Stochastische Prozesse

Eine Folge von Zufallsgrößen $X_t, t = 1, 2, ...$
heißt stochastischer Prozeß

Beispiel: Zufällige Sequenz:
Identisch und unabhängig verteilte X_t.

Prozeß:	X_1	X_2	X_3	X_4	X_5	X_6	X_7	.	.
Realisierung:	A	C	T	A	T	A	A	.	.

Prozeß:	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	.	.
Realisierung:	0	0	1	2	1	2	3	.	.

Definition von Y:

$$Y_t = \left\{ \begin{array}{ll} 3 & (X_t = A, Y_{t-1} = 2) \text{ oder } Y_{t-1} = 3 \\ 2 & X_t = A, Y_{t-1} = 1 \\ 1 & X_t = T \\ 0 & \text{sonst} \end{array} \right\}$$

Der Prozeß $Y_t, t = 1, 2, ...$ entdeckt das Stopkodon TAA.

$P(Y_t = 3) = P(\text{Die Sequenz } X_1 X_2 ....X_t \text{ enthält ein } TAA)$

Neues Problem:
Die Verteilung von Y_t hängt von der Realisierung von Y_{t-1} ab.

Prozeß:	X_1	X_2	X_3	X_4	X_5	X_6	X_7	.	.
Realisierung:	A	C	T	A	T	A	A	.	.

Prozeß:	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	.	.
Realisierung:	0	0	1	2	1	2	3	.	.

Definition von Y:

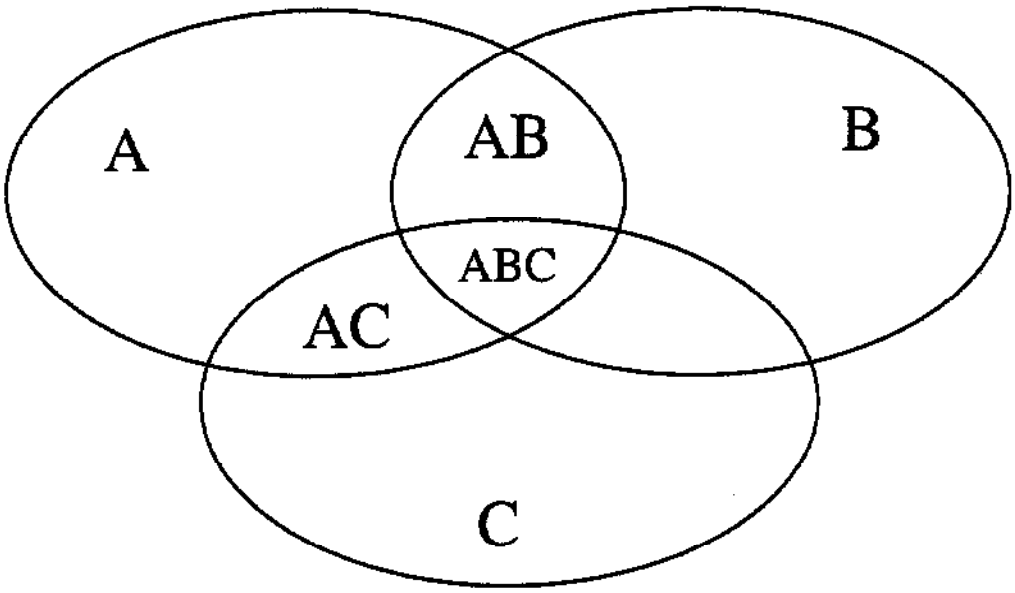
$$Y_t = \left\{ \begin{array}{ll} 3 & (X_t = A, Y_{t-1} = 2) \text{ oder } Y_{t-1} = 3 \\ 2 & X_t = A, Y_{t-1} = 1 \\ 1 & X_t = T \\ 0 & \text{sonst} \end{array} \right\}$$

Verteilung von Y_t bei bekanntem Y_{t-1}:

		Y_t			
		0	1	2	3
Y_{t-1}	0	$\frac{3}{4}$	$\frac{1}{4}$	0	0
	1	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0
	2	$\frac{2}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$
	3	0	0	0	1

Anfangswahrscheinlichkeit Y_0 : $P(Y_0 = 0) = 1$

Bedingte Wahrscheinlichkeit



$$P(AB) \leq P(A) \text{ weil } AB \subseteq A \quad \Rightarrow \quad 0 \leq \frac{P(AB)}{P(A)} \leq 1$$

Definition: $$P(B \mid A) = \frac{P(AB)}{P(A)}$$

heißt bedingte Wahrscheinlichkeit für B wenn A bereits eingetreten ist.

Neues Wahrscheinlichkeitsmaß
Definiert auf allen Teilmengen von A
Teilalgebra von Ereignissen mit $A = \Omega$

Formel: $P(AB) = P(A)\, P(B \mid A)$

Die Randverteilung von Y_t:

Beispiel: $P(Y_t = 3) = ?$

Unterscheide die Ereignisse und Wahrscheinlichkeiten:

$$P(Y_t = 3 \mid Y_{t-1} = 0) = p_{03} = 0$$
$$P(Y_t = 3 \mid Y_{t-1} = 1) = p_{13} = 0$$
$$P(Y_t = 3 \mid Y_{t-1} = 2) = p_{23} = 1/4$$
$$P(Y_t = 3 \mid Y_{t-1} = 3) = p_{33} = 1$$

$$P(Y_{t-1} = 0) = p_0^{t-1}$$
$$P(Y_{t-1} = 1) = p_1^{t-1}$$
$$P(Y_{t-1} = 2) = p_2^{t-1}$$
$$P(Y_{t-1} = 3) = p_3^{t-1}$$

$$P(Y_{t-1} = 0, Y_t = 3) = p_0^{t-1} p_{03} = 0$$
$$P(Y_{t-1} = 1, Y_t = 3) = p_1^{t-1} p_{13} = 0$$
$$P(Y_{t-1} = 2, Y_t = 3) = p_2^{t-1} p_{23} = (1/4) p_2^{t-1}$$
$$P(Y_{t-1} = 3, Y_t = 3) = p_3^{t-1} p_{33} = p_3^{t-1}$$

$$P(Y_t = 3) = (1/4) p_2^{t-1} + p_3^{t-1} = p_3^t$$

Allgemein:

$$P(Y_t = 3) = \sum_{k=0}^{3} p_k^{t-1} p_{k3}$$

$$P(Y_t = j) = \sum_{k=0}^{3} p_k^{t-1} p_{kj} = p_j^t$$

t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10	t=11
1	3/4	11/16	43/64								
0	1/4	4/16	16/64								
0	0	1/16	4/64								
0	0	0	1/64								

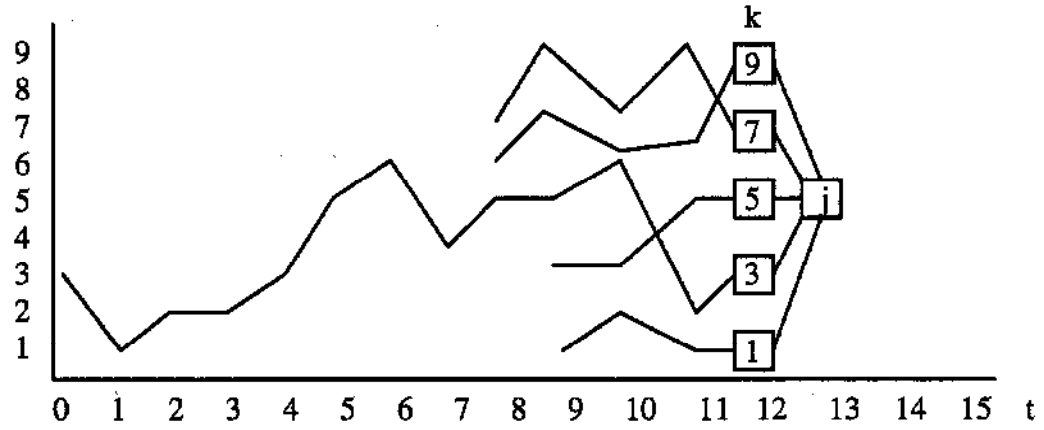
Markoff- Modell der Ordnung 1

$x_0, x_1, x_2, \ldots, x_T$ diskrete Zufallsgrößen mit Werten 1,2,....,n

$P(x_0 = i) = p_{0i}$ Anfangswahrscheinlichkeiten

$P(x_t = j \mid x_{t-1} = i) = p_{ij}$ Übergangswahrscheinl.

Zustandsraum (Trajektorien)



p_{ij}^n n-Schritt Übergangswahrscheinlichkeiten ($p_{ij}^1 = p_{ij}$)

$p_{ij}^{n+1} = \sum_k p_{ik}^n p_{kj} = \sum_k p_{ik} p_{kj}^n$ Rekursionsformel

$\sum_i p_{0i} p_{ij}^t = P(x_t = j)$ Randverteilungen

Durchgangswahrscheinlichkeit

Übergangsmatrix: $\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

Zustandsverteilung: $p_t' = p_0' \mathbf{P}^t$

$$p_0 = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}, p_1 = \begin{pmatrix} p_3 \\ p_1 \\ p_2 \end{pmatrix}, p_2 = \begin{pmatrix} p_2 \\ p_3 \\ p_1 \end{pmatrix}, p_3 = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Stationäre Verteilung

$$\begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix} (p_1\ p_2\ \ldots\ p_k) = \begin{pmatrix} p_1\ p_2\ \ldots\ p_k \\ p_1\ p_2\ \ldots\ p_k \\ p_1\ p_2\ \ldots\ p_k \\ p_1\ p_2\ \ldots\ p_k \\ p_1\ p_2\ \ldots\ p_k \\ p_1\ p_2\ \ldots\ p_k \end{pmatrix}$$

Satz von Frobenius für positive Matrizen

$p_{ij} > 0$ für alle i und $j \Rightarrow \lim_{n\to\infty} P^n = 1p'$ für einen Vektor p

Die Verteilung p is heißt stationäre Verteilung und hat die Eigenschaft

$$p' = p'P^m \text{ für alle } m$$

Beweis:

$$1p' = \lim_{n\to\infty} P^n = \lim_{n\to\infty} P^{n+m} = \left(\lim_{n\to\infty} P^n\right)P^m = 1p'P^m$$

$$1p' = 1p'P^m \Rightarrow p' = p'P^m$$

Wichtige Eigenschaften:

$p'P = p'$ Matrix P hat linken Eigenvektor p'

$P1 = 1$ Matrix P hat rechten Eigenvektor 1

Positive Vektoren: $(x_1, x_2, ..., x_n)' \quad x_i > 0$

z.B. Wahrscheinlichkeitsvektoren

$$(p_1, p_2, ..., p_n)' \quad p_i > 0 \quad \Sigma p_i = 1$$

Erwartungswertabschätzung:

$$\min_i x_i < \sum_i p_i x_i < \max_i x_i$$

Gleichheit gilt gdw alle x_i gleich sind

Übergangsmatrizen

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{n1} & \cdot & \cdots & p_{nn} \end{pmatrix} \quad p_{ij} > 0 \quad \sum_j p_{ij} = 1 \text{ für alle } i$$

$$Px = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{n1} & \cdot & \cdots & p_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_j p_{1j} x_j \\ \sum_j p_{2j} x_j \\ \cdot \\ \cdot \\ \cdot \\ \sum_j p_{nj} x_j \end{pmatrix} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(1)} \end{pmatrix} = x^{(1)}$$

Dann gilt

$$\min_i x_i < x_k^{(1)} < \max_i x_i \text{ für alle } k = 1, ..., n$$

$$\min_i x_i < \min_k x_k^{(1)} < \max_k x_k^{(1)} < \max_i x_i$$

wiederholte Anwendung liefert

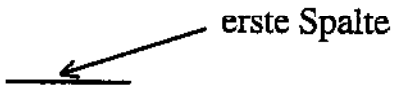
$$P^m x = P x^{(m-1)} \to \alpha 1 = \alpha \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \text{wenn } m \to \infty$$

Anwendung auf jede Spalte von P liefert

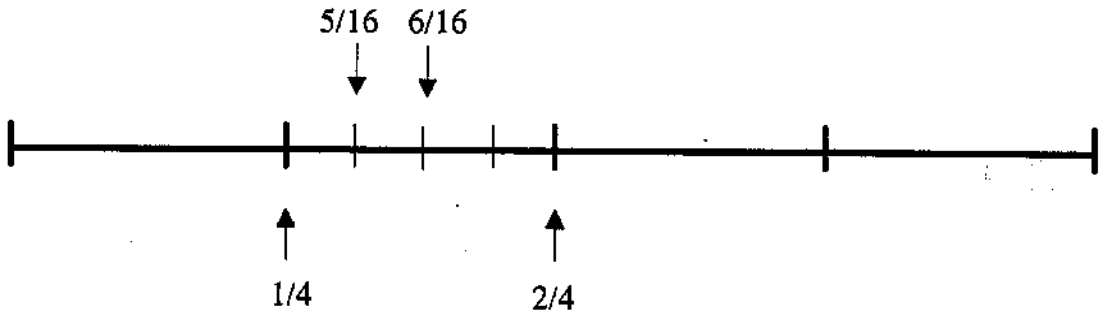
$$P^m P \to (\alpha_1 1, \alpha_2 1, ..., \alpha_n 1) = 1\alpha' = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (\alpha_1 \ \alpha_2 \ \ldots \ \alpha_n)$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (\alpha_1 \ \alpha_2 \ \ldots \ \alpha_n) = \begin{pmatrix} \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \\ \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \\ \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \\ \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \\ \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \\ \alpha_1 \ \alpha_2 \ \ldots \ \alpha_n \end{pmatrix}$$

Beispiel:



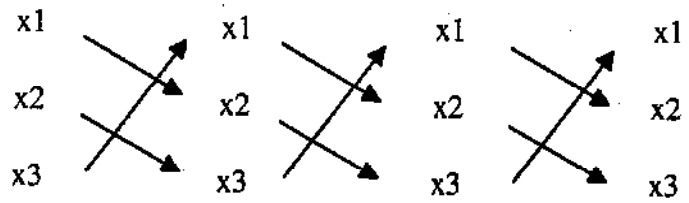
$$\frac{1}{4}\begin{pmatrix}1&2&1\\1&1&2\\2&1&1\end{pmatrix}\frac{1}{4}\begin{pmatrix}1\\1\\2\end{pmatrix}=\frac{1}{16}\begin{pmatrix}5\\6\\5\end{pmatrix}$$



$$\frac{1}{4}\begin{pmatrix}1&2&1\\1&1&2\\2&1&1\end{pmatrix}\frac{1}{16}\begin{pmatrix}5\\6\\5\end{pmatrix}=\frac{1}{64}\begin{pmatrix}22\\21\\21\end{pmatrix}$$

C $\min_i x_i^{(n)} = \sum_{k=1}^{n}\left(\frac{1}{4}\right)^k \to \sum_{k=1}^{\infty}\left(\frac{1}{4}\right)^k = \frac{1}{3}$

$$\begin{pmatrix}p_1\\p_2\\p_3\end{pmatrix}\quad\begin{pmatrix}p_3\\p1\\p_2\end{pmatrix}\quad\begin{pmatrix}p_2\\p_3\\p_1\end{pmatrix}\quad\begin{pmatrix}p_1\\p_2\\p_3\end{pmatrix}$$



Übergangsmatrix: $\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

Zustandsverteilung: $p'_t = p'_0 \mathbf{P}^t$

$$p_0 = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}, p_1 = \begin{pmatrix} p_3 \\ p_1 \\ p_2 \end{pmatrix}, p_2 = \begin{pmatrix} p_2 \\ p_3 \\ p_1 \end{pmatrix}, p_3 = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Aufgabenblatt 2:

Aufgabe 1 (Matrizenmultiplikation)

a) Berechnen Sie $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = ?$

b) Beweisen Sie $(A+B)C = AC + BC$

c) Beweisen Sie $(AB)C = A(BC)$

d) Welche Matrix X hat die Eigenschaft $AX = A$ für alle A?

