Statistik

– Vorlesung 2 –

Günter Duckeck

J. Elmsheuser

Sommersemester 2008

Inhalt:

- Charakterisierung von Daten
- Wichtige Verteilungen

Überblick

- Einführung, Beispiele, Wahrscheinlichkeiten, Bayes Theorem
- **Beschreibung von Daten und Verteilungen**
- Monte Carlo Methoden
- Fehler
- Parameterschätzung
- Likelihood- und χ^2-Methode
- Hypothesentest
- Wahrscheinlichkeiten und Vertrauensintervalle
- Klassifizierung und Neuronale Netze

- Optimierung & Parametrisierung
- Datamining

1 Charakterisierung der Daten

- “Daten–Arten”
- Mittelwert
- Breite
- Korrelation

1.1 “Daten–Arten”

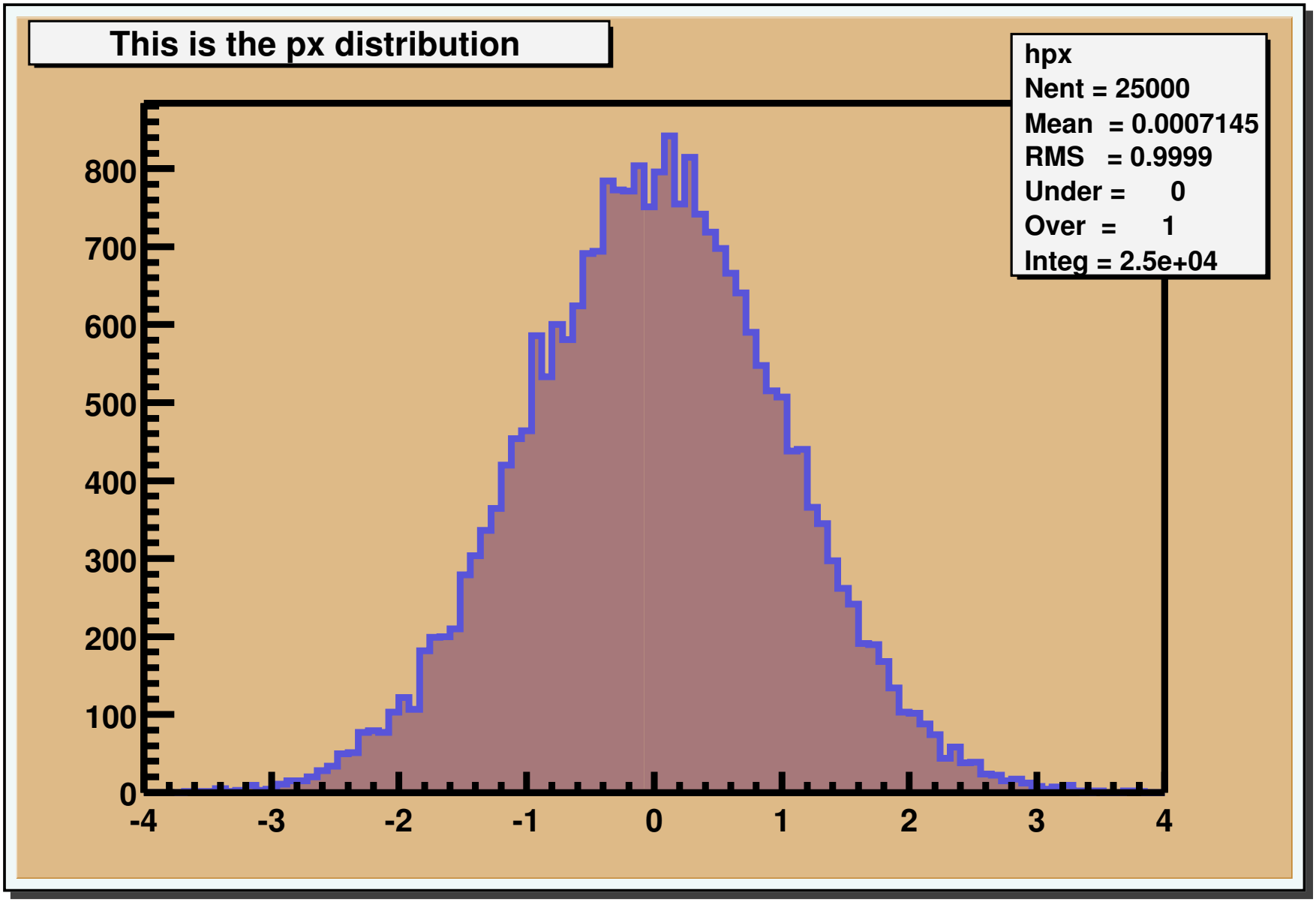
Unterscheidung nach

- **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
  - **diskrete** Werte, z.B. Anzahl (Teilchen im Ereignis, Personen im Fahrzeug, Münzen in der Börse)
  - **kontinuierliche** Werte, z.B. Masse des Teilchens, Größe einer Person, Intensität einer Quelle.

Datenvisualisierung

Häufigkeitsverteilung: Eine oder mehrere Grössen werden wiederholt gemessen. Darstellung in ein- oder mehr-dimensionalen ‘Histogrammen’:

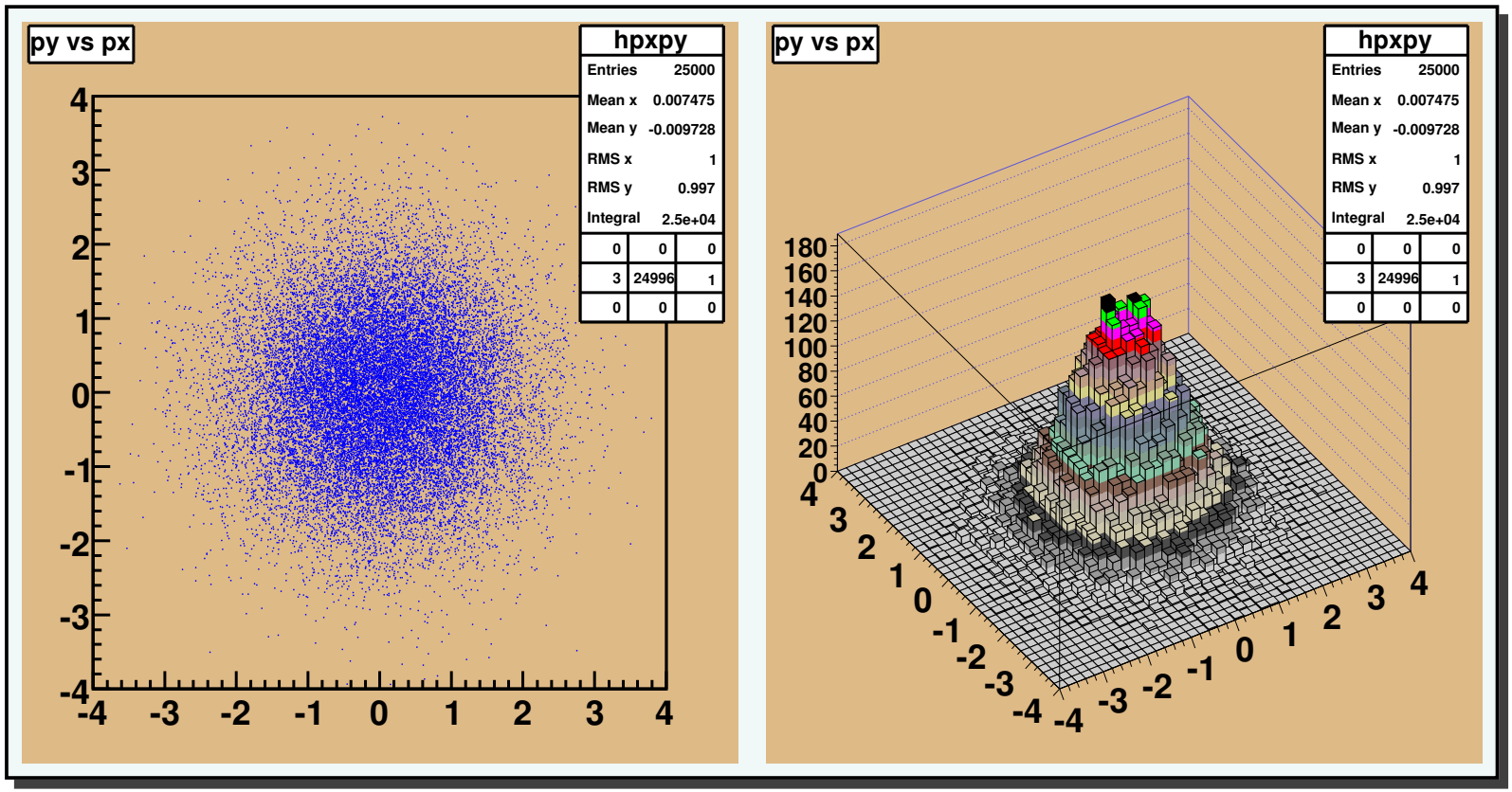
- Bereich `xlow` bis `xhigh` unterteilt in `nchannel` Intervalle.
- `Root: TH1F * h1 = new TH1F("h1","mytitle",nchannel, xlow, xhig)`
- Jede Messung `x` wird in das Histogram gefüllt:
  `Root: h1->Fill(x)`
  $\Rightarrow$ Kanal-Nr: $nchannel\,\frac{x - x_{low}}{x_{hig} - x_{low}}$ Inhalt wird um **1** erhöht.
- Darstellung $\Rightarrow$ Einträge pro Intervall.


This is the px distribution
hpx
Nent = 25000
Mean = 0.0007145
RMS = 0.9999
Under = 0
Over = 1
Integ = 2.5e+04
800
700
600
500
400
300
200
100
0
-4
-3
-2
-1
0
1
2
3
4

Analog Erweiterung auf **2 Dimensionen**:

- Bereich `xlow` bis `xhigh` unterteilt in `nx` Intervalle.
  Bereich `ylow` bis `yhigh` unterteilt in `ny` Intervalle.
  $\Rightarrow$ `nx*ny` Segmente
- `Root: TH2F * h2 = new TH2F("h2","mytitle",nx, xlow, xhig, ny, ylow, yhigh)`
- Jedes Paar `(x, y)` wird in das Histogram gefüllt:
  `Root: h2->Fill(x, y)`
  $\Rightarrow$ Kanal-Nr: $(nx\frac{x - x_{low}}{x_{hig} - x_{low}}, ny\frac{y - y_{low}}{y_{hig} - y_{low}})$; Inhalt wird um **1** erhöht.

2D-Darstellung als Scatter-Plot, Legoplot, Kontouren, uvm.



1.2 Mittelwert

Gebräuchlichste Art Daten $x_1, x_2, ..., x_N$ 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:

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

Harmonisches Mittel:

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

Median

“Wert in der Mitte”: Die Hälfte der Werte ist kleiner, die andere Hälfte größer.

