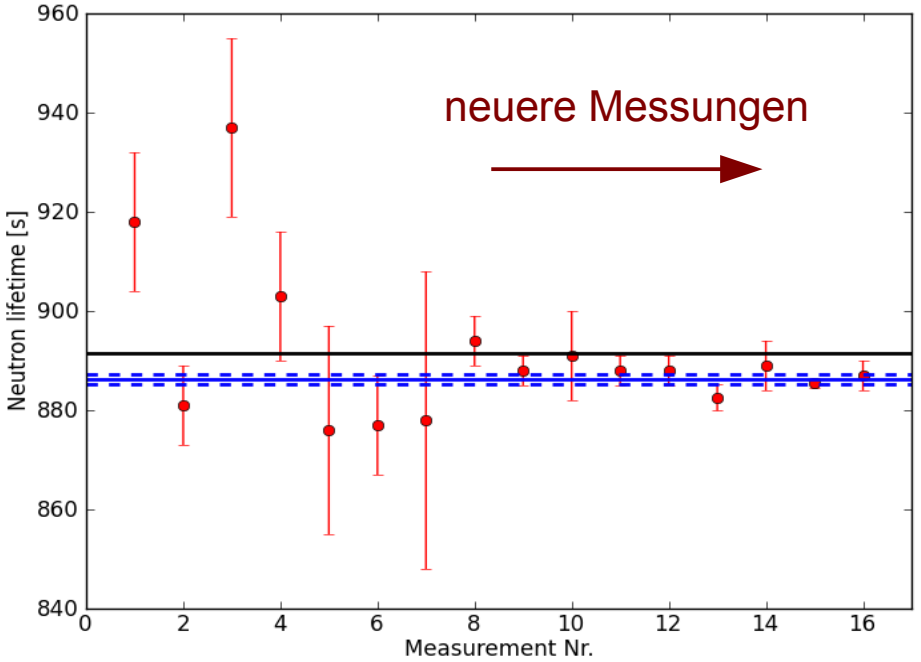
x_i : Intervallzentren

n_i : Anzahl Einträge

Mittelwert des Histogramms = gewichteter Mittelwert der Intervallzentren

Einfacher und gewichteter Mittelwert

Beispiel: 16 Messungen der Lebensdauer des Neutrons



arithmetischer Mittelwert

gewichteter Mittelwert (mit Unsicherheit)

nlife.py
nlife.dat

# t[s]	dt[s]
918	14
881	8
937	18
903	13
876	21
877	10
878	30
894	5
888	3
891	9
888	3
888	3
882.6	2.7
889	5
885.4	1.0
887	3

[Quelle: Particle Data Group]

- **arithmetischer Mittelwert aller Messungen: $1/16 \times \Sigma\, t_i = 891.4$ s**
- **aber: neuere Messungen präziser als ältere → sollten mehr Gewicht haben**
- **gewichteter Mittelwert: $1 / \Sigma(1/\sigma_i^2) \times \Sigma\,(t_i/\sigma_i^2) = (886.3 \pm 0.9)$ s**

pylab: Befehl `average()` kann gewichtete Mittelwerte berechnen

Unsicherheit auf gewichtetem Mittelwert

Gewichteter Mittelwert von N Messungen x_i mit Messunsicherheiten σ_i

$$w_i = 1/\sigma_i^2 \quad \Rightarrow \quad \bar{x} = \frac{1}{\sum_{i=1}^{N} \left(1/\sigma_i^2\right)} \times \sum_{i=1}^{N} \left(x_i/\sigma_i^2\right)$$

- **Messungen voneinander unabhängig: benutze Gaußsche Fehlerfortpflanzung**

$$\sigma_{\bar{x}} = \sqrt{\sum_{i=1}^{N} \left(\frac{\partial \bar{x}}{\partial x_i} \cdot \sigma_i\right)^2} \quad \Rightarrow \quad \sigma_{\bar{x}} = \frac{\sqrt{\sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2} \cdot \sigma_i\right)^2}}{\sum_{i=1}^{N} \left(1/\sigma_i^2\right)} = \frac{1}{\sqrt{\sum_{i=1}^{N} \left(1/\sigma_i^2\right)}}$$

- **Spezialfall: Messunsicherheit auf allen Messungen gleich, d.h. $\sigma_i = \sigma$ für alle i**

