Statistische Methoden der Datenanalyse

Wintersemester 2012/2013

Albert-Ludwigs-Universität Freiburg

Prof. Markus Schumacher, Dr. Stan Lai

Physikalisches Institut Westbau 2 OG

Markus.Schumacher@physik.uni-freiburg.de

stan.lai@cern.ch

http://terascale.physik.uni-freiburg.de/lehre/ws_1213/statmethoden_ws1213

Kapitel 1

Beschreibung von Daten

Deskriptive Statistik

Kapitel 0

Beschreibung von Daten

Deskriptive Statistik

Datensatz und Typen von Daten

Messung liefert Stichprobe von Messdaten vom Umfang n: x_i i=1,n

Ziel: Information extrahieren
Auditorium erklären

Mittel: übersichtliche Darstellung → Graphik
Bestimmung weniger charakteristischer Zahlen

Typen von Daten:

numerische:
- diskrete: abzählbar viele z.B: natürliche oder ganze Zahlen
- kontinuierliche: meist reelle Zahlen (Größe, Gewicht, …)

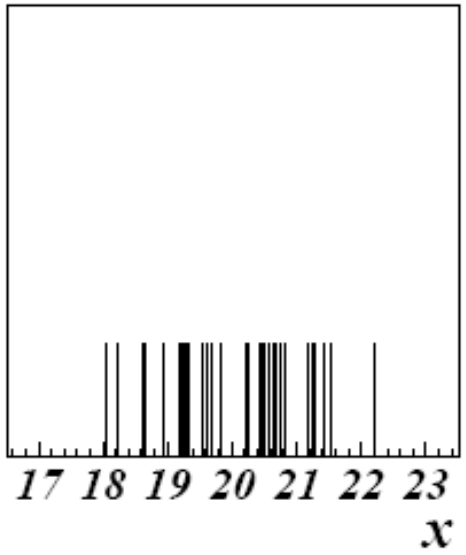
qualitative:
Farben, Sorten, ja/nein, …

Darstellung von Daten

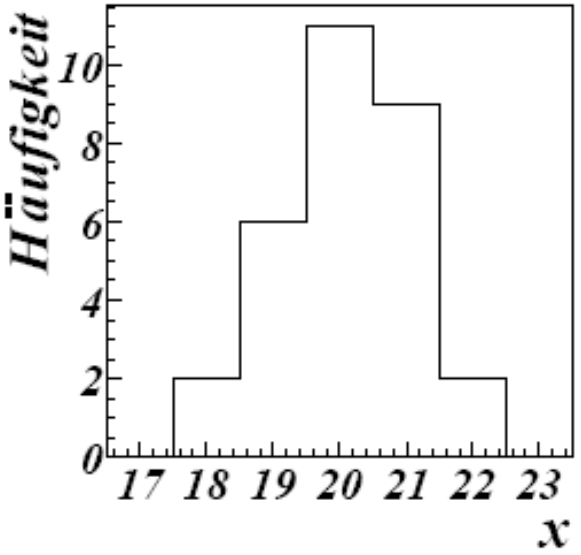
Tabelle

x[00] = 18.21	x[15] = 20.21
x[01] = 20.63	x[16] = 19.6
x[02] = 19.24	x[17] = 19.69
x[03] = 20.24	x[18] = 19.3
x[04] = 19.21	x[19] = 20.49
x[05] = 21.25	x[20] = 19.62
x[06] = 19.69	x[21] = 20.67
x[07] = 20.73	x[22] = 20.56
x[08] = 20.43	x[23] = 18.62
x[09] = 18.05	x[24] = 20.46
x[10] = 21.52	x[25] = 18.64
x[11] = 21.41	x[26] = 22.22
x[12] = 18.95	x[27] = 21.16
x[13] = 19.83	x[28] = 19.54
x[14] = 20.81	x[29] = 21.27

Balkendiagramm



Histogramm



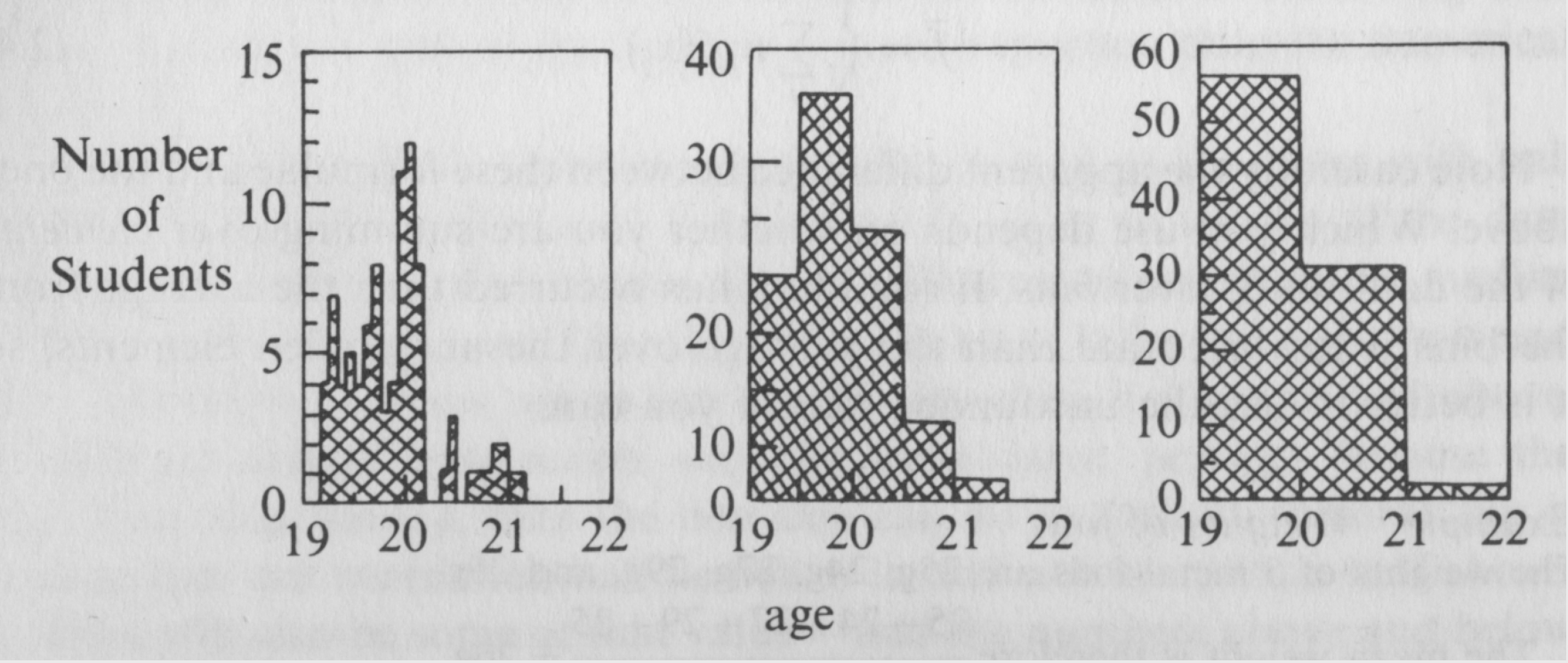
In Naturwissenschaften: meistens Histogramme (auch mehrdimensionale)