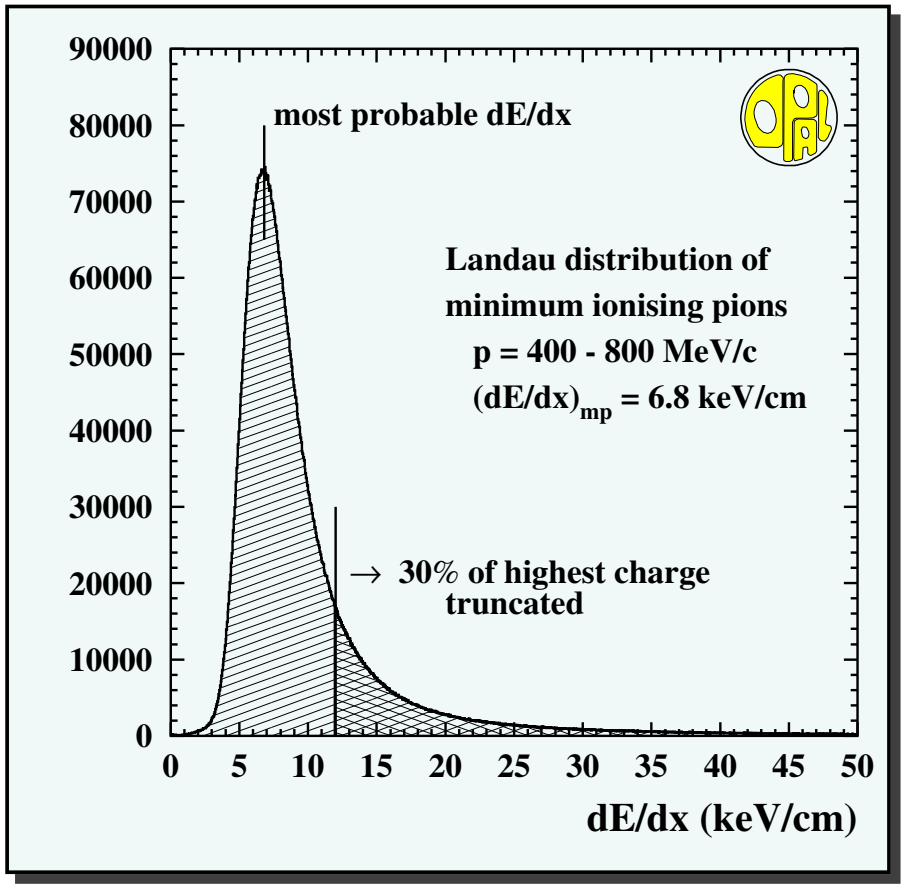
Sehr nützlich für Datensätze, bei denen die zugrundeliegende Verteilung unklar ist oder lange Schwänze hat oder die Reihenfolge wichtiger ist als die numerische Grösse.

Am einfachsten zu bekommen durch Sortieren des Datensatzes, dann

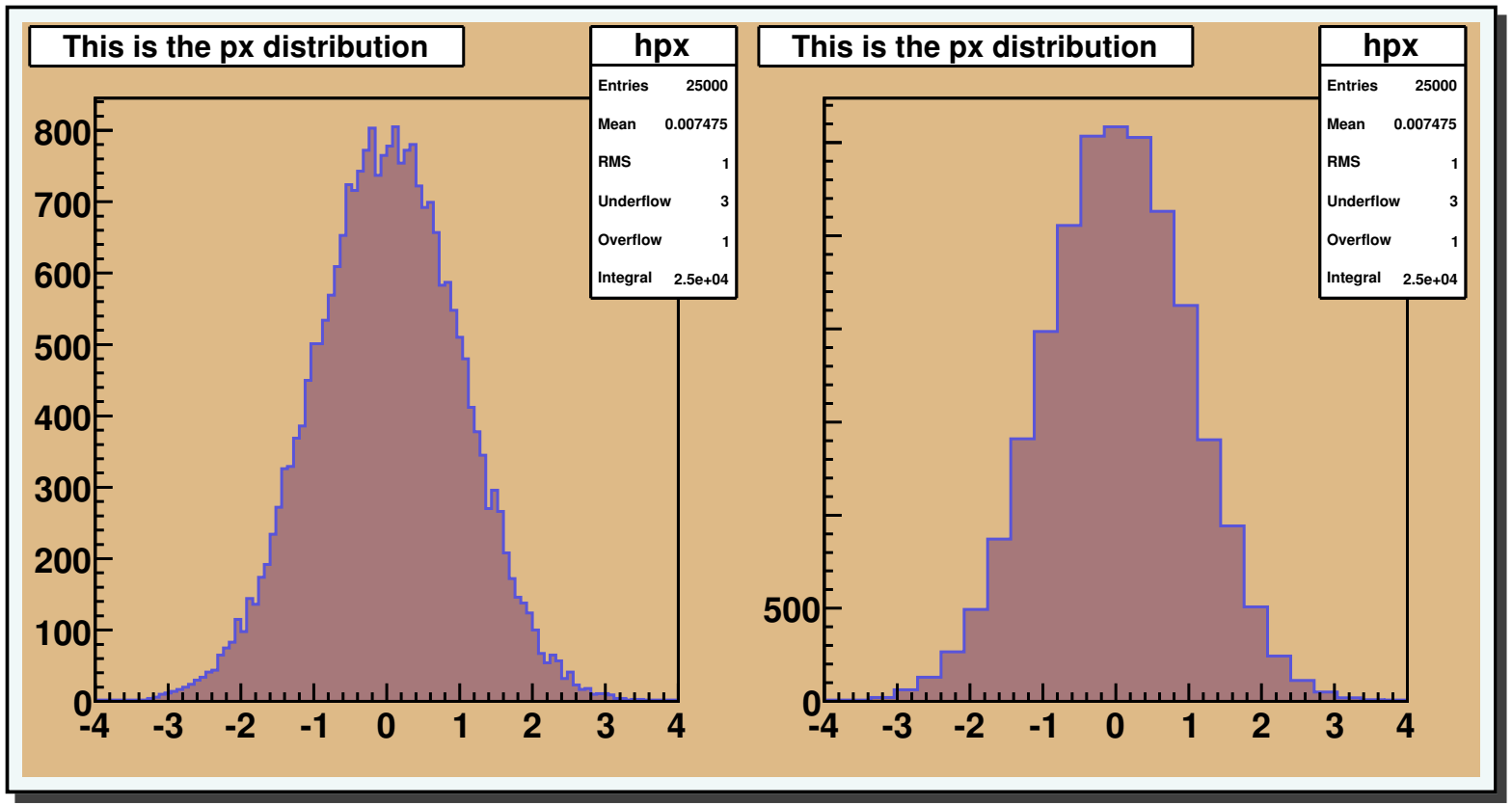
$$Median = x_{N/2}$$

truncated mean ist weitere Variante: Dabei wird der Mittelwert nur aus einem eingeschränkten Bereich bestimmt, wegen langen Schwänzen, asymmetrischen Beiträgen, etc.

