Statistische Methoden der Datenanalyse

Wintersemester 2012/2013

Albert-Ludwigs-Universität Freiburg

Prof. Markus Schumacher, Dr. Stan Lai

Physikalisches Institut Westbau 2 OG

E-Mail: Markus.Schumacher@physik.uni-freiburg.de

stan.lai@cern.ch

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

Anknüpfung an letzte Vorlesung

Axiomatische Definition:

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$

bedingte Wahrscheinlichkeit:
A gegeben *B* (mit *P(B) ≠ 0)*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes-Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Kapitel 3

Wahrscheinlichkeitsdichtefunktionen

Zufallsvariablen und Wkt.dichtefunktion

Eine Zufallsvariable (ZV) ist eine numerische Charakeristik, die einem Element der Grundgesamtheit zugeordnet wird.

ZV können diskret oder kontinuierlich sein.

Annahme: das Ergebnis eines Experimentes ist kontinuierliche ZV *x*
Wkt. Dafür, genau den Wert x_i zu messen ist beliebig klein.

$$\text{Wahrscheinlichkeit } x \text{ im Intervall } [x, x+dx] \text{ zu erhalten } = f(x)dx.$$

$\rightarrow f(x) =$ Wahrscheinlichkeitsdichtefunktion (WDF)

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \quad x \text{ muss irgendwo sein}$$

