Beschreibung von Daten und Verteilungen

Blockkurs für Bachelor-Studenten - Einführung - 10 April 2018

Sascha Mehlhase (LMU München)

Worum soll es heute gehen?

- Charakterisierung von Daten
  - Daten-Arten
  - Mittelwert(e)
  - Breite
  - Korrelation(en)

- wichtige Verteilungen
  - Gleichverteilung
  - Binomialverteilung
  - Poissonverteilung
  - Gauss- oder Normalverteilung
  - Zentraler Grenzwert-Satz

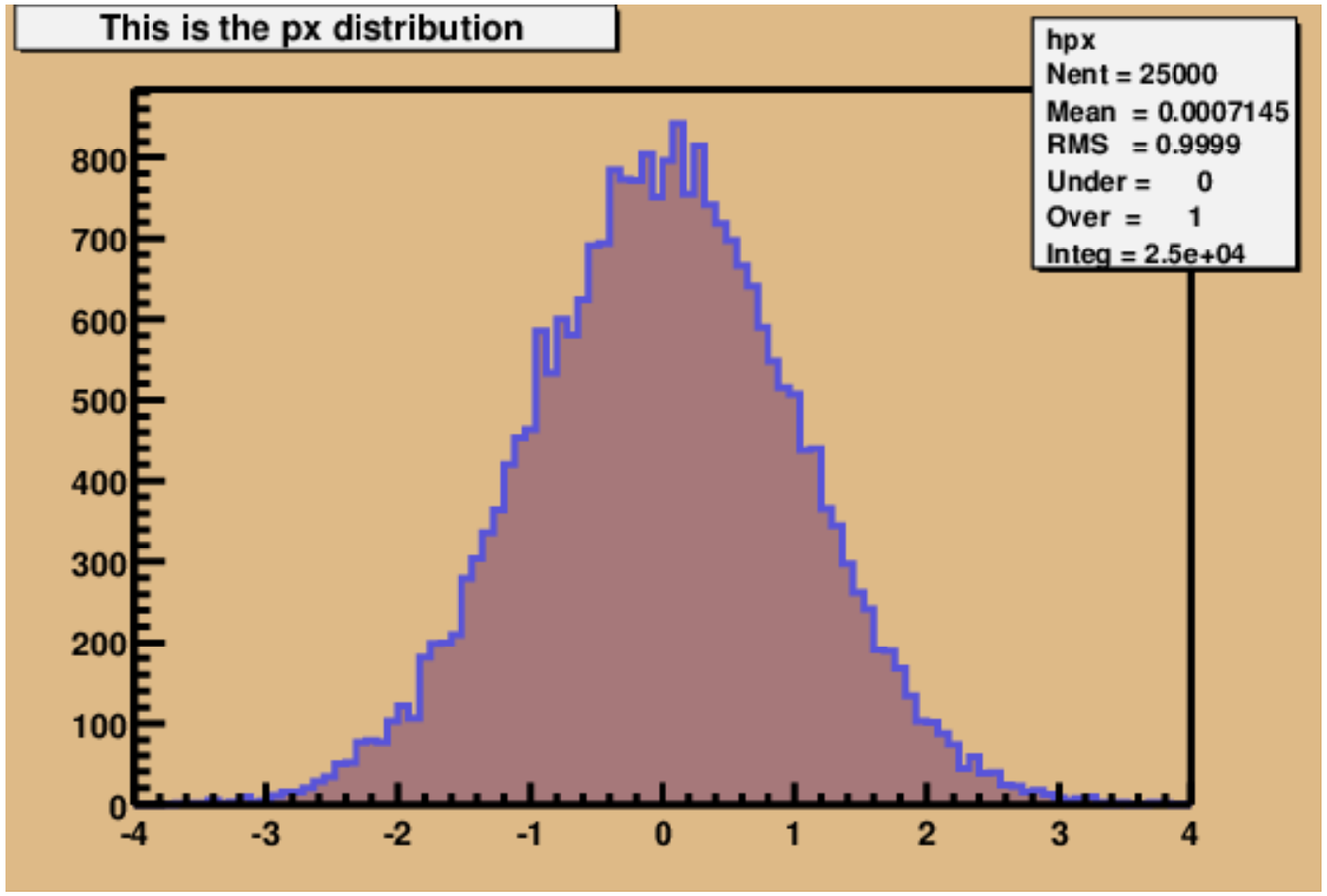
Charakterisierung von Daten

- Daten-Arten
  - Unterscheidung nach
    - **qualitativ**
      - z.B. Kategorien, Farben, Befinden, Emotionen, …
      - durchaus häufig: medizinische Studien, Meinungsumfragen, Marktforschung, …, aber unangenehm zu behandeln, mathematisch schwer fassbar.

    - **quantitativ**
      - numerische Werte, überwiegender Datentyp in Physik, im folgenden Beschränkung darauf.
        Weitere Unterteilung in
        - **diskret**e Werte - z.B. Anzahl (Teilchen im Ereignis, Personen im Fahrzeug, Münzen in der Börse)
        - **kontinuierlich**e Werte - z.B. Masse des Teilchens, Größe einer Person, Intensität einer Quelle

Charakterisierung von Daten

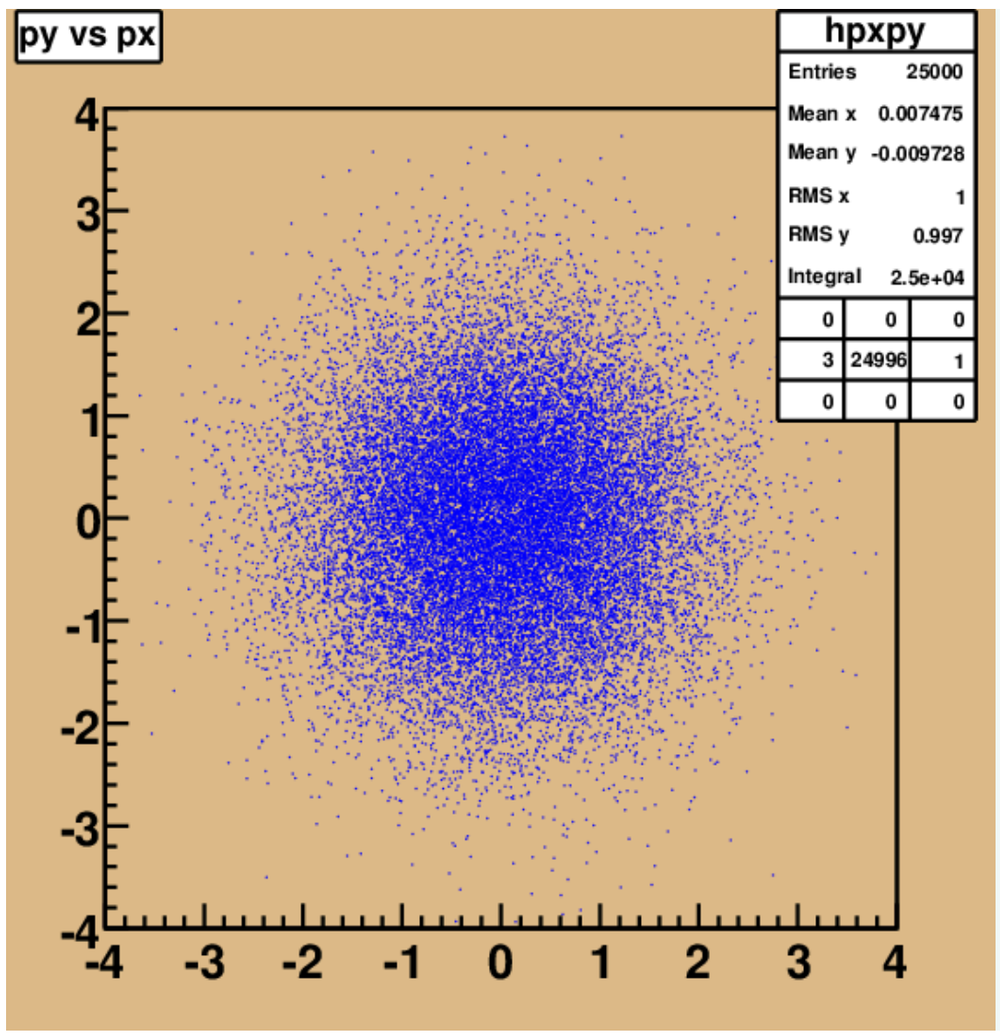
- Datenvisualisierung
  - Häufigkeitsverteilung
    - eine oder mehrere Größen werden wiederholt gemessen.
      Darstellung in ein- oder mehr-dimensionalen ‘Histogrammen’.
    - Bereich `xlow` bis `xhigh` unterteilt in `nchannel` Intervalle
      ```
      TH1F *h1 = new TH1F("h1", "name", nchannel, xlow, xhigh);
      ```
    - Jede Messung x wird in das Histogramm gefüllt
      ```
      h1->Fill(x);
      ```
      Inhalt von Kanal `nchannel*(x-xlow)/(xhigh-xlow)` um 1 erhöht