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{\sum_{i=1}^{N} 1/\sigma^2}} = \frac{1}{\sqrt{N \times 1/\sigma^2}} = \frac{\sigma}{\sqrt{N}}$$

nächste Woche

aber aufgepasst: Gaußsche Fehlerfortpflanzung gilt nur, wenn die Messungen voneinander unabhängig sind

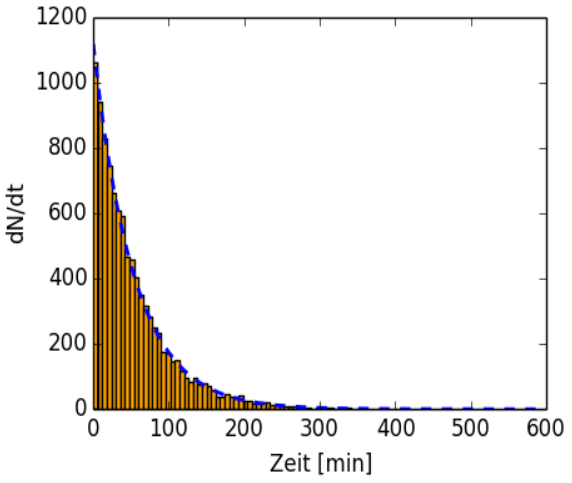
Mittelwert histogrammierter Daten

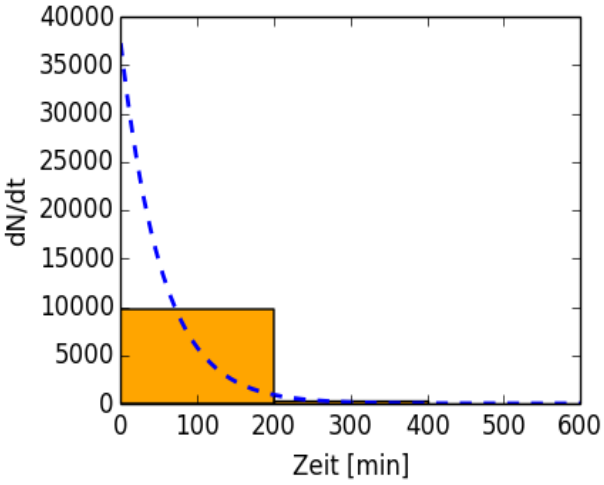
Beispiel: 10'000 exponentialverteilte Messwerte (Zerfallszeiten radioaktiver Quelle)

expohist.py

```
#!/usr/bin/env python
from pylab import *
#
# generiere exponentialverteilte Werte
#
N = 10000
meantrue = 53.7
tmeas = exponential(meantrue,N)
#
# Mittelwert der Verteilung
#
meanmeas = mean(tmeas)
#
# histogrammiere die Werte
#
tmin = 0 ; tmax = 600 ; nbins = 100
ni,ti,patch = hist(tdata,nbins,(tmin,tmax))
#
# Mittelwert des Histogramms
#
tbin = ti[0:-1]+ti[1:])/2.0
meanhist = dot(ni,tbin) / sum(ni)
```

- **"wahrer" Mittelwert: 53.7 min**
- **Mittelwert der Messwerte: 53.6 min**
- **Mittelwert eines Histogramms mit**
  - 100 Intervallen: 53.6 min
  - 10 Intervallen: 59.0 min
  - 3 Intervallen: 105. min





bei zu groß gewählter Intervallbreite geht Information verloren !

Vorlesungsprogramm

- **Einführung, Messunsicherheiten, Darstellung von Messdaten**
- **Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik**
  - Mittelwert, Standardabweichung, Kovarianz und Korrelation
- **Fehlerfortpflanzungsgesetz**
- **Wahrscheinlichkeitsverteilungen**
  - diskrete Verteilungen, kontinuierliche Verteilungen
  - zentraler Grenzwertsatz
- **Monte-Carlo Methode**
- **Wahrscheinlichkeitsverteilungen II**
  - Faltung zweier Verteilungen
  - Verteilungen zweier Variablen
- **Stichproben und Schätzfunktionen**
  - Maximum-Likelihood Methode
  - Methode der kleinsten Quadrate
- **Interpretation von Messergebnissen**
  - Konfidenzintervalle, Testen von Hypothesen