Für diskrete Zufallsvariable mit Ergebnissen x_i mit z.B. $i = 1, 2, \ldots$

$$P(x_i) = p_i$$ Wahrscheinlichkeits-Massenfunktion

$$\sum_i P(x_i) = 1$$ *x* muss einen der möglichen Werte annehmen

Approximative WDF aus Histogramm

WDF = Histogramm mit

unendlich vielen Einträgen

(unendlich großes Sample),

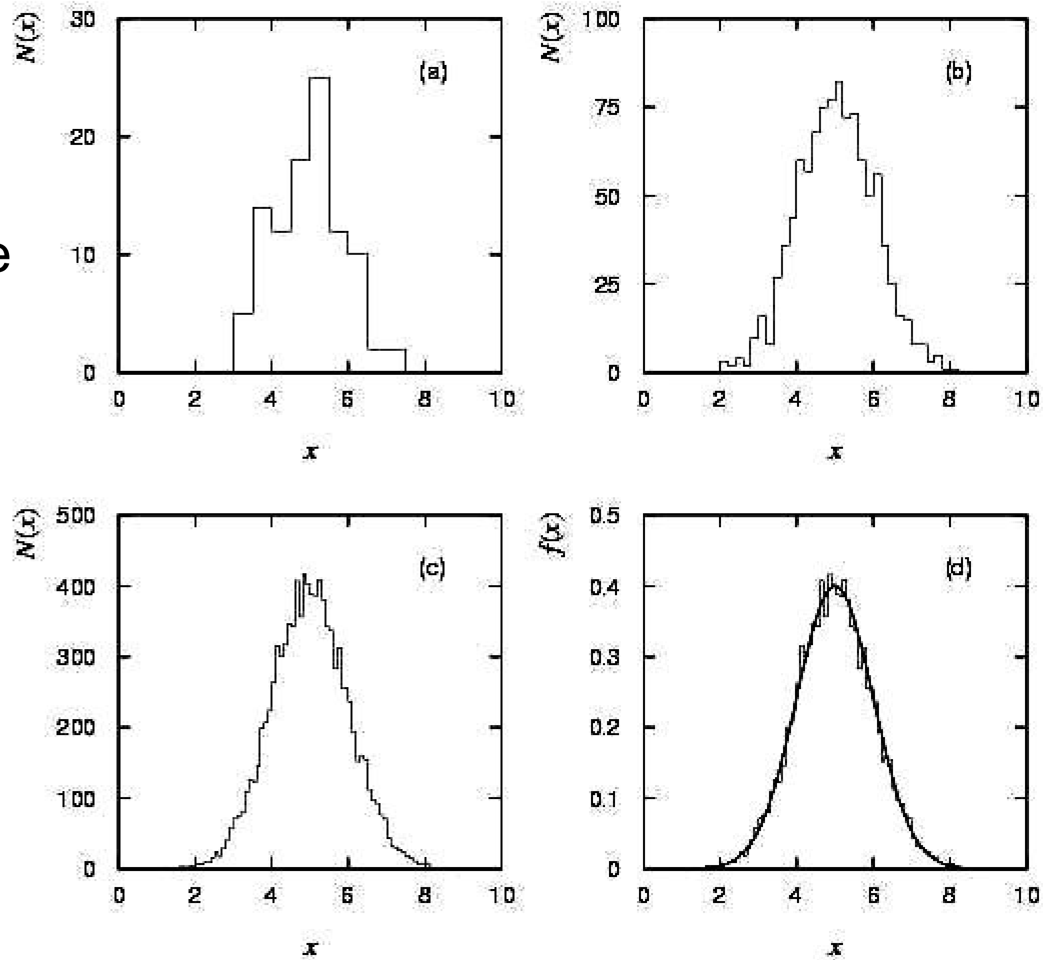
und verschwindender (0) Binbreite

normiert auf “eins”

$$f(x) = \frac{N(x)}{n\Delta x}$$

$n =$ number of entries

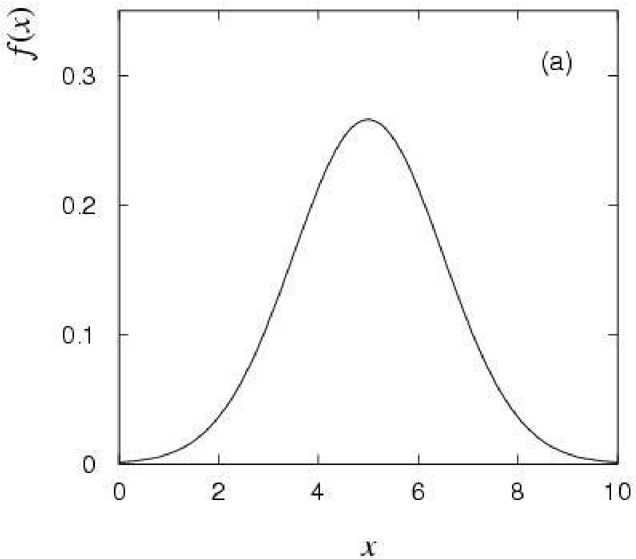
$\Delta x =$ bin width

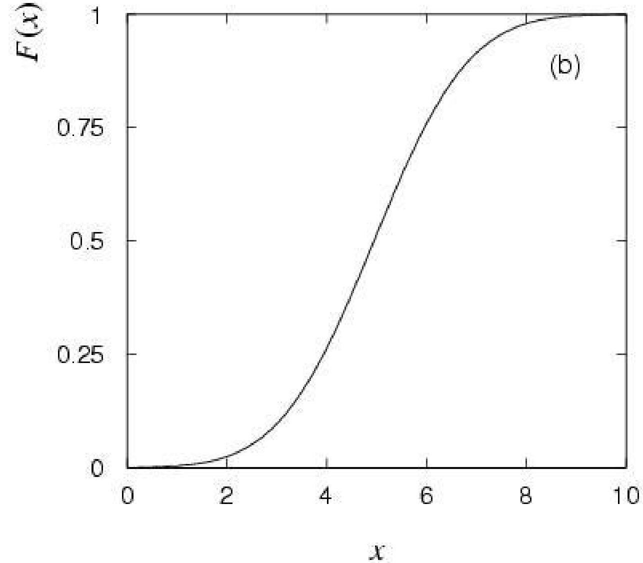


Kumulativerteilung

Wahrscheinlichkeit ein Ergebnis kleiner oder gleich *x* zu erhalten

$$\int_{-\infty}^{x} f(x')\,dx' \equiv F(x)$$ Kumulativerteilung





Alternativ kann WDF über $$f(x) = \frac{\partial F(x)}{\partial x}$$ definiert werden