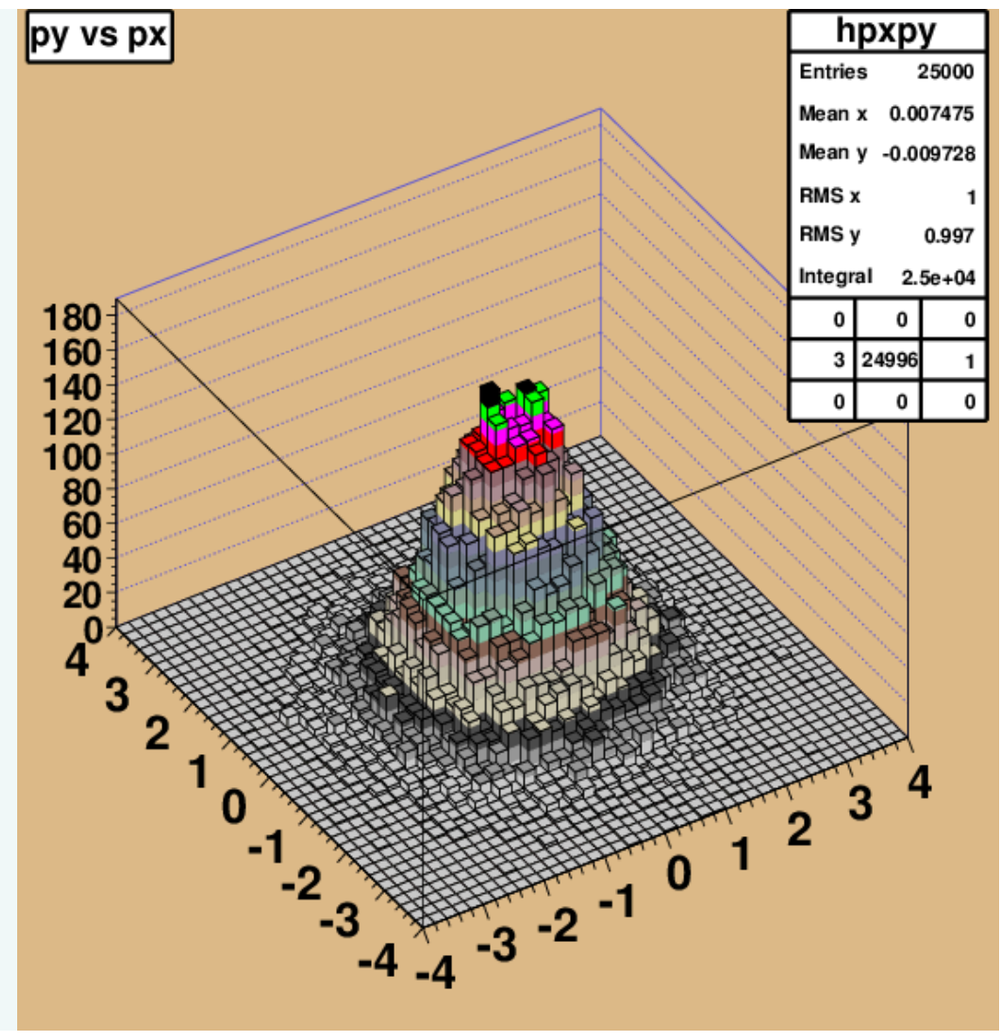


Charakterisierung von Daten

- Datenvisualisierung
  - Häufigkeitsverteilung
    - analoge Erweiterung auf zwei Dimensionen

```
TH2F *h2 = new TH2F("h2", "mytitle", nx, xlow, xhig, ny, ylow, yhigh);
h2->Fill(x, y);
```





Charakterisierung von Daten

- Mittelwert
  - Gebräuchlichste Art Daten x1, x2, …, xN mit einer einzelnen Größe zu beschreiben ist das **arithmetische Mittel**:

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

Sinnvoll in vielen Fällen, aber nicht die einzige Art:

  - geometrisches Mittel:

$$\sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$$

  - harmonisches Mittel:

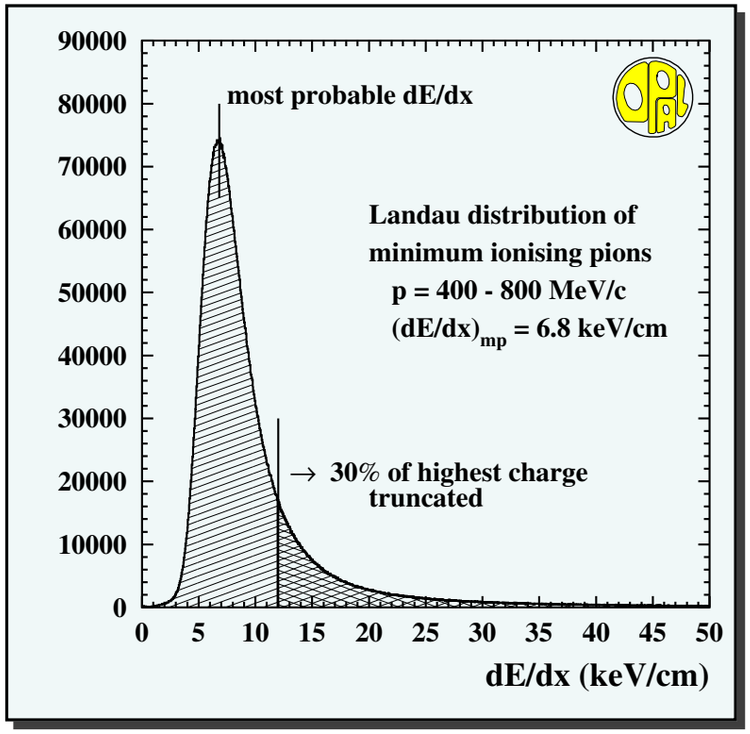
$$\left(\frac{N}{1/x_1 + 1/x_2 + \ldots + 1/x_N} \right)^{-1}$$