Andere Darstellungsoptionen: Säulendiagramme, Kuchendiagram, …

Wahl der Bin-Breiten

Kriterien; - statistische Fluktuationen in Bins klein → große Breiten
- keine Struktur/ Eigenschaft "verstecken" → kleine Breiten

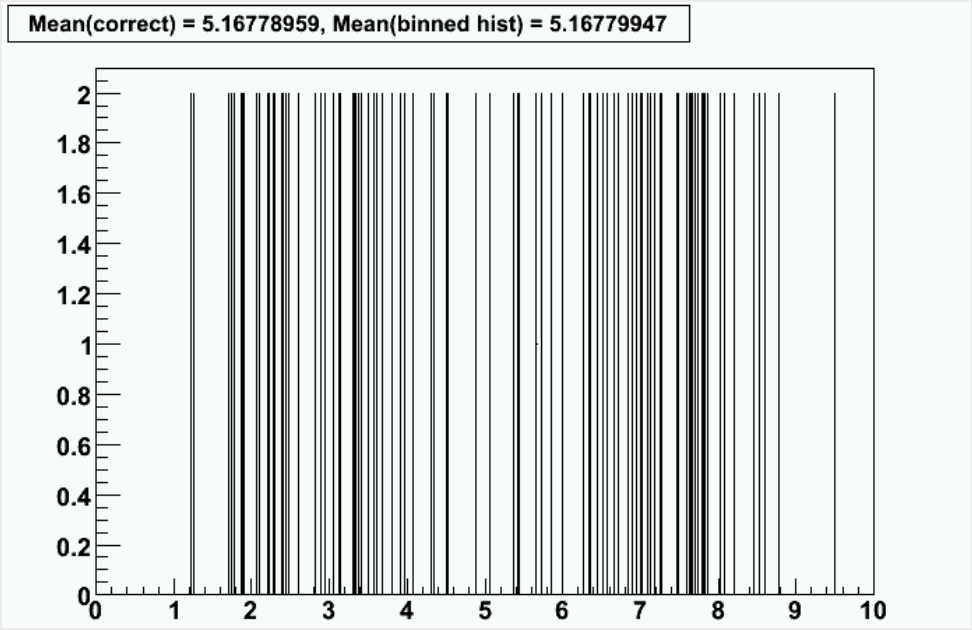
Faustregel: > 5 bis 10 Einträge pro Bin



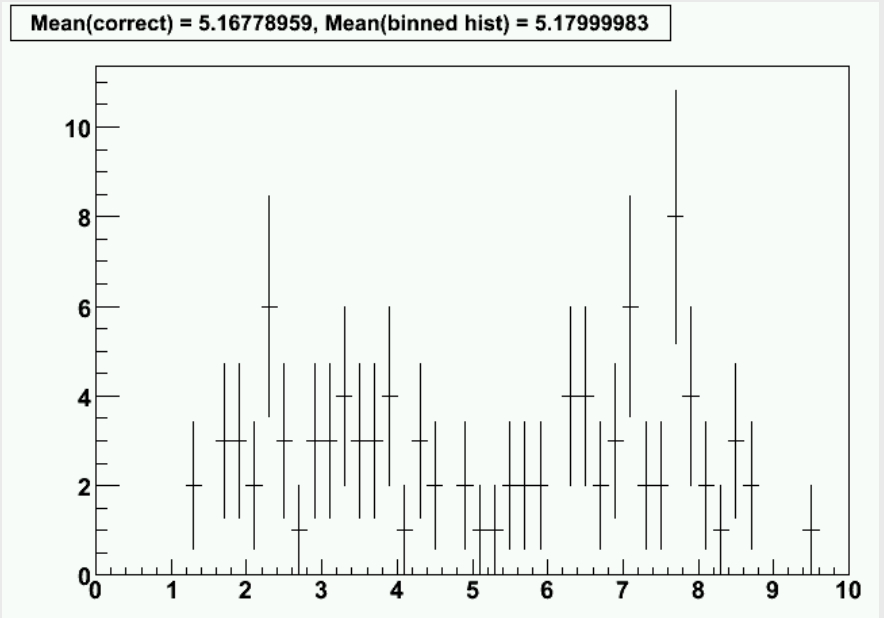
Wahl der Bin-Breiten (II)

Beispiel:
Stichprobe vom Umfang 100 gezogen aus Gauss(μ=3,σ=1)+Gauss(μ=7,σ=1)
Mittelwert aus Einzelwerten: Mean (correct)
Mittelwert aus Binmittelwerten und Binhäufigkeit: Mean (binned hist)