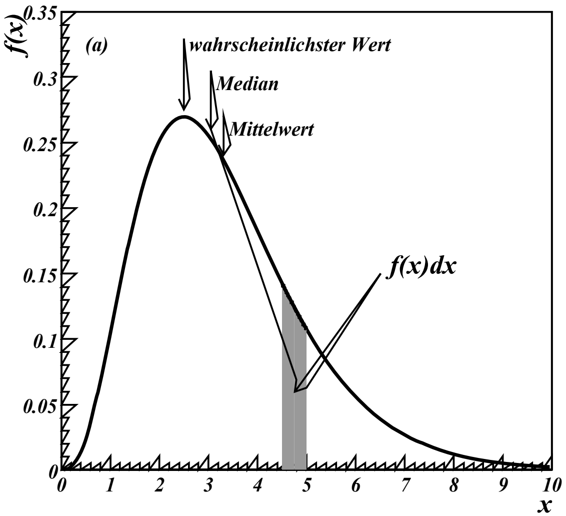
Für diskrete ZV gilt $$F(x) = \sum_{x_i \le x} P(x_i)$$

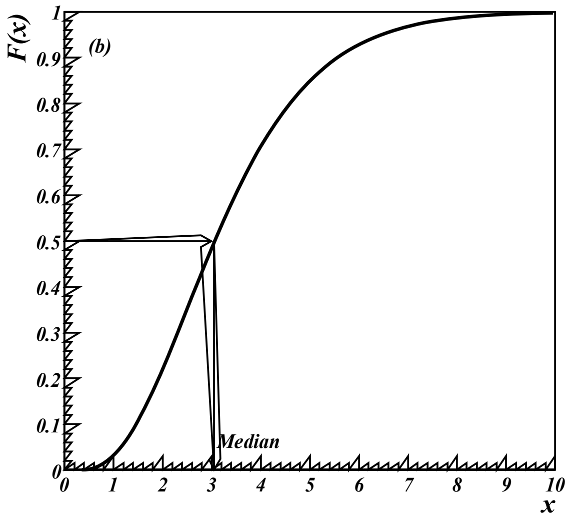
Quantile

Das Quantil x_α ist der Wert der ZV x für den die Wahrscheinlichkeit, dass die Messung einen Wert kleiner x_α ergibt gerade α ist.

$$F(x_\alpha) = \int_{-\infty}^{x_\alpha} f(x)dx = \alpha \Rightarrow x_\alpha = F^{-1}(\alpha).$$

Das Quantil zu α=0.5 ist der Median $x_{1/2}$





Erwartungswerte

Betrachte kontinuierliche Zufallsvariable *x* with WDF *f* (*x*).

Definiere den Erwartungswert (arithmetischen.Mittelwert) als

$$E[x] = \int x\, f(x)\, dx$$

Notation (meist): $E[x] = \mu$

Achtung:

E[ ] ist keine Zufallsvariable, keine Funktion der Stichprobe

ist ein linearer Operator auf der Funktion z.B. a(x)=x

$$E[\alpha x + \beta] = \alpha E[x] + \beta$$

ist eine Charakteristik der Wahrscheinlichkeitsdichtefunktion

ist abhängig von den Parametern der WDF

Erwartungswerte

Für ein Funktion $y(x)$ with WDF $g(y)$, gilt

$$E[y] = \int y\, g(y)\, dy = \int y(x) f(x)\, dx$$ (äquivalent)

n-tes algebraisches Moment (n=1 ergibt Mittelwert):

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x) dx = \mu'_n$$

n-tes zentrales Moment (n>1):

$$E[(x-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^n f(x) dx = \mu_n$$

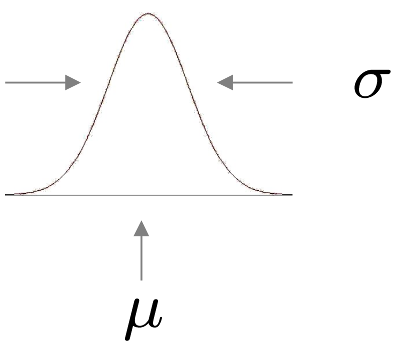
Varianz = 2tes zentrales Moment:

$$V[x] = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \qquad V[x] = \sigma^2$$

$$V[x] = E[(x-\mu)^2] = E[x^2] - 2E[x]\mu + \mu^2 = E[x^2] - \mu^2.$$

Standardabweichung: $\sigma = \sqrt{\sigma^2}$

$$\sigma_x = \sqrt{V[x]},$$



Erwartungswerte (III)