Charakterisierung von Daten

- Median
  - “Wert in der Mitte”: Die Hälfte der Werte ist kleiner, die andere Hälfte größer.
  - Sehr nützlich für Datensätze, bei denen die zugrundeliegende Verteilung unklar ist oder lange Ausläufer hat oder die Reihenfolge wichtiger ist als die numerische Größe.
  - Am einfachsten zu bekommen durch Sortieren des Datensatzes, dann

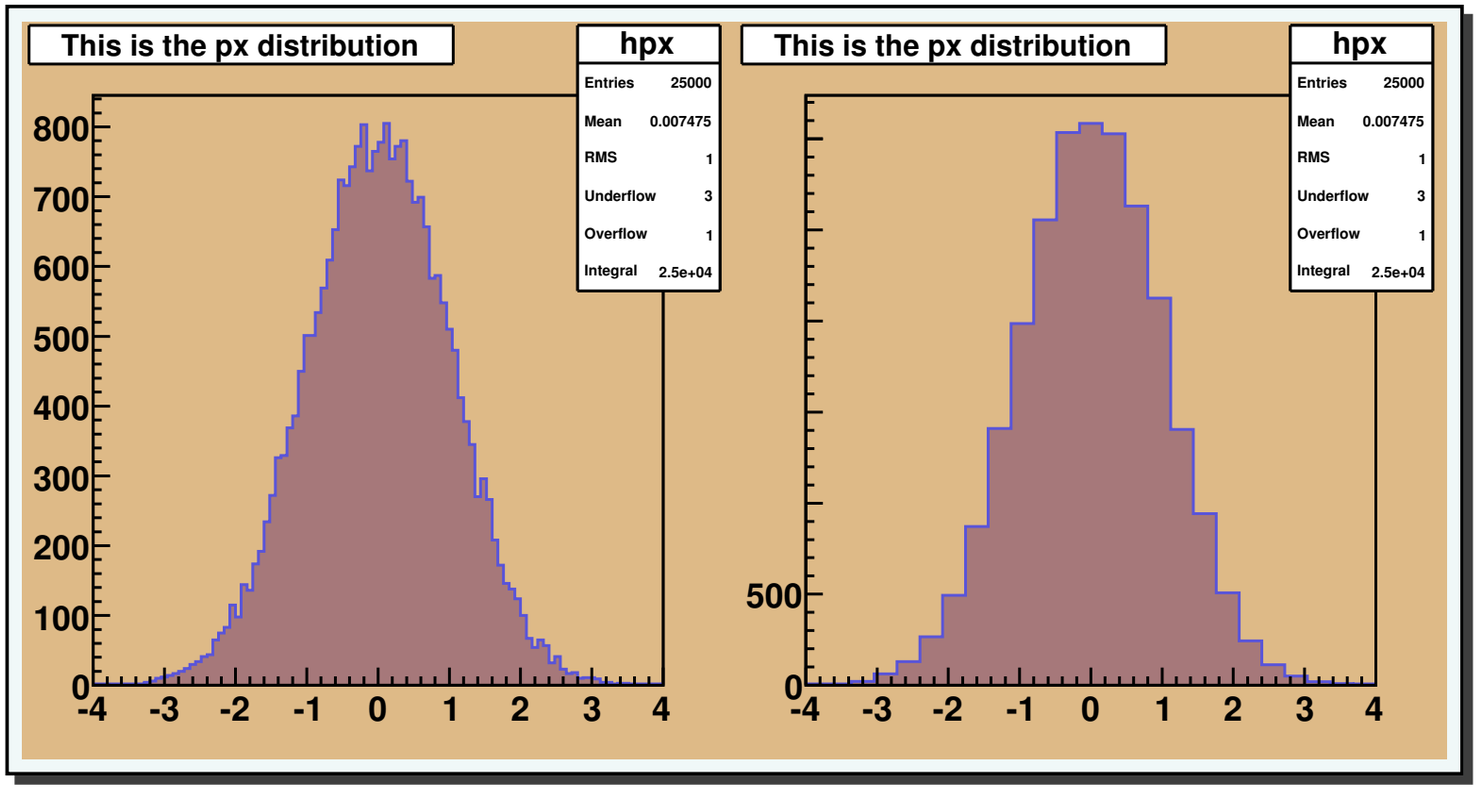
$$\boxed{Median = x_{N/2}}$$

- truncated mean
  - Mittelwert nur aus einem eingeschränkten Bereich bestimmt
  - Wegen langen Ausläufern, asymmetrischen Beiträgen, etc.
  - Ein Beispiel aus Teilchenphysik ist die Messung des spezifischen Energieverlusts, typisch werden die höchsten 30% verworfen.



Charakterisierung von Daten

- wahrscheinlichster Wert
  - bei “gebinnten” Datensätzen
  - x_{max} vom Bin mit höchstem Eintrag
  - ungebräuchlich da sehr schwankend und abhängig von Ereigniszahl und Anzahl Bins



Charakterisierung von Daten

- “Breite"
  - Nach Mittelwert die 2. wichtige Größe zur Charakterisierung eines Datensatzes.
  - Am gebräuchlichsten die **Varianz**

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

  - bzw. die **Standardabweichung**

$$\sigma(x) = \sqrt{V(x)} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum x_i^2}{N} - \bar{x}^2}$$

Charakterisierung von Daten

- “Breite"
  - Weitere Möglichkeiten zur Charakterisierung der Breite:

$$\boxed{\frac{\sum |x_i - \bar{x}|}{N}}$$

    ungebräuchlich, mathematisch sehr unangenehme Eigenschaften

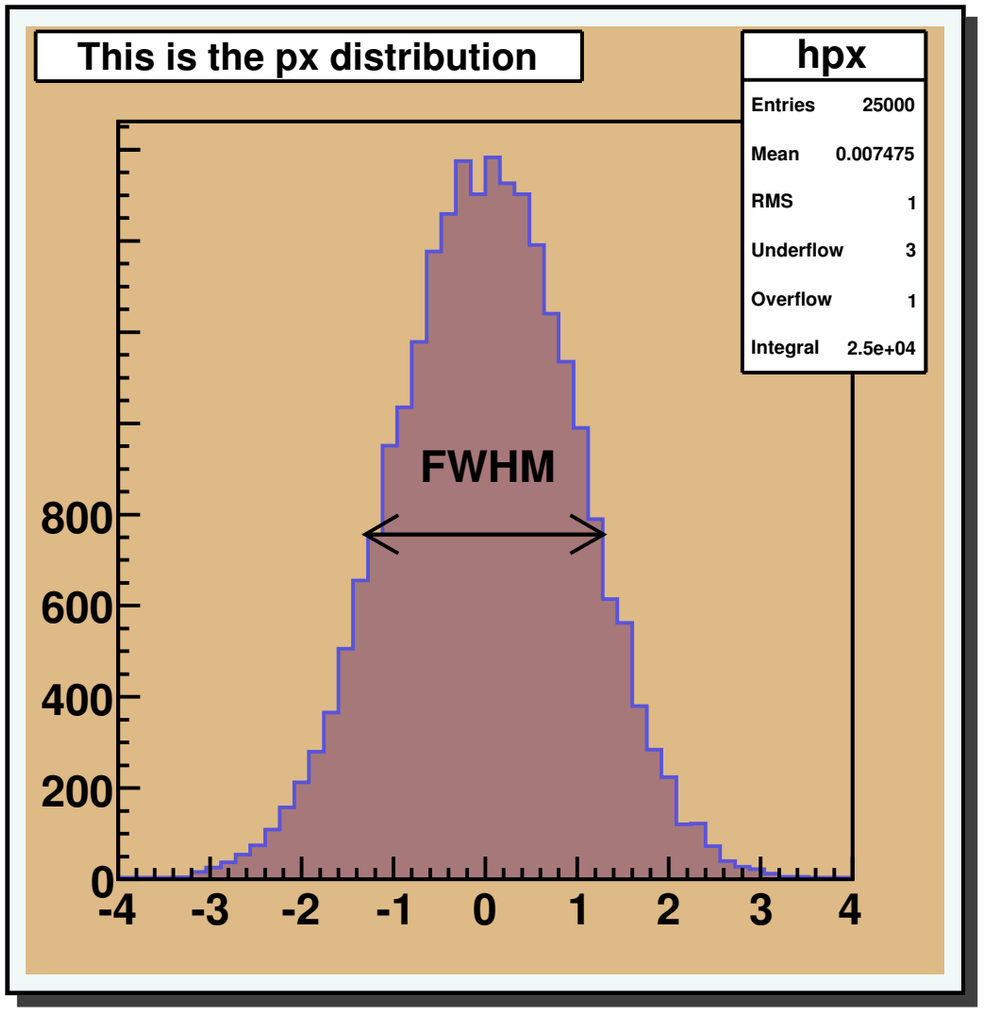
  - Oft nützlich die sogenannten Quantile:
    - unteres Quartil: 25 % der Werte sind kleiner
    - oberes Quartil: 75 % der Werte sind kleiner
    - Breite als Differenz zwischen oberem und unterem Quartal
    - oder beliebige Percentile: XX % Percentil = XX % der Werte sind kleiner