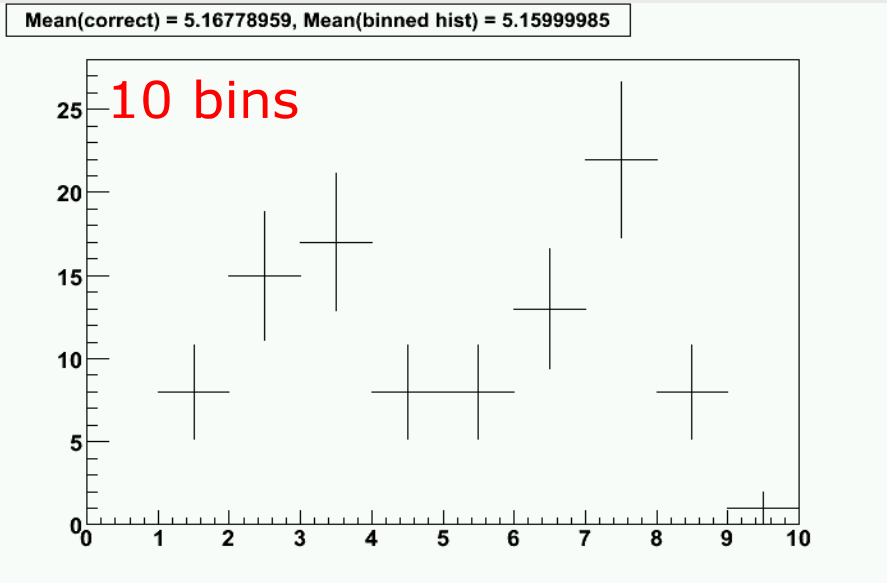
Kein Binning/Balkendiagramm

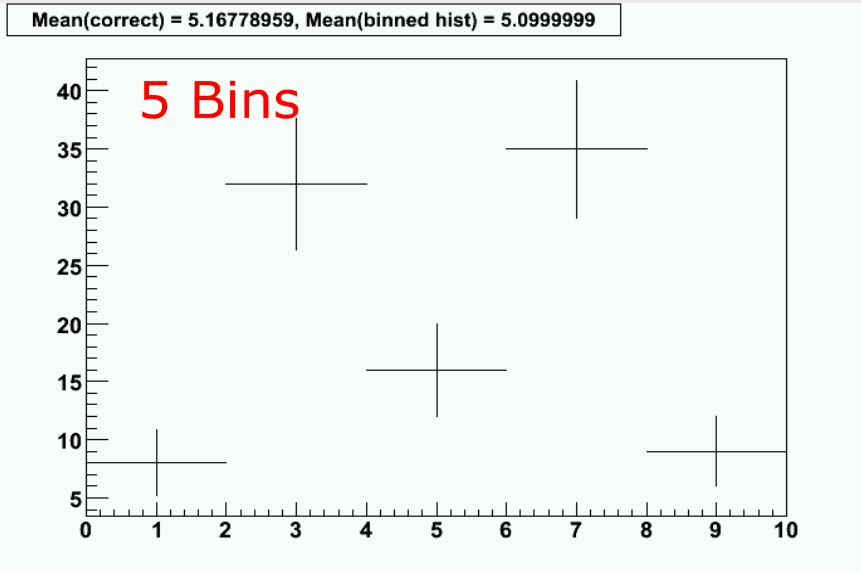


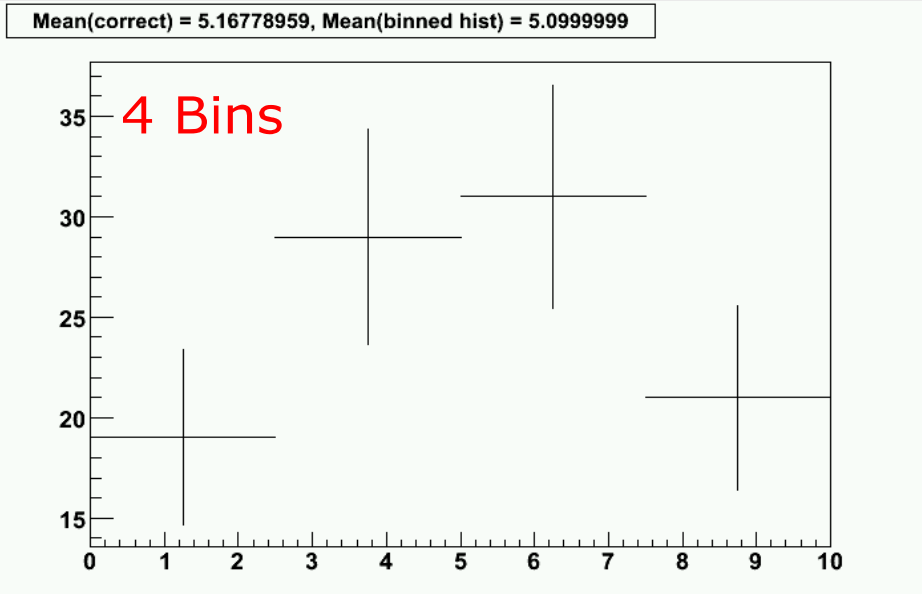
50 Bins (Breite = 0.2)

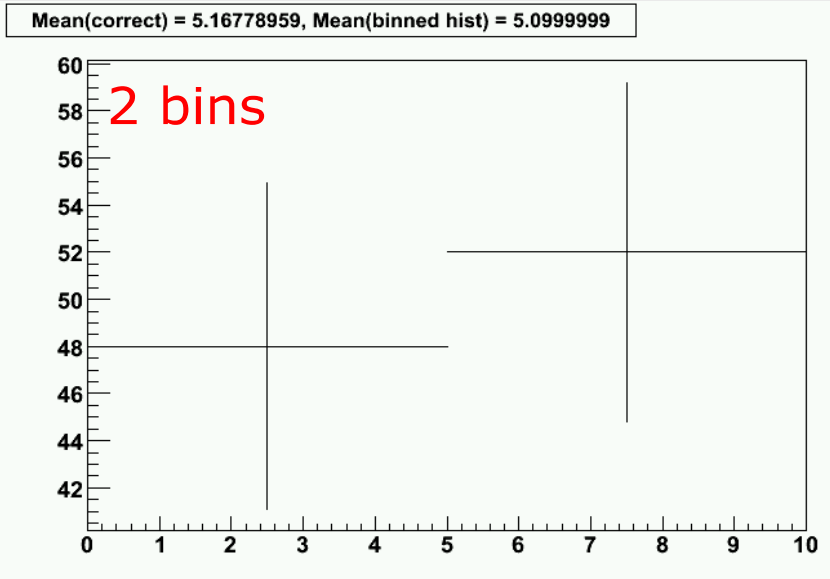


Wahl der Bin-Breiten (III)









Charakteristische Größen: “Schwerpunkt”