Aufgabe 2: Ungleichungen für den Mittelwert

Beweise für gegebene $x_i \geq 0$ und Wahrscheinlichkeiten p_i die Ungleichungen $\min_i x_i \leq \sum_i p_i x_i \leq \max_i x_i$

Aufgabe 3: Übergangsmatrizen (alle Zeilensummen sind 1)

a) Ist das Produkt zweier Übergangsmatrizen eine Übergangsmatrix?

b) Betrachte die Übergangsmatrix

$$P = \begin{pmatrix} 1-p & p & 0 \\ 1-p & 0 & p \\ 1 & 0 & 0 \end{pmatrix}$$ und zeige für $p < 1$, daß $P^{\infty} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

Aufgabe 4: Indikatorprozesse

Konstruiere einen Markoff Prozeß erster Ordnung, der das Auftreten eines Stopkodons (TAA oder TAG oder TGA) entdeckt.

Aufgabe 5: Konvolution

Es seien X und Y zwei diskrete unabhängig verteilte Zufallsgrößen mit $P(X = i) = p_i, P(Y = i) = q_i, i = 1, ..., n.$
Geben Sie einen Algorithmus zur Berechnung der Verteilung von X+Y an.

EIKE MEERBACH

MUSTERLÖSUNG ÜBUNGSBLATT 1

1a) SINN: DIE ENTSTEHENDEN ELEMENTE SIND ERZEUGER (DURCH VEREINIGUNGEN) EINER σ-ALGEBRA.

$$(A+A^c)(C+C^c)(G+G^c)(T+T^c) =$$

$$(AC+AC^c)(AG+AG^c)(AT+AT^c)+(A^cC+A^cC^c)(A^cG+A^cG^c)(A^cT+A^cT^c)$$

$$=(\emptyset+AC^c)(\emptyset+AG^c)(\emptyset+AT^c)+(C+A^cC^c)(G+A^cG^c)(T+A^cT^c)$$

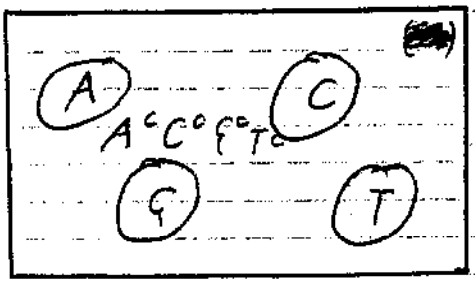
$$= A\cdot A\cdot A+(CG+CA^cG^c)(CT+CA^cT^c)+(A^cC^cG+A^cA^cC^cG^c)(A^cC^cT+A^cA^cT^c)$$

$$= A+CA^cG^cCA^cT^c+(G+A^cC^cG^c)(T+A^cC^cT^c)$$

$$= A+C+GT+GA^cC^cT^c+A^cC^cG^cT+A^cC^cG^cA^cC^cT^c$$

$$= A+G+\emptyset+G+T+A^cC^cG^cT^c$$

$$= \emptyset+A+C+G+T+A^cC^cG^cT^c.$$



1b) DIE ALGEBRA KANN NUN AUS VEREINIGUNGEN GEBILDET WERDEN. BEACHTE: $A^cC^cG^cT^c$ IST EREIGNIS, ABER NICHT MÖGLICH.

DIE ALGEBRA BESTEHT DANN AUS DEN FOLGENDEN 16 ELEMENTEN:

$\{\emptyset, A, C, G, T, A\cup C, A\cup G, A\cup T, C\cup G, C\cup T, G\cup T,$
$A\cup C\cup G, A\cup C\cup T, A\cup G\cup T, C\cup G\cup T, A\cup C\cup G\cup T\}$

1c) DIE UNABHÄNGIGKEIT KANN DURCH TABELLE ÜBERPRÜFT WERDEN.

BEACHTE: $A_1\cap A_2=\emptyset \Rightarrow P(A_1\cup A_2)=P(A_1)+P(A_2)$

SO IST ZUM BEISPIEL

$$P((A\cup C\cup T)\cap(C\cup T))=P(A\cap C\cup T)=P(C)+P(T).$$

	A	C	G	T	A∪C	A∪G	A∪T	C∪G	C∪T	G∪T	A∪C∪G	A∪C∪T	A∪G∪T	T∪C∪G	A∪C∪G∪T
A	1/4		0	0	0	1/4	1/4	1/4	0	0	0	1/4	1/4	1/4	0	1
C	1/16			0	0	1/4	0	0	1/4	1/4	0	1/4	1/4	0	1/4	1
G	1/16	1/16			0	0	1/4	0	1/4	0	1/4	1/4	0	1/4	1/4	1
T	1/16	1/16	1/16			0	0	1/4	0	1/4	1/4	0	1/4	1/4	1/4	1
A∪C	2/16	2/16	2/16	2/16						0	2/4	2/4	1/4	1/4		2/4
A∪G	2/16	2/16	2/16	12/816					0		2/4	1/4	2/4	1/4		2/4
A∪T	1/8	1/8	1/8	1/8				0			1/4	2/4	2/4	1/4		2/4
C∪G	1/8	1/8	1/8	1/8			1/4				2/4	1/4	2/4	2/4		3/4
C∪T	1/8	1/8	1/8	1/8		1/4					1/4	2/4	1/4	2/4		3/4
G∪T	1/8	1/8	1/8	1/8	1/4						1/4	1/4	2/4	2/4		3/4
A∪C∪G	3/16	3/16	3/16	3/16	2/8	2/8	2/8	2/8	2/8	2/8		2/4	2/4	2/4		3/4
A∪C∪T	3/16	3/16	3/16	3/16	3/8	3/8	2/8	3/8	2/8	3/8	9/16		2/4	2/4		3/4
A∪G∪T	3/16	3/16	3/16	3/16	2/8	2/8	2/8	2/8	2/8	2/8	9/16	9/16		2/4		3/4
T∪C∪G	3/16	3/16	3/16	3/16	2/8	2/8	2/8	2/8	2/8	2/8	9/16	9/16	9/16			3/4
A∪C∪G∪T	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4		
∅	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

← P(·∩·)

P(·)P(·) →

DANN SIND ALSO UNABHÄNGIG $\Omega = A \cup C \cup G \cup T$, $\emptyset$ VON ALLEN ANDEREN EREIGNISSEN, UND JEWEILS:

$\{A \cup C\}, \{A \cup G\}; \{A \cup T\}, \{A \cup C\}, \ldots$ (12 PAARE).

MUSTERLÖSUNG – ÜBUNGSBLATT 2 (AUSZÜGE) EIKE MEERBACH

1b) ZWEI $(n\times m)$-MATRIZEN SIND GLEICH, WENN ALLE IHRE EINTRÄGE ÜBEREINSTIMMEN, D.H. $A=B \Leftrightarrow a_{ij}=b_{ij}\ \forall i,j\in\{1,\dots,n\}$.

Sei $(A+B)\cdot C = D$; $A\cdot C + B\cdot C = E$ und d_{ij} ein beliebiger Eintrag aus D:

$$d_{ij} = \sum_{k=1}^{m}(a_{ik}+b_{ik})\cdot c_{kj} = \sum_{k=1}^{m} a_{ik}\cdot c_{kj} + b_{ik}\cdot c_{kj}$$

$$= \sum_{k=1}^{m} a_{ik}\cdot c_{ik} + \sum_{k=1}^{m} b_{ik}\cdot c_{kj} = e_{ij}\ .$$

c) WIE OBEN $(AB)C = D$; $A(BC)=E$; d_{ij} BELIEBIG:

$$d_{ij} = \sum_{k=1}^{m}\Big(\sum_{l=1}^{m} a_{il}\,b_{lk}\Big)c_{kj} = \sum_{l=1}^{m}\sum_{k=1}^{m} a_{il}\,b_{lk}\,c_{kj}$$

$$= \sum_{l=1}^{m} a_{il}\Big(\sum_{k=1}^{m} b_{lk}\,c_{kj}\Big) = e_{ij}\ .$$

d) ES SEI A EINE $(n\times m)$-MATRIX

X SEI $(\delta_{ij})_{m\times m}$, d.h. $x_{ij} = \begin{cases} 0 & \text{für } i\neq j \\ 1 & \text{für } i=j \end{cases}$.

$\Rightarrow$ c_{ij} SEI BELIEBIGER EINTRAG IN AX:

$$\Rightarrow c_{ij} = \sum_{k=1}^{m} a_{ik}\,x_{kj} = a_{ij} \Rightarrow AX = A.$$

DIE GESUCHTE MATRIX IST ALSO DIE EINHEITSMATRIX.

2.) $$\min_i x_i = \sum_j p_j \min_i x_i \le \sum_j p_j x_j \le \sum_j p_j \max_i x_i = \max_i x_i\ .$$

3b) DURCH VOLLSTÄNDIGE INDUKTION LÄSST SICH DIE FORM DER MATRIZEN P^n BEWEISEN:

i) $$P^{2n} = \begin{pmatrix} x_{11} & 0 & 0 \\ x_{21} & p^{2n} & 0 \\ x_{31} & 0 & p^{2n} \end{pmatrix}$$

ii) $$P^{2n+1} = \begin{pmatrix} x_{11} & 0 & 0 \\ x_{21} & 0 & p^{2n+1} \\ x_{31} & p^{2n+1} & 0 \end{pmatrix}$$

DIE EINTRÄGE AN DEN STELLEN x_{11-31} INTERESSIEREN UNS DABEI NICHT.

(3b) Beispiel c) I_A: $n=1$ $\quad P^2 = \begin{pmatrix} x_{11} & 0 & 0 \\ x_{21} & p^2 & 0 \\ x_{31} & 0 & p^2 \end{pmatrix}$

(durch Nachrechnen prüfen)

I_V:

$$P^{2n} = \begin{pmatrix} x_{11} & 0 & 0 \\ x_{21} & p^{2n} & 0 \\ x_{31} & 0 & p^{2n} \end{pmatrix}$$

$I_S: n \to n+1 \quad P^{2(n+1)} = P^{2n+2} = P^{2n} \cdot P^2$

$$= \begin{pmatrix} x_{11} & 0 & 0 \\ x_{21} & p^{2n} & 0 \\ x_{31} & 0 & p^{2n} \end{pmatrix} \begin{pmatrix} \hat{x}_{11} & 0 & 0 \\ \hat{x}_{21} & p^2 & 0 \\ \hat{x}_{31} & 0 & p^2 \end{pmatrix} = \begin{pmatrix} \tilde{x}_{11} & 0 & 0 \\ \tilde{x}_{21} & p^{2n+2} & 0 \\ \tilde{x}_{31} & 0 & p^{2n+2} \end{pmatrix}$$ q.e.d.

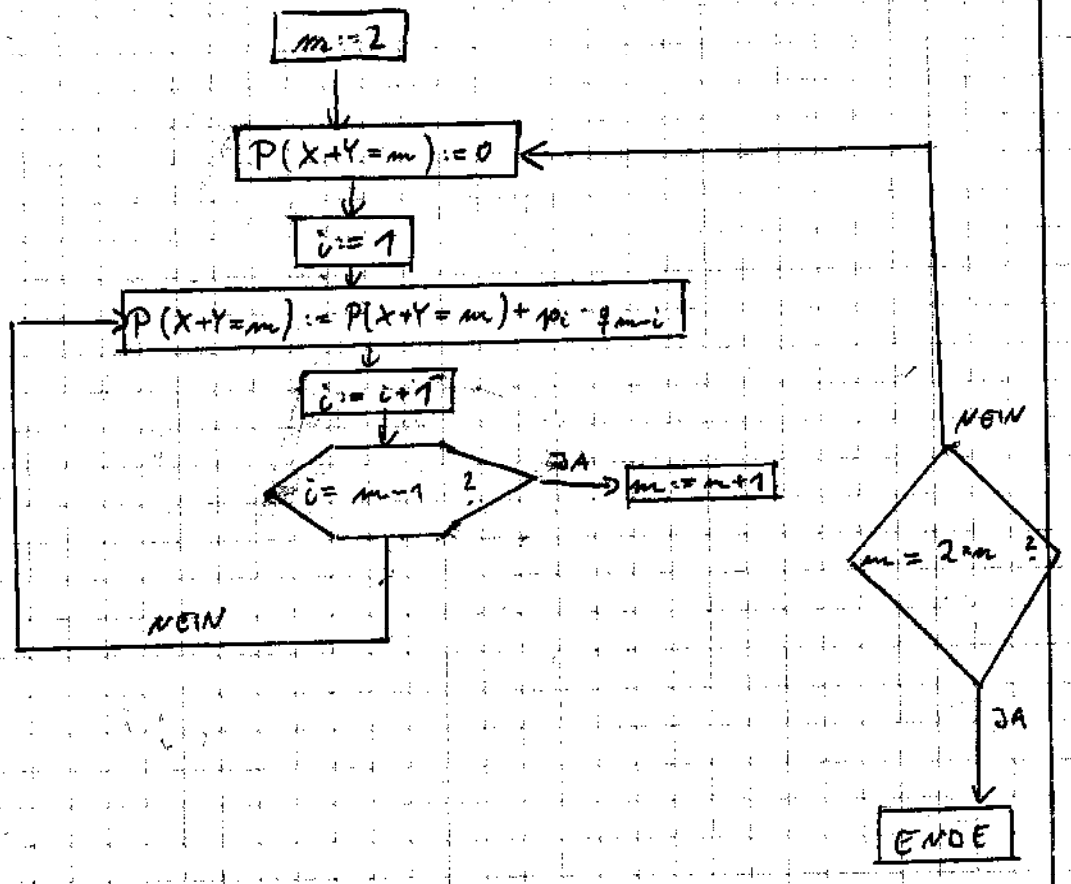
c) wird genauso gezeigt.

Nun gilt $\lim_{n\to\infty} p^n = 0$, da $|p|<1$

Da alle P^n Übergangsmatrizen sind (siehe 2a) müssen also die Einträge in x_{n-1} gegen 1 konvergieren.

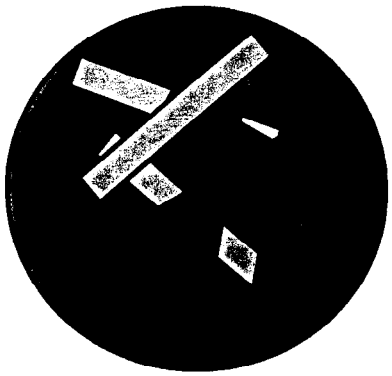
$$\Rightarrow \lim_{2n\to\infty} P^{2n} = \lim_{2n\to\infty} P^{2n+1} = \lim_{n\to\infty} P^n = P^\infty = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

5)



(Vanessa Walt)

Bayes'sche Formel



Merkmal A (verborgen)
Gewinn, Niete

Merkmal B (beobachtbar)
Frabe: gelb, grün, rot

Stichprobenraum, Verteilung von (A,B) ?

		A		
		1	2	
B	1	n_{11}	n_{12}	$n_{.1}$
	2	n_{21}	n_{22}	$n_{.2}$
	3	n_{31}	n_{32}	$n_{.3}$
		$n_{1.}$	$n_{2.}$	$n_{..}$

		A		
		1	2	
B	1	p_{11}	p_{12}	$p_{.1}$
	2	p_{21}	p_{22}	$p_{.2}$
	3	p_{31}	p_{32}	$p_{.3}$
		$p_{1.}$	$p_{2.}$	1

$P(A=i,B=j)=p_{ij}=n_{ij}/n_{..}$ gemeinsame Verteilung von A und B

$P(A=i)=p_{i.}=n_{i.}/n_{..}$ Randverteilung von A

$P(B=j)=p_{.j}=n_{.j}/n_{..}$ Randverteilung von B

$P(A=i \mid B=j)=n_{ij}/n_{.j}$ Verteilung von A bei bekanntem B

$P(B=j \mid A=i)=n_{ij}/n_{i.}$ Verteilung von B bei bekanntem A

$$P(A=i \mid B=j) = \frac{P(A=i,B=j)}{P(B=j)} = \frac{P(B=j \mid A=i)\,P(A=i)}{P(B=j)}$$

$$= \frac{P(B=j \mid A=i)\,P(A=i)}{\sum_k P(B=j \mid A=k)\,P(A=k)}$$

Übungsaufgaben

Assoziativgesetze:

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributivgesetze:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

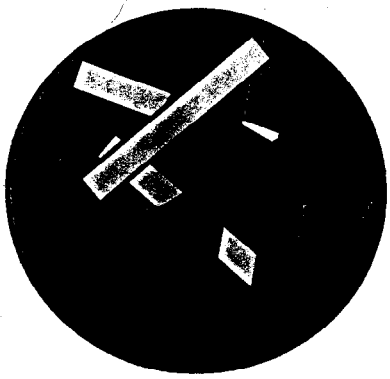
Komplementbildung:

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Losbeispiel: Merkmal A (Gewinn, Niete)
Merkmal B (Frabe: gelb, grün, rot)

Daten:

		A		
		1	2	
	1	10	20	
B	2	60	30	
	3	50	40	



Mit welchen Wahrscheinlichkeiten gewinnt man, wenn ein Los mit der Farbe 1, 2 oder 3 gezogen wird ?