Charakterisierung von Daten

- “Breite"
  - Bei “gebinnten” Datensätzen ist die FWHM (= full width half maximum) eine robuste Alternative:

$$FWHM = x^{high}_{max/2} - x^{low}_{max/2}$$

  - Vorteil: Beschränkt sich auf zentralen Teil der Verteilung, lange asymmetrische Ausläufer sind unkritisch.



Charakterisierung von Daten

- Skew
  - Naheliegend nach Mittelwert und Varianz auch höhere Potenzen zu betrachten:

$$\boxed{Skew = \frac{\sum(x_i - \bar{x})^3}{N\sigma^3}}$$

    Faktor $1/\sigma^3$ macht Skew dimensionslos

  - nützlich zur Charakterisierung der Asymmetrie, positiv bei Ausläufern nach rechts und v.v., verschwindet bei symmetrischen Verteilungen.

Charakterisierung von Daten

- Beschreibung mehrerer Variablen
  - Oft werden gleich mehrere Werte x_i, y_i, z_i, … pro “Ereignis” aufgenommen, z.B. Impuls und Richtung, Körpergröße und Gewicht, Abschlussnote in Theorie und Experimentalphysik, …
  - Neben Mittelwert und Standardabweichung der einzelnen Größen ist die (Un-) Abhängigkeit ein wichtiges Kriterium.
  - Angelehnt an die Varianz einer einzelnen Größe ist die **Kovarianz** zweier Größen gegeben durch:

$$\boxed{cov(x,y) = \frac{1}{N}\sum(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \overline{xy} - \overline{x} \cdot \overline{y}}$$

  - Wenn hohe (tiefe) Werte von x oft mit hohen (tiefen) Werten von y vorkommen ist die Kovarianz positiv, bei entgegengesetztem Verhalten negativ bei unabhängigen Größen verschwindet sie.

Charakterisierung von Daten

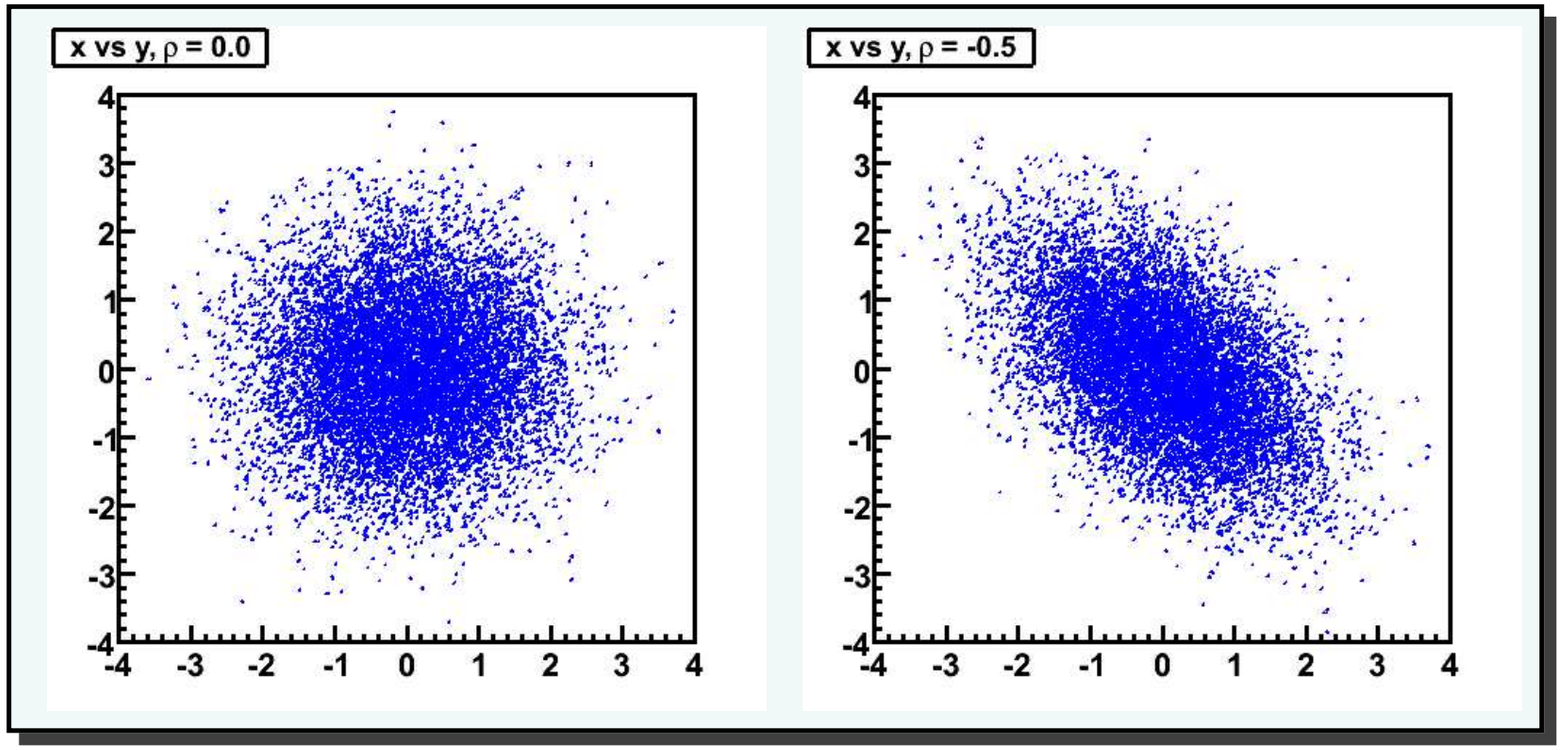
- Beschreibung mehrerer Variablen
  - Ein besseres Mass für die Abhängigkeit zweier Variablen ist die **Korrelation** ρ