Beispielprogramme im Verzeichnis

`/disk/puma/da/vorl/stat`

Maße für die Breite einer Verteilung

Mittlere Abweichung der Messwerte vom Mittelwert

$$\frac{1}{N} \cdot \sum_{i=1}^{N} |x_i - \bar{x}|$$

- unschöne mathematische Behandlung (z.B. beim Bilden von Ableitung)

Statistiker: Varianz der Verteilung

$$V(x) \equiv \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

Herleitung: Uebungen

- okay bzgl. mathematischer Behandlung
- aber: andere Einheit als Messgröße

Physiker: Standardabweichung der Verteilung

$$\sigma_x \equiv \sqrt{V(x)} = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2} = \sqrt{\overline{x^2} - \bar{x}^2}$$

Standardabweichung

Aufgepasst: zwei Definitionen der "Standardabweichung"!

$$\sigma_x \equiv \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \overline{x})^2}$$

$$s_x \equiv \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \overline{x})^2}$$

in pylab: `std(x,0)` `std(x,1)`

- **Definition mit 1 / *N* ist die Standardabweichung der gemessenen Verteilung**
- **Definition mit 1 / (*N*-1) gibt einen Schätzwert für die Standardabweichung der "wahren" Verteilung, die gemessen werden soll**
- **Unterschied für große *N* vernachlässigbar, nicht aber für kleine *N***
- **deshalb wichtig: immer angeben, welche Definition Sie verwenden**

keine Angst,
wird in ein paar Wochen
hoffentlich klar ...

Standardabweichung einer Verteilung und Unsicherheit auf ihrem Mittelwert

Standardabweichung σ_x der Verteilung

- ist bestimmt durch die Streuung der einzelnen Messwerte um den Mittelwert
- ist ein Maß für die Messunsicherheit auf den einzelnen Messungen
- hängt nicht von der Zahl der Messungen ab

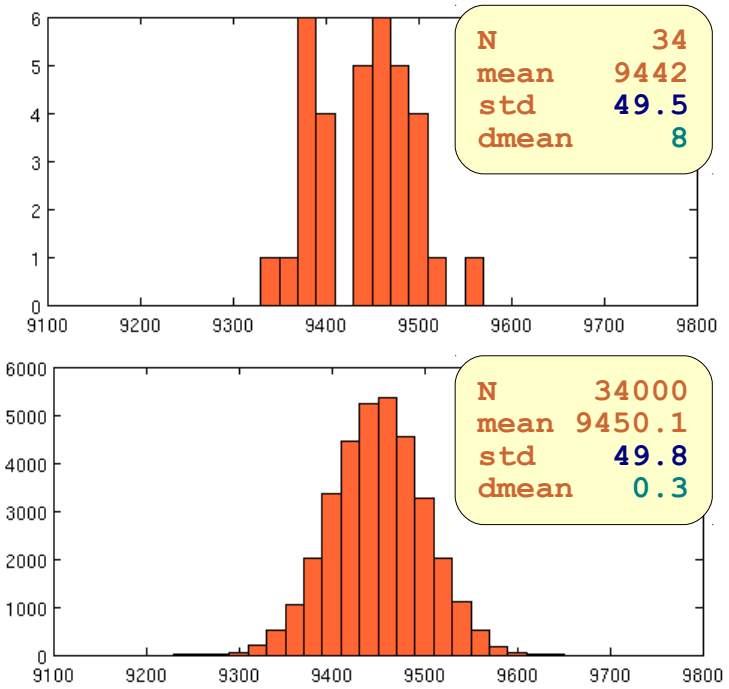
Unsicherheit auf dem Mittelwert der Verteilung

- ist umso kleiner, je kleiner die Streuung der Messwerte ist
- nimmt mit zunehmender Anzahl Messungen ab

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

vgl. Folie 6

- Beispiel: Verteilung gaußverteilter Zufallszahlen, erzeugt mit $\mu = 9450$ und $\sigma = 50$