“Schwerpunkt”

meistens: arithmetischer Mittelwert

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

Bei Einteilung in j Bins mit
n_j Einträgen und Binzentralwerten x_j
Achtung: Informationsverlust durch Binning → ungenauer Mittelwert

$$\bar{x} = \frac{1}{N}\sum_j n_j x_j$$

seltener: harmonisches

$$\frac{N}{1/x_1 + 1/x_2 + 1/x_3 + \cdots + 1/x_N}$$

geometrisches Mittel

$$\sqrt[N]{x_1 x_2 x_3 \cdots x_N}$$

Charakteristische Größen: “Schwerpunkt” (II)

Modalwert (Mode):
wahrscheinlichster, häufigster Wert in der Stichprobe

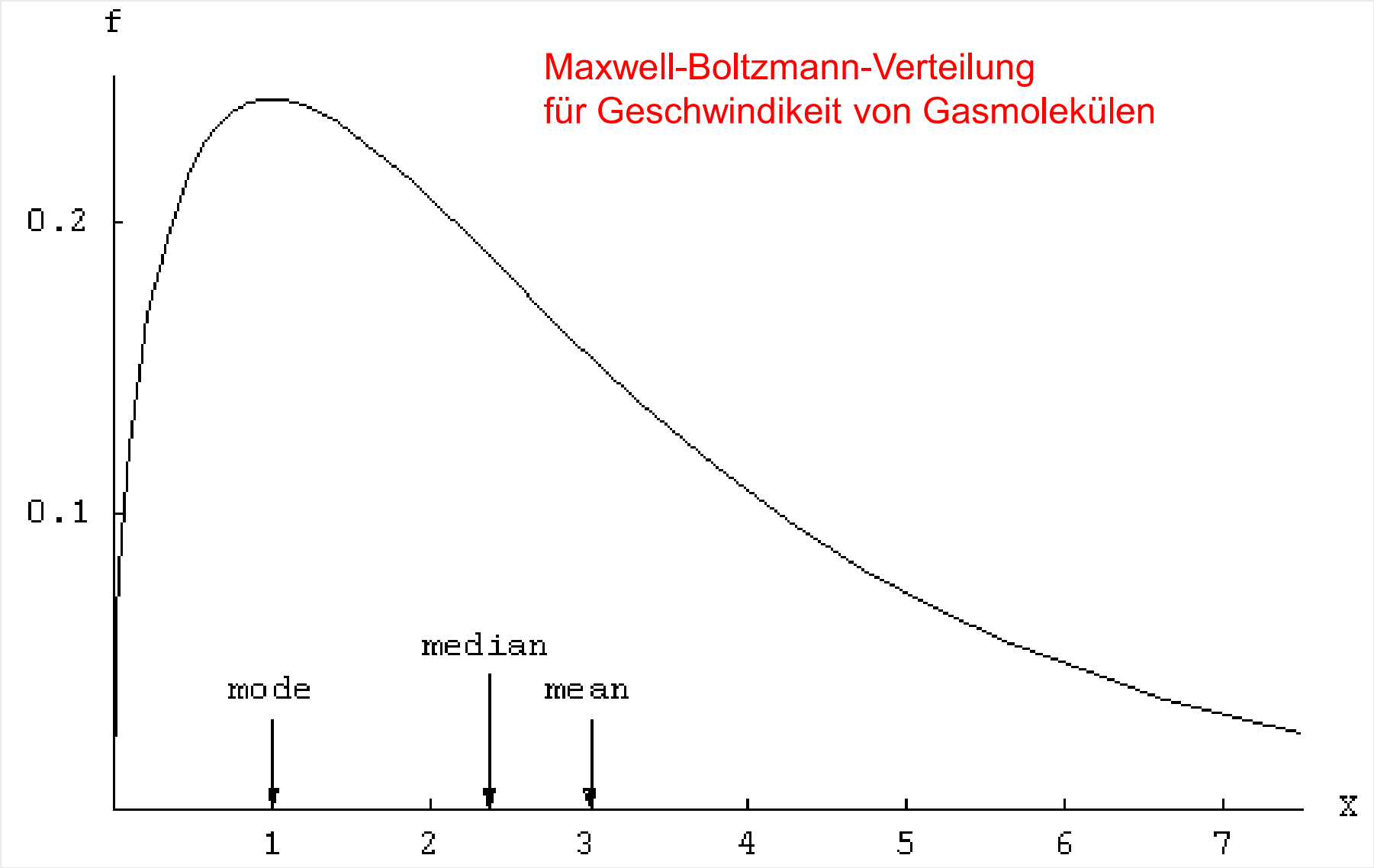
Median:
50% der Daten oberhalb und 50% der Daten unterhalb

Stichprobenumfang n ungerade:
(n-1)/2 unterhalb, (n-1)/2 oberhalb Median = “(n-1)/2+1”te Wert

Stichprobenumfang n gerade:
Median = Mittelpunkt zwischen “n/2”ten und “(n/2)+1”ten Wert

Bei Histogrammen keine einfache Angabe möglich

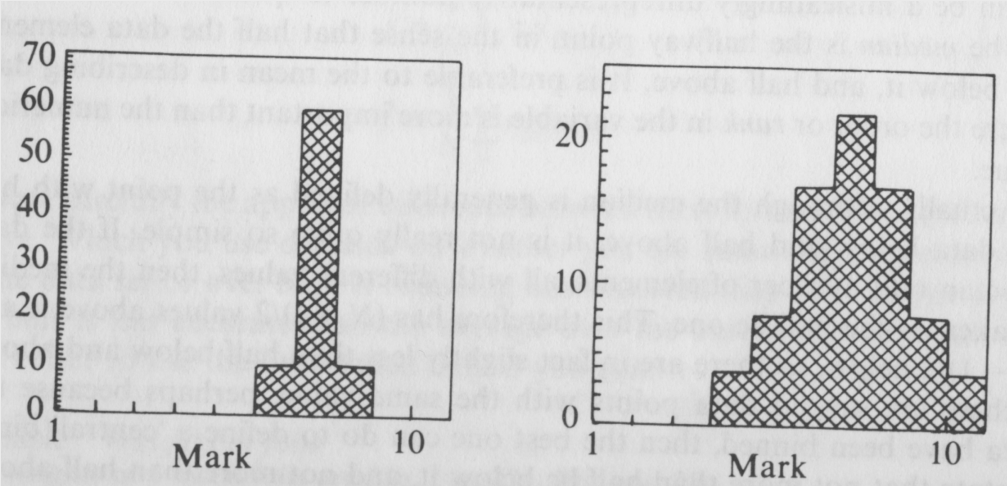
Mittelwert, Median und Modalwert (Mode)