Ein Beispiel aus Teilchenphysik ist die Messung des spezifischen Energieverlusts, typisch werden die höchsten 30% verworfen.



Bei “binned” Datensätzen auch noch *wahrscheinlichster Wert*, d.h. x_{max} von Bin mit höchstem Eintrag. Eher ungebräuchlich da sehr schwankend und abhängig von Ereigniszahl und Anzahl Bins.



1.3 “Breite” eines Datensatzes

Nach Mittelwert die 2. wichtige Grösse zur Charakterisieung 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**

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

Weitere Möglichkeiten zur Charakterisierung der Breite:

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

eher ungebräuchlich, mathematisch sehr unangenehme Eigenschaften

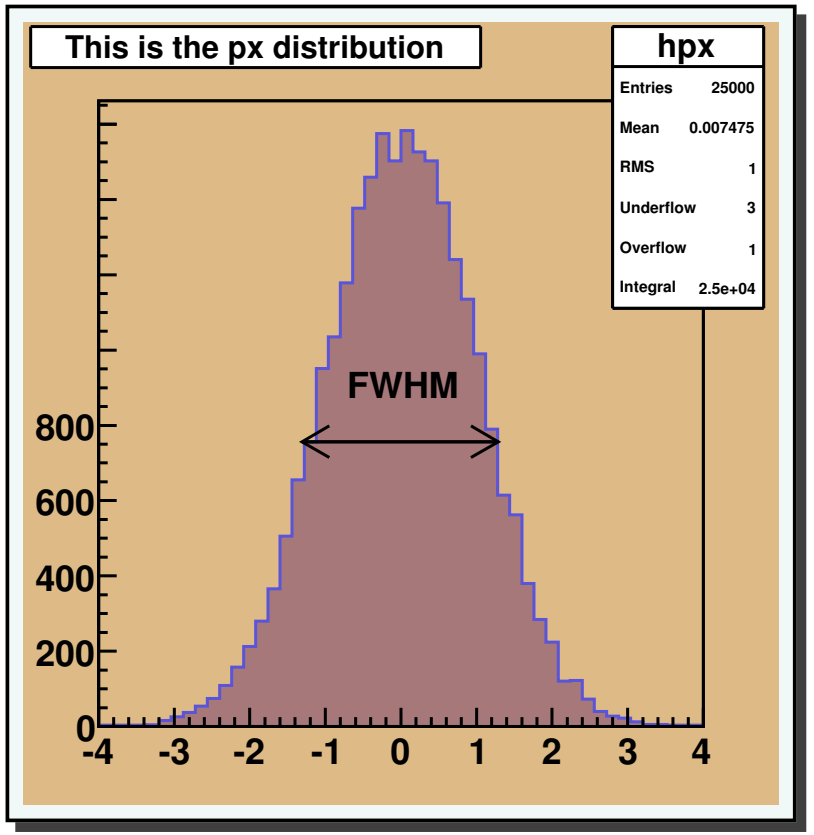
Oft nützlich die sogenannten Quantile:

- unteres Quartil: 25 % der Werte sind kleiner.
- oberes Quartil: 75 % der Werte sind kleiner.
- Breite als Differenz zwischen oberem und unterem Quartil.
- oder beliebige Percentile: XX % Percentil = XX % der Werte sind kleiner

Bei “binned” 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 Schwänze sind unkritisch.



Höhere Potenzen

Naheliegend nach Mittelwert und Varianz auch höhere Potenzen zu betrachten:

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

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

Skew ist nützlich zur Charakterisierung der Asymmetrie, positiv bei Schwänzen nach rechts und v.v., verschwindet bei symmetrischen Verteilungen.

1.4 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:

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

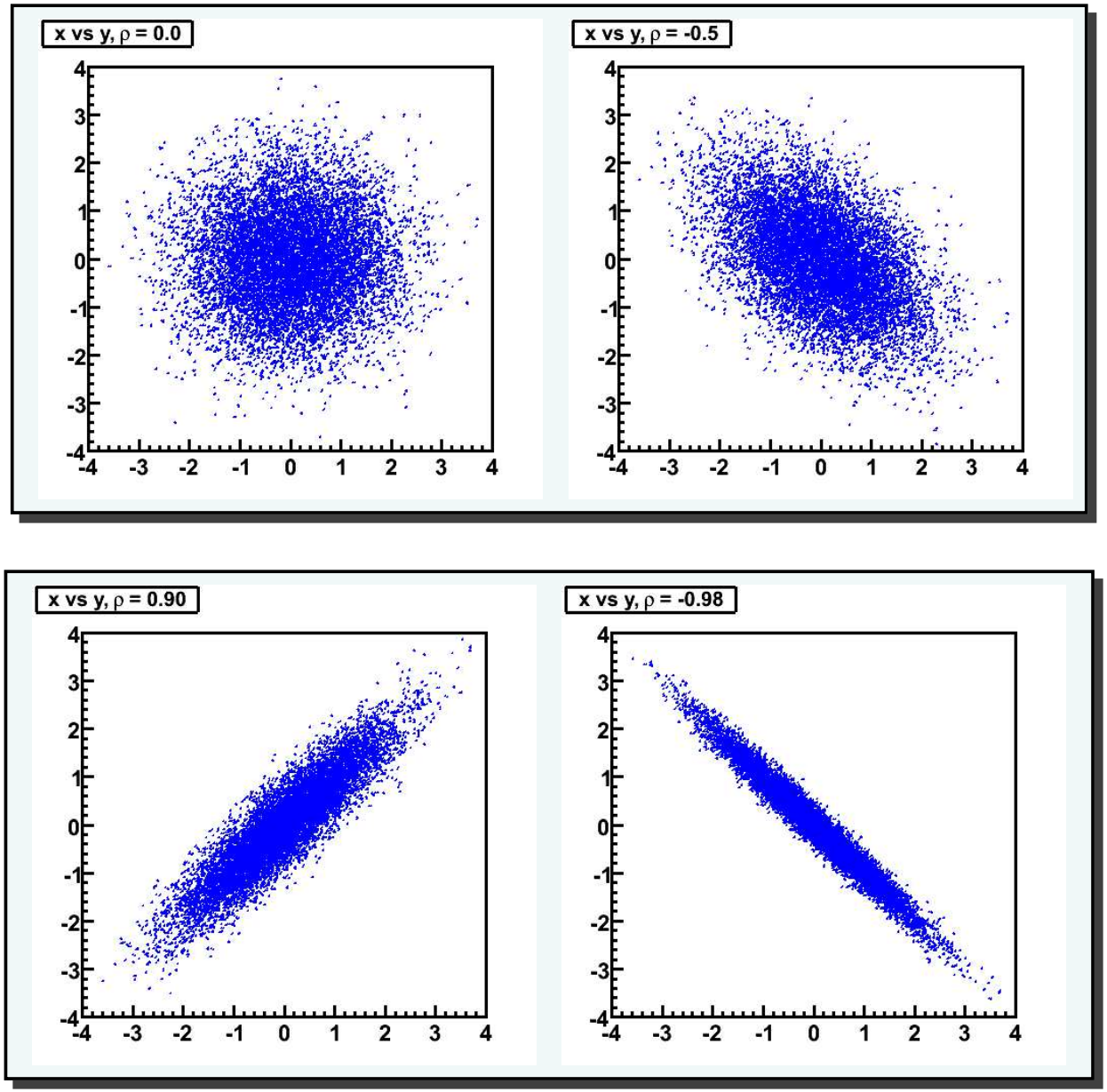
Wenn hohe (tiefe) Werte von x oft mit hohen (tiefen) Werten von y vorkommen ist die Kovarianz positiv, bei entgegengesetztem Verhalten negativ bei unabhängigen Grössen verschwindet sie.

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.


x vs y, ρ = 0.0
x vs y, ρ = -0.5
x vs y, ρ = 0.90
x vs y, ρ = -0.98

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 Korrelationsmatrix sind symmetrische $N \times N$ Matrizen.

1.5 Zusammenfassung

Charakterisierung von Daten durch

- Mittelwert $\bar{x}$ und
- Varianz $V(x)$ bzw Standardabweichung σ sind Standard
- abhängig von Verteilung der Daten gegebenenfalls auch alternative Grössen wie *Median, truncated mean, Quantile, FWHM* sinnvoll

Bei mehr als einer Variable ausserdem noch *Kovarianz* bzw. *Korrelation* oder allgemein $N \times N$ *Kovarianzmatrix* bei N Messgrössen.

2 Wichtige Verteilungen

Mittelwert und Varianz für Verteilungen als Integral über die Wahrscheinlichkeitsdichte (**pdf**) 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: $\sigma(x) = \sqrt{V(x)}$

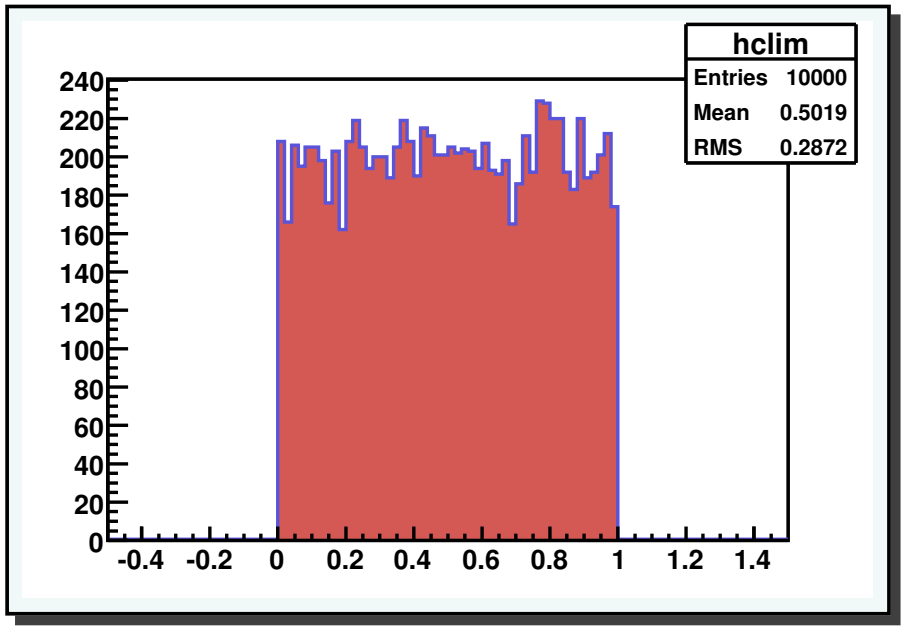
2.1 Gleichverteilung

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 Zufallszahl-Generatoren, Simulationen und leicht zum Üben:

Für eine Gleichverteilung in $[0,1]$ ist der Mittelwert 0.5 und die Varianz $\sigma^2 = 1/12$



2.2 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

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

solcher Möglichkeiten gibt, jede mit gleicher Wahrscheinlichkeit.