Standardabweichung einer histogrammierten Verteilung

Histogramm mit *N* Intervallen

Folie 9 Folie 4

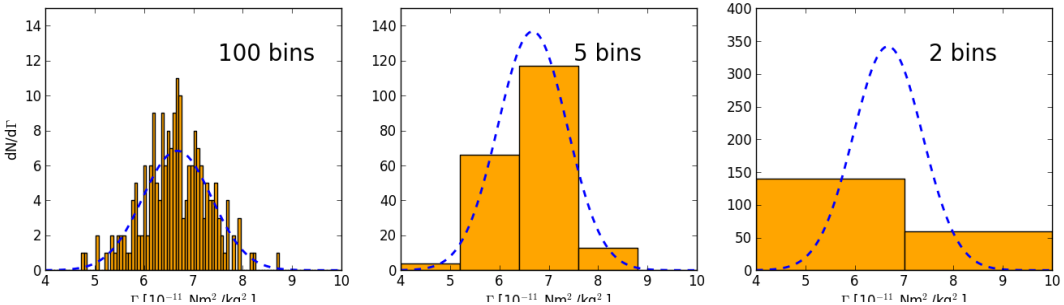
$$\sigma_x = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{\frac{\sum_{i=1}^{N} n_i \cdot x_i^2}{\sum_{i=1}^{N} n_i} - \left(\frac{\sum_{i=1}^{N} n_i \cdot x_i}{\sum_{i=1}^{N} n_i}\right)^2}$$

x_i : Intervallzentren

n_i : Anzahl Einträge

Beispiel: 200 Messungen der Gravitationskonstante (s. letzte Woche)

- Standardabweichung der Verteilung: $0.66 \times 10^{-11} \mathrm{N \cdot m^2/kg}$
- Standardabweichung des Histogramms mit
  - 50 Intervallen: 0.67
  - 5 Intervallen: 0.71
  - 2 Intervallen: 1.23



wieder: Informationsverlust bei zu groß gewählter Intervallbreite !

Vorlesungsprogramm

- **Einführung, Messunsicherheiten, Darstellung von Messdaten**
- **Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik**
  - Mittelwert, Standardabweichung, Kovarianz und Korrelation
- **Fehlerfortpflanzungsgesetz**
- **Wahrscheinlichkeitsverteilungen**
  - diskrete Verteilungen, kontinuierliche Verteilungen
  - zentraler Grenzwertsatz
- **Monte-Carlo Methode**
- **Wahrscheinlichkeitsverteilungen II**
  - Faltung zweier Verteilungen
  - Verteilungen zweier Variablen
- **Stichproben und Schätzfunktionen**
  - Maximum-Likelihood Methode
  - Methode der kleinsten Quadrate
- **Interpretation von Messergebnissen**
  - Konfidenzintervalle, Testen von Hypothesen

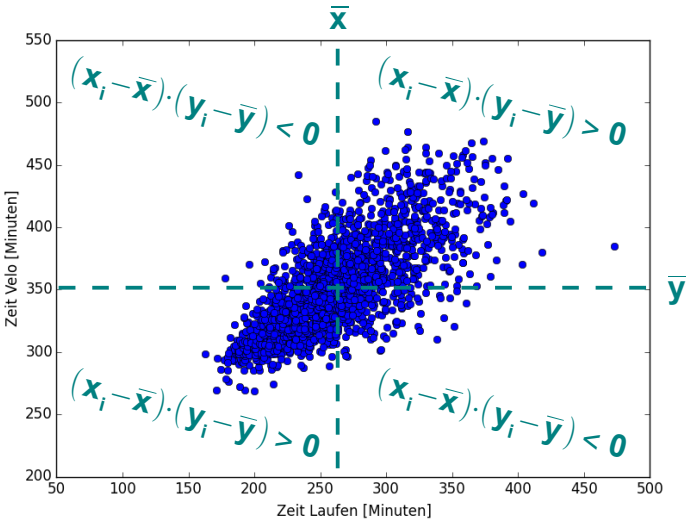
Beispielprogramme im Verzeichnis

`/disk/puma/da/vorl/stat`

Korrelation und Kovarianz

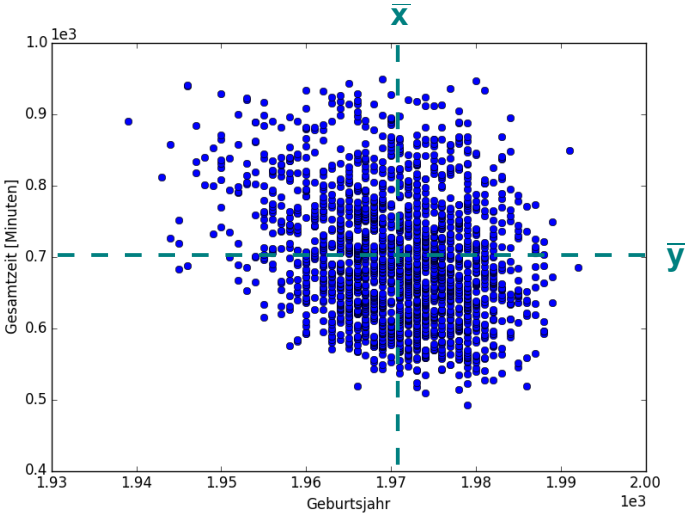
Betrachte statistischen Zusammenhang zwischen zwei Zufallsvariablen