Charakteristische Größen: “Streuung/Breite”

2 Verteilungen mit gleichem Mittelwert aber unterschiedlicher Streuung

→2. charakteristische Zahl: Streuung/Breite/Dispersion



Erster Versuch: mittlerer Abstand vom Mittelwert

kein sinnvolles Ergebnis

$$\frac{1}{N}\sum_i (x_i - \bar{x}) = \frac{1}{N}\sum_i x_i - \frac{1}{N}\sum_i \bar{x}$$
$$= \bar{x} - \bar{x}$$
$$= 0.$$

Charakteristische Größen: “Streuung/Breite” (II)

Varianz = mittlere quadratische Abweichung vom arithmetischem Mittel

$$V = \frac{1}{N}\sum_i (x_i - \bar{x})^2$$

kleine Rechung liefert …

$$V(x) = \overline{x^2} - \bar{x}^2$$

$$V(x) = \frac{1}{N}\sum x_i^2 - \left(\frac{1}{N}\sum x_i\right)^2$$

Standardabweichung (gleiche Einheit wie Messwerte)

$$\sigma = \sqrt{V(x)}$$

$$\sigma = \sqrt{\overline{x^2} - \bar{x}^2}$$

In Praxis: meist Standardabweichung da gleiche Einheit wie Messwerte
In Theorie: oft Varianz da mathematisch einfacher zu handhaben

Charakteristische Größen: “Streuung/Breite” (II)

Andere Definition von Varianz und Standardabweichung:

$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}.$$

Varianz verändert durch Besselkorrektur N/(N-1)

Erste Definition liefert keinen erwartungstreuen Schätzer der Varianz
Erläuterung später. Für große n ist Unterschied marginal.

Weitere Größen für die Breite:

FWHM = volle Breite auf halber Höhe

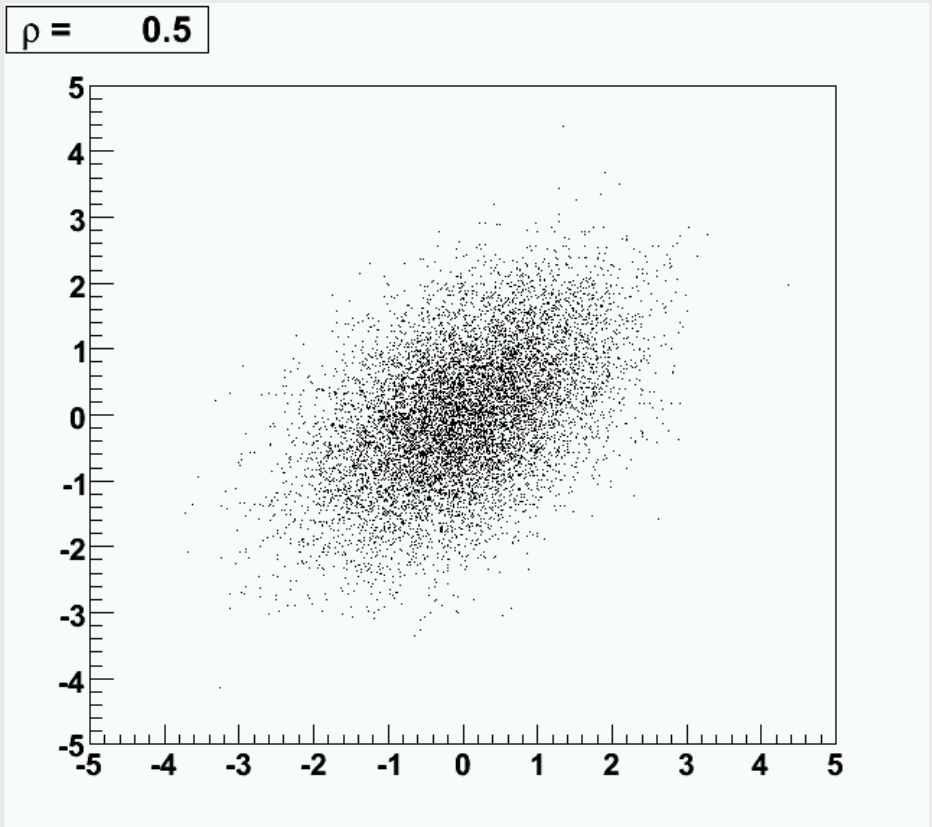
mittlere absolute Abweichung vom Mittelwert