Also ist die Wahrscheinlichkeit insgesamt

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

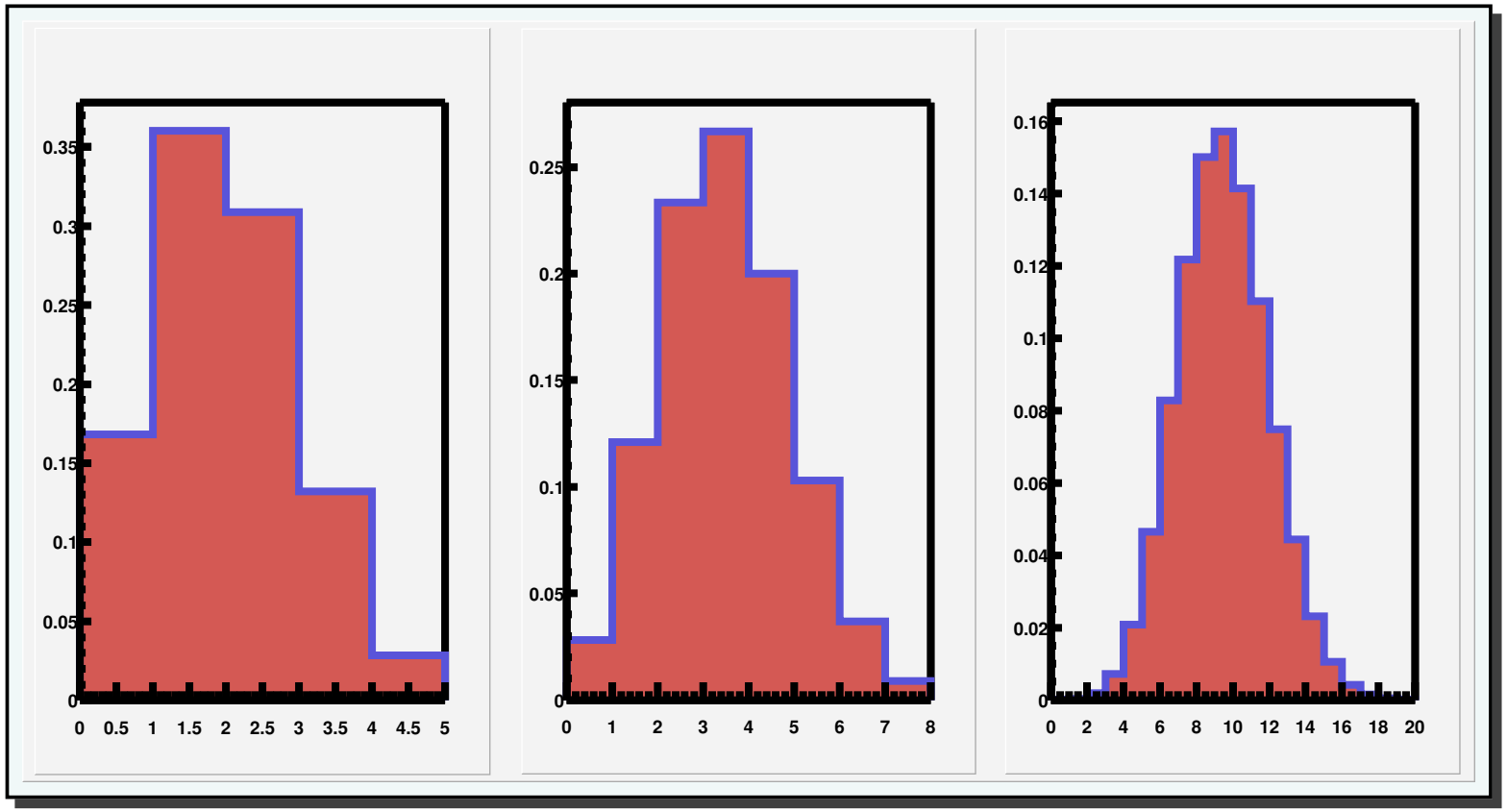
Ganz analog kann man für den radioaktiven Zerfall vorgehen: Gesucht ist die Wahrscheinlichkeit k Zerfälle in einer Zeit T zu beobachten, bei N Kernen mit Zerfallskonstante λ .

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

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

Mittelwert und Varianz der Binomialverteilung sind:

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



2.3 Poissonverteilung

Die Poissonverteilung ist der Grenzfall der Binomialverteilung für $n \to \infty,\ p \to 0,\ np = const$.

Am Beispiel des radioaktiven Zerfalls lässt sich das gut veranschaulichen:

- Die Intervalle Δt werden immer kleiner, also
  $n \to \infty,\ p = \lambda\, N\, \Delta t \to 0,\ np = \lambda\, N\, T$.

