



Kapitel 3: Wahrscheinlichkeitstheorie



CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Prof. Dr. Fred Hamker

Department of Computer Science

Wahrscheinlichkeitstheorie 1

Wahrscheinlichkeitstheorie

$$\Omega = \{ \text{[1][2][3][4][5][6]} \}$$

$$H = \{ \text{[1][2]} \}$$

$$P(H) = \frac{|H|}{|\Omega|} = \frac{|\{ \text{[1][2]} \}|}{|\{ \text{[1][2][3][4][5][6]} \}|}$$

Inhalt:

- Verbundwahrscheinlichkeit
- Summenregel
- Produktregel
- Bayes Theorem
- Apfel/Orange Beispiel
- A priori- und A posteriori Wahrscheinlichk.
- Beispiele
- Bayesscher Wahrscheinlichkeitsbegriff
- Wahrscheinlichkeitsdichte
- Bayes Theorem für kontinuierliche Variablen
- Erwartungswert
- Varianz und Kovarianz
- Wahrscheinlichkeitsdichteverteilungen

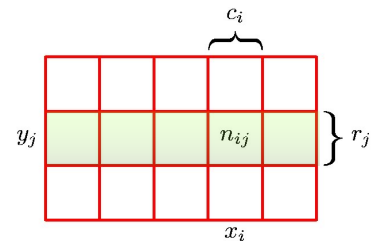
Bedingte Wahrscheinlichkeit

Betrachte die **bedingte Wahrscheinlichkeit (conditional probability)**

$P(X = x_i | Y = y_j)$ ist die bedingte Wahrscheinlichkeit für $X = x$, wenn $Y = y$ beobachtet wurde

$$P(X = x_i | Y = y_j) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

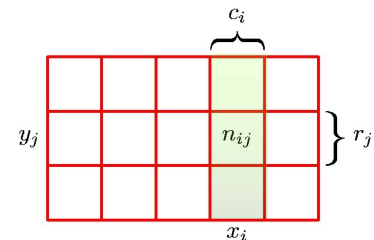
$$P(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$



$P(Y = y_j | X = x_i)$ ist die bedingte Wahrscheinlichkeit für $Y = y$, wenn $X = x$ beobachtet wurde

$$P(Y = y_j | X = x_i) = \frac{P(X = x, Y = y)}{P(X = x)}$$

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



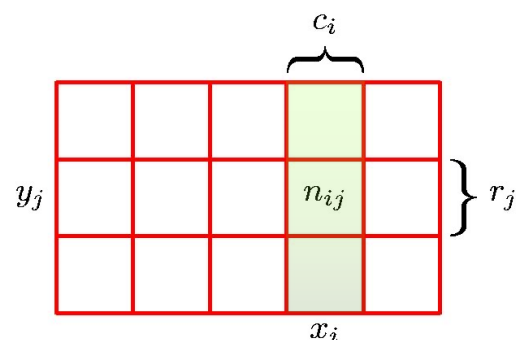
Produktregel

Betrachte die **bedingte Wahrscheinlichkeit (conditional probability)**

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

und die Randwahrscheinlichkeit

$$P(X = x_i) = \frac{c_i}{N},$$



dann ist die Verbundwahrscheinlichkeit das **Produkt** aus bedingter Wahrscheinlichkeit und Randwahrscheinlichkeit

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(Y = y_j | X = x_i) \cdot P(X = x_i)$$

Verallgemeinerung:

$$P(X_1, \dots, X_n) = P(X_n | X_1, \dots, X_{n-1}) \cdot \dots \cdot P(X_2 | X_1) \cdot P(X_1)$$

Bayes' Theorem

Summenregel: $P(X) = \sum_Y P(X, Y)$

Produktregel: $P(X, Y) = P(Y|X) \cdot P(X)$

Aus der Produktregel und der Symmetrieeigenschaft $P(X, Y) = P(Y, X)$ erhält man unmittelbar einen Zusammenhang zwischen den bedingten Wahrscheinlichkeiten

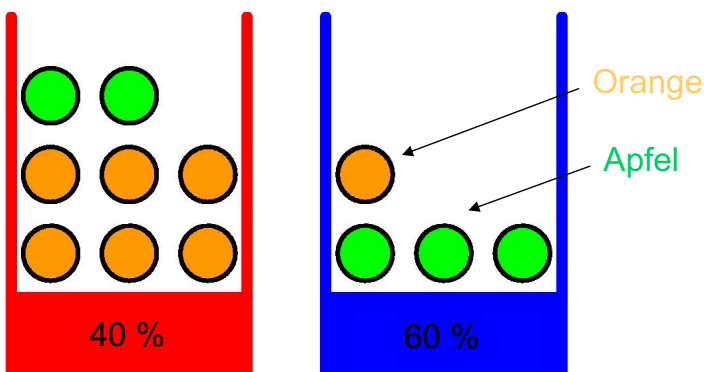
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$