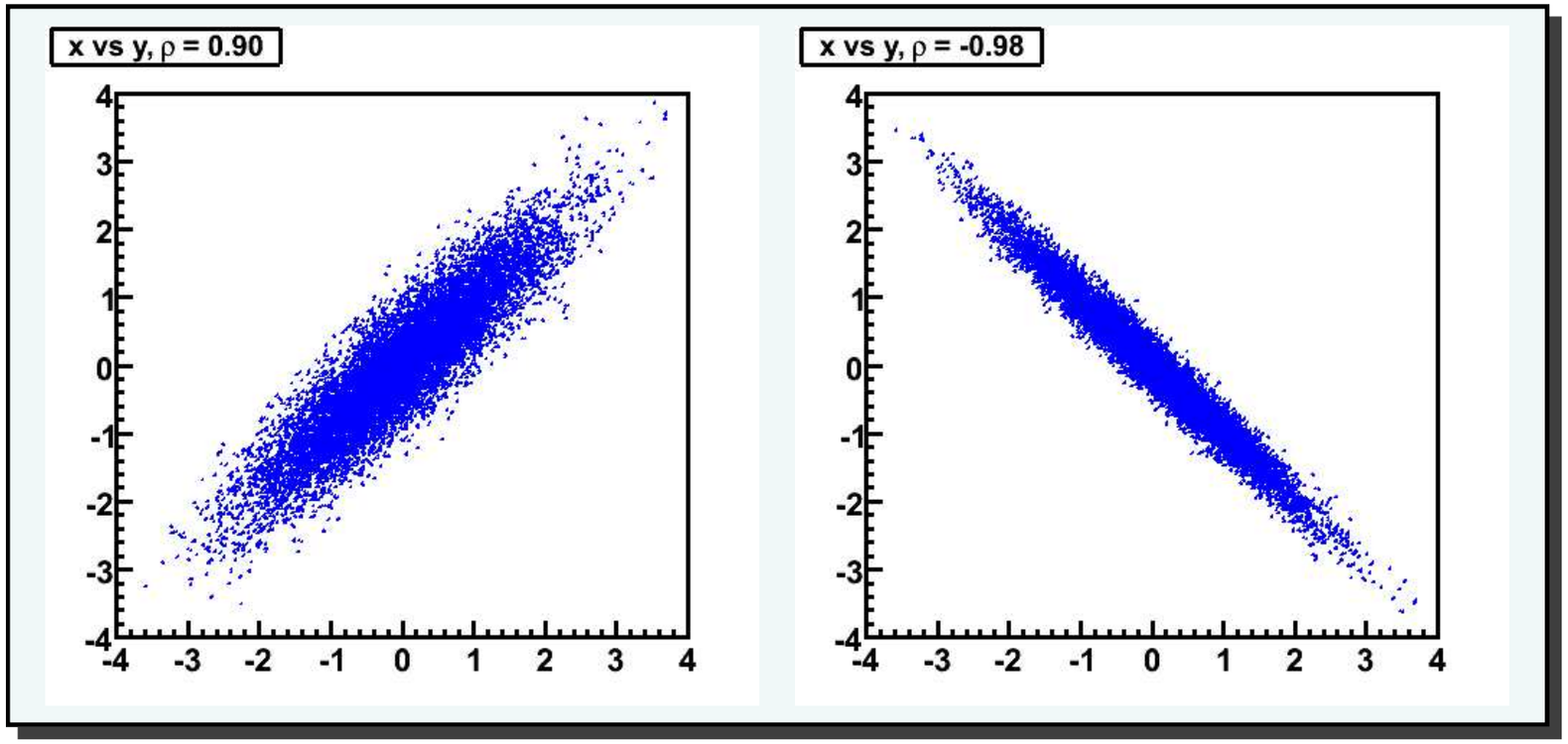
$$\rho \equiv \frac{cov(x,y)}{\sigma(x)\sigma(y)}$$

wobei immer gilt: $-1 \leq \rho \leq 1$.

$\rho = \pm 1$ heisst volle Korrelation, die Werte von y sind durch x bestimmt (oder v.v.) und enthalten keine zusätzliche Information.

Charakterisierung von Daten





Charakterisierung von Daten

- Beschreibung mehrerer Variablen
  - Bei mehr als zwei Variablen Erweiterung auf **Kovarianzmatrix**

$$V_{ij} = cov(x_{(i)}, y_{(i)})$$

bzw **Korrelationsmatrix**

$$\rho_{ij} = \frac{cov(x_{(i)}, y_{(i)})}{\sigma_i \sigma_j}$$

Kovarianzmatrix bzw. Korrelatioxsmatrix sind symmetrische N×N-Matrizen.

Wichtige Verteilungen

- Mittelwert und Varianz für Verteilungen als Integral über die Wahrscheinlichkeitsdichte (pdf - probability density function) anstatt über Summe der Einzel-Messungen:

Mittelwert: $\bar{x} = \int_{-\infty}^{+\infty} x \cdot p(x)\, dx,$

Varianz: $V(x) = \int_{-\infty}^{+\infty} (x - \bar{x})^2 \cdot p(x)\, dx,$

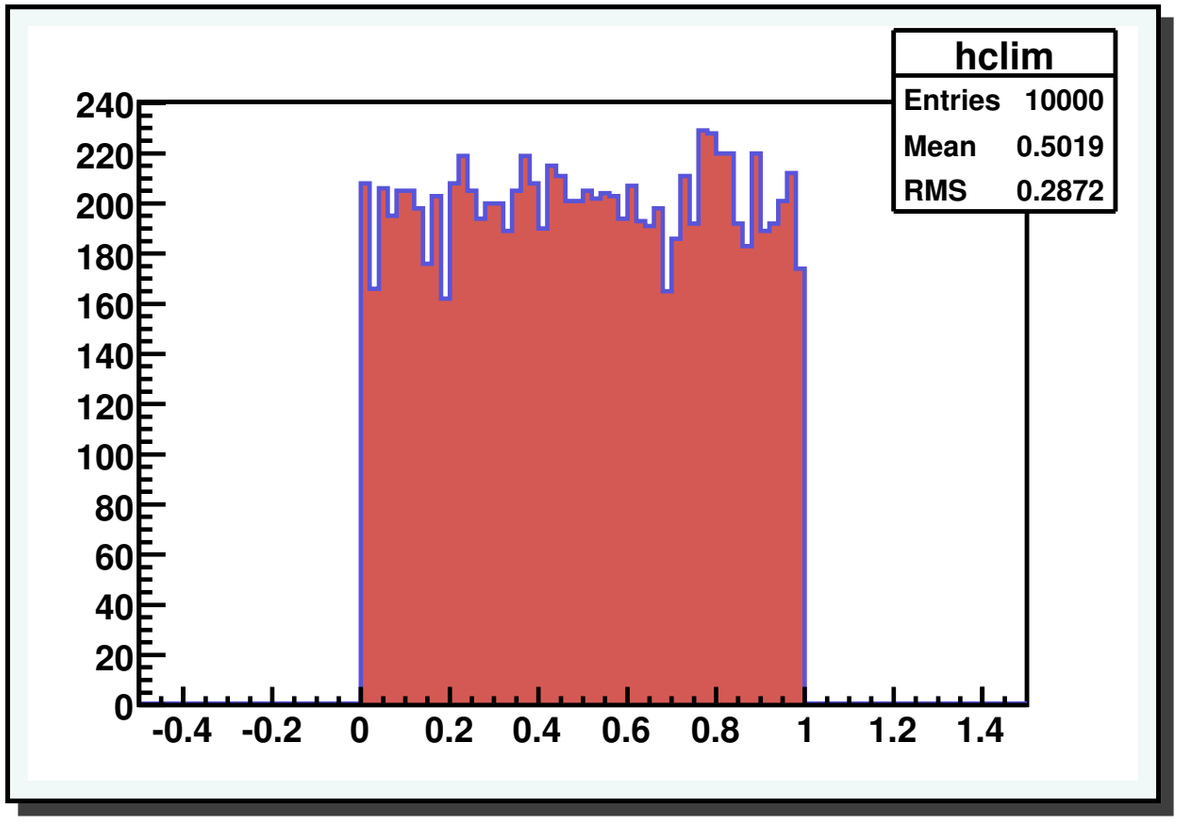
Standardabweichung: $\sigma(x) = \sqrt{V(x)}$