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

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

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

- d.h. insgesamt

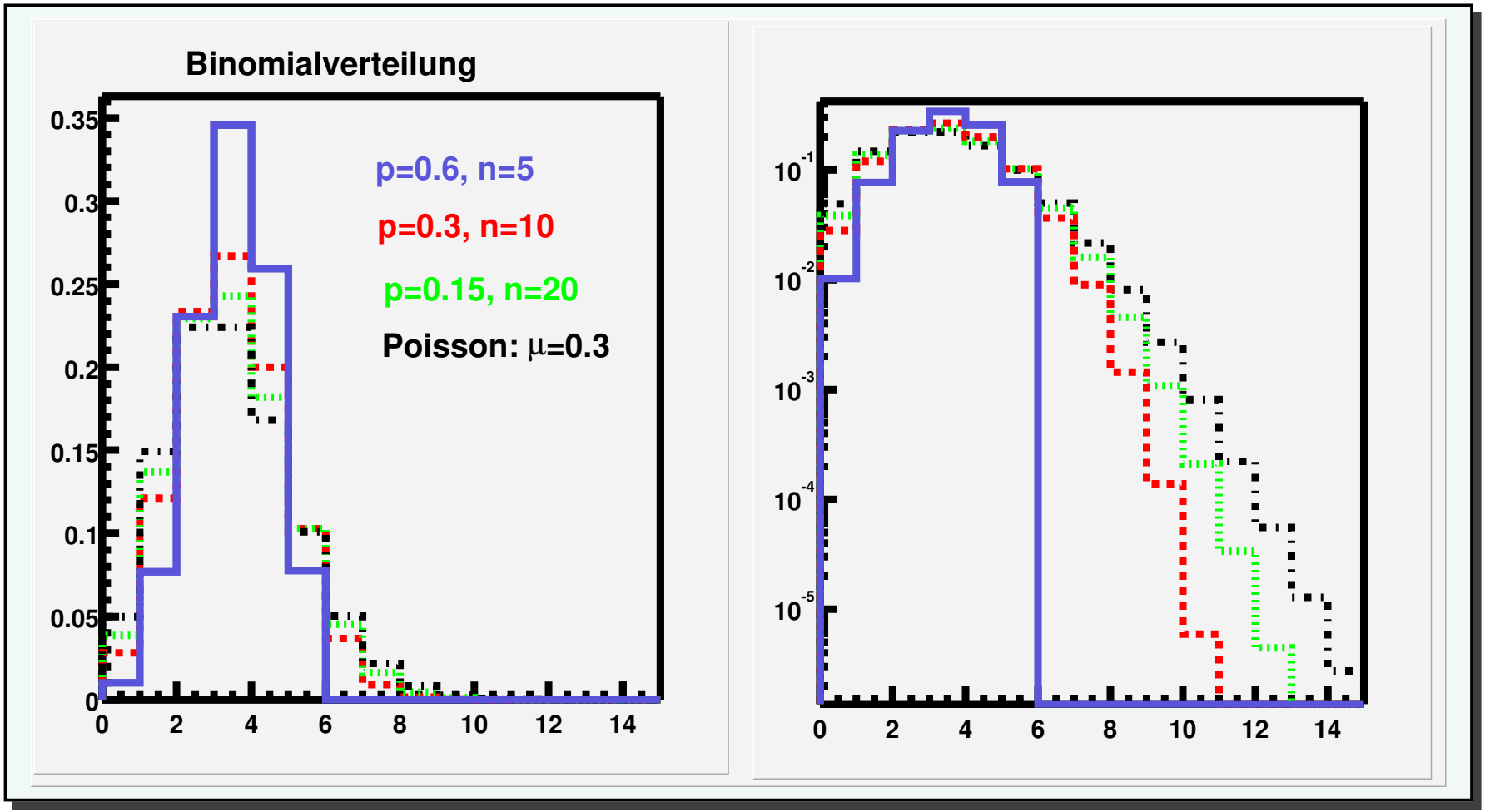
$$P(k) = \frac{\mu^k e^{-\mu}}{k!}$$

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.


Binomialverteilung
p=0.6, n=5
p=0.3, n=10
p=0.15, n=20
Poisson: μ=0.3
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 2 4 6 8 10 12 14
10^-1
10^-2
10^-3
10^-4
10^-5
0 2 4 6 8 10 12 14

Poisson–Statistik 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.

2.4 Gauss– oder Normalverteilung

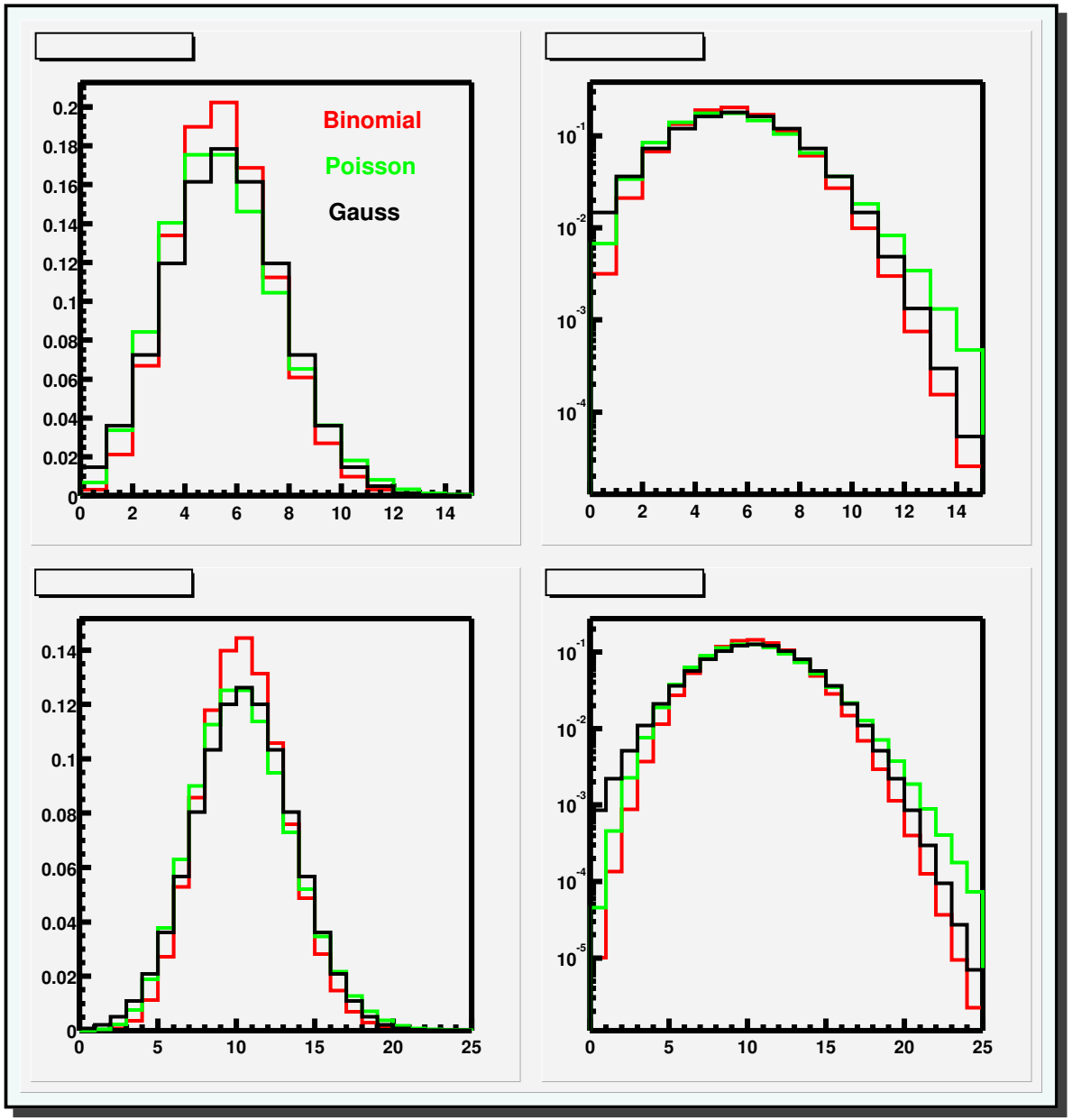
Das ist die wichtigste Verteilung in der Statistik