Der Nenner normiert die Wahrscheinlichkeit auf eins und lässt sich berechnen mit

$$P(X) = \sum_Y P(X|Y)P(Y)$$

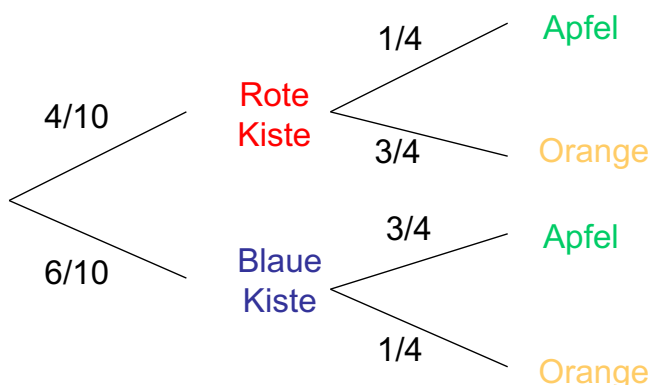
Apfel/Orange Beispiel



- Wähle zufällig eine Kiste
- Entnehme zufällig eine Frucht
- Lege die Frucht zurück

$$P(B = r) = 4/10$$

$$P(B = b) = 6/10$$



$$P(F = a|B = r) = 1/4$$

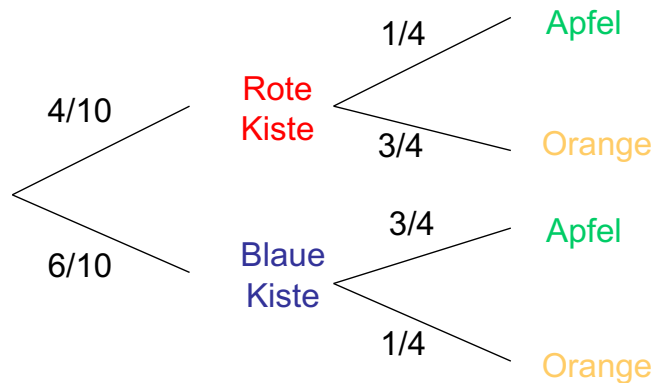
$$P(F = o|B = r) = 3/4$$

$$P(F = a|B = b) = 3/4$$

$$P(F = o|B = b) = 1/4$$

Apfel/Orange Beispiel

Wie hoch ist die Wahrscheinlichkeit einen Apfel zu bekommen ?



$$P(F = a) = P(B = r) \cdot P(F = a|B = r) + P(B = b) \cdot P(F = a|B = b)$$

$$= \frac{4}{10} \cdot \frac{1}{4} + \frac{6}{10} \cdot \frac{3}{4} = \frac{11}{20}$$

Angenommen wir bekommen eine Orange. Wie hoch ist die Wahrscheinlichkeit, dass sie aus der roten Kiste kommt?

$$P(B = r|F = o) = ?$$

Apfel/Orange Beispiel

Angenommen wir bekommen eine Orange. Wie hoch ist die Wahrscheinlichkeit, dass sie aus der roten Kiste kommt?

$P(B = r) = \frac{4}{10}$	$P(F = a B = r) = \frac{1}{4}$	$P(F = a B = b) = \frac{3}{4}$
$P(B = b) = \frac{6}{10}$	$P(F = o B = r) = \frac{3}{4}$	$P(F = o B = b) = \frac{1}{4}$
$P(F = o) = \frac{9}{20}$		

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)} = \frac{\frac{3}{4} \cdot \frac{4}{10}}{\frac{9}{20}} = \frac{2}{3}$$

A priori- und A posteriori-Wahrscheinlichkeit

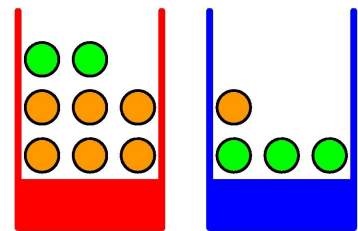
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \longrightarrow P(Y|X) \propto P(X|Y) \times P(Y)$$

A posteriori W. \propto bedingte W. x A priori W.

posterior probability \propto likelihood x prior probability

Unabhängigkeit

$$P(X, Y) = P(Y) \cdot P(X)$$



Add

Beispiel

Beispiel: h Hypothese: Jemand hat Malaria $P(h) = 0,0001$

e Ereignis: Jemand hat hohes Fieber

Regel: $e \rightarrow h$

Wir suchen die bedingte Wahrscheinlichkeit dafür, dass jemand Malaria hat unter der Bedingung, dass er hohes Fieber hat

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} = \frac{0,75 \cdot 0,0001}{0,14} \approx 0,000535$$

Beispiel

Ein bestimmter Test habe eine Fehlerquote von 5% nicht entdeckten Erkrankten und von 10% fälschlich als krank Diagnostizierten.

Sie wurden damit positiv getestet, als erkrankt eingestuft. Wie hoch ist die Wahrscheinlichkeit, dass sie tatsächlich erkrankt sind?