Wichtige Verteilungen

- **Gleichverteilung**
  - Die einfachste Verteilung, alle Werte in einem gewissen Intervall sind gleich wahrscheinlich:

$$p(x) = \frac{1}{b-a} \; \forall\, x \in [a,b],\; 0\; sonst$$

  - Wichtig für Glücksspiele aller Art, Grundlage für Zufallszahlgeneratoren, Simulationen und leicht zum Üben: Für eine Gleichverteilung in `[0,1]` ist der Mittelwert `0.5` und die Varianz σ^2=`1/12`



Wichtige Verteilungen

- **Binomialverteilung**
  - Die Binomialverteilung beschreibt Experimente bei denen jedes Einzelexperiment nur zwei mögliche Ergebnisse hat.
  - Einfachstes Beispiel ist der Wurf einer Münze. Gesucht ist z.B. die Wahrscheinlichkeit bei n Würfen k mal Kopf zu bekommen.
  - Eine spezielle Möglichkeit diese Resultat zu erzielen ist in den ersten k Würfen jeweils Kopf zu erhalten (p^k) und in den folgenden $n-k$ jeweils Zahl $(1-p)^{n-k}$. D.h. die Wahrscheinlichkeit ist $p^k(1-p)^{n-k}$. Aus der Kombinatorik kann man folgern, daß es

$$\boxed{\binom{n}{k} = \frac{n!}{(n-k)!k!}}$$

solcher Möglichkeiten gibt, jede mit gleicher Wahrscheinlichkeit.

Wichtige Verteilungen

- **Binomialverteilung**
  - Also ist die Wahrscheinlichkeit insgesamt

$$P(k \times \text{Kopf}) = p^k (1-p)^{n-k} \binom{n}{k}$$

  - Ganz analog kann man für den radioaktiven Zerfall vorgehen: Gesucht ist die Wahrscheinlichkeit k Zerfälle in einer Zeit T zu beobachten, bei N Kernen mit Zerfallskonstante λ.

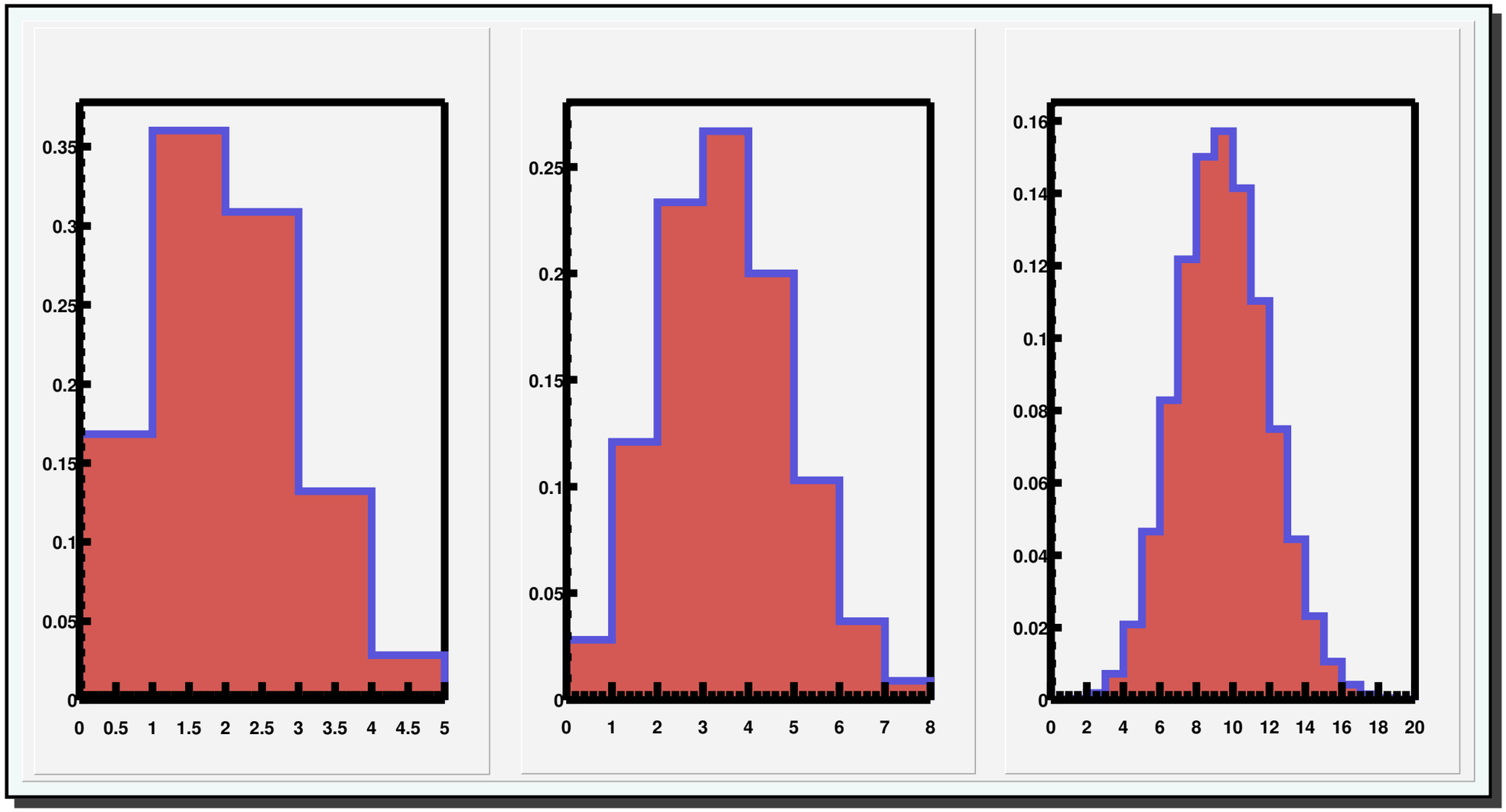
    Dazu unterteilt man die Zeit T in n kleine Intervalle Δt . Die Wahrscheinlichkeit einen Zerfall in Δt zu beobachten ist $p=\lambda N \Delta t$, wobei Δt so klein sein soll, dass $\lambda N \Delta t \ll 1$ ist.
    Wie beim Münzenwurf folgt dann die Binomialverteilung

$$P(k) = p^k (1-p)^{n-k} \binom{n}{k}$$

Wichtige Verteilungen

- **Binomialverteilung**
  - Mittelwert und Varianz der Binomialverteilung sind:

$$\bar{x} = np, \quad \sigma^2 = np(1-p)$$



Wichtige Verteilungen

- **Poissonverteilung**
  - Die Poissonverteilung ist der Grenzfall der Binomialverteilung für $n \to \infty$, $p \to 0$, $np = const$.
  - Am Beispiel des radioaktiven Zerfalls gut zu veranschaulichen:
    - Die Intervalle Δt werden immer kleiner, also $n \to \infty$, $p = \lambda N \Delta t \to 0$, $np = \lambda N T$.