- **positive Korrelation:** Wert einer Variablen nimmt im Mittel zu, wenn der Wert der anderen zunimmt
- **negative Korrelation:** Wert einer Variablen nimmt im Mittel ab, wenn der Wert der anderen zunimmt



Kovarianz für N Wertepaare (x_1,y_1), …, (x_N,y_N)

$$cov(x,y) \equiv \frac{1}{N}\cdot\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) = \overline{xy}-\bar{x}\,\bar{y}$$



- cov (x,y) > 0 für positive Korrelation
- cov (x,y) < 0 für negative Korrelation
- cov (x,y) = 0 wenn keine Korrelation

Nachteil: Wert für cov (x,y) ≠ 0 hängt von den für x und y gewählten Einheiten ab

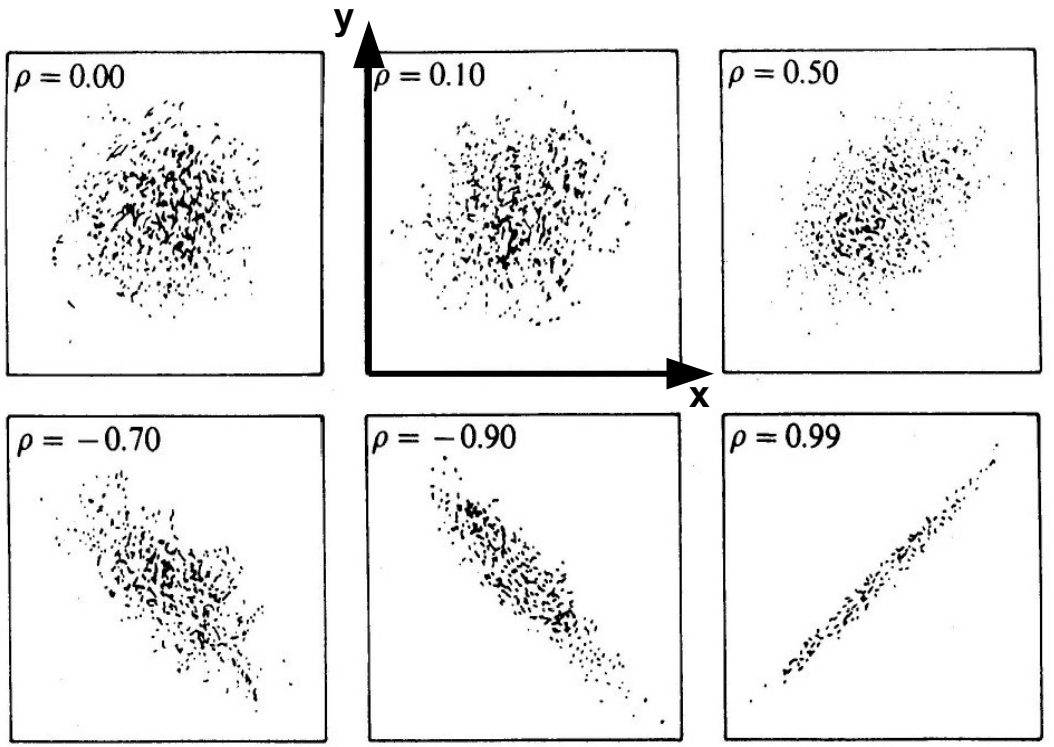
Korrelationskoeffizient

Einheitenloses, normiertes Maß für Korrelation zweier Zufallsvariablen

$$\rho \equiv \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}$$

$-1 \le \rho \le 1$

- $\rho = 0$: keine Korrelation
- $\rho > 0$: positive Korrelation
- $\rho < 0$: negative Korrelation
- $\rho = \pm 1$: vollständige Korrelation,
  Wert von x_i legt Wert von y_i fest und umgekehrt