Claverie and Bougueleret (1986)

Most frequent 8-tuples:

TTTTTTTT	ATATTTTT	TTAATTTT	TTATATTT
CTTTTTTT	TAATTTTT	ATAATTTT	ATATATTT
ATTTTTTT	AAATTTTT	AAAATTTT	TAATATTT
TCTTTTTT	TTTCTTTT	TTTTCTTT	AAATATTT
TATTTTTT	TTTGTTTT	TTTTGTTT	TTTAATTT
AATTTTTT	TTTATTTT	TTTTATTT	AATAATTT
TTCTTTTT	ATTATTTT	ATTTATTT	AAAAATTT
TTGTTTTT	TATATTTT	TATTATTT	TTTTTCTT
TTATTTTT	AATATTTT	AATTATTT	TTTTTGTT
TTTTTATT	ATTTTATT	TATTTATT	AATTTATT

Dinucleotide composition

AA:	1506	AC:	881	AG:	988	AT:	1123
CA:	903	CC:	1021	CG:	368	CT:	1254
GA:	831	GC:	681	GG:	727	GT:	633
TA:	1259	TC:	963	TG:	788	TT:	1723

total: 15649 nucleotides

Sea urchin complete mitochondrial genome

$P(A|A)$, $P(A|C)$, ... $\}$ 16x P_{ij}

$P_0(A)$, $P_0(C)$, $P_0(G)$, $P_0(T)$ ← Anfangswahrscheinlichkeit

$P(A,C) = P(A) \cdot P(C|A)$ (P_0, P_{ij})

$P(A,C,G) = P_{0i} \cdot P_{ij} \cdot P_{jk}$

Bayes decision rules (M. Borodovsky)

General formula:

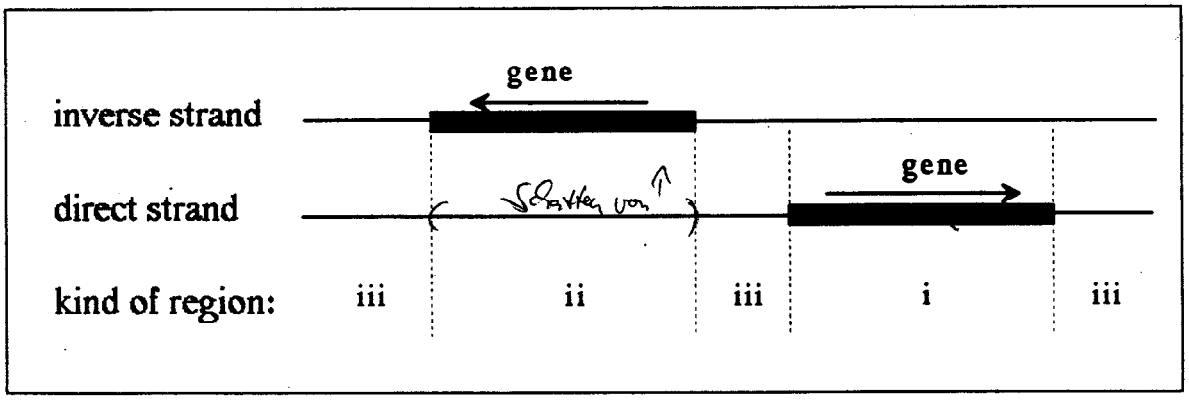
$$P(C \mid S) = \frac{P(S \mid C) * P(C)}{P(S \mid C) * P(C) + P(S \mid N) * P(N)}$$

Individual reading frames:

$$P(C_i \mid S) = \frac{P(S \mid C_i) * P(C_i)}{P(S \mid C_i) * P(C_i) + P(S \mid N) * P(N)}$$

Simultaneous consideration of all reading frames:

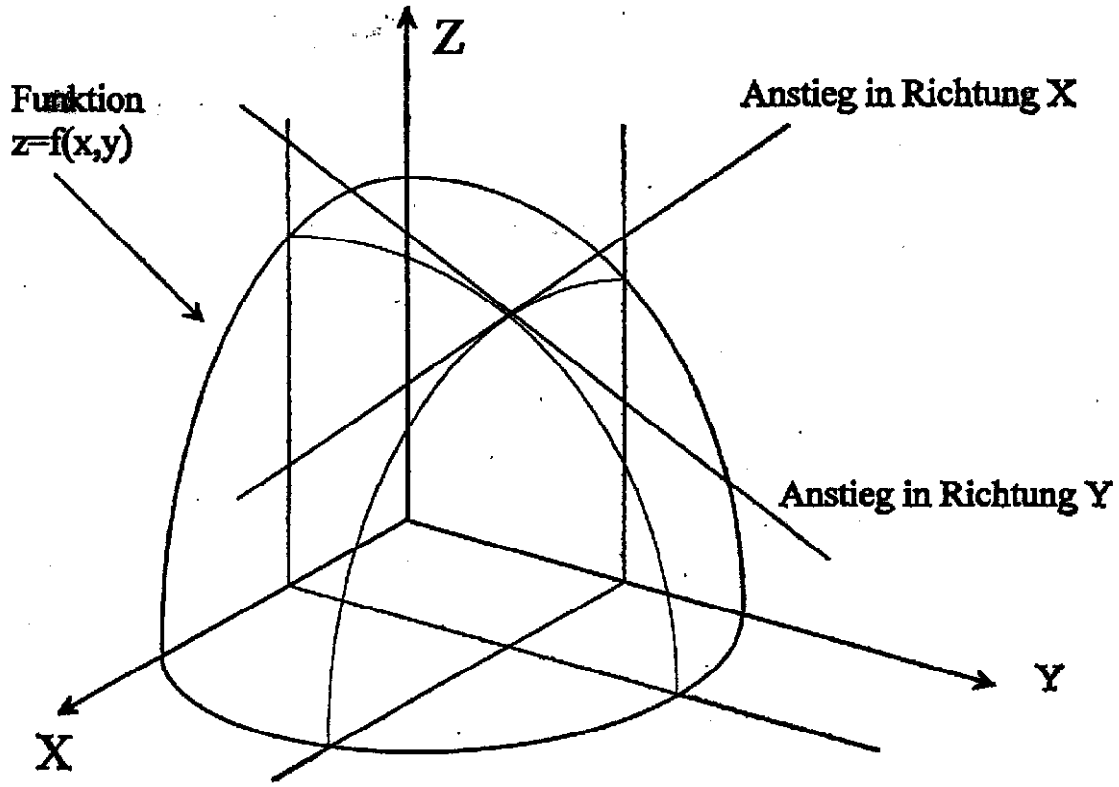
$$P(C_i \mid S) = \frac{P(S \mid C_i) * P(C_i)}{P(S \mid N) * P(N) + \sum_{i=1}^{3} P(S \mid C_i) * P(C_i)}$$



Simultaneous consideration of strand and anti strand

$$P(C_i \mid S) = \frac{P(S \mid C_i) * P(C_i)}{P(S \mid N) * P(N) + \sum_{i=1}^{3} P(S \mid C_i) * P(C_i) + \sum_{i=1}^{3} Q(S \mid C_i) * Q(C_i)}$$

$$Q(C_i \mid S) = \frac{Q(S \mid C_i) * Q(C_i)}{P(S \mid N) * P(N) + \sum_{i=1}^{3} P(S \mid C_i) * P(C_i) + \sum_{i=1}^{3} Q(S \mid C_i) * Q(C_i)}$$

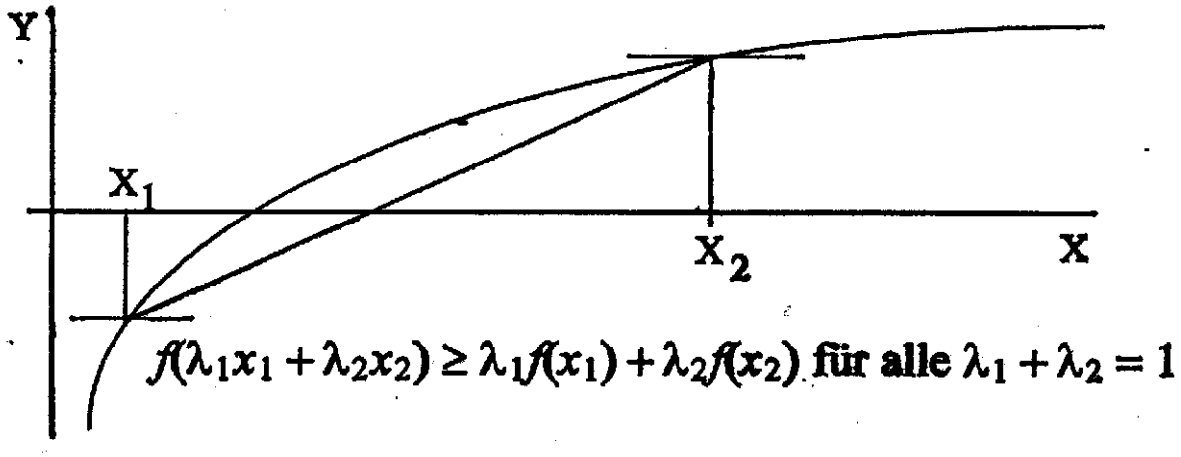


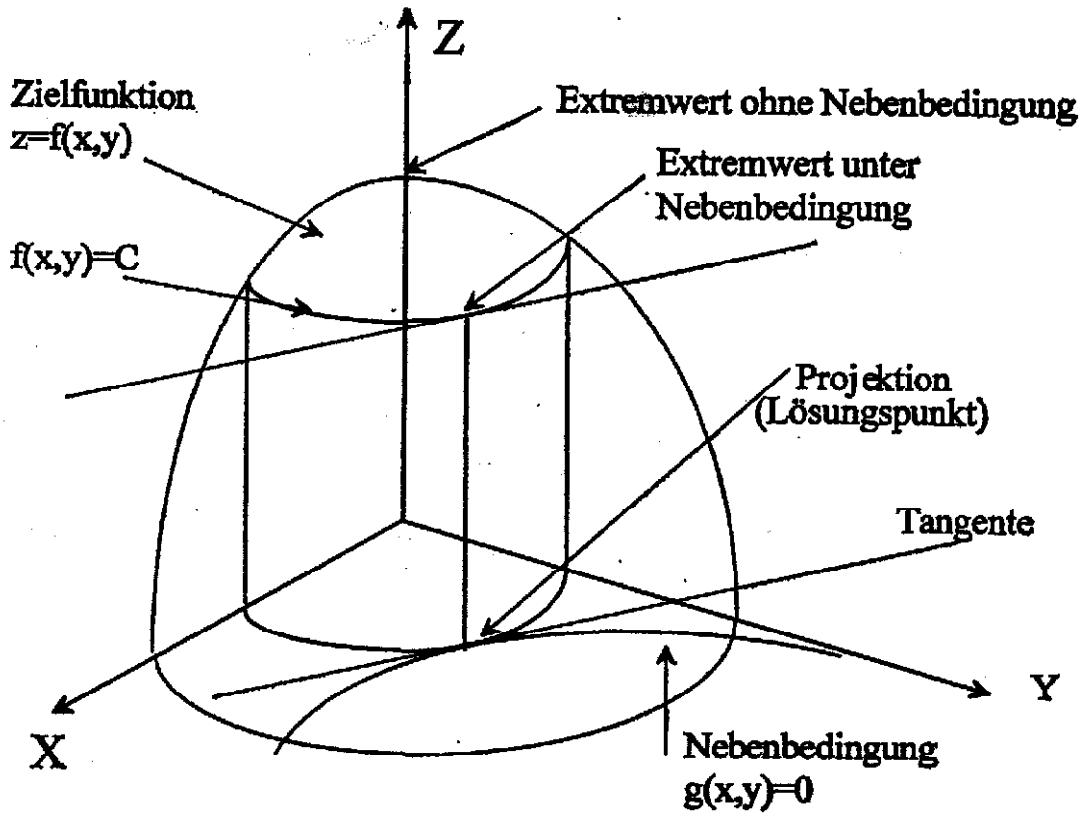
Die Ableitung einer Funktion von mehreren Variablen ($z = f(x,y)$) nach einer speziellen Variablen (x oder y) heißt partielle Ableitung

$$f_x = \frac{\partial f(x,y)}{\partial x} \quad , \quad f_y = \frac{\partial f(x,y)}{\partial y}$$

Sie liefert den Anstieg einer Funktion in Richtung der ausgewählten Variablen.

Satz: Liegt in einem Punkt ein lokales Maximum oder Minimum vor, so verschwinden dort alle partiellen Ableitungen.
Für konvexe/konkave Funktionen folgt aus dem Verschwinden aller partieller Ableitungen das Vorliegen eines absoluten Maximums/Minimums.





Beobachtung:

Wenn C das Maximum von f(x,y) unter g(x,y)=0 ist, dann hat f(x,y)=C im Lösungspunkt die gleiche Ableitung wie g(x,y)=0.

Durch implizites Differenzieren folgt:

$g(x,y) = x^2 + 1 - y = 0$ (implizite Darstellung) $\Big| \frac{dy}{dx} = -\frac{\frac{\partial g(x,y)}{\partial x}}{\frac{\partial g(x,y)}{\partial y}}$

$$\frac{dy}{dx} = -\frac{f_x}{f_y} \quad \text{mit } f_x = \frac{\partial f(x,y)}{\partial x} \text{ und } f_y = \frac{\partial f(x,y)}{\partial y}$$

$$\frac{dy}{dx} = -\frac{g_x}{g_y} \quad \text{mit } g_x = \frac{\partial g(x,y)}{\partial x} \text{ und } g_y = \frac{\partial g(x,y)}{\partial y}$$

Gleichung: Ableitung von g nach x

$$\frac{f_x}{f_y} = \frac{g_x}{g_y} \Rightarrow \begin{matrix} f_x = -\lambda g_x \\ f_y = -\lambda g_y \end{matrix} \quad \text{mit } \lambda = \frac{f_y}{g_y} = \frac{f_x}{g_x}$$

Langrange-Funktion:

$$f(x,y) + \lambda g(x,y) \Rightarrow \begin{matrix} f_x + \lambda g_x = 0 \\ f_y + \lambda g_y = 0 \end{matrix}$$

Schätzung von Parametern

Sequenz S = ACGAGTCATTCGAATCGT.....

Markoff Modell 0-ter Ordnung

$$P(S) = p(A)p(C)p(G)p(A).....$$

$$\ln P(S) = \ln p(A) + \ln p(C) + \ln p(G) + \ln p(A).....$$

$$\ln P(S) = \sum n_i \ln p_i$$

Aufgabe: Für welche Parameter p_i hat die beobachtete Sequenz S die größte Wahrscheinlichkeit?

$$\max_{0 \le p_i \le 1, \sum p_i = 1} \sum n_i \ln p_i = \sum n_i \ln \hat{p}_i$$

Die $\hat{p}_i$ heißen Maximum Likelihood Schätzungen der Parameter p_i

Lösung: Lagrange Funktion: $\sum n_i \ln p_i + \lambda(\sum p_i - 1)$

Ableitung nach p_k: $\frac{n_k}{p_k} + \lambda = 0 \Rightarrow n_k = -\lambda p_k$

Summation über k $\Rightarrow N = -\lambda$

Ergebnis: $p_k = \frac{n_k}{N}$

Markoff Modell 1-ter Ordnung

$$\ln P(S) = \ln p_{0A} + \ln p_{AC} + \ln p_{CG} + \ln p_{GA} + ...$$

$$\ln P(S) = \ln p_{oA} + \sum n_{ij} \ln p_{ij}$$

Aufgabe: $$\max_{0 \le p_{ij} \le 1, \sum_j p_{ij} = 1} \sum_i \left(\sum_j n_{ij} \ln p_{ij} \right) = \sum_i \left(\sum_j n_{ij} \ln \hat{p}_{ij} \right)$$

Lösung: $\hat{p}_{ij} = \frac{n_{ij}}{n_{i.}}$

Ungleichung der Informationstheorie

Seien a_i und b_i positive Zahlen mit $\sum b_i \le \sum a_i$.