$$\frac{1}{N}\sum_i |x_i - \bar{x}|$$

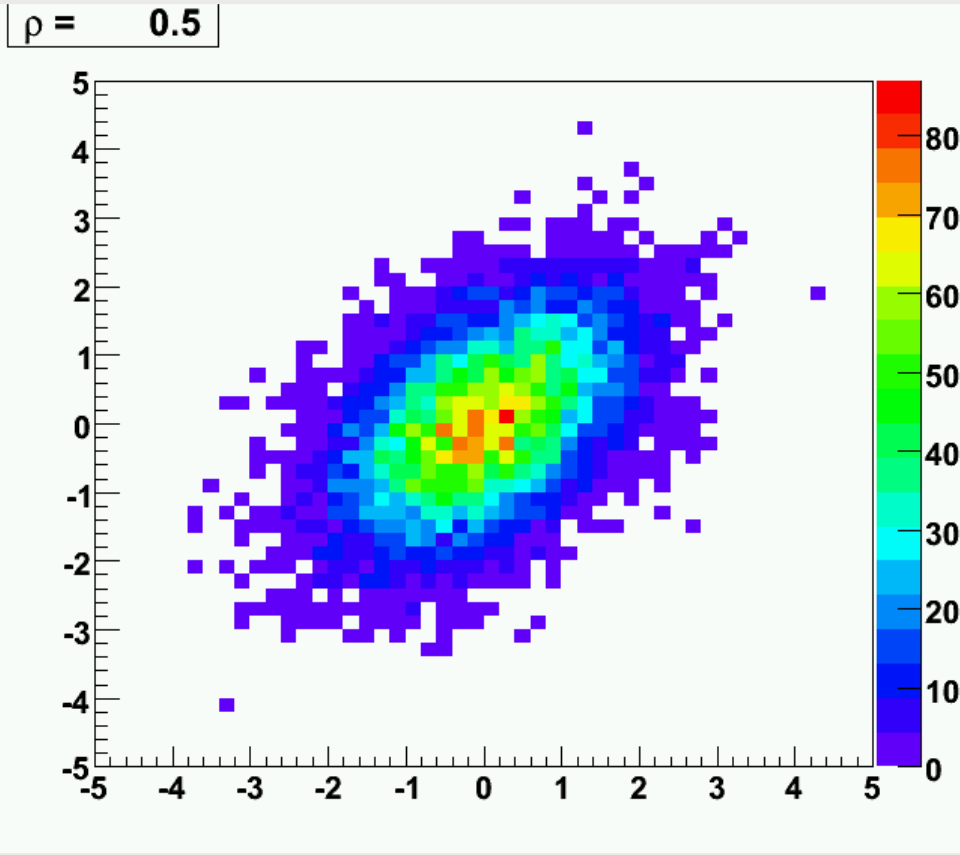
2 Observablen: Darstellungsoptionen

Teilweise besteht ein Messung aus mehreren Observablen
z.B: Größe und Gewicht, Noten in 2 Fächern, …

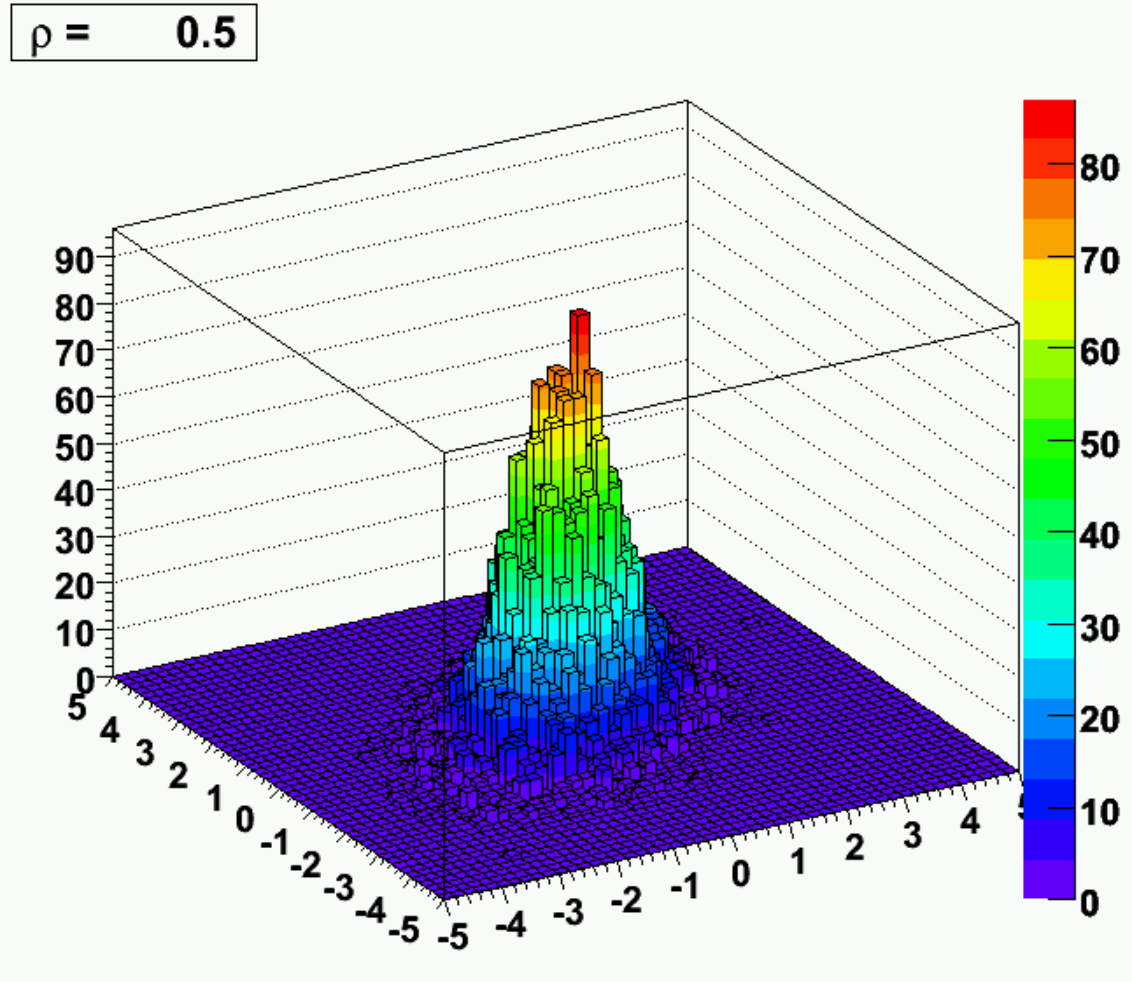
“Scatter-Plot”



Kontur-Histogramm



2 Observablen: Darstellungsoptionen (II)



3-dimensionale Darstellung als “Lego-Plot”

Kovarianz für 2 Observable

Zusätzliche Frage: Beziehung/Relation zwischen Messgrößen?

→ Kovarianz:

$$\begin{aligned}\operatorname{cov}(x,y) &= \frac{1}{N}\sum_i (x_i-\bar{x})(y_i-\bar{y}) \\ &= \overline{(x-\bar{x})(y-\bar{y})} \\ &= \overline{xy} - \bar{x}\,\bar{y}.\end{aligned}$$

$$\operatorname{cov}(x,x) = V(x)$$

wenn für x > Mittelwert(x), bevorzugt y > Mittelwert(y) dann cov(x,y) > 0
< <

Beispiele: Größe und Gewicht: Kovarianz > 0
Fitness und Gewicht: Kovarianz < 0
Größe und IQ: Kovarianz = 0

Korrelation für 2 Observable

Kovarianz hat teilweise komische Einheit
(z.B.für Größe und Gewicht: kg m)
→Interpretation schwierig