Gesucht: Wahrscheinlichkeit erkrankt zu sein, wenn positives Testergebnis: $P(E|TestPos)$

$$P(E|TestPos) = \frac{P(TestPos|E)P(E)}{P(TestPos)}$$

Wahrscheinlichkeit erkrankt zu sein: $P(E)$

Wahrscheinlichkeit eines positiven Testergebnisses, wenn erkrankt: $P(TestPos|E) = 100\% - 5\% = 95\%$

Wahrscheinlichkeit eines positiven Testergebnisses: $P(TestPos)$

Beispiel

Gesucht: Wahrscheinlichkeit erkrankt zu sein, wenn positives Testergebnis: $P(E|TestPos)$

$$P(E|TestPos) = \frac{P(TestPos|E)P(E)}{P(TestPos)}$$

$$P(TestPos) = P(TestPos|E)P(E) + P(TestPos|\bar{E})P(\bar{E})$$

$$P(\bar{E}) = 1 - P(E)$$

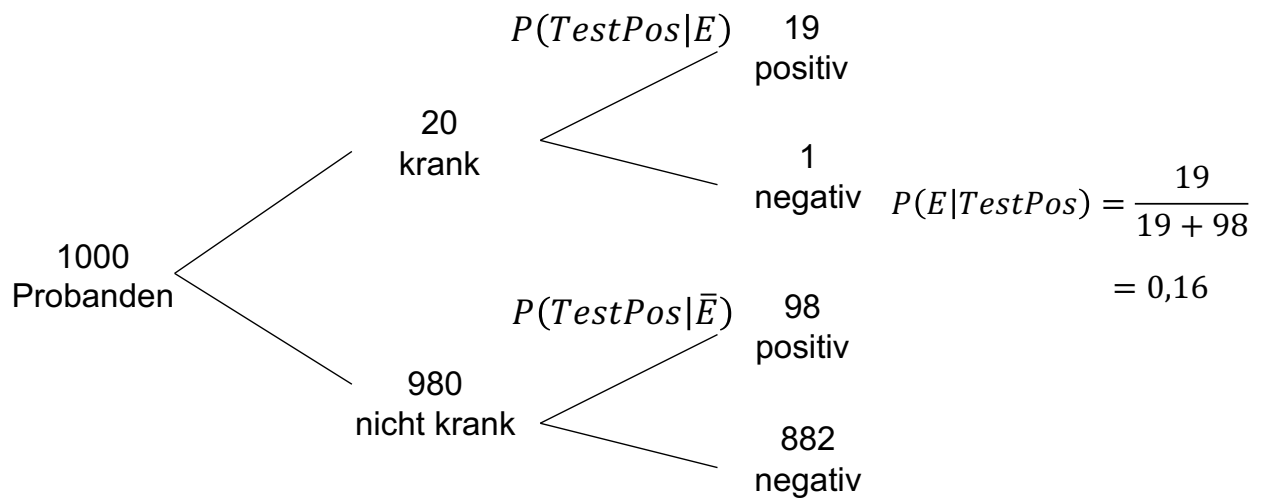
Wahrscheinlichkeit eines positiven Testergebnisses, wenn erkrankt: $P(TestPos|E) = 100\% - 5\% = 95\%$

Wahrscheinlichkeit eines positiven Testergebnisses, wenn nicht erkrankt: $P(TestPos|\bar{E}) = 10\%$

$$P(E|TestPos) = \frac{0,95 \cdot P(E)}{0,95 \cdot P(E) + 0,1 \cdot (1 - P(E))}$$

Beispiel

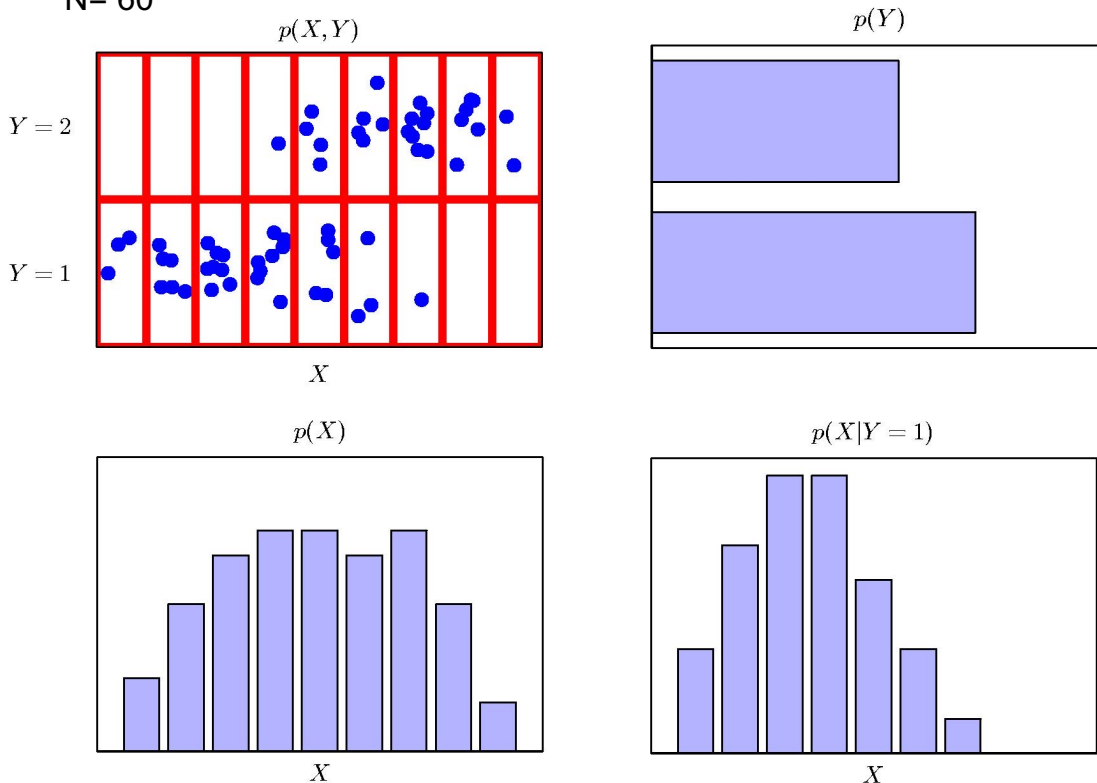
Wahrscheinlichkeit erkrankt zu sein: $P(E) = 2\%$