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

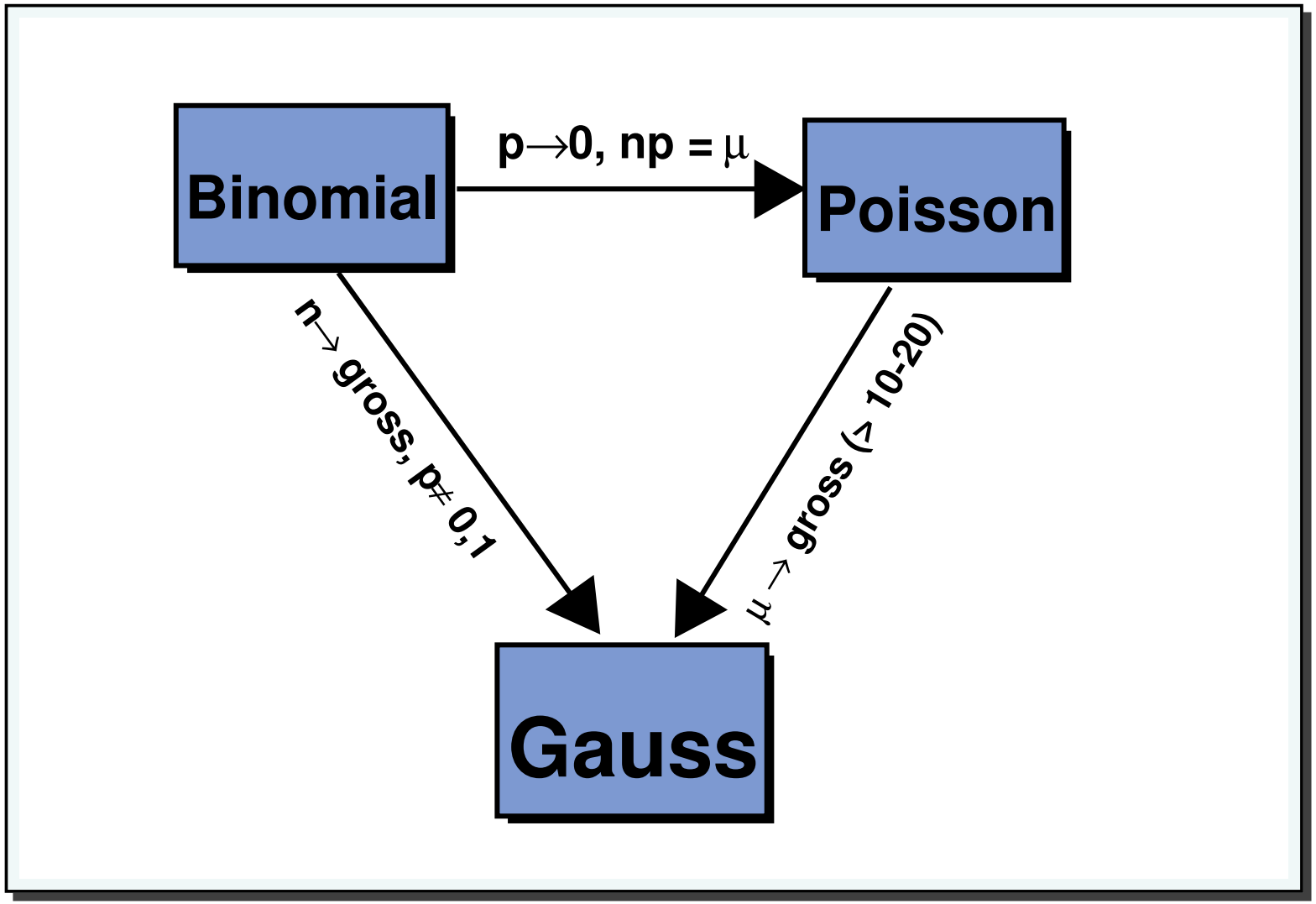
Der Mittelwert ist μ und die Varianz σ^2 .

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 .


Binomial
Poisson
Gauss

Standardverteilungen



2.5 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

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

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.

$\Rightarrow$ Demo und Übungen

2.6 Zusammenfassung

Binomial-, Poisson- und Gauss-Verteilung sind Standard–verteilungen, insbesondere zur Beschreibung von **Zählexperimenten**.

Die *Gauss-Verteilung* hat überragende Bedeutung wegen dem
Zentralen Grenzwertsatz:
Eine Summe von Zufallsgrössen nähert sich der Gauss-Verteilung an, unabhängig von der Verteilung der ursprünglichen Zufallsgrössen.

Einige weitere Verteilungen in der nächsten Vorlesung.

Ausführliche Liste in **“Hand-book on Statistical Distributions”**
(`http://www.physto.se/~walck/`)