Weitere zentrale Momente (selten gebraucht):

$$\boxed{\text{Schiefe } \gamma = \mu_3/\sigma^3}$$

misst Asymmetrie der Verteilung:
positiv (negativ) für Ausläufer nch rechts (links)

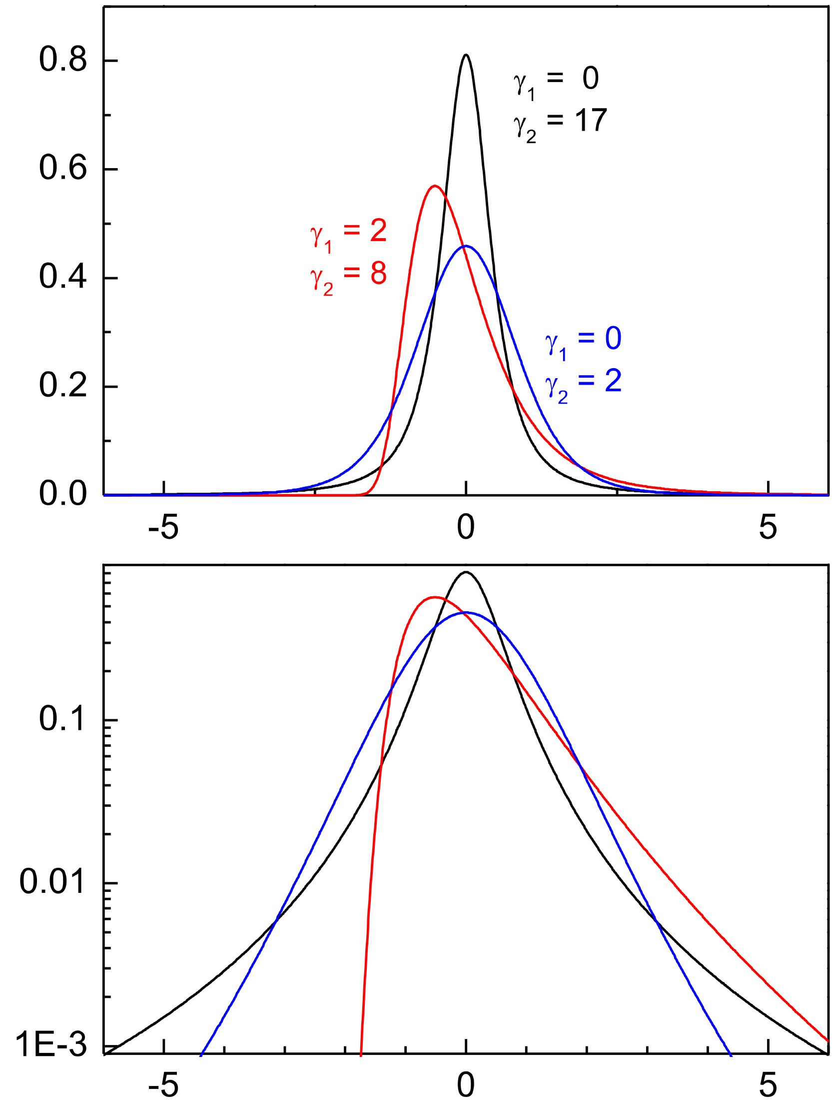
$$\boxed{\text{Wölbung oder Kurtosis } \kappa = \mu_4/\sigma^4 - 3}$$

misst Anteil im Zentrum zum Anteil in Ausläufern
relativ zur Gauss-/Normalverteilung

0: für Gaussverteilung
positiv: für mehr Ausläufer als Gaussverteilung bei gleicher Varianz
negativ: für weniger Ausläufer als Gaussverteilung bei gleicher Varianz