Definition der Korrelation/ des Korrelationskoeffizienten

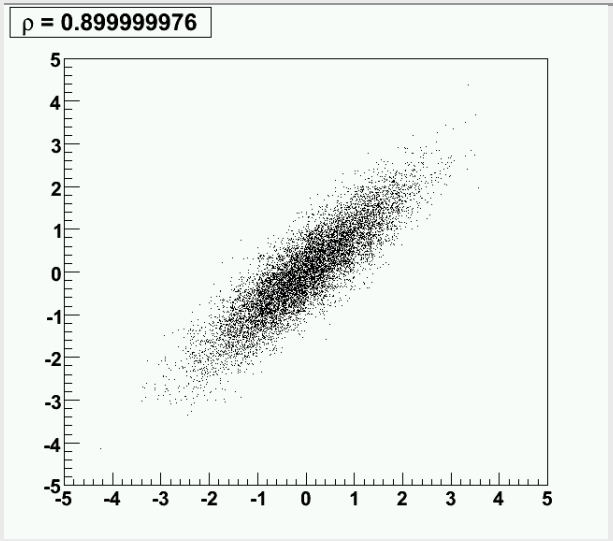
$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

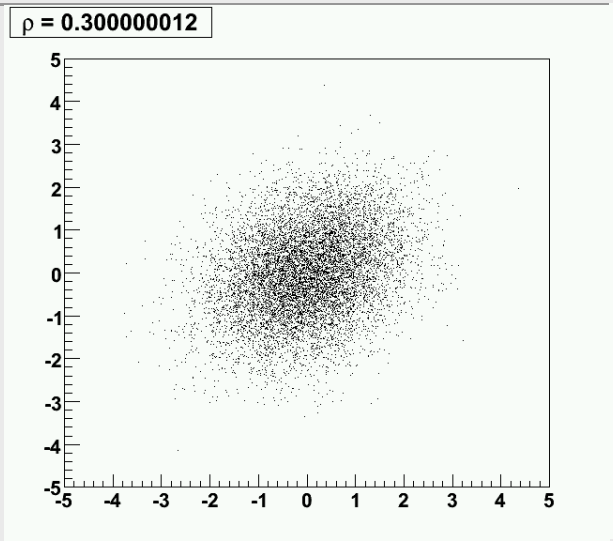
$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}.$$

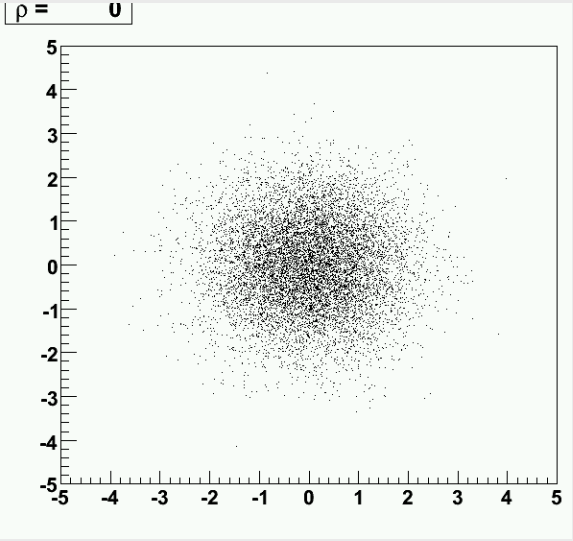
Korrelation zwischen 0 und 1

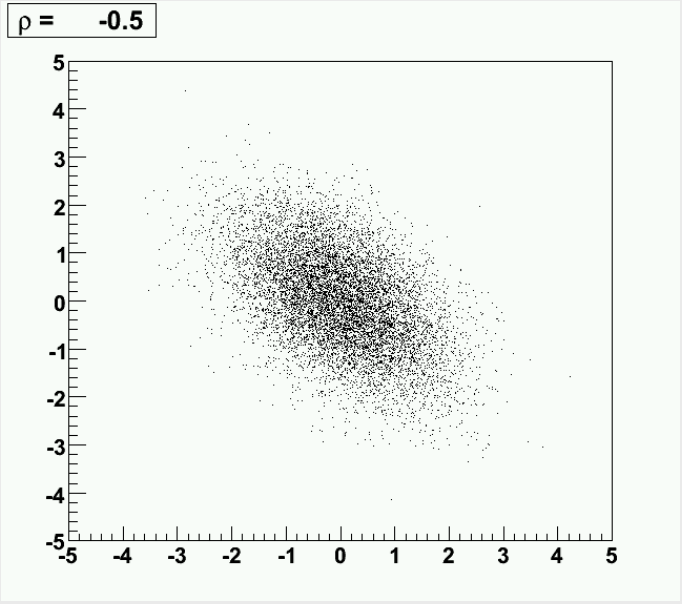
= 0: unkorrelliert
= ±1: vollständig korreliert
d.h. Kenntnis von x liefert Kenntnis von y