$$P(k) = \left(\frac{\lambda N T}{n}\right)^k \left(1 - \frac{\lambda N T}{n}\right)^{n-k} \frac{n!}{(n-k)!k!}$$

    - Mit $\Delta t \to 0$ bzw. $n \to \infty$ folgt

$$\left(1 - \frac{\lambda N T}{n}\right)^n \to e^{-\lambda N T}\,, \quad \left(1 - \frac{\lambda N T}{n}\right)^{-k} \to 1\,, \quad \frac{n!}{(n-k)!} \to n^k$$

Wichtige Verteilungen

- **Poissonverteilung**
  - d.h. insgesamt

$$P(k) = \frac{\mu^k e^{-\mu}}{k!}$$

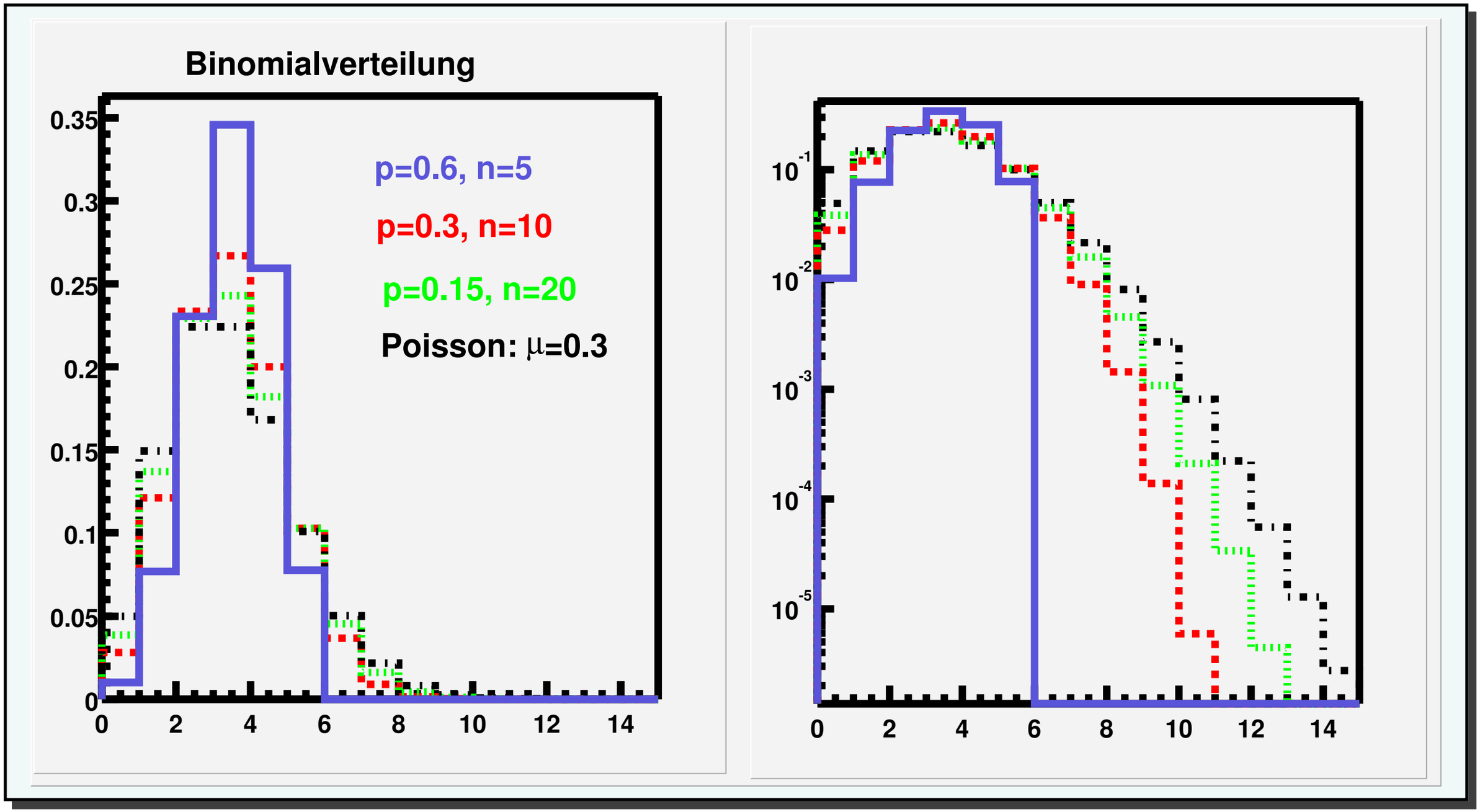
mit $\mu = \lambda N T$.

  - Für Mittelwert und Varianz der Poissonverteilung erhält man:

$$\bar{x} = \mu, \quad \sigma^2 = \mu$$

In der Praxis wird die Binomialverteilung schon für “kleine” $n \approx 10-20$ durch eine entsprechende Poissonverteilung gut beschrieben.

Wichtige Verteilungen



Wichtige Verteilungen

Poissonstatistik bei den alten Preussen

Ein Klassiker in alten Statistikbüchern ist die Statistik der preussischen Armee zu tödlichen Unfällen durch Huftritte pro Armee-Corps und Jahr.

Über 20 Jahre und für 10 Corps wurden 122 Todesfälle gezählt (in 200 Corps-Jahren). Das ergibt $\mu = 122/200 = 0.61$.

N-Todesfälle	0	1	2	3	4	5	6
Beobachtete Corps-Jahre	109	65	22	3	1	0	0
Erwartete Corps-Jahre	108.7	66.3	20.2	4.1	0.6	0.07	0.01

Perfekte (fast zu gute) Übereinstimmung mit Poisson-Vorhersage.