Schiefe $\gamma = \mu_3/\sigma^3$

$$
\begin{aligned}
\gamma_1 &= E\left[(x-\mu)^3\right]/\sigma^3 \\
&= E\left[x^3 - 3\mu x^2 + 3\mu^2 x - \mu^3\right]/\sigma^3 \\
&= \left\{E(x^3) - 3\mu\left[E(x^2) - \mu E(x)\right] - \mu^3\right\}/\sigma^3 \\
&= \frac{E(x^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}\,.
\end{aligned}
$$



Wölbung oder Kurtosis $\kappa = \mu_4/\sigma^4 - 3$

Alle Verteilungen haben
Mittelwert = 0 und Varianz =1

Binomial-Verteilung

Betrachte *N* unabhängige Messungen/Experimente (Bernoulli-Versuche):

Ausgang jedes Experimentes ist entweder ”Erfolg” oder “Misserfolg”
Die Wahrscheinlichkeit für Erfolg (Misserfolg) sind p (1-p)

Definiere diskrete ZV als Anzahl der Erfolge n ($0 \leq n \leq N$).

Wkt. für eine spezielle Reihenfolge von n Erfolgen und (N-n) Misserfolgen ist:

$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

Aber Anordnung ist unwichtg →Kombinatorik es gibt $\dfrac{N!}{n!(N-n)!}$

Möglichkeiten (Permutationen) *n* Erfolge in *N* Versuchen einzuordnen.
Geamtwkt. für *n* Erfolge ist die Summe der Wkt. für jede Permutation.

Binomial-Verteilung (2)

Die Binomialverteilung lautet:

$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Zufalls-
variable

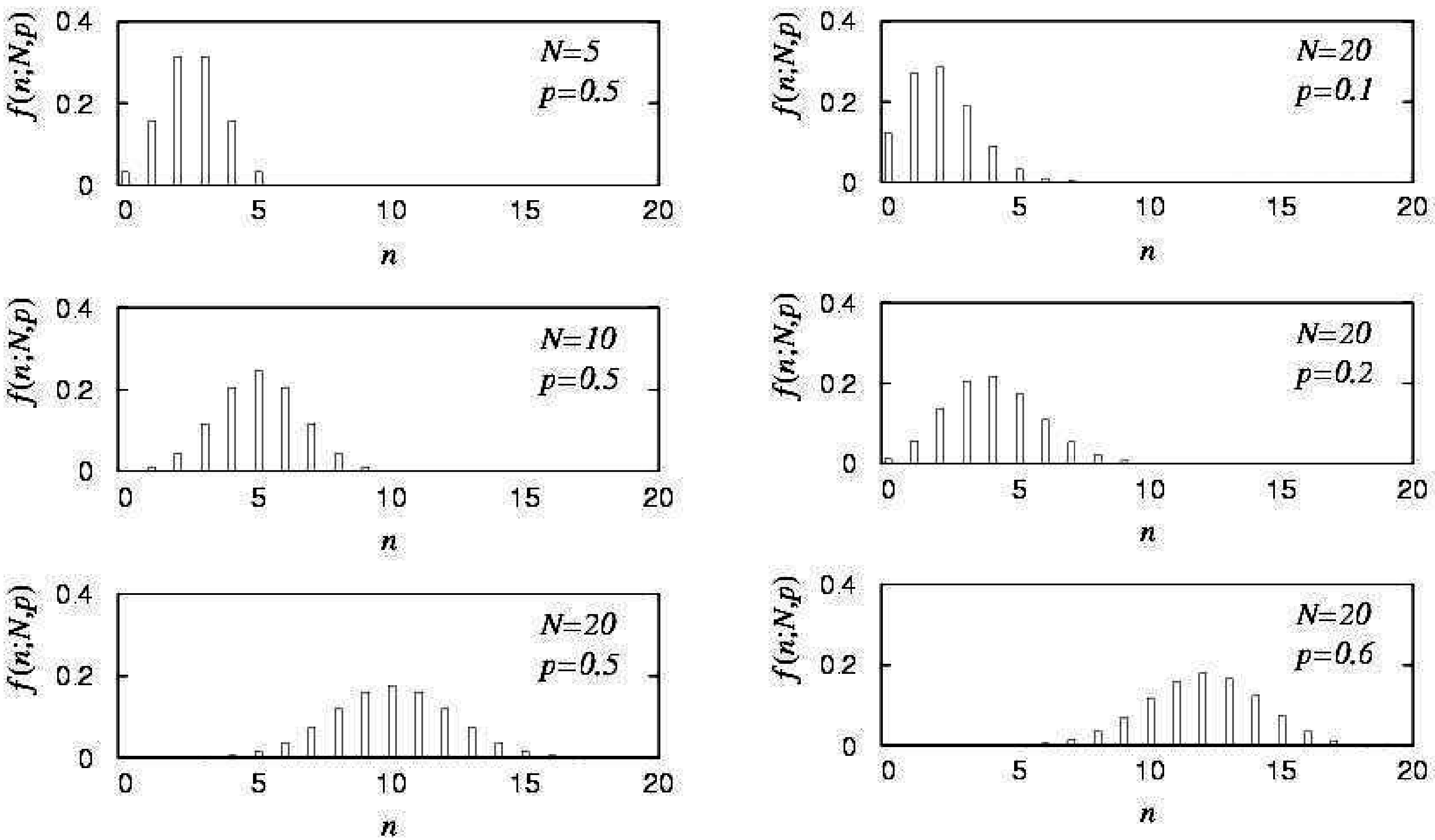
Parameter

Für den Erwartungswert und die Varianz findet man:

$$E[n] = \sum_{n=0}^{N} n f(n; N, p) = Np \qquad V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Die charakteristische Funktion lautet: $\varphi_n(k; p, N) = (p(\exp(ik) - 1) + 1)^N$

Verteilungen für verschiedene Parameter p und Anzahl der Versuche N:



Binomial-Verteilung: Beispiele

Nachweis eines Objektes mit Detektor
Ansprechwahrscheinlichkeit eines Detektors $p=0.95$

Bedingung: mindestes Treffer in drei Detektorlagen für Rekonstruktion der Spur
Frage: wieviel Kammern werden benötigt um >99% effizient zu sein?

3 Kammern: $P(k=3;0.95,3) = 0.95^3 = 0.857$
4 Kammern: $P(k>=3;0.95,4)=0.171+0.815=0.986$
5 Kammern: $P(k>=3;0.95,5)=0.021+0.204+0.774=0.999$

Wahrscheinlichkeit für Erfolg p
Wiviele Versuche damit mindestens 1 Erfolg mit Wkt >= α?
$P(k>=1;p,n) >= \alpha$ oder $1-P(K=0;p,n)>=\alpha$
$P(k=0)<=1-\alpha$
$(1-p)^N<=1-\alpha$
$N>=\ln(1-\alpha)/\ln(1-p)$

Binomial-Verteilung: Beweise

Normierung:

$$\sum_{n=0}^{N} f(n;N,p) = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$
$$= (p + (1-p))^N$$
$$= 1.$$

Erwartungswert: $E[n]$

$$E[n] = \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$
$$= \sum_{n=1}^{N} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$
$$= Np \sum_{n=1}^{N} \frac{(N-1)!}{(n-1)!((N-1)-(n-1))!} p^{(n-1)} (1-p)^{(N-1)-(n-1)}$$
$$= Np \sum_{n=0}^{N-1} \frac{(N-1)!}{(n)!((N-1)-(n))!} p^n (1-p)^{(N-1)-n}$$
$$= Np$$