Dann gilt $$\sum a_i \ln \frac{b_i}{a_i} \le 0$$

und Gleichheit gilt genau dann wenn $a_i = b_i$ für alle i.

Beweis: Taylor-Entwicklung der Logarithmusfunktion liefert für $x > 0$

$$\ln x = (x-1) - \frac{(x-1)^2}{2y^2} \text{ mit } y \in (1,x)$$

Für $x = \frac{b_i}{a_i}$ folgt

$$\ln \frac{b_i}{a_i} = \left(\frac{b_i}{a_i} - 1.\right) - \frac{\left(\frac{b_i}{a_i} - 1\right)^2}{2y_i^2}, \quad y_i \in (1, \frac{b_i}{a_i})$$

$$\ln \frac{b_i}{a_i} = \frac{1}{a_i}(b_i - a_i) - \frac{(b_i - a_i)^2}{2y_i^2 a_i^2}, \quad y_i \in (1, \frac{b_i}{a_i})$$

$$\sum a_i \ln \frac{b_i}{a_i} = \sum (b_i - a_i) - \sum \frac{(b_i - a_i)^2}{2y_i^2 a_i} \le 0$$

Die zweite Summe verschwindet genau dann wenn $a_i = b_i$ und dann verschwindet auch die erste.

Anwendung: $$a_i = \frac{n_i}{N}, b_i = p_i, \sum a_i = 1, \sum b_i = 1$$

$$\Rightarrow \quad \sum \frac{n_i}{N} \ln p_i \le \sum \frac{n_i}{N} \ln \frac{n_i}{N}$$

Übungsaufgaben

1. Beweisen Sie die folgenden Mengenrelationen.

a) Assoziativgesetze: $(A \cup B) \cup C = A \cup (B \cup C)$
$(A \cap B) \cap C = A \cap (B \cap C)$

b) Distributivgesetze: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

c) Komplementbildung: $(A \cup B)^c = A^c \cap B^c$
$(A \cap B)^c = A^c \cup B^c$

2. Losbeispiel: Merkmal A (Gewinn 1, Niete 2)
Merkmal B (Frabe: gelb 1, grün 2, rot 3)

		A		
		1	2	
	1	10	20	
B	2	60	30	
	3	50	40	

Die entsprechenden Anzahlen von Losen sind in links stehender Tabelle gegeben.

Mit welchen Wahrscheinlichkeiten gewinnt ein Los der Farbe 1, 2 oder 3 ?

3. Es sei $N = (n_{ij})$ die Matrix der Dinukleotidanzahlen in einer Sequenz und $P = (n_{ij}/n_{i.})$ die Matrix der geschätzten Übergangswahrscheinlichkeiten. Zeige, daß der Zeilenvektor $p' = ((n_{i.} + n_{.i})/2n_{..})$ für großes $n_{..}$ näherungsweise die Eigenvektorbedingung $p'P = p'$ erfüllt.

4. Eine Funktion heißt konvex, wenn

$$f(\lambda_1 x_1 + \lambda_2 x_2) \geq \lambda_1 f(x_1) + \lambda_2 f(x_2) \text{ für alle } \lambda_1 + \lambda_2 = 1, \lambda_1 \geq 0, \lambda_2 \geq 0$$

Zeigen Sie für zwei solche Funktionen $f(x)$ und $g(x)$, daß ihre Summe auch konvex ist.

Aufgabe 5: (freiwillig)
Schreiben Sie ein Programm zur Berechnung der in Aufgabe 3 definierten Größen. Schreiben Sie ein Programm zur Erzeugung von zufälligen Sequenzen nach einem Markoff Modell.

Momente von Worthäufigkeiten

X_n Anzahl des Auftretens eines Wortes in einer zufälligen Sequenz der Länge n
Diskrete Zufallsgröße

$P(X_n = i) = p_i^{(n)}$ Kann rekursiv berechnet werden

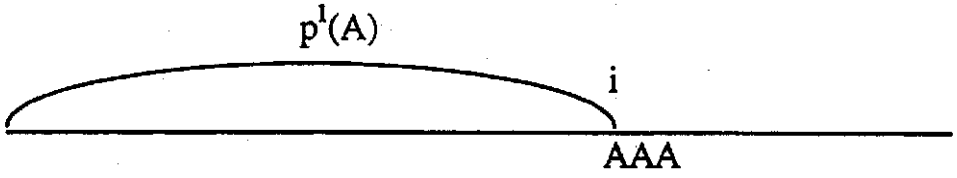
```
Wort          AAA
Sequenz       ACGTAAACGTGCTATGCAAAAGTCAATATGCAATAAA
Wortprozeß    10001230000001000123300012010001201 23
Zählprozeß    00000011111111111112333333333333333 34
```

Kleffe & Langbecker (1990)
Exact computation of pattern probabilities in random sequences generated by Markov chains. Comp. Applic. Biosci. 6, 347-353

Berechnung von $E(X_n)$

Indikatorfunktion: $I_i = \begin{cases} 1 & \text{Wort tritt in Sequenzposition i auf} \\ 0 & \text{Wort tritt in Sequenzposition i nicht auf} \end{cases}$

$$E(I_i) = P(I_i = 1) = p_i(W)$$



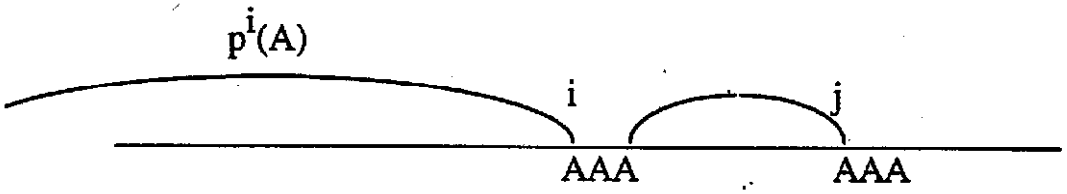
$k = l(W)$ Wortlänge

$$X_n = \sum_{i=1}^{n-k+1} I_i \quad \Rightarrow \quad E(X_n) = \sum_{i=1}^{n-k+1} p_i(W) \quad \text{häufig} \quad \underline{(n-k+1)p(W)}$$

Berechnung von $E(X_n^2)$

$$E\left(\sum_{i=1}^{n-k+1} I_i\right)^2 = \sum_{i,j=1}^{n-k+1} E(I_i I_j) = \sum_{i,j=1}^{n-k+1} P(I_i = 1 \cap I_j = 1)$$

$$\sum_{i=1}^{n-k+1} p_i(W) + 2 \sum_{i<j}^{n-k+1} P(I_i = 1 \cap I_j = 1) = E(X_n) + 2 \sum_{i<j}^{n-k+1} p_{ij}(W) p_i(W)$$



$$P(I_i = 1 \cap I_j = 1) = P(I_j = 1 \mid I_i = 1) P(I_i = 1) = p_{ij}(W) p_i(W)$$

Wortüberlappungen: AAA: AAAA AAAAA
ATA: ATATA

Es gibt höchstens k-1 überlappende Wörter W_λ , $\lambda \in \{1, ..., k-1\}$

Doppelsumme zerfällt

$$\sum_{i<j}^{n-k+1} p_{ij}(W) p_i(W) = \sum_{\lambda} \sum_{i=1}^{n-k+1-\lambda} p_i(W_\lambda) p_{\lambda}(W) + \sum_{i<j-k+1}^{n-k+1} p_{ij}(W) p_i(W)$$

$$= \sum_{\lambda} E(n(W_\lambda)) + \sum_{i<j-k+1}^{n-k+1} p_{ij}(W) p_i(W)$$

Kleffe & Borodovsky (1992) First and second moment of counts of words in random texts generated by Makov chains. Comp. Applic. Biosci. Vol.8 No. 5 433-441

Gemischte Momente: $E(X_n Y_n) = \sum_{i,j} P(I_{Xi} = 1 \cap I_{Yj} = 1)$

Lösungen

Aufgabe 2:

```
ACGTACGTACGTACGTACGTACGTACGTACGTACGTCGTACGTACGTAAA
12345678901234567890123456789012345678901234567890
```

$$I_i = \left\{ \begin{array}{ll} 1 & \text{wenn ACG an der Stelle } i \text{ auftritt} \\ 0 & \text{sonst} \end{array} \right\}$$

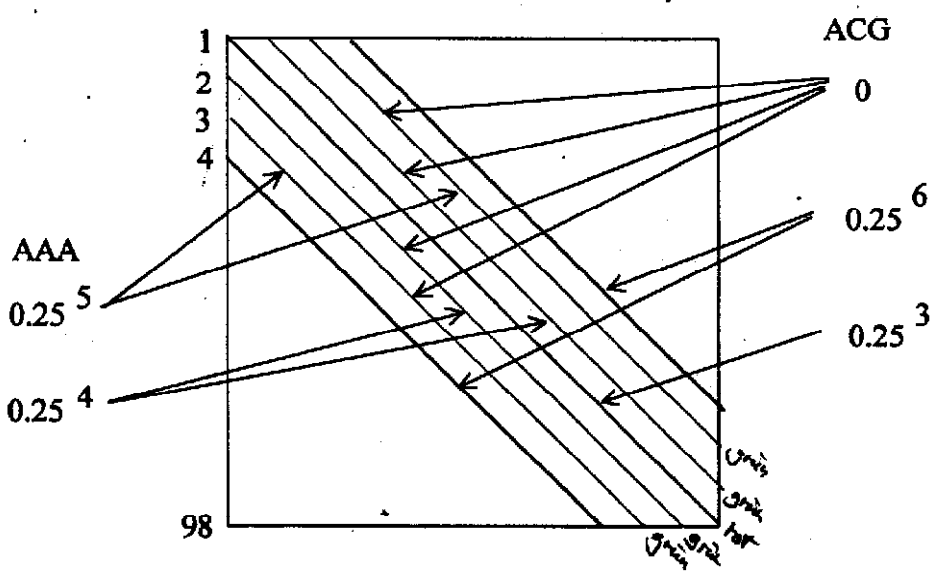
$$E(I_i) = P(I_i = 1) = \left(\frac{1}{4}\right)^3$$

$X = \sum_i I_i$ ist die die gewünschte Zufallsgröße

$$E(X) = 98 \left(\frac{1}{4}\right)^3 = 1,53125$$

$$V(X) = E(X - EX)^2 = E(X^2) - (E(X))^2$$

$$E(X^2) = E\left(\sum_i I_i\right)^2 = E\sum_{ij} I_i I_j = \sum_{ij} P(I_i = 1 \text{ and } I_j = 1) = \sum_{ij} p_{ij}$$



Sequenz: **A A C T G C A T**

WS: $p_{0A}\ p_{AA}\ p_{AC}\ p_{CT}\ p_{TG}\ p_{GC}\ p_{CA}\ p_{AT} \cdots$

$$\ln P(S) = c + \sum n_{ij} \ln p_{ij}$$

$$\ln \frac{P_H(S)}{P_A(S)} = c + \sum n_{ij} \ln \left(\frac{p_{ij}^H}{p_{ij}^A} \right) = c + \sum n_{ij} c_{ij}$$

Anz. der beobachteten Dinukleotide

Der Likelihood Quotient ist eine lineare Funktion der Worthäufigkeiten in der Sequenz. Seine Verteilung ist eindeutig bestimmt durch die Verteilung der n_{ij}.

Verteilung von $N = (n_{ij})$ bei bekanntem Anfangselement i_0:

$$\ln P(N \mid i_0) = \ln n(i_0, N) + \sum_{i,j} n_{ij} \ln p_{ij}$$

$n(i_0, N)$ ist Anzahl der Sequenzen, die mit Buchstaben i_0 beginnen und Dinukleotidhäufigkeiten $N = (n_{ij})$ besitzen.
Jede dieser Sequenzen endet mit einem eindeutig bestimmten Buchstaben j_0.

$$\ln n(i_0, N) = \ln F(i_0, N) + \sum_i \ln \left(\frac{n_{i.}!}{\prod_j n_{ij}!} \right) \qquad \text{(Satz von Whittle)}$$

$F(i_0, N)$ ist der $(j_0, i_0)te$ Kofaktor der Matrix $R = (\delta_{ij} - \frac{n_{ij}}{n_{i.}})$

Patrick Billingsley (1961)
Statistical methods in Markov chains. Ann. Math.Statist. Vol 82 12-40

Zwei Erwartungswerte:

$$E_H\left(\ln \frac{P_H(S)}{P_A(S)} \right) = m_H = c + \sum E_H(n_{ij})\, c_{ij}$$
$$E_A\left(\ln \frac{P_H(S)}{P_A(S)} \right) = m_A = c + \sum E_A(n_{ij})\, c_{ij}$$

Zwei Varianzen:

Co-variance

$$D_H\left(\ln \frac{P_H(S)}{P_A(S)} \right) = \sigma_H = \sum_{ij} \sum_{kl} c_{ij}\, cov_H(n_{ij}, n_{kl})\, c_{kl}$$
$$D_A\left(\ln \frac{P_H(S)}{P_A(S)} \right) = \sigma_A = \sum_{ij} \sum_{kl} c_{ij}\, cov_A(n_{ij}, n_{kl})\, c_{kl}$$

Binomiale Verteilung

$B(1,p)$: $X \in (0,1)$ ist ZG mit $P(X=1) = p$, $P(X=0) = 1-p$

$$E(X) = 1 * p + 0 * (1-p) = p$$

$$V(X) = EX^2 - (EX)^2 = p - p^2 = p(1-p)$$

$B(n,p)$: $Y = X_1 + X_2 + ... + X_n$, $X_i \sim B(1,p)$ unabhängig

$$E(Y) = np \qquad V(Y) = n * p * (1-p)$$

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

Induktionsbeweis:

n-1	
11001001100101011 10	1
11001001101101011 10	0

$$\frac{(n-1)!}{(k-1)!\,(n-k)!}$$

$$\frac{(n-1)!}{k!\,(n-1-k)!}$$

Summe: $\frac{n!}{k!(n-k)!}$

$$P(Y=0) = (1-p)^n \qquad P(Y=n) = p^n$$

Poisson Verteilung

$$Y \sim B(n,p) \quad \Rightarrow \quad P(Y=0) = (1-p)^n \underset{n\to\infty}{\to} 0$$

Physikalische Prozesse:

$$\begin{array}{ll} p & \text{sehr klein} \\ n & \text{sehr groß} \end{array} \quad \Rightarrow \quad p = \frac{\alpha}{n}$$

$$P(Y=0) = \left(1-\frac{\alpha}{n}\right)^n \underset{n\to\infty}{\to} e^{-\alpha}$$ Eulersche Konstante

allgemein:

$$P(Y=k) = \frac{n!}{k!(n-k)!}\left(\frac{\alpha}{n}\right)^k\left(1-\frac{\alpha}{n}\right)^{n-k} \underset{n\to\infty}{\to} \frac{\alpha^k}{k!}e^{-\alpha}$$

$$E(Y) = n\left(\frac{\alpha}{n}\right) \to \alpha$$

$$V(Y) = n\left(\frac{\alpha}{n}\right)\left(1-\frac{\alpha}{n}\right) \to \alpha$$

Eine Zufallsgröße mit dieser Grenzverteilung
heißt Poisson verteilt mit Parameter α !

Beweis: Mit $b_k(n) = \frac{n!}{k!(n-k)!}\left(\frac{\alpha}{n}\right)^k\left(1-\frac{\alpha}{n}\right)^{n-k}$ gilt

$$\frac{b_{k+1}(n)}{b_k(n)} = \left(\frac{n-k}{k+1}\right)\left(\frac{\alpha}{n}\right)\left(1-\frac{\alpha}{n}\right)^{-1} \to \frac{\alpha}{k+1}$$

Worthäufigkeiten

Wahrscheinlichkeit für das Auftreten eines Wortes an einer festen Sequenzstelle:

$$P(ATG) = p(A)p(T)p(G) = p_{ATG}$$

Frage: Mit welcher Wahrscheinlichkeit tritt es x mal auf ?