Wichtige Verteilungen

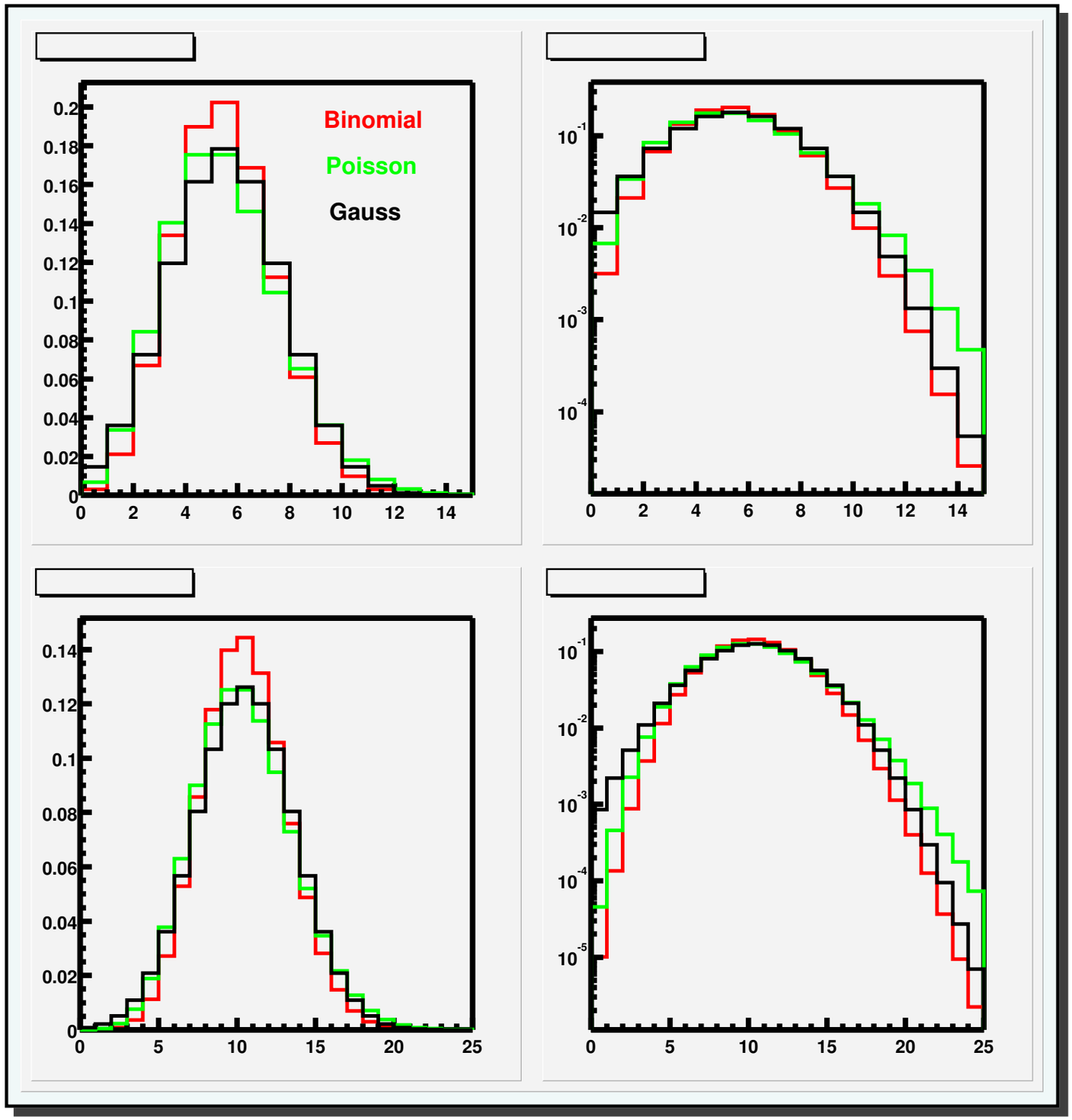
- **Gauss- oder Normalverteilung**
  - Das ist die wichtigste Verteilung in der Statistik

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Der Mittelwert ist μ und die Varianz σ^2 .

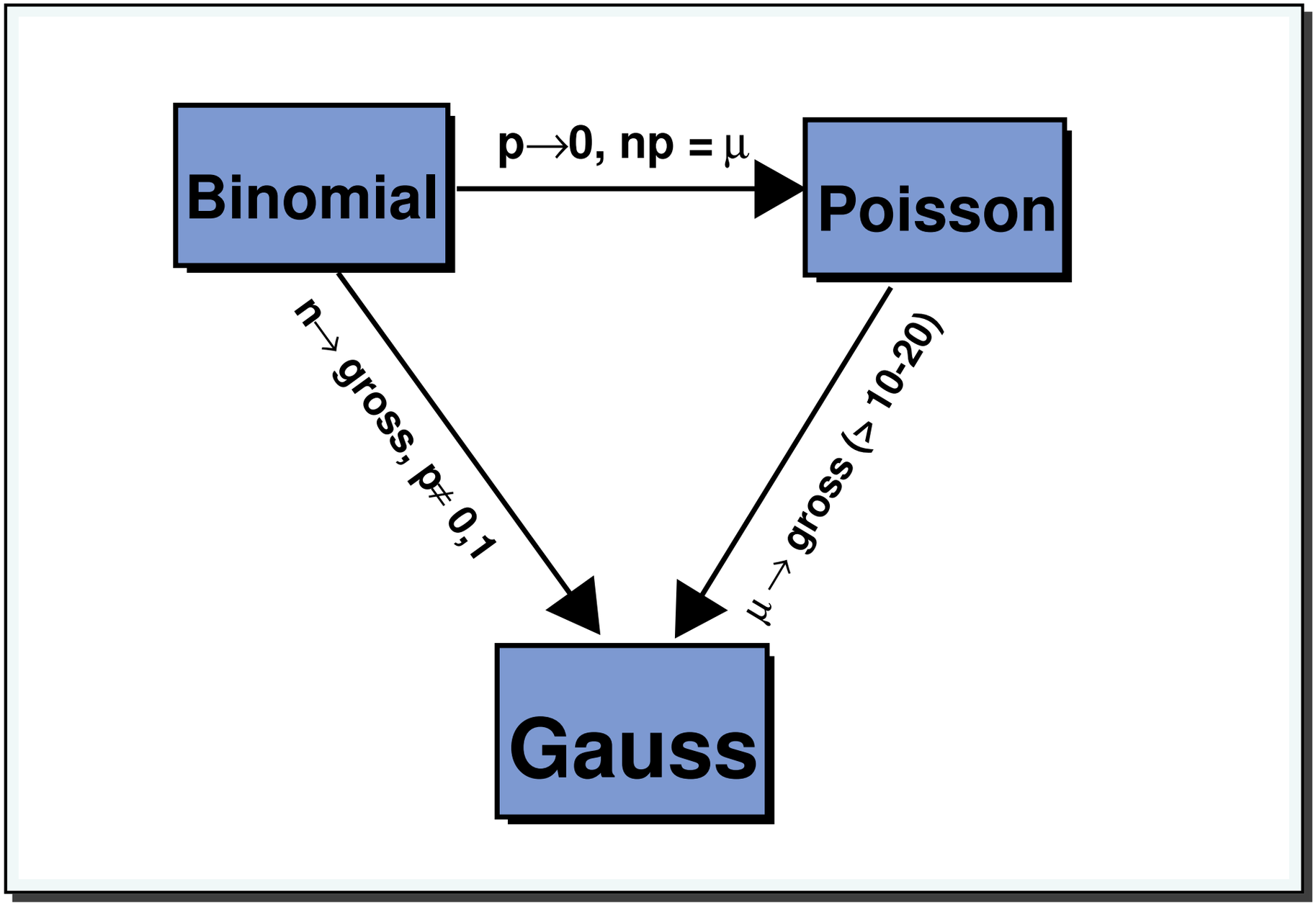
Die Poissonverteilung geht für grosse μ in die Gaussverteilung über, wobei schon für $\mu \approx 10$ die Gaussverteilung eine brauchbare Näherung ist.
Analog geht die Binomialverteilung in die Gaussverteilung über für grosse n und np.

Wichtige Verteilungen



Wichtige Verteilungen

- **Standardverteilungen**



Wichtige Verteilungen

- **Zentraler Grenzwert-Satz**
  - Das wichtigste Theorem in der Statistik, es besagt:

    Für eine Menge von unabhängigen Zufallsvariablen x_i mit Mittelwert μ und Varianz σ^2 nähert sich die Größe

$$\boxed{y = \frac{\sum x_i}{n}}$$

    für große n einer Gaussverteilung mit Mittelwert μ und Varianz σ^2/n an.

    Dabei spielt die zugrundeliegende Verteilung der x_i keine Rolle; auch wenn sie z.B. aus der Gleichverteilung oder der Exponentialverteilung stammen, ist ihr Mittelwert y normalverteilt.