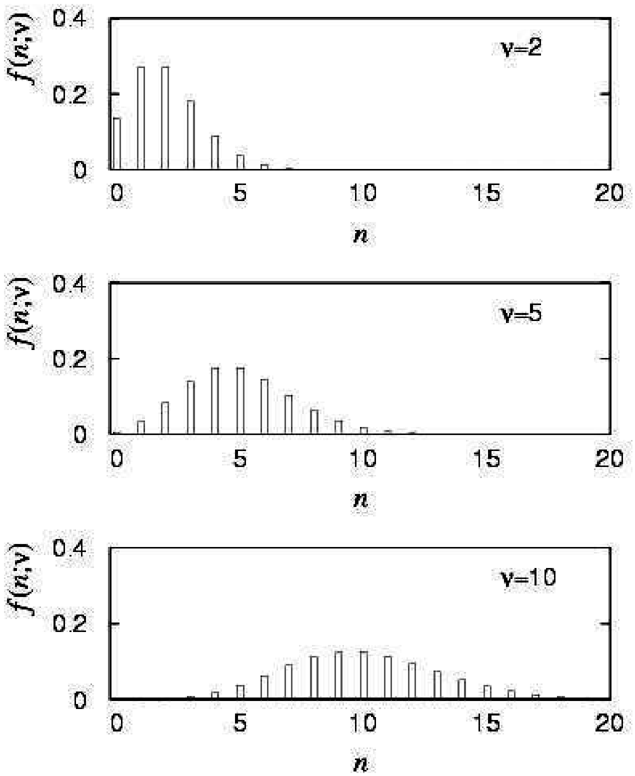
Betrachte Grenzfall der Binomialverteilung mit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \,. \quad \text{konstant}$$

$\to$ *n* folgt der Poissonverteilung:

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu} \qquad (n \geq 0)$$

$$E[n] = \nu\,, \qquad V[n] = \nu\,.$$



Example: number of scattering events *n* with cross section σ found for a fixed integrated luminosity, with

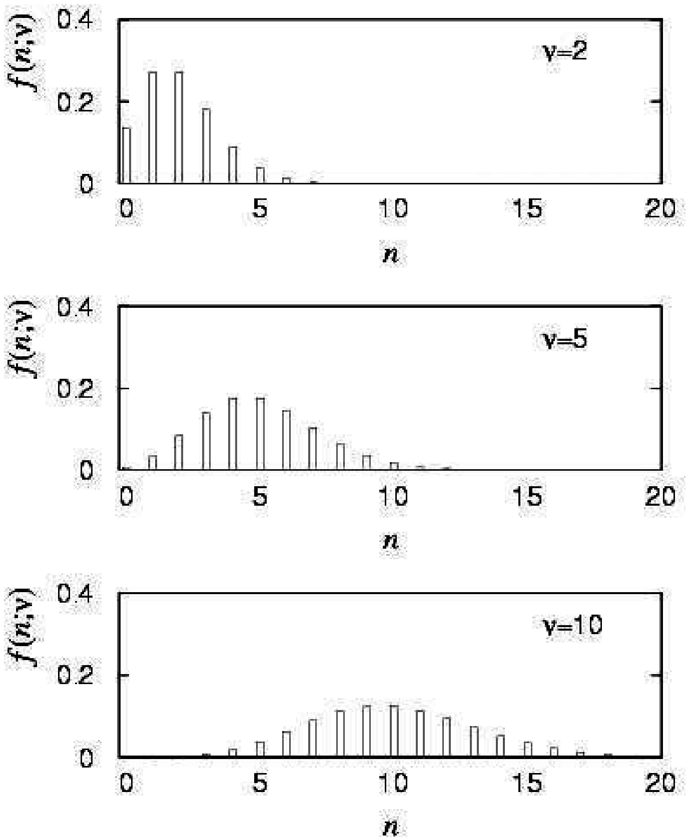
$$\nu = \sigma \int L\, dt \,.$$

Poissonverteilung: Herleitung 1

Betrachte Grenzfall der Binomialverteilung mit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \text{ . konstant}$$

$\to$ *n* folgt der Poissonverteilung:



$$f(n;\nu) = \frac{\nu^n}{n!} e^{-\nu} \qquad (n \geq 0)$$

$$E[n] = \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} \exp(-\lambda) = \lambda,$$

$$V[n] = \sum_{n=0}^{\infty} (n-\lambda)^2 \frac{\lambda^n}{n!} \exp(-\lambda) = \lambda.$$