$$P(E|TestPos) = \frac{0,95 \cdot P(E)}{0,95 \cdot P(E) + 0,1 \cdot (1 - P(E))} = \frac{0,95 \cdot 0,02}{0,95 \cdot 0,02 + 0,1 \cdot 0,98} = 0,16$$

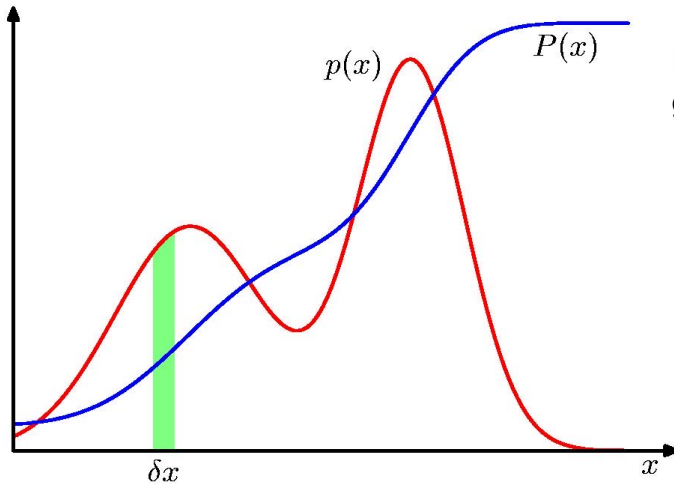
Wahrscheinlichkeitsverteilungen

N= 60



Wahrscheinlichkeitsdichte

Bisher haben wir lediglich die Wahrscheinlichkeiten von *diskreten* Ereignissen betrachtet. Hier führen wir die Wahrscheinlichkeit bezüglich *kontinuierlicher* Variablen ein.



Für die Wahrscheinlichkeitsdichte $p(x)$ gilt:

$$P = p(x)\delta x \quad \delta x \rightarrow 0$$

$$P(x \in (a, b)) = \int_a^b p(x)dx$$

$$p(x) \geq 0$$

$$P(z) = \int_{-\infty}^z p(x)dx \quad \text{Kumulative Verteilungsfunktion}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Bayes' Theorem für kontinuierliche Variablen

Summenregel

$$p(x) = \int_y p(x, y)dy$$

Produktregel

$$p(x, y) = p(y|x) \cdot p(x)$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Bayesscher Wahrscheinlichkeitsbegriff

Wahrscheinlichkeiten werden typischerweise im Sinne der relativen Häufigkeiten von zufälligen Ereignissen betrachtet (frequentistischer Wahrscheinlichkeitsbegriff).

Nach dem Bayesschen Wahrscheinlichkeitsbegriff werden keine Zufallsvariablen benötigt. Jede Information oder Messwert kann mit einer vernünftigen Einschätzung (*engl. belief*) interpretiert werden. Demnach stellen Wahrscheinlichkeiten ein Maß für die *Unsicherheit* dar.

Erwartungswert

Häufig interessiert man sich für bestimmte Kenngrößen einer Verteilung.

Erwartungswert:

diskret

$$E[F] = \sum_x P(x)f(x)$$

kontinuierlich

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

Approximation des Erwartungswerts durch eine endliche Stichprobengröße N :

$$E[F] \approx \frac{1}{N} \sum_{n=1}^N f(x_n)$$

Bedingter Erwartungswert:

$$E(f|y) = \sum_x P(x|y)f(x)$$

Varianz

Die Varianz ist ein Maß für die Streuung der Daten um den Erwartungswert:

$$\text{var}[f] = E[(f(x) - E[f(x)])^2]$$

$$\text{var}[f] = E[f^2(x)] - E[f(x)]^2$$

Aha!

$$\text{var}[f] = E[f^2(x) - 2f(x) \cdot E[f(x)] + E[f(x)]^2]$$

$$= E[f^2(x)] - 2E[f(x)] \cdot E[f(x)] + E[f(x)]^2$$

$$= E[f^2(x)] - E[f(x)]^2$$

Kovarianz

Bei zwei Zufallsvariablen x und y gibt die Kovarianz an, inwieweit die Variablen gleichartig variieren (linear). Die Kovarianz ist gleich Null, wenn die Variablen unabhängig sind oder der Zusammenhang besondere Nichtlinearitäten aufweist (z.B. U-förmig).

$$\text{cov}[x, y] = E[(x - E[x])(y - E[y])]$$

$$\text{cov}[x, y] = E[xy] - E[x] E[y]$$

Die Kovarianz gibt die Richtung einer Beziehung zwischen zwei Variablen an, über die Stärke des Zusammenhangs wird aber keine Aussage getroffen.

Kovarianz - Korrelationskoeffizient

Um einen Zusammenhang vergleichbar zu machen, muss die Kovarianz standardisiert werden. Man erhält dann den Korrelationskoeffizienten r .

$$r = \text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{Var}(x) \cdot \text{Var}(y)}} \in [-1, 1]$$

+1: perfekter linearer Zusammenhang,
0: kein linearer Zusammenhang,
-1: perfekter gegensätzlicher linearer Zusammenhang.

Schätzung:

$$\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

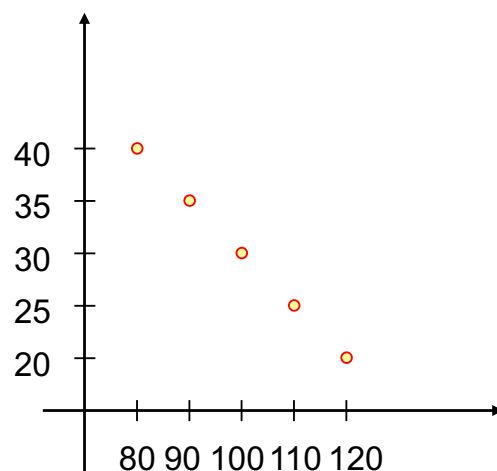
$$\text{Var}(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Beispiel 1 – Negative Korrelation

x	y
80	40
90	35
100	30
110	25
120	20



$$\bar{x} = 100$$

$$\text{Var}(x) = 250$$

$$\text{cov}(x, y) = -125$$

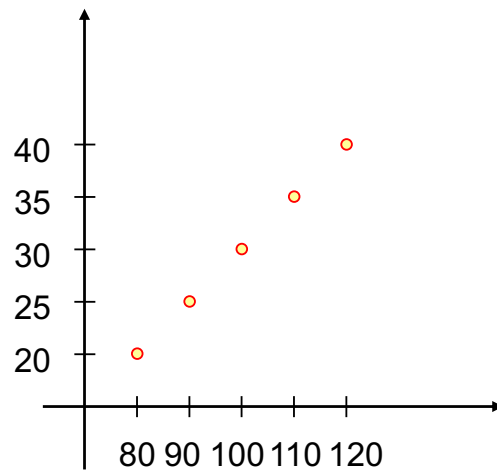
$$\bar{y} = 30$$

$$\text{Var}(y) = 62,5$$

$$r = -1$$

Beispiel 2 – Positive Korrelation

x	y
80	20
90	25
100	30
110	35
120	40



$$\bar{x} = 100$$

$$\text{Var}(x) = 250$$

$$\text{cov}(x, y) = 125$$

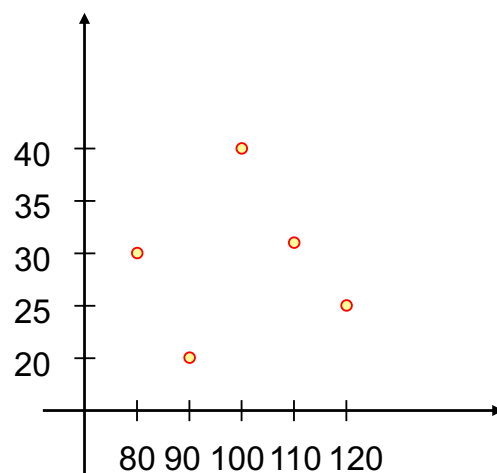
$$\bar{y} = 30$$

$$\text{Var}(y) = 62,5$$

$$r = 1$$

Beispiel 3 – keine Korrelation

x	y
80	30
90	20
100	40
110	31
120	25



$$\bar{x} = 100$$

$$\text{Var}(x) = 250$$

$$\text{cov}(x, y) = 2,5$$

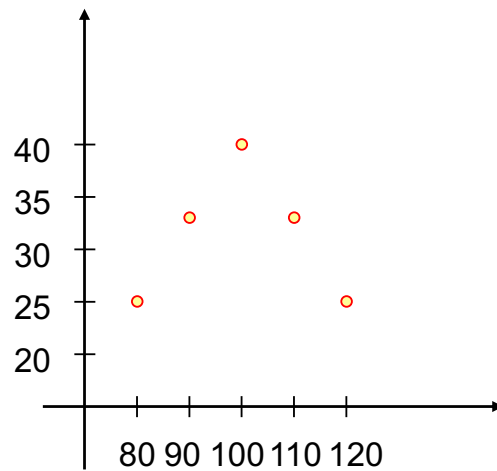
$$\bar{y} = 29,2$$

$$\text{Var}(y) = 55,7$$

$$r \approx 0,02$$

Beispiel 4 – keine (lineare) Korrelation!!

x	y
80	25
90	33
100	40
110	33
120	25



$$\bar{x} = 100$$

$$\text{Var}(x) = 250$$

$$\text{cov}(x, y) = 0$$

$$\bar{y} = 31,2$$

$$\text{Var}(y) = 40,2$$

$$r = 0$$

Wahrscheinlichkeitsdichteverteilungen

Wahrscheinlichkeitsdichteverteilungen können parametrisch oder nicht-parametrisch sein.

In der Praxis ist die Verteilung in der Regel nicht bekannt. Es ist daher erforderlich, diese zu schätzen.

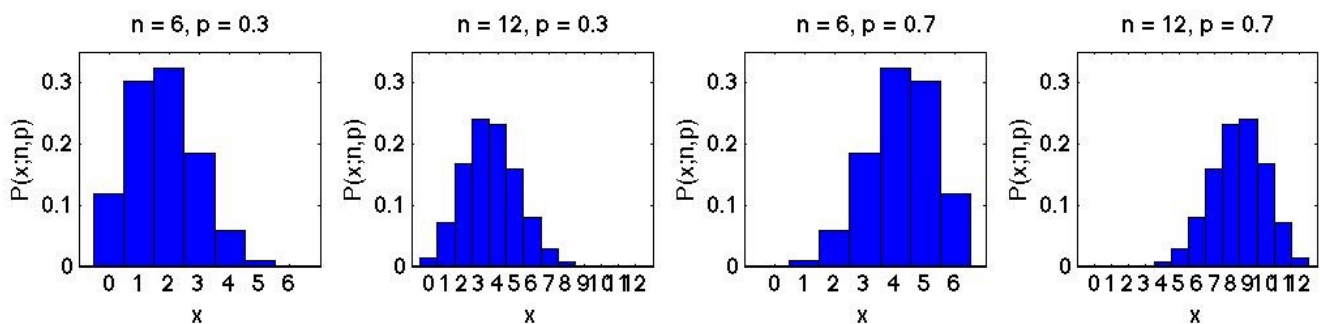
Hier zunächst einige parametrische Wahrscheinlichkeitsdichteverteilungen:

Binomialverteilung

- Diskrete Wahrscheinlichkeitsverteilung
- Beschreibt den wahrscheinlichen Ausgang einer Folge von gleichartigen Versuchen, die jeweils nur zwei mögliche Ergebnisse haben.
- Gewünschtes Ergebnis eines Versuches hat die Wahrscheinlichkeit p
- Zahl der Versuche ist n

Die Binomialverteilung gibt an, mit welcher Wahrscheinlichkeit sich insgesamt x Erfolge einstellen.

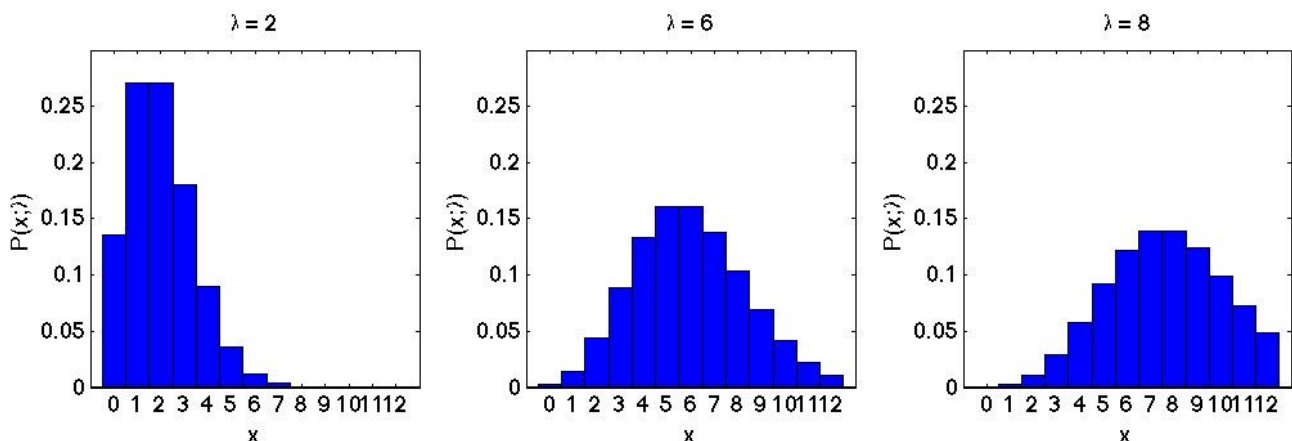
$$P(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$



Poissonverteilung

Die Poissonverteilung ist ebenfalls eine diskrete Wahrscheinlichkeitsverteilung. Sie ist die Grenzverteilung der Binomialverteilung bei sehr kleinem p und sehr großem Stichprobenumfang (n).

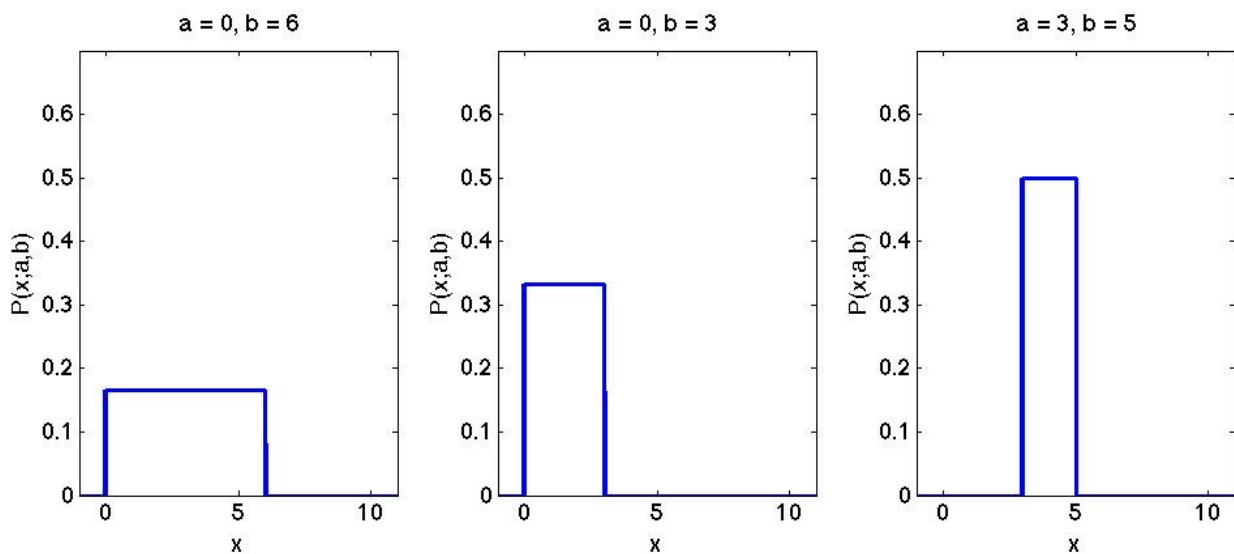
$$P(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$



Gleichverteilung

Bei der Gleichverteilung ist die Dichte über dem Wertebereich konstant.

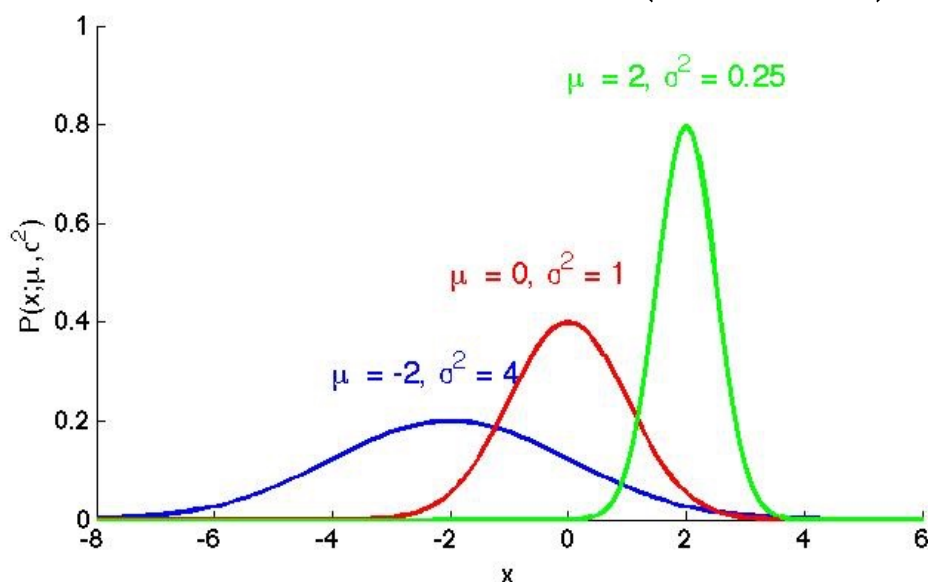
$$p(x; a, b) = \frac{1}{b - a}$$



Normalverteilung

Die besondere Bedeutung der Normalverteilung beruht auf dem zentralen Grenzwertsatz, der besagt, dass eine Summe von n unabhängigen, identisch verteilten Zufallsvariablen im Grenzwert normalverteilt ist.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



Nicht-parametrische Verteilungen

- Parametrische Wahrscheinlichkeitsdichteverteilungen sind sehr starr bzgl. der Form der Verteilung.
- Nicht-parametrische Verteilungen machen wenig Annahmen bzgl. der Form einer Verteilung.
- Nicht-parametrische Verteilungen benötigen allerdings in der Regel eine größere Stichprobe.

Histogramm

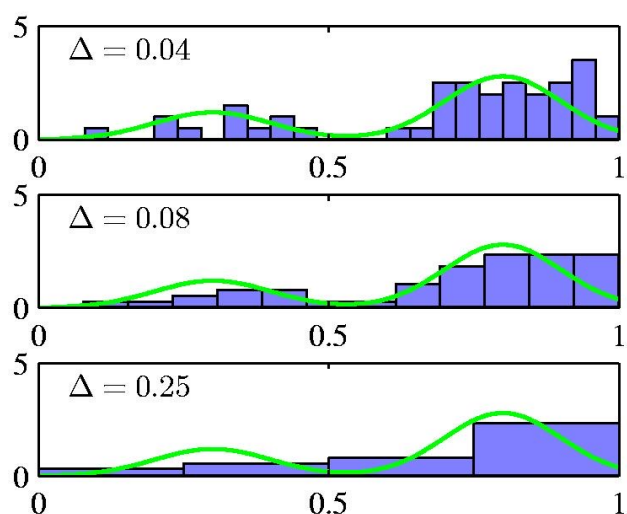
Teile den Raum in Intervalle gleicher Größe auf, und zähle die Anzahl der Datenpunkte in den jeweiligen Intervallen aus.

$$p_i = \frac{n_i}{N\Delta_i}$$

Δ_i : Größe des Intervalls

n_i : Anzahl der Datenpunkte in Δ_i

N : Anzahl der Datenpunkte



Nachteil: Verteilung ist nicht stetig.

Parzen Fenster / Kerndichtschätzer

Überlagerung von Kernen, die abhängig von der Stichprobenrealisation positioniert werden.

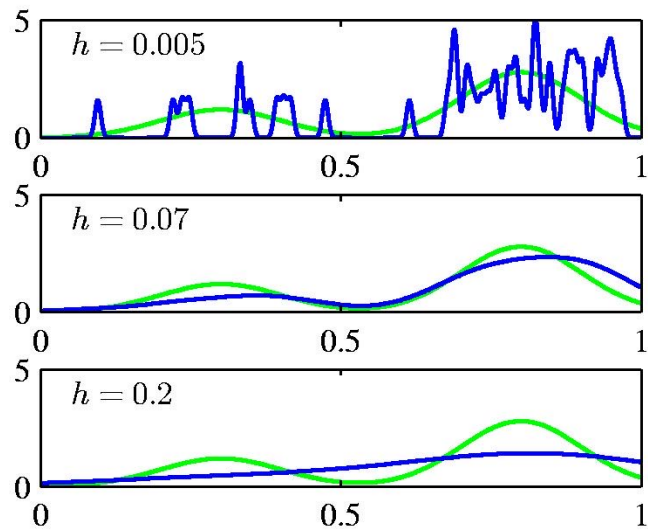
$$p(x) = \frac{K}{NV}$$

K : Anzahl der Datenpunkte im Volumen V

N : Anzahl der Datenpunkte

V konst. Bestimme K aus Daten durch Verwendung einer Fensterfunktion (Gewichtung):

$$p(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{\|x-x_n\|^2}{2h^2}}$$



Wähle h so klein, wie es die Daten erlauben!

K-nächste Nachbarn

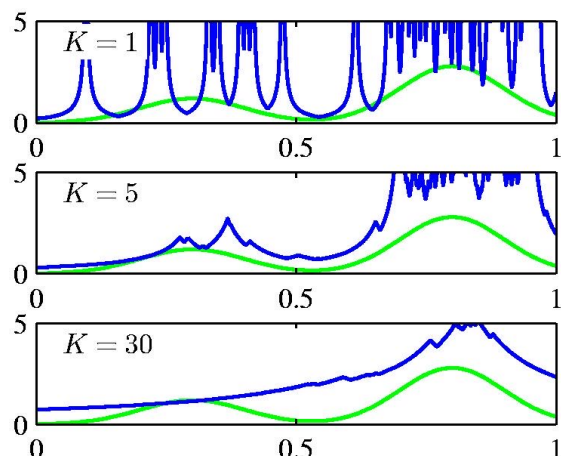
Bestimme für jede Position x das Volumen V so groß, dass es genau K Datenpunkte enthält.

$$p(x) = \frac{K}{NV}$$

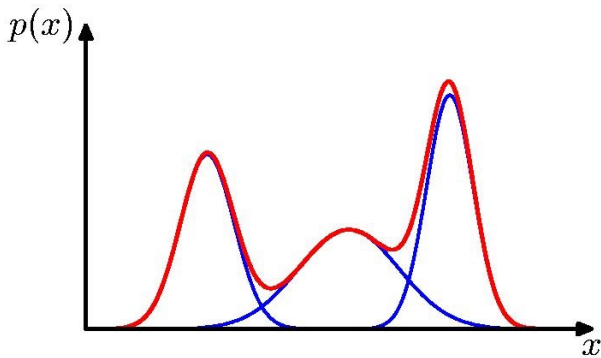
K : Anzahl der Datenpunkte im Volumen V

N : Anzahl der Datenpunkte

K konst. Bestimme V aus Daten:



Mischung von Gaussverteilungen



Jede beliebige Verteilung $p(x)$ lässt sich als Mischung von Gaussverteilungen approximieren.

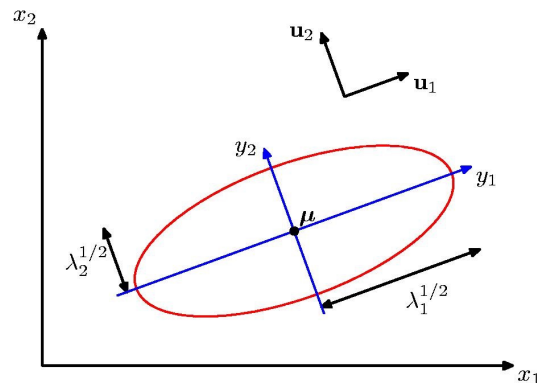
$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k)$$

Multivariate Gaussverteilung:

$$N(x|\mu, \Sigma) =$$

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right]$$

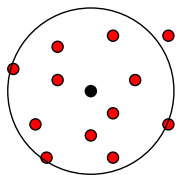
Σ Kovarianzmatrix



Multivariate Normalverteilungen

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right]$$

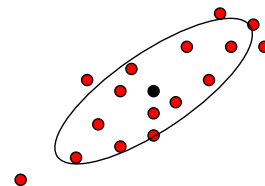
Σ Kovarianzmatrix



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Mischung von Gaussverteilungen

$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k)$$

Mischkoeffizienten: π_k

$$\sum_{k=1}^K \pi_k = 1$$

Schätzung der Parameter, siehe EM Algorithmus
in Kapitel "Schätzverfahren".