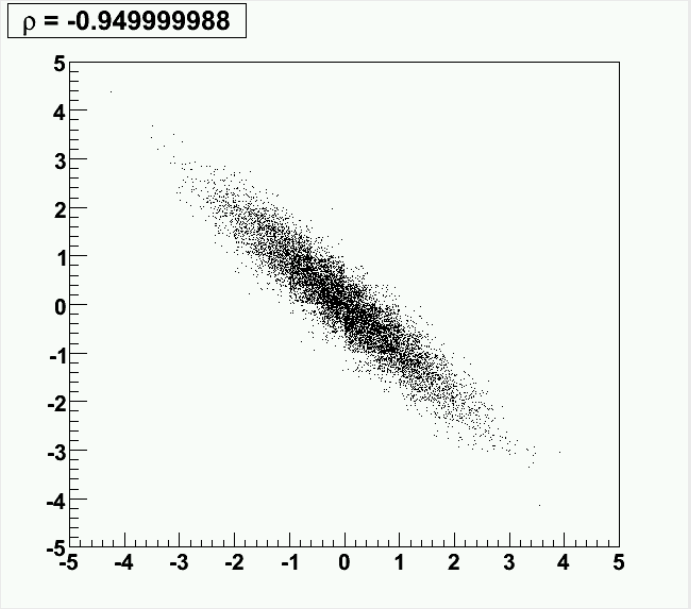
Beispiel für Korrelationen



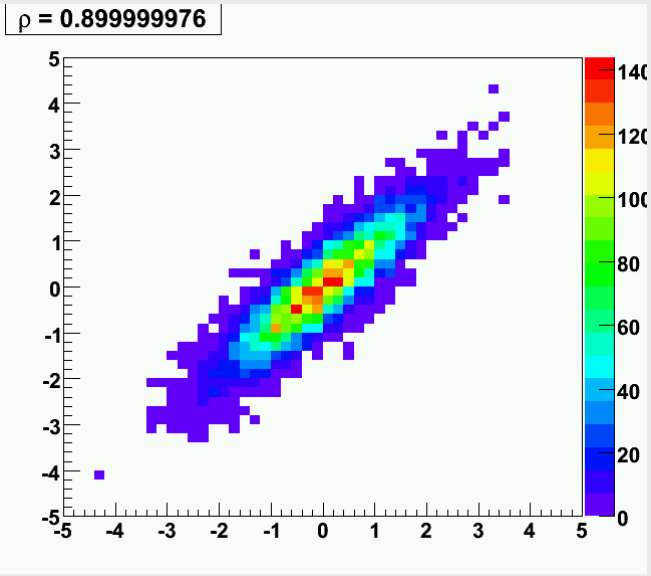


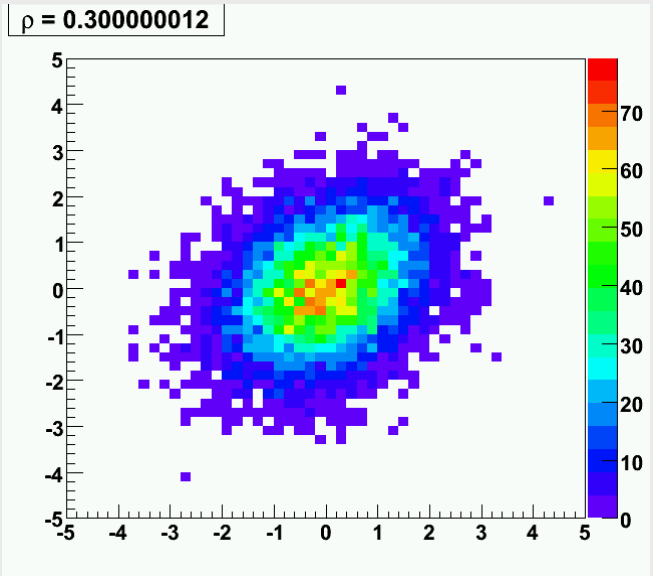


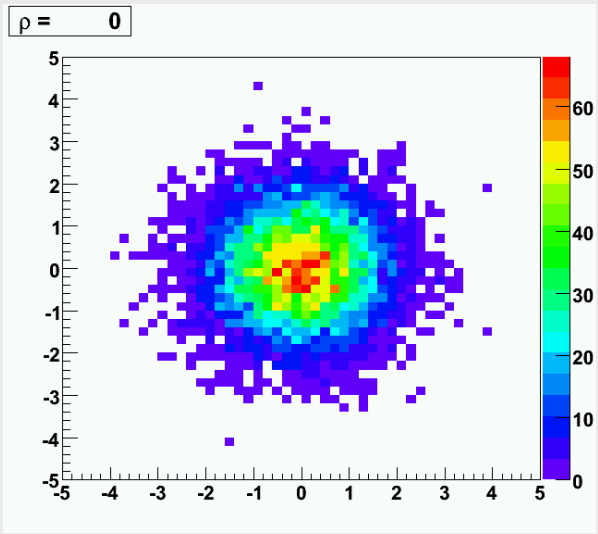


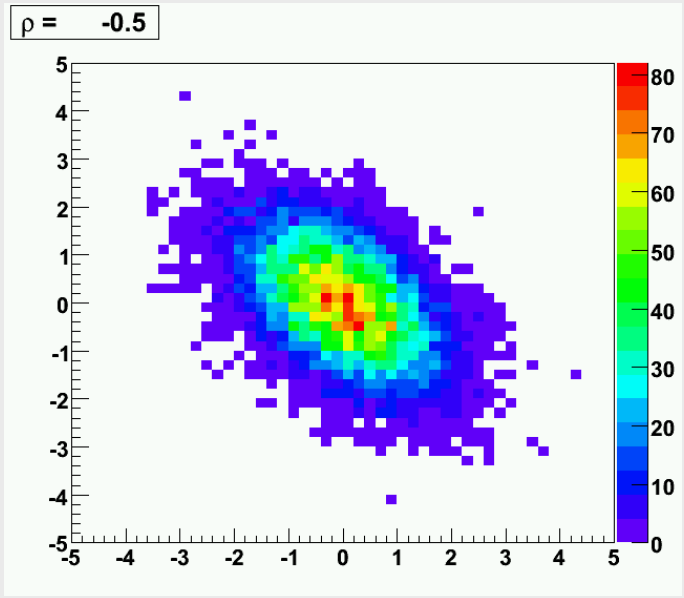


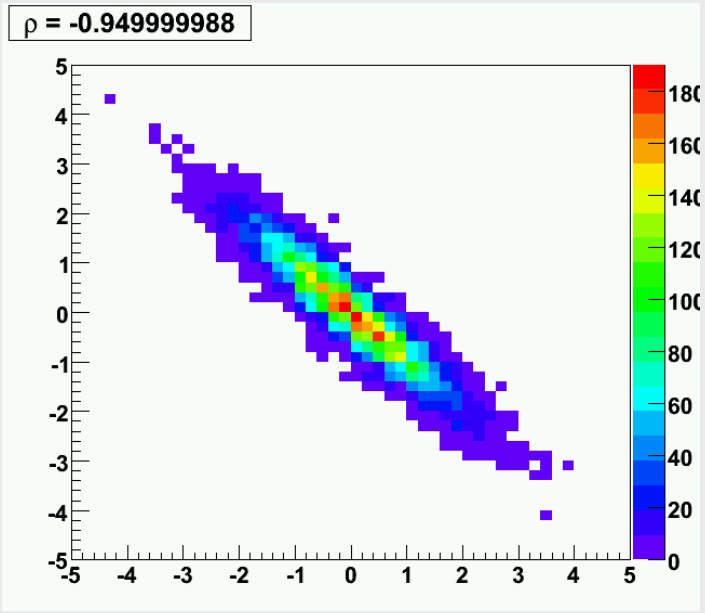
Beispiel für Korrelationen (II)











Beispiel für Korrelationen (II)