Effektive Sequenzlänge: N

Anzahl des Auftretens eines Wortes: WC

Binomialansatz:

$$P(WC = x) = \frac{N!}{x!(N-x)!} p_{ATG}^{x} (1 - p_{ATG})^{N-x}$$

In den meisten Fällen wird p_{ATG} aus einer Stichprobe geschätzt.

Poissonansatz (seltene Wörter):

$$P(WC = x) = \frac{\exp(-\alpha)}{x!} \alpha^{x}$$

Wahrscheinlichkeit für das Auftreten eines Wortes an einer festen Stelle:

$$P(ATGATG) = p(A)p(T)p(G)p(A)p(T)p(G) = p_{ATGATG}$$

Effektive Sequenzlänge: N

Poissonparameter (Intensität): $\alpha = N\, p_{ATGATG}$

Sicherheitsschranken

Sei X eine diskrete Zufallsgröße mit Werten $0, 1, 2, 3, \ldots . n$

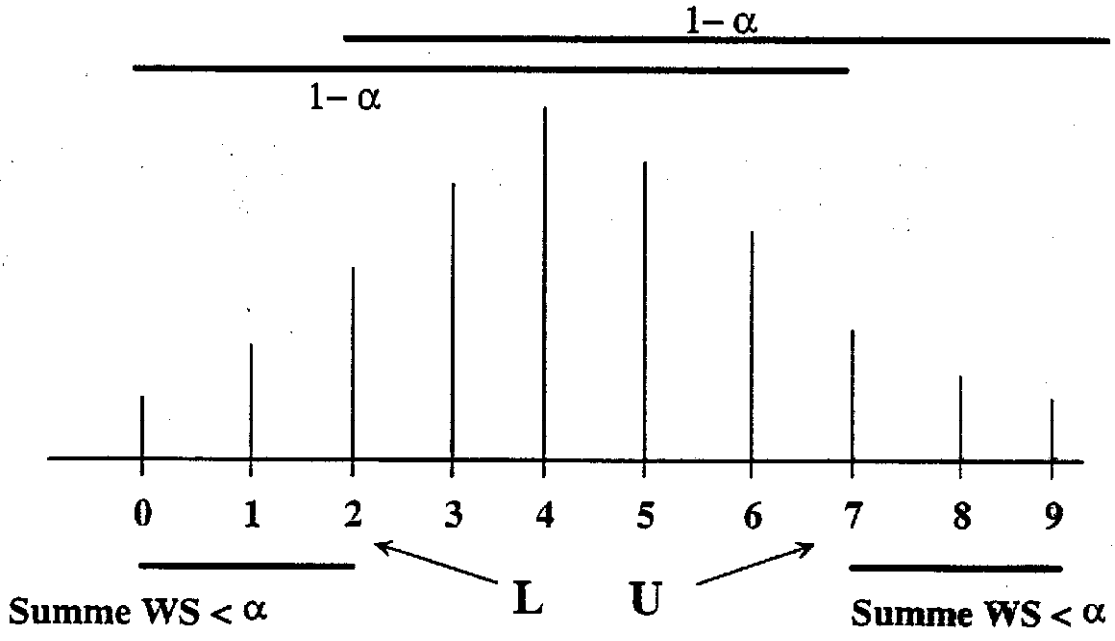
$$P(X = k) = p_k, \qquad \sum_{k=0}^{n} p_k = 1$$

Das größte L mit der Eigenschaft $\sum_{k=L}^{n} p_k > 1 - \alpha$

heißt untere Sicherheitsschranke zum Niveau α

Das kleinste U mit der Eigenschaft $\sum_{k=0}^{U} p_k > 1 - \alpha$

heißt obere Sicherheitsschranke zum Niveau α



Upper and lower 2.5% confidence limits for binomial counts as function of expectation (EX) and sample size N in comparison with the Poisson distribution. The binomial probability is EX/N.

N: Size of binomial sample

EX	100		200		300		400		500		600		700		800		900		1.000		Poisson	
1	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
2	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6
3	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8
4	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9
5	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11
6	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12
7	2	13	2	13	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14
8	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15
9	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16
10	5	17	4	17	4	17	4	18	4	18	4	18	4	18	4	18	4	18	4	18	4	18
20	12	29	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30
30	21	40	20	41	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42
40	31	51	29	52	29	53	29	53	29	53	28	53	28	53	28	53	28	54	28	54	28	54
50	40	61	38	63	38	64	37	64	37	65	37	65	37	65	37	65	37	65	37	65	37	65
60	50	70	48	74	47	75	46	75	46	76	46	76	46	76	46	76	46	76	46	76	45	77
70	61	80	57	84	56	86	55	86	55	87	55	87	55	87	55	87	55	87	55	87	54	88
80	72	89	67	95	65	96	65	97	64	97	64	98	64	98	64	98	64	98	64	98	63	99
90	84	96	76	105	75	107	74	108	73	108	73	108	73	109	73	109	73	109	73	109	72	110

Upper and lower 2.5% confidence limits for binomial counts as function of expectation (EX) and sample size N in comparison with the Poisson distribution. The binomial probability is EX/N.

N: Size of binomial sample

EX	1.000		2.000		3.000		4.000		5.000		6.000		7.000		8.000		9.000		10.000		Poisson	
1	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
2	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6
3	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8	0	8
4	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9	1	9
5	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11
6	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12
7	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14	2	14
8	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15	3	15
9	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16	4	16
10	4	18	4	18	4	18	4	18	4	18	4	18	4	18	4	18	4	18	4	18	4	18
20	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30	12	30
30	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42	20	42
40	28	54	28	54	28	54	28	54	28	54	28	54	28	54	28	54	28	54	28	54	28	54
50	37	65	37	65	37	65	37	65	37	65	37	65	37	65	37	65	37	65	37	65	37	65
60	46	76	46	76	45	76	45	77	45	77	45	77	45	77	45	77	45	77	45	77	45	77
70	55	87	54	88	54	88	54	88	54	88	54	88	54	88	54	88	54	88	54	88	54	88
80	64	98	63	99	63	99	63	99	63	99	63	99	63	99	63	99	63	99	63	99	63	99
90	73	109	72	110	72	110	72	110	72	110	72	110	72	110	72	110	72	110	72	110	72	110
100	82	120	81	121	81	121	81	121	81	121	81	121	81	121	81	121	81	121	81	121	81	121
200	176	226	174	228	174	228	173	228	173	229	173	229	173	229	173	229	173	229	173	229	173	229
300	272	330	269	333	268	334	268	334	268	334	267	335	267	335	267	335	267	335	267	335	267	335
400	370	431	365	436	364	438	363	439	363	439	363	439	362	439	362	440	362	440	362	440	361	441
500	469	532	462	539	460	541	459	542	459	543	458	543	458	544	458	544	458	544	458	544	457	545
600	570	631	560	641	557	644	556	646	555	646	555	647	554	647	554	648	554	648	554	648	552	649
700	671	729	658	743	655	747	653	748	652	749	652	750	651	751	651	751	651	751	650	751	649	753
800	775	825	757	844	753	849	751	851	750	852	749	853	748	854	748	854	747	854	747	855	745	857
900	881	919	856	945	851	950	849	953	847	955	846	956	845	956	845	957	845	957	844	957	842	960

Multinomiale Verteilung

(identical independent distributed)

$X_i, i = 1, ..., N$ i.i.d. vektorwertige diskrete Zufallsgrößen

$$X_i \in \left\{ \begin{pmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ . \\ 0 \end{pmatrix}, ..., \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 1 \end{pmatrix} \right\} = \{e_1, ..., e_m\} \subset R^m$$

$$P(X_i = e_j) = p_j \quad j = 1, ..., m \quad i = 1, ..., N$$

Definition:

$$Y = X_1 + X_2 + ... + X_N = \begin{pmatrix} n_1 \\ n_2 \\ . \\ . \\ n_m \end{pmatrix} \sim M(N, m)$$

heißt multinomial verteilt mit Parametern N und m.

Beispiel:
Zufallssequenz
der Länge N:

$$\begin{pmatrix} n_A \\ n_C \\ n_G \\ n_T \end{pmatrix} \sim M(N, 4)$$

Aufgabenblatt 4

1. Es sei 1 ein Vektor mit allen Komponenten gleich 1 und D eine Diagonalmatrix. Was bewirken die Matrizenmultiplikationen $DA, AD, 1'A, A1$ für die Zeilen und Spalten einer Matrix A?

2. Es seien X_i $i = 1, ..., n$ unabhängig und identisch verteilte Indikatorvariablen.
   Berechnen Sie die Wahrscheinlichkeit $P(\max_i X_i = 1)$.

3. Eine zufällige Sequenz S der Länge 10 wird mit einem Markoff-Prozeß der Ordnung 0 mit den Wahrscheinlichkeiten p_A, p_C, p_G, p_T erzeugt.
   Es seien n_{AA}, n_{AC} die Anzahlen der Dinukleotide AA bzw. AC in dieser Sequenz. Berechnen Sie $E(n_{AA}), E(n_{AC}), V(n_{AA}), V(n_{AC}), Cov(n_{AA}, n_{AC})$.

4. Berechnen Sie die Verteilungen von n_{AA} und n_{AC} unter den Annahmen von Aufgabe 3.

5. (freiwillig) Schreiben Sie ein Programm zur Lösung von Aufgabe 4.

Multinomiale Verteilung

X_i , $i = 1, ..., N$ i.i.d. vektorwertige diskrete Zufallsgrößen

$$X_i \in \left\{ \begin{pmatrix} 1 \\ 0 \\ . \\ . \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ . \\ 0 \end{pmatrix}, ..., \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 1 \end{pmatrix} \right\} = \{e_1, ..., e_m\} \subset R^m$$

$$P(X_i = e_j) = p_j \quad j = 1, ..., m \quad i = 1, ..., N$$

Definition:

$$Y = X_1 + X_2 + ... + X_N = \begin{pmatrix} n_1 \\ n_2 \\ . \\ . \\ n_m \end{pmatrix} \sim M(N, m)$$

heißt multinomial verteilt mit Parametern N und m.

Beispiel:
Zufallssequenz
der Länge N:

$$\begin{pmatrix} n_A \\ n_C \\ n_G \\ n_T \end{pmatrix} \sim M(N, 4)$$

Momente der multinomialen Verteilung

$$E(X_i) = \begin{pmatrix} p_1 \\ p_2 \\ . \\ . \\ p_m \end{pmatrix} = \boldsymbol{p} \quad p_j = P(X_i = e_j) \;\Rightarrow\; E(Y) = N\boldsymbol{p}$$

$$Cov(X_i) = E(X_i X_i') - E(X_i)E(X_i)' = E(X_i X_i') - \boldsymbol{p}\boldsymbol{p}'$$

$$E(X_i X_i') = p_1 e_1 e_1' + ... + p_1 e_1 e_1' = \begin{pmatrix} p_1 & & & & \\ & p_2 & & & \\ & & . & & \\ & & & . & \\ & & & & p_m \end{pmatrix}$$

$$Cov(X_i) = Diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}' \quad , \quad Cov(Y) = N(Diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}')$$

Verteilung: $P(Y = \begin{pmatrix} n_1 \\ . \\ . \\ n_m \end{pmatrix}) = \frac{N!}{n_1!,\ldots,n_m!}\, p_1^{n_1} \ldots p_m^{n_m}$

```
123332211321231232123123233211232131112213
**333****3***3**3***3**3*33****3**3******3
```

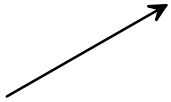
#Möglichkeiten: $\dfrac{N!}{n_3!(N-n_3!)}$

Ersetze * durch 1 oder 2 ergibt $\dfrac{(N-n_3)!}{n_1!\, n_2!}$ Möglichkeiten

$$\frac{N!}{n_3!(N-n_3!)} \quad \frac{(N-n_3)!}{n_1!\, n_2!} = \frac{N!}{n_1!\, n_2!\, n_3!}$$

Sequenzmotive

ACGTTCACGTGGTAACTATGCGTATATCA



Start Codon

Problem: ATG tritt nicht nur in der Rolle des Start Codons auf !
Können wir richtige und falsche ATG unterscheiden ?

Schlüssel: Sequenzumgebung, Verteilung der angrenzenden Basen

ACGTTCACGGCTAGCATATGCGTATATCA
CGTAATCCGTGGTAACTATGCGTATAAAC
AACGTATTCAGAAACCGATGCCCCTATCA
ACGTTCACGTGGTAACTATGCGCGTAACA

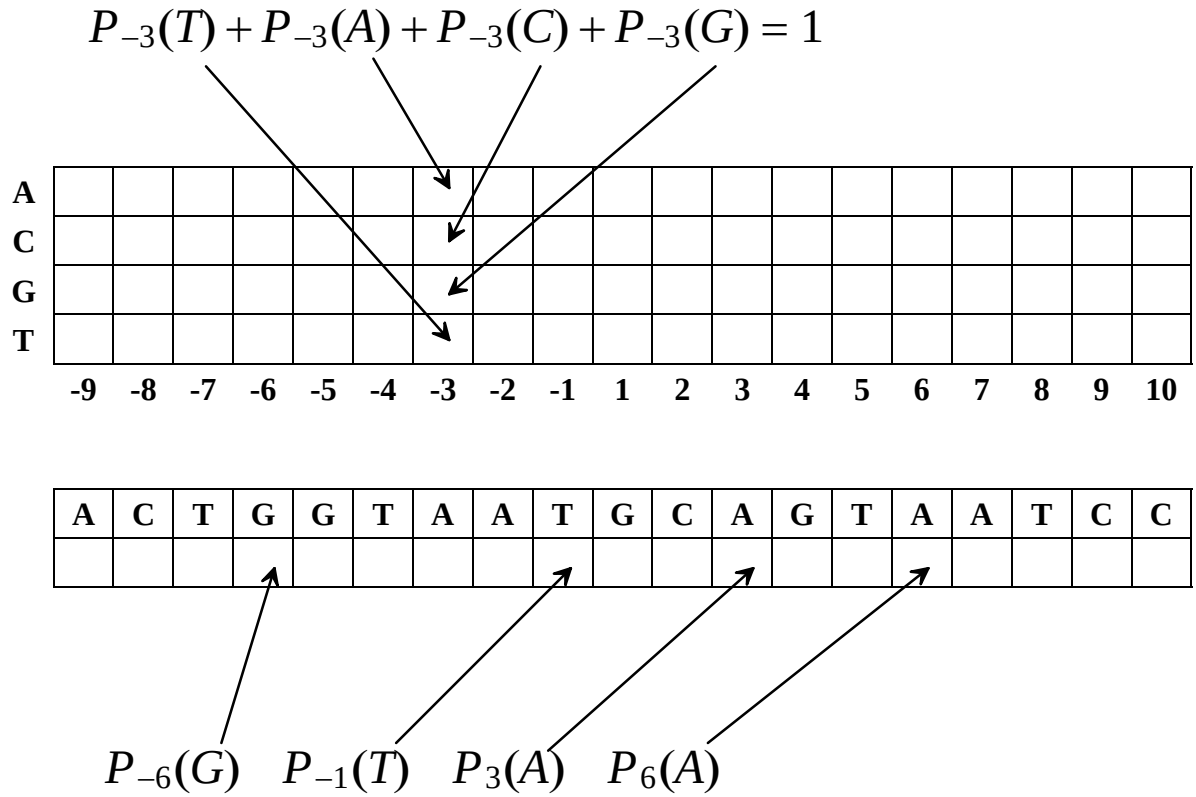
Stichprobe vom Umfang N

ACGTACCGGTACAAACTATGAAAATATCA

Positionen -1,-2,-3,.... Positionen 1,2,3,.....

In jeder Spalte können wir die Verteilung der Basen auf Abweichungen von der "normalen Basenverteilung" in einer Sequenz untersuchen.

Bewertung von Vorhersageverfahren (Entscheidungsregeln)



Beobachtungen: **Sequenzstellen**

2 Attribute: **echtes Stopkodon** **(H) Hypothese**
falsches Stopkodon **(A) Alternative**

Entscheidungsregel: **f(S) , S = Sequenzstelle**

wir entscheiden: **H wenn f(S) > C**
A wenn f(S) $\leq$ C

Stichprobenzusammensetzung:

AP (actual positive)	**Anzahl der H-Stellen**
AN (actual negative)	**Anzahl der A-Stellen**
PP	**Anzahl der Stellen mit f(S) > C (H)**
PN	**Anzahl der Stellen mit f(S) $\leq$ C (A)**

	Hyp.	Alt.	
Pos	**TP**	**FP**	**PP**
Neg	**FN**	**TN**	**PN**
	AP	**AN**	Stichproben-umfang

TP (true positve) Anzahl der Fälle, in denen echt positive Stellen als solche erkannt werden

FP (false positive) Anzahl der Fälle, in denen echt negative Stellen als positive erkannt werden

FN (false negative) Anzahl der Fälle, in denen echt positive Stellen als negative erkannt werden

TN (true negative) Anzahl der Fälle, in denen echt negative Stellen als solche erkannt werden

Sensibilität: **Sn = TP/AP**
Anteil der richtig erkannten positiven Stellen

Spezifiät: **Sp = TP/PP**
Anteil der richtig erkannten positiven Stellen
an den als positiv klassifizierten Stellen

Korrelationskoeffizient:

$$C = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}}$$

Snyder and Stormo (1993) Nucleic Acids Res. 21, 607-613

C = 1 Genau dann wenn alle Vorhersagen richtig sind
C = -1 Genau dann wenn alle Vorhersagen falsch sind
C = 0 Wenn die Vorhersagen keinen Zusammenhang mit der Wahrheit erkennen lassen (unkorrelierte Zufallsgrößen)

Modell: $X = \begin{cases} 1 \text{ unter } H \\ 0 \text{ unter } A \end{cases}$ $Y = \begin{cases} 1 \text{ Vorhersage ist } H \\ 0 \text{ Vorhersage ist } A \end{cases}$

Stichprobe:

Ereignis	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Beobachtung: X	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0
Beobachtung: Y	1	0	1	1	1	1	0	0	0	1	1	1	1	0	0

Unabhängig verteilte Paare (X_i, Y_i) von korrelierten Zufallsgrößen

Häufigkeitstabelle:

		Y	
		0	**1**
X	**0**	**TN**	**FP**
	1	**FN**	**TP**

Gewöhnlicher Stichprobenkorrelationskoeffizient:
(Pearson Product Moment Correlation Coefficient, 1900)

$$\rho = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}}$$

Unser Spezialfall:

$$\begin{aligned} \sum (X_i - \overline{X})^2 &= \sum X_i^2 - N\overline{X}^2 &&= AP - AP^2/N \\ &= AP(1 - AP/N) &&= N^{-1}AP(N - AP) \\ &= N^{-1}AP * AN \end{aligned}$$

$$\sum (Y_i - \overline{Y})^2 = N^{-1}PP * PN$$

$$\sum (X_i - \overline{X})(Y_i - \overline{Y})^2 = N^{-1}(N * TP - AP * PP)$$

Matrizenoperationen

$\mathbf{A}, \mathbf{B} \in M(n \times n)$; $\mathbf{C} \in M(n \times r)$; $u, v \in M(n \times 1)$; $\alpha \in \mathbb{R}$.

i) Addition von Matrizen gleicher Grösse:

$$\mathbf{A} + \mathbf{B} = \mathbf{D} \in M(m \times n) \text{ mit } d_{ij} = a_{ij} + b_{ij}.$$

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & & \vdots \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1n} \\ b_{21} & & \vdots \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & & \vdots \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + bmn \end{pmatrix}$$

ii) Multiplikation zweier Matrizen:

$$\mathbf{A} \cdot \mathbf{C} = \mathbf{E} \in M(m \times r) \text{ und } e_{ij} = \sum_{k=1}^{n} a_{ik} c_{kj}.$$

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k} c_{k1} & \dots & \sum_{k=1}^{n} a_{1k} c_{kn} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} a_{mk} c_{k1} & \dots & \sum_{k=1}^{n} a_{mk} c_{kn} \end{pmatrix}$$

Merke $(m \times n) \cdot (n \times r) = (m \times r)$

Im allgemeinen gilt: $\mathbf{F} \cdot \mathbf{G} \neq \mathbf{G} \cdot \mathbf{F}$!! $(\mathbf{F}, \mathbf{G} \in M(n \times n))$

iii) Multiplikation von Matrix und Skalar:

$$\alpha \cdot \mathbf{A} = \mathbf{A} \cdot \alpha = \mathbf{H}, \text{ mit } h_{ij} = \alpha \cdot a_{ij}$$

$$\alpha \cdot \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{m1} & \dots & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \alpha \cdot a_{11} & \alpha \cdot a_{12} & \dots & \alpha \cdot a_{1n} \\ \alpha \cdot a_{21} & \alpha \cdot a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \alpha \cdot a_{m1} & \dots & \dots & \alpha \cdot a_{mn} \end{pmatrix}$$

Bemerkungen:

a) Mit $\mathbf{I_n} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} \in M(n \times n)$ gilt:

$$\forall\, \mathbf{F} \in M(n \times n) : \mathbf{F} \cdot \mathbf{I_n} = \mathbf{I_n} \cdot \mathbf{F} = \mathbf{F}.$$

b) $\mathbf{F} \in M(n \times n)$ invertierbar

$$:\Leftrightarrow \exists\, \mathbf{F^{-1}} \in M(n \times n) : \mathbf{F^{-1}F} = \mathbf{FF^{-1}} = \mathbf{I_n}.$$

c) Das **Transponierte** einer Matrix $\mathbf{A}$, wird mit $\mathbf{A^T}$ oder $\mathbf{A'}$ bezeichnet:

$$\mathbf{A^T} = \widetilde{\mathbf{A}} \in M(n \times m), \text{ mit } \tilde{a_{ij}} = a_{ji}.$$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{m1} & \dots & \dots & a_{mn} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{1n} & \dots & \dots & a_{mn} \end{pmatrix}$$

- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\alpha \mathbf{A})^T = \alpha \mathbf{A}^T$
- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\mathbf{AC})^T = \mathbf{C}^T \mathbf{A}^T$
- $\mathbf{A}$ symetrisch $:\Leftrightarrow \mathbf{A}^T = \mathbf{A}$.

Man beachte:

- $\mathbf{F}v = s \in M(n \times 1), \quad s_i = \sum_{k=1}^{n} f_{ik} v_k$
- $w^T \mathbf{A} = t \in M(1 \times n), \quad t_i = \sum_{k=1}^{n} f_{ki} v_k$
- $u^T v = \beta \in \mathbb{R}$
- $uv^T = \mathbf{G} \in M(n \times n)$

Starten und Beenden von *Matlab*

Das **Starten** erfolgt in drei Schritten:

1. Einloggen an einen *Unix*-Rechner, z.B.:
   math3.math.fu-berlin.de
   euklid.math.fu-berlin.de.
2. Aufrufen einer Console.
3. *Matlab* durch Eingabe von `matlab` starten.

Sehr wichtig ist das richtige Beenden von *Matlab*.

- Niemals *Matlab* durch das Schließen des Consolefensters beenden!
- Immer mit `quit` oder CONTROL+Z das Programm offiziell beenden.
- Laufende Prozesse (Endlosschleifen!) können durch CONTROL+C abgebrochen werden.

Vorteile von *Matlab*

1. Keine Initialisierung, Kategorisierung oder Deklarierung von Variablen.
2. Variablennamen können nahezu beliebig gewählt werden.
   Achtung: Zwischen GROSS- und kleinSCHREIBUNG wird Unterschieden.
3. Kompilierungsprobleme entfallen.
4. Der Programmcode aller eingebauten Funktionen ist anzeigbar.
5. Viele Möglichkeiten graphischer Darstellungen.
6. Viele Matrizenfunktionen sind vorhanden.
7. *MATLAB* IST EINFACH!

Nach dem Start

Am *Matlab*prompt können wir nun:

- Variablen benennen/ Matrizen, Strings oder Vektoren eingeben etc.
- Funktionen aufrufen.
- Eigene Programme aufrufen.
- Den *Matlab*editor aufrufen (edit).
- Die *Matlab*hilfe aufrufen (helpdesk).

Eingabe von Matrizen, Vektoren, Skalaren und Strings

Wie geben wir folgende Objekte in *Matlab* ein?

$$5,\ \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix},\ \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix},\ \text{“ACGT”}$$

```
                    < M A T L A B >
          Copyright 1984-2000 The MathWorks, Inc.
               Version 6.0.0.88 Release 12
                       Sep 21 2000


  To get started, select "MATLAB Help" from the Help menu.

>> a=5

a =

     5

>> A=[0.5 0.5 0; 0 0.5 0.5; 0 0 1]

A =

    0.5000    0.5000         0
         0    0.5000    0.5000
         0         0    1.0000

>> v=[1/3; 1/3; 1/3]

v =

    0.3333
    0.3333
    0.3333

>> v=[1/3 1/3 1/3]'

v =

    0.3333
    0.3333
    0.3333
```

```
>> s='ACGT'

s =

ACGT

>>
```

' ist also der Operator für das Transponieren.

Phantasievolle Variablennamen helfen die Übersicht zu behalten! Z.B.:

```
> Faktor=5

Faktor =

     5

>> Startverteilung=[1/3 1/3 1/3]';
>> Uebergangsmatrix=[0.5 0.5 0;0 0.5 0.5;0 0 1];
>> nucleotids='ACGT'

nucleotids =

ACGT

>>
```

Elementare Rechenoperationen

Die elementaren Rechenoperationen sind nahezu intuitiv handhabbar:

```
>> B=[1 0 0;0 0 1;0 1 0];
>> A*B

ans =

    0.5000         0    0.5000
         0    0.5000    0.5000
         0    1.0000         0

>> C=A+B

C =

    1.5000    0.5000         0
         0    0.5000    1.5000
         0    1.0000    1.0000

>> u=A*v

u =

    0.3333
    0.3333
    0.3333

>>
```

Die Hauptschwierigkeit besteht darin, auf die richtigen Dimensionen der beteiligten Objekte zu achten!

Zugriff auf einzelne Elemente von Matrizen

Mit A(*Zeile*,*Spalte*) kann auf einzelne Elemente einer Matrix, bzw. eines Vektors zugegriffen werden:

```
>> A(2,3)

ans =

    0.5000

>>
```

: ist ein Laufindex, mit ihm werden ganze Zeilen oder Spalten angesprochen:

```
>> A(2,:)

ans =

         0    0.5000    0.5000

>>
```

Es können auch Anfangs- und Endwerte für den Laufindex angegeben werden:

```
>> A(2,1:2)

ans =

         0    0.5000

>> A(1:2,1:2)

ans =

    0.5000    0.5000
         0    0.5000

>>
```

Strings werden dabei wie Vektoren behandelt:

```
>> nucleotids(2)

ans =

C

>>
```

Elementares Programmieren

- Erstellen von *Matlab*programmen mit jedem Texteditor!
- Abspeichern im Arbeitsverzeichnis unter "*name*.m".

if-Abfragen

Wenn meine Bedingung gilt, tue die weiteren Anweisungen.
Die wichtigsten Operatoren, für die Bedingungen sind:
==,~=,<,>,<=,>=.
Beispiel:

```
if b~=0
       c=a/b
    else
    c=0
end
```

for-Schleifen

Wiederhole folgende Anweisungen so und so oft.
Beispiel:

```
for i=1:10
         s=s+i
end
```

while-Schleifen

Tue meine Anweisung, solange diese Bedingung erfüllt ist.

while-Schleifen werden solange durchlaufen, bis eine Bedingung erfüllt ist.
(Hervorragend zur Erzeugung von Endlosschleifen geeignet!!)

Beispiel:

```
e = 1;
while (1+e) > 1
    e = e/5;
end
e = e*2
```

Spezielle Matrizen

Nützliche Funktionen zur Erzeugung von Matrizen:

rand, eye und zeros:

eye(n,m) erzeugt eine $n \times m$ Matrix, die auf der Diagonalen Einsen als Eintrag hat, ansonsten nur Nullen.

rand(n,m) füllt eine $n \times m$ Matrix mit zufälligen Einträgen zwischen 0 und 1. Beachte `x=rand` erzeugt eine einzelne Zufallszahl!

zeros(n,m) erzeugt eine Matrix, die nur Nullen enthällt.

Beispiele:

```
>> eye(3)

ans =

     1     0     0
     0     1     0
     0     0     1

>> eye(2,5)

ans =

     1     0     0     0     0
     0     1     0     0     0

>> eye(size(A))

ans =

     1     0     0
     0     1     0
     0     0     1
```

```
>> rand(4,2)

ans =

    0.3941    0.1122
    0.5030    0.4433
    0.7220    0.4668
    0.3062    0.0147

>> zeros(3)

ans =

     0     0     0
     0     0     0
     0     0     0

>>
```

Programmbeispiel

Die zufällige Erzeugung einer Hundertersequenz aus Nucleotiden:

```
laenge=100;
nukleotide='ACGT';
Sequenz='';
for i=1:laenge
        r=rand;
                    Sequenz(i)=nukleotide(ceil(4*r));
                    i=i+1;
end
Sequenz
```

Übungen I

Gegeben seien die folgenden Objekte:

$$\mathbf{A} = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 0.9 & 0.1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 1 & 0 \end{pmatrix}, v = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}, u = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} a = 7$$

1. Berechne per Hand und kontrolliere über *Matlab* folgende Ausdrücke:

$$\mathbf{B}u, \quad \mathbf{AB}, \quad \mathbf{BA}, \quad u^T\mathbf{B}, \quad u^Tv, \quad uv^T, \quad a\mathbf{B}a\mathbf{A}$$

2. Setze in $\mathbf{B}$ b_{11} auf 1 und b_{22} auf 2!
   Bringe *Matlab* dazu, die 2. Zeile von $\mathbf{B}$ mit der 3. Spalte von $\mathbf{A}$ zu multiplizieren (ohne diese als neue Vektoren einzugeben!).
   $(b_{21}b_{22}b_{23}) \cdot \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix}$

3. Ist $\mathbf{A}$ eine primitive Matrix ($:\Leftrightarrow \exists n \in \mathbb{N} : \mathbf{A}^n > 0$, d.h. jeder Eintrag von $\mathbf{A}^n$ ist grösser 0)?
   Wenn ja, kannst Du ungefähr die stationäre Verteilung angeben (Erinnerung: was war mit $\mathbf{A}^\infty$?)?

4. Setze a_{44} auf 1, was ist zu beobachten?

5. Was bewirkt `sum(A)`?

6. Wie kann ich die Zeilensummen von $\mathbf{A}$ berechnen lassen?

7. Schaue in der Hilfe von Matlab nach, was `trace(A)` bewirkt!

8. Gebe `spy(A)`, `surf(A)` und `surf(rand(30))` und `surf(eye(30))` ein!

9. Starte den Editor mit `edit`.
   Schreibe ein kurzes Programm, welches erst die Matrizen $\mathbf{A}$ und $\mathbf{B}$ definiert, und dann $\mathbf{A} + \mathbf{B}$ rechnet!
   Speichere es ab (es wird unter *Name*.m abgespeichert) und führe es am Matlabprompt aus (`name`)!
   So einfach ist Programmieren unter *Matlab*!!

Das Zusammensetzen von Matrizen

In *Matlab* ist es ganz einfach Matrizen zu kombinieren, so seien z.B. folgende Matrizen gegeben:

$$\mathbf{A} = \begin{pmatrix} 5 & 7 & 8 \\ 3 & 4 & 3 \\ 2 & 1 & 0 \end{pmatrix}; \mathbf{M} = \begin{pmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \ (=\mathsf{eye}(3,2)) \ ; \ \mathbf{E} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \ (=\mathsf{ones}(3,1))$$

Diese Matrizen lassen sich unter *Matlab* z.B. so zusammensetzen:

```
>> A=[5 7 8;3 4 3;2 1 0];
M=[-1 -1 -1;-1 -1 -1;-1 -1 -1];
Z=eye(3,2);
E=ones(3,1);
C=[A,M;Z,E,rand(3);[6 5 4 3 2 1]]

C =

    5.0000    7.0000    8.0000   -1.0000   -1.0000   -1.0000
    3.0000    4.0000    3.0000   -1.0000   -1.0000   -1.0000
    2.0000    1.0000         0   -1.0000   -1.0000   -1.0000
    1.0000         0    1.0000    0.4860    0.4565    0.4447
         0    1.0000    1.0000    0.8913    0.0185    0.6154
         0         0    1.0000    0.7621    0.8214    0.7919
    6.0000    5.0000    4.0000    3.0000    2.0000    1.0000
```

Es lassen sich auch gezielte Bereiche einer Matrix überschreiben:

```
>> C(2:4,4:6)=A

C =

    5.0000    7.0000    8.0000   -1.0000   -1.0000   -1.0000
    3.0000    4.0000    3.0000    5.0000    7.0000    8.0000
    2.0000    1.0000         0    3.0000    4.0000    3.0000
    1.0000         0    1.0000    2.0000    1.0000         0
         0    1.0000    1.0000    0.8913    0.0185    0.6154
         0         0    1.0000    0.7621    0.8214    0.7919
    6.0000    5.0000    4.0000    3.0000    2.0000    1.0000
```

Beispiel:
Programm zur Erzeugung einer Übergangsmatrix mit zwei Clustern.

```
% Erzeugen meiner Matrizen, die zusammengefuegt werden
Z=zeros(3);
R1=rand(3);
R2=rand(3);

% Die Zeilensummen werden auf 1 normiert
   s1=sum(R1');
   s2=sum(R2');
  S1=diag(1./s1);
  S2=diag(1./s2);
normR1=S1*R1;
normR2=S2*R2;

% Die Matrixzusammengesetzt und eine Bruecke zwischen
% den Bloecken gesetzt(durch austauschen von Elementen)
Uebergangsmatrix=[normR1, Z; Z, normR2];
   a=Uebergangsmatrix(3,3); b=Uebergangsmatrix(4,4);
Uebergangsmatrix(3,3)=0; Uebergangsmatrix(4,4)=0;
Uebergangsmatrix(3,4)=a; Uebergangsmatrix(4,3)=b
```

Diese Programm liefert z.B. folgendes Ergebnis:

```
Uebergangsmatrix =

    0.5020    0.2568    0.2412         0         0         0
    0.2026    0.7812    0.0162         0         0         0
    0.2771    0.3479         0    0.3750         0         0
         0         0    0.2509         0    0.5201    0.2289
         0         0         0    0.2689    0.3225    0.4087
         0         0         0    0.4201    0.0935    0.4864

>> sum(Uebergangsmatrix')

ans =

     1     1     1     1     1     1

>>
```

Dateien unter *Matlab*

Wir unterscheiden 3 Datei-Typen:

- Daten-Dateien
- Script-Dateien
- Funktions-Dateien

i) Daten-Dateien

Abspeichern von Sessions:

save *Name*
Z.B.:

```
save Probesitzung
```

Die Session wird unter Probesitzung.mat gespeichert.
Geladen wird sie wieder mit

```
load Probesitzung
```

Abspeichern von Variablen:

```
save Name Variable1 Variable2....
load Name
```

Zum Beispiel: save interessanteMatrizen A B a

Einlesen von Matrizen

```
D=load('matrix.dat')
```

matrix.dat ist eine Datei der Form:

```
1 2 3
4 5 6
7 8 9
```

ii) Script-Dateien

Speichern eine Liste von Anweisungen unter *name.m* ab.
Ein Aufruf mit name führt zur Ausführung.

iii) Funktions-Dateien

Kennzeichnend ist die erste Zeile:

function *Ausgabevariable/vektor* =NAME(*Eingabeparameter*)
‘

Beispiele:
function p=Wkeit(Sequenz)
function [a,b]=auswertung(c,d,e)

Aufruf der Funktion:
auswertung(5,7,8)
f=auswertung(5,7,8)
[f,g]=auswertung(5,7,8)

Beispiele für Funktions-Dateien:

1.) Nullstellen eines quadratischen Polynoms

```
function [n1,n2] = rnPol2(a,b,c)

% Ausgabe von der Nullstellen eines quadratischen Polynoms
% der Form ax^2+bx+c.

if a==0
    error('Es handelt sich um kein quadratisches Polynom!')
end

n1=-b/(2*a)+((b/(2*a))^(2)-c/a)^(0.5)
n2=-b/(2*a)-((b/(2*a))^(2)-c/a)^(0.5)
```

2.) Rekursive Funktionen - die Fibonacci-Zahlen:

```
function fn = fibonacci(n)

% Bei Eingabe einer natuerlichen Zahl n, wird die entsprechende
% Fibonacci-Zahl ausgegeben.

if n==0
   fn=1;
elseif n==1
   fn=1;
else
   fn=fibonacci(n-1)+fibonacci(n-2);
```

3. Wahrscheinlichkeit einer (kurzen) Sequenz.

```
function p=Sequenz(s)

% fuer eine ACGT-Sequenz wird die Wahrscheinlichkeit ausgerechnet.
% Dabei werden die AnfangsWkeiten und UebergangsWkeiten
% extern herbeigeholt.
% Zugrunde liegt ein Markov-Prozess 1.Ordnung

% Laden der Dateien in denen die gueltigen Buchstaben, Uebergangswahr-
% scheinlichkeiten und Anfangsverteilung abgelegt sind.
Buchstaben=load('Buchstaben.dat')
Uebergangsmatrix=load('ACGTMatrix.dat');
Anfangswahrscheinlichkeiten=load('ACGTVektor.dat');

% Laenge der Sequenz
l=size(s);
% Mit diesen Vektor greifen wir auf die Uebergangsmatrix zu
z=zeros(size(Buchstaben));
% Laenge der for-Schleifen
c=size(z);
x=0

% p wird entsprechend des ersten Buchstaben mit der
%Anfangswahrscheinlichkeit belegt.
if l==0
    error('In s ist keine Sequenz abgelegt!.')
end

a=s(1);
for i=1:c
```

```
   if Buchstaben(i)==a
      p=Anfangswahrscheinlichkeiten(i);
      x=i;
   end
end
if x==0
   error('Unbekannter Buchstabe in der Sequenz')
end

% Nun wird die Anfangswahrscheinlichkeit entsprechend den
% nachfolgenden Buchstaben multipliziert
for j=2:l
   v=z;u=z;u(x)=1; % Damit greifen wir auf die UeMatrix zu.
   a=s(j);
   x=0;
     for i=1:c
        if Buchstaben(i)==a
           v(i)=1;
           x=i;
        end
     end
   if x==0
      error('Unbekannter Buchstabe in der Sequenz!')
   end
% Hier wird auf das Element ij der UeMatrix zugegriffen
 p=p+u'*Uebergangsmatrix*v;
end
```

Übungen II

Gegeben seien:

$$\mathbf{A} = \begin{pmatrix} 0.3 & 0 & 0.7 & 0 \\ 0 & 0.3 & 0 & 0.7 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}; \mathbf{B} = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.4 & 0 & 0 & 0.6 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix}; \mathbf{C} = \begin{pmatrix} 0.2 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.2 & 0.8 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{pmatrix}$$

1. Welche dieser Matrizen ist Primitiv?
   - Bestimme näherungsweise die stationären Verteilungen der primitiven Matrizen!
   - Gib eine Interpretation für die stationäre Verteilung einer Übergangsmatrix.
   - (Zusatz: Bisher hatten wir Primitivität nur als formale Definition, fällt Dir eine Interpretation ein, d.h. was kann man über einen Markovprozess sagen dessen Übergangsmatrix primitiv ist?)
2. Berechne für eine der oben herausgefundenen stationären Verteilungen (π): $\pi^T\mathbf{B}$, $\mathbf{B}\pi$, $\pi^T\mathbf{C}$, $\mathbf{C}\pi$, $\pi^T\mathbf{C}\pi$, $\pi^T\mathbf{B}\pi$.
3. Schreibe ein kurzes Programm, welches:
   - auf zwei Matrizen $\mathbf{A}$ und $\mathbf{B}$ zugreift (Annahme: diese sind definiert und nicht leer!).
   - diese bei gleicher Grösse addiert.
   - sonst eine mit Nullen ergänzt, so dass sie dasselbe Format haben, dann addiert.

Übungen II

Gegeben seien:

$$\mathbf{A} = \begin{pmatrix} 0.3 & 0 & 0.7 & 0 \\ 0 & 0.3 & 0 & 0.7 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}; \mathbf{B} = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.4 & 0 & 0 & 0.6 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix}; \mathbf{C} = \begin{pmatrix} 0.2 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.2 & 0.8 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{pmatrix}$$

1. Welche dieser Matrizen ist Primitiv?
   - Bestimme näherungsweise die stationären Verteilungen der primitiven Matrizen!
   - Gib eine Interpretation für die stationäre Verteilung einer Übergangsmatrix.
   - (Zusatz: Bisher hatten wir Primitivität nur als formale Definition, fällt Dir eine Interpretation ein, d.h. was kann man über einen Markovprozess sagen dessen Übergangsmatrix primitiv ist?)
2. Berechne für eine der oben herausgefundenen stationären Verteilungen (π): $\pi^T\mathbf{B}$, $\mathbf{B}\pi$, $\pi^T\mathbf{C}$, $\mathbf{C}\pi$, $\pi^T\mathbf{C}\pi$, $\pi^T\mathbf{B}\pi$.
3. Schreibe ein kurzes Programm, welches:
   - auf zwei Matrizen $\mathbf{A}$ und $\mathbf{B}$ zugreift (Annahme: diese sind definiert und nicht leer!).
   - diese bei gleicher Grösse addiert.
   - sonst eine mit Nullen ergänzt, so dass sie dasselbe Format haben, dann addiert.

Satz von Taylor

$f(x)$ habe stetige Ableitungen der Ordnungen 1,...,n+1 im Intervall $[a,b]$. Dann existiert ein $c \in (a,b)$ mit

$$f(b) = f(a) + \sum_{i=1}^{n} \frac{f^{(i)}(a)}{i!}(b-a)^i + \frac{f^{(n+1)}(c)}{(n+1)!}(b-a)^{n+1}$$

$f^{(i)}(a)$ ist die i-te Ableitung von $f(x)$ an der Stelle a. Die Stelle c hängt von a und b ab.

Beweis: $$P_n(x) = f(a) + \sum_{i=1}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

heißt Taylor-Polynom der Funktion $f(x)$ der Ordnung n und an der Stelle a. Es gilt

$$P_n(a) = f(a),\ P_n^{(i)}(a) = f^{(i)}(a),\ i = 1,...,n$$

Das Gleiche gilt für die Funktion

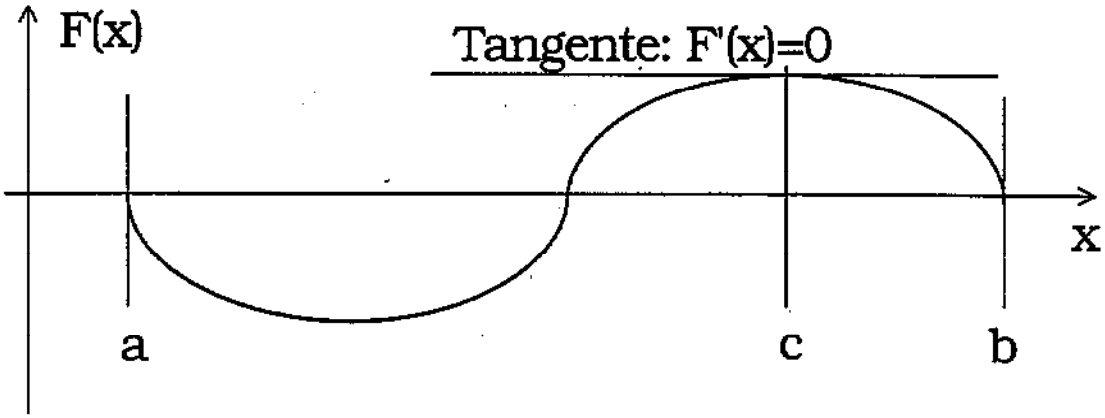
$$Q_n(x) = P_n(x) + K(x-a)^{n+1}$$

Die Konstante K kann nun so bestimmt werden, daß zusätzlich $Q_n(b) = f(b)$ gilt. Das Problem ist die Berechnung von K, ohne $f(b)$ zu kennen.

Hier hilft der Satz von Rolle!

$$F(x) = f(x) - Q_n(x)$$

$F(a) = 0,\ F(b) = 0,\ F(x)$ hat n+1 stetige Ableitungen.



$$
\begin{array}{llll}
F(a)=0, \quad F(b)=0 & \Rightarrow & F'(c_1)=0, & c_1 \in (a,b) \\
F'(a)=0, \quad F'(c_1)=0 & \Rightarrow & F''(c_2)=0, & c_2 \in (a,c_1) \\
F''(a)=0, \quad F''(c_2)=0 & \Rightarrow & F'''(c_3)=0, & c_3 \in (a,c_2) \\
\cdot \qquad\qquad \cdot & \Rightarrow & \cdot & \cdot \\
\cdot \qquad\qquad \cdot & \Rightarrow & \cdot & \cdot \\
\cdot \qquad\qquad \cdot & \Rightarrow & \cdot & \cdot \\
F^{(n)}(a)=0 \quad F^{(n)}(c_n)=0 & \Rightarrow & F^{(n+1)}(c_{n+1})=0, & c_{n+1} \in (a,c_n)
\end{array}
$$

Aus der Gleichung $F^{(n+1)}(c)=0$ wird die Konstante K bestimmt!

$$F^{(n+1)}(x)=f^{(n+1)}(x)-K(n+1)!$$

$$f^{(n+1)}(c)-K(n+1)!=0 \;\Rightarrow\; K=\frac{f^{(n+1)}(c)}{(n+1)!}$$

$$Q_n(x)=P_n(x)+\frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$$

Für $x=b$ gilt $Q_n(b)=f(b)$ und damit die Behauptung.

Restglied: $R = f^{(n+1)}(c)\,\frac{(x-a)^{(n+1)}}{(n+1)!},\ c \in (a,x)$

$f^{(n+1)}(c)$: ist beschränkt

$\frac{(x-a)^{(n+1)}}{(n+1)!}$: strebt gegen 0 für festes x und a

$f(x) = f(a) + \sum_{i=1}^{\infty} \frac{f^{(i)}(a)}{i!}(x-a)^i$ heißt Taylor-Entwicklung

Beispiel 1:

$$f(x) = e^x \Rightarrow f^{(n+1)}(x) = e^x \Rightarrow R \to 0 \quad \forall x$$

Taylor-Entwicklung an der Stelle $x = 0$. (Maclaurin Reihe)

$$e^x = 1 + \sum_{i=1}^{\infty} \frac{x^i}{i!} = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Beispiel 2:

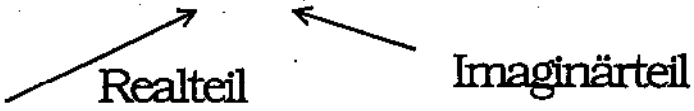
$$\begin{aligned} f(x) &= \sin(x) & f'(x) &= \cos(x) \\ f''(x) &= -\sin(x) & f^{(3)}(x) &= -\cos(x) \\ f^{(4)}(x) &= \sin(x) & f^{(5)}(x) &= \cos(x) \\ &\vdots & &\vdots \\ f^{(2k)}(x) &= (-1)^k \sin(x) & f^{(2k+1)}(x) &= (-1)^k \cos(x) \\ f^{(2k)}(0) &= 0 & f^{(2k+1)}(0) &= (-1)^k \end{aligned}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$$

Eulers Formel

Komplexe Zahlen: $z = a + ib,\ i^2 = -1$



$$z_1 + z_2 = a_1 + a_2 + i(b_1 + b_2)$$
$$z_1 z_2 = (a_1 + ib_1)(a_2 + ib_2) = a_1 a_2 - b_1 b_2 + i(a_1 b_2 + b_1 a_2)$$

Definition: $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$

$$e^{ix} = \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} = 1 + \frac{ix}{1!} + \frac{i^2x^2}{2!} + \frac{i^3x^3}{3!} + \frac{i^4x^4}{4!} + \ldots.$$

$$= 1 + \frac{ix}{1!} - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \ldots.$$

$$= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots\ldots\right) + i\left(\frac{x}{1!} - \frac{x^3}{3!} + \ldots.\right)$$

Eulers Formel $e^{ix} = \cos(x) + i\sin(x)$

$e^{i\pi} = \cos(\pi) + i\sin(\pi) = 1$

Stirlings Formel $n! \approx n^n e^{-n}\sqrt{2\pi n}$

$$\lim_{n\to\infty} \frac{(x-a)^n}{n!} = 0\ \forall x, a$$

k-dimensionale Taylorsche Formel

M sei offenen Menge des R^k, $f(x_1,...x_k)$ n+1 mal stetig differenzierbar auf M und a und b zwei Punkte in M, deren Verbindungsstrecke ganz in M liegt. Dann existiert ein Punkt c auf dieser Verbindungsstrecke mit

$$\begin{aligned} f(b) = f(a) &+ \frac{1}{1!}\sum_{i_1} D_{i_1} f(a)(b_{i_1} - a_{i_1}) \\ &+ \frac{1}{2!}\sum_{i_1 i_2} D_{i_1 i_2} f(a)(b_{i_1} - a_{i_1})(b_{i_2} - a_{i_2}) \\ &\vdots \\ &+ \frac{1}{n!}\sum_{i_1 \ldots i_n} D_{i_1 \ldots i_n} f(a)(b_{i_1} - a_{i_1})\ldots(b_{i_n} - a_{i_n}) \\ &+ \frac{1}{(n+1)!}\sum_{i_1 \ldots i_{n+1}} D_{i_1 \ldots i_{n+1}} f(c)(b_{i_1} - a_{i_1})\ldots(b_{i_n} - a_{i_{n+1}}) \end{aligned}$$

Wichtigste Anwendung: Näherungsformel für Varianzen

Taylor Reihe 1. Ordnung:

$$f(x) \approx f(a) + \sum_{i_1} D_{i_1} f(a)(x_{i_1} - a_{i_1})$$

$$f(x) - Ef(x) \approx f(a) - Ef(x) + \sum_{i_1} D_{i_1} f(a)(x_{i_1} - a_{i_1})$$

Gilt für $a = Ex$ ungefähr $f(a) = Ef(x)$ so gilt

$$\begin{aligned} E(f(x) - Ef(x))^2 &\approx \sum_{i_1 i_2} D_{i_1} f(a) D_{i_2} f(a)\, E(x_{i_1} - a_{i_1})(x_{i_2} - a_{i_2}) \\ &\approx \sum_{i_1 i_2} D_{i_1} f(a) D_{i_2} f(a)\, cov(x_{i_1}, x_{i_2}) \end{aligned}$$

Aufgabenblatt 5

Aufgabe 1: Berechne den Wert für e durch Reihenentwicklung der Funktion e^x mit einer Genauigkeit von $\pm 10^{-4}$ (+ Begründung)

Aufgabe 2: Entwickle die Funktion $\ln(x)$ an der Stelle $x = 1$ in eine Taylor-Reihe.

Aufgabe 3: Finde die Maclaurin-Reihe für $\frac{1}{1-x}$ und $\frac{1}{1+x}$

Aufgabe 4: Verwende die Definition $e^{ix} = \cos(x) + i\sin(x)$ und beweise $e^{ix}e^{iy} = e^{i(x+y)}$, $e^{-ix} = \frac{1}{e^{ix}}$ unter Verwendung der Additionstheoreme. (für cos + sin)

Aufgabe 5: Beweise

$$\ln(\sum_i p_i) = \ln\left(\sum_i \exp(C + \ln(p_i))\right) - C$$

Maclaurin Reihe von $E(e^{itX})$

Ableitungen:

$$\begin{aligned} f(t) &= E(e^{itX}) & f(0) &= 1 \\ f'(t) &= iE(Xe^{itX}) & f'(0) &= iE(X) \\ &\vdots & &\vdots \\ f^{(k)}(t) &= i^k E(X^k e^{itX}) & f^{(k)}(t) &= i^k E(X^k) \end{aligned}$$

Maclaurin Reihe mit Restglied:

$$f(t) = 1 + \sum_{k=1}^{n} \frac{i^k E(X^k) t^k}{k!} + \frac{i^{n+1} E(X^{n+1} e^{icX}) t^{n+1}}{(n+1)!}, c \in (0, t)$$

Restgliedabschätzung: $\left| i^{n+1} E(X^{n+1} e^{icX}) \right| \leq E\left| X^{n+1} \right|$

Spezialfall n=2:

$$f(t) = 1 + iE(X)t - \frac{1}{2}E(X^2)t^2 - \frac{1}{6}iE(X^3 e^{icX})t^3$$

$$\left| iE(X^3 e^{icX}) \right| \leq E\left| X^3 \right|$$

Für $E(X) = 0$ und $var(X) = 1$ folgt

$$f(t) = 1 - \frac{1}{2}t^2 - \frac{1}{6}iE(X^3 e^{icX})t^3 = 1 - \frac{t^2}{2}(1 - \varepsilon(t))$$

$$\lim_{t \to 0} \varepsilon(t) = 0$$

Zentraler Grenzwertsatz

(Normalverteilung)

Es seien $X_i \sim (0,1), i = 1,...,n$ unabhängig und identisch verteilte ZG mit $E(X_i) = 0,\ var(X_i) = 1$. Dann gilt

$$\lim_{n\to\infty} P\left(a < \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx$$

Beweis:

Die charakteristische Funktion von $\frac{1}{\sqrt{n}}X_i$ sei $f(\frac{t}{\sqrt{n}})$.

Dann ist $\left(f(\frac{t}{\sqrt{n}})\right)^n$ die c.F. von $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$

Mit

$$\left(f(\frac{t}{\sqrt{n}})\right)^n = \left(1 - \frac{t^2}{2n}(1 - \varepsilon(\frac{t}{\sqrt{n}}))\right)^n \to e^{-\frac{t^2}{2}}$$

erhalten wir die c.F. der ZG $\lim_{n\to\infty} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$.

Gegenrechnung: $t = i\theta \Rightarrow t^2 = -\theta^2$

$$\begin{aligned}
\phi(t) = \phi(i\theta) &= E(e^{-\theta X}) \\
&= \frac{1}{\sqrt{2\pi}}\int e^{-\theta x} e^{-\frac{x^2}{2}}dx \\
&= e^{\frac{\theta^2}{2}}\frac{1}{\sqrt{2\pi}}\int e^{-\frac{x^2+2\theta x+\theta^2}{2}}dx \\
&= e^{\frac{\theta^2}{2}}\frac{1}{\sqrt{2\pi}}\int e^{-\frac{(x+\theta)^2}{2}}dx \\
&= e^{\frac{\theta^2}{2}} = e^{-\frac{t^2}{2}}
\end{aligned}$$

Beispiel: Binomial →Normal

$Y_i:\ E(Y_i)=p,\ var(Y_i)=pq,\ q=1-p,\ i=1,...,n$

$S_n=Y_1+...+Y_n$ hat Binomialverteilung

Betrachte $X_i=\frac{Y_i-p}{\sqrt{pq}}:\ E(X_i)=0,\ var(X_i)=1$

$$\Rightarrow \quad \lim_{n\to\infty} P\left(a<\frac{S_n-np}{\sqrt{npq}}\le b\right)=\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx$$

Anwendung:

$$P(np+a\sqrt{npq}<S_n\le np+b\sqrt{npq})\approx\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx$$

Setze $k-1=np+a\sqrt{npq},\ k=np+b\sqrt{npq}$

$$\Rightarrow P(S_n=k)\approx\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx$$

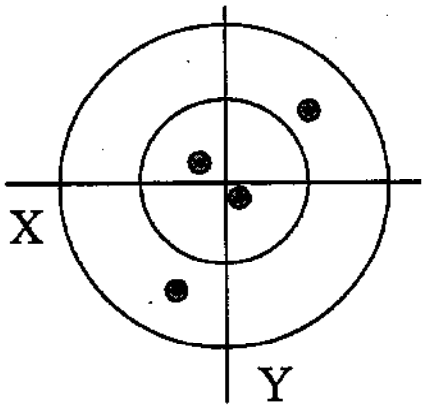
Näherungsformel für

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

Stirlings Formel ergibt

$$P(S_n=k)\approx\frac{1}{\sqrt{2\pi}}\frac{n^{n+\frac{1}{2}}}{k^{k+\frac{1}{2}}(n-k)^{n-k+\frac{1}{2}}}p^kq^{n-k}$$

Normalverteilung



Herschels Hypothese

(a) Randverteilungen $P(x)$ und $Q(y)$ haben stetige Dichten $p(x)$ und $q(y)$

(b) Die gemeinsame Dichtefunktion $f(x,y)$ hängt nur von $r=\sqrt{x^2+y^2}$ ab.

(c) X und Y sind unabhängig verteilt.

$$\Rightarrow f(x,y)=f(x)f(y)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}$$

Beweis: $$f(x,y)=p(x)q(y)=s(r),\ r^2=x^2+y^2$$

$$x=0, y\geq 0 \Rightarrow q(y)=\frac{s(y)}{p(0)} \Rightarrow f(x,y)=\frac{p(x)s(y)}{p(0)}=s(r)$$

$$y=0, x\geq 0 \Rightarrow \frac{p(x)s(0)}{p(0)}=s(x)\Rightarrow p(x)=\frac{s(x)p(0)}{s(0)}$$

$$f(x,y)=\frac{s(x)p(0)}{s(0)}\frac{s(y)}{p(0)}=\frac{s(x)s(y)}{s(0)}=s(r)$$

Mit $g(x)=s(x)/s(0)$ und $f(x)=\ln(g(x))$ folgt:

$$g(x)g(y)=g(r),\ r^2=x^2+y^2$$
$$f(x)+f(y)=f(r),\ r^2=x^2+y^2$$

Wiederholtes Anwenden dieser Gleichung liefert

$$f(r)=f(x_1)+\ldots+f(x_k),\ r^2=x_1^2+\ldots+x_k^2$$

$$k=n^2, x_i=x\ \forall k \Rightarrow f(nx)=n^2f(x)$$
$$x=1 \Rightarrow f(n)=n^2f(1)$$

Wenn m eine ganze Zahl ist folgt mit $x = \frac{m}{n}$

$$n^2 f(\frac{m}{n}) = f(m) = m^2 f(1) \Rightarrow f(\frac{m}{n}) = f(1)\left(\frac{m}{n}\right)^2$$

Also gilt $f(x) = cx^2$ mit $c = f(1)$ für alle rationalen Zahlen. Aus der Stetigkeit von $f(x)$ folgt diese Eigenschaft für alle x.

$$f(x) = \ln g(x) = \ln \frac{s(x)}{s(0)} = \ln \frac{p(x)}{p(0)} = cx^2 \Rightarrow p(x) = p(0)e^{cx^2}$$

Für eine Dichtefunktion muß c negativ sein, wir setzen $c = -\frac{1}{2\sigma^2}$. Dann liefert Integration

$$p(0)\int e^{-\frac{x^2}{2\sigma^2}} dx = p(0)\sigma \int e^{-\frac{y^2}{2}} dy = p(0)\sigma\sqrt{2\pi} = 1$$

$$\Rightarrow p(0) = \frac{1}{\sigma\sqrt{2\pi}} \Rightarrow p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

Weiter folgt $p(x) = q(x)$ und wegen der Unabhängigkeit von X und Y

$$f(x,y) = p(x)p(y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Verteilung heißt 2-dimensionale Normalverteilung mit Erwartungswert 0 und Kovarianzematrix $\sigma^2 I$

Determinante

Für eine *nxn* Matrix *A* ist die Determinate definiert durch

$$|A| = \sum_{\Phi} (-1)^{\phi(j_1,\ldots,j_n)} \prod_{i=1}^{n} a_{ij_i}$$

Anz. der Vertauschungen der Cols. Jede Vertauschung: $\phi = -1$

Φ ist die Menge aller Permutationen $j_1, \ldots, j_n$ der Indices $1, \ldots, n$ und $\phi(j_1, \ldots, j_n)$ ist die Anzahl der Vertauschungen, die zur betrachteten Permutation führen.

Beispiel: $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$

Spaltenpermutation: $1\ 2 \Rightarrow a_{11}a_{22}$

$2\ 1 \Rightarrow a_{12}a_{21}, \phi(2,1) = 1$

Beispiel: $\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$

$$= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{11}a_{32}a_{23}$$

Permutation	ϕ	Produkt
1 2 3	0	$a_{11}a_{22}a_{33}$
1 3 2	1	$a_{11}a_{23}a_{32}$
2 1 3	1	$a_{12}a_{21}a_{33}$
2 3 1	2	$a_{12}a_{23}a_{31}$
3 1 2	2	$a_{13}a_{21}a_{32}$
3 2 1	1	$a_{13}a_{22}a_{31}$

Auswär
Auswär

Eigenschaften von Determinanten

a) Die Determinante ist ein Polynom n-ter Ordnung in den Elementen der Matrix *nxn* Matrix A.

b) Die Determinate ist Null wenn alle Elemente einer Zeile Null sind.

c)
$$\begin{vmatrix} a_{11} & \cdots & a_{1n} & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & 0 \\ a_{01} & \cdots & a_{0n} & 1 \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}$$

d) Multipliziert man alle Element einer Zeile mit c so wird die Determinante das c-Fache.

e) Für eine Matrix A mit zwei gleiche Zeilen folgt $|A| = 0$.

f)
$$\begin{vmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ a_{21} & & a_{2n} \\ \cdot & & \\ \cdot & & \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & & a_{2n} \\ \cdot & & \\ \cdot & & \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} + \begin{vmatrix} b_{11} & \cdots & b_{1n} \\ a_{21} & & a_{2n} \\ \cdot & & \\ \cdot & & \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}$$

g) Für eine Dreiecksmatrix gilt

$$\begin{vmatrix} a_{11} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ \cdot & & \cdot & 0 & 0 \\ \cdot & & & \cdot & 0 \\ a_{n1} & \cdots & & & a_{nn} \end{vmatrix} = a_{11}a_{22}\ldots a_{nn}$$

h) $|A| = |A'|$, $|cA| = c^n |A|$

Integraltransformation

Es sei X ein Vektor und Ω ein Bereich, auf dem die Fuktion $f(X)$ definiert ist, und $Y = T(X)$ eine eineindeutige Transformation. Dann gilt

$$\int_{\Omega} f(X)dX = \int_{T(\Omega)} f(T^{-1}(Y))\left|\frac{\partial X}{\partial Y}\right| dY$$

$\left|\frac{\partial X}{\partial Y}\right| = \left|\frac{\partial x_i(Y)}{\partial y_j}\right|$ heißt Funktionaldeterminante.

Für $Y = AX + C$ folgt

$$\int_{\Omega} f(X)dX = \int_{\{Y=AX+C:\, X\in\Omega\}} f(A^{-1}(Y-C))\left|A^{-1}\right| dY$$

Mit der Dichte der p-dimensionalen Normalverteilung $N_p(0, I)$ also $f(X) = (2\pi)^{-p/2} e^{-\|X\|^2/2}$ folgt

$$\int_{\Omega} f(X)dX = (2\pi)^{-p/2}\left|A^{-1}\right| \int_{\{Y=AX+C:\, X\in\Omega\}} e^{-(Y-C)'(A^{-1})'A^{-1}(Y-C)/2} dY$$

Die Funktion $f(Y) = (2\pi)^{-p/2}\left|A^{-1}\right| e^{-(Y-C)'(A^{-1})'A^{-1}(Y-C)/2}$ heißt Dichte der Normalverteilung $N_p(C, AA')$ mit Erwartungswert C und Kovarianzmatrix AA'.

Zufallsvektoren

Erwartungswert:

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad EY=\begin{pmatrix} Ey_1 \\ Ey_2 \\ \cdot \\ \cdot \\ \cdot \\ Ey_n \end{pmatrix} \quad Ea'Y=a'EY=a'm, \quad \forall a$$

Kovarianzmatrix:

$$E(Y-EY)(Y-EY)'=\Big(E(y_i-Ey_i)(y_j-Ey_j) \Big)=(\sigma_{ij})$$

$$\begin{aligned} V(a'Y) &= E(a'Y-a'EY)^2=E(a'(Y-EY))^2 \\ &= E(a'(Y-EY)(Y-EY)'a) \\ &= a'E(Y-EY)(Y-EY)'a \end{aligned}$$

i) $D=\Big(\sigma_{ij} \Big) \geq 0$ Kovarianzmatrix

ii) $\sigma_{ij}^2 \leq \sigma_{ii}\sigma