[aus: Barlow, Statistics]

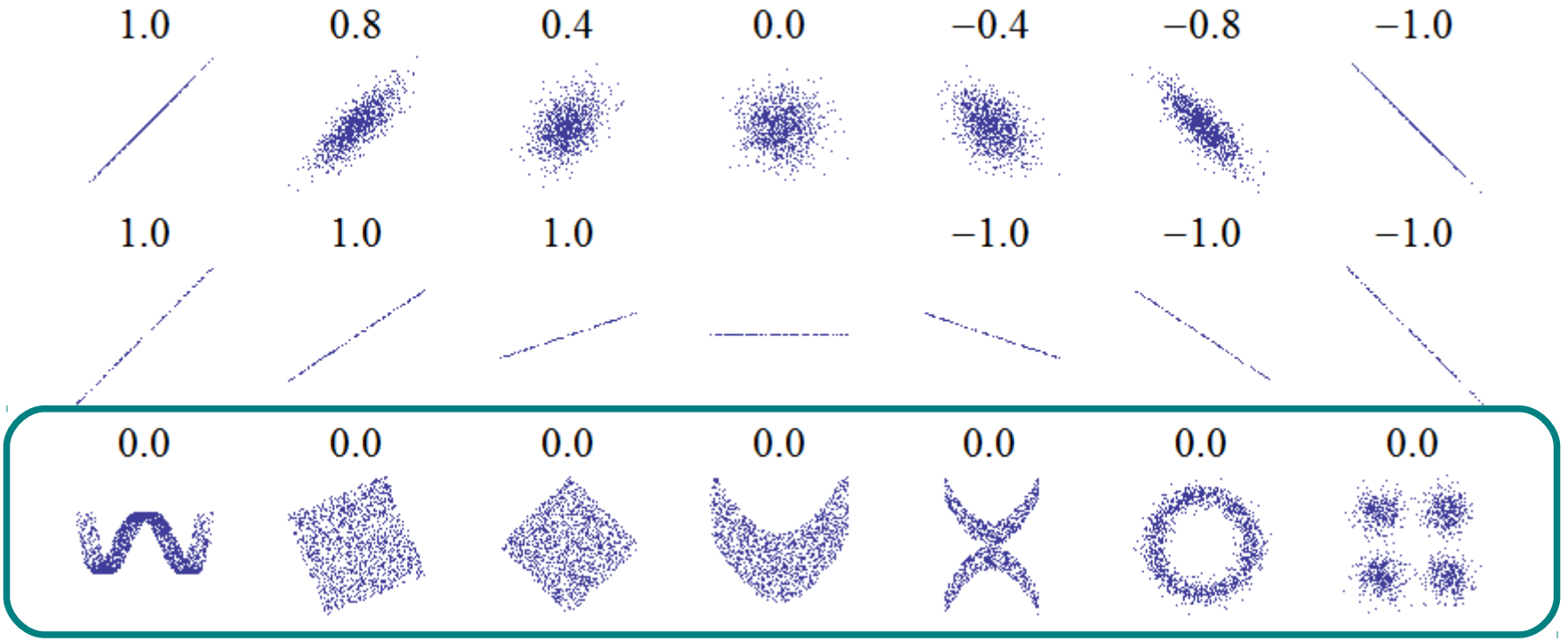
Korrelationskoeffizient

Einheitenloses, normiertes Maß für Korrelation zweier Zufallsvariablen

$$\rho \equiv \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}$$

$-1 \leq \rho \leq 1$

- **aber: aufgepasst bei nicht-linearen Zusammenhängen**



[von: wikipedia.de]

Zusammenfassung

- **(arithmetischer) Mittelwert einer Verteilung:**

einfach: $\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$ gewichtet: $\bar{x} = \frac{\sum_{i=1}^{N} x_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}$

- **Standardabweichung einer Verteilung:**

$$\sigma_x = \sqrt{V(x)} = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2} = \sqrt{\overline{x^2} - \bar{x}^2}$$

- **Unsicherheit auf dem Mittelwert einer Verteilung:**

einfach: $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$ gewichtet: $\sigma_{\bar{x}} = \sqrt{\frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}}$

- **linearer Korrelationskoeffizient zweier Variablen:**

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \qquad (-1 \le \rho \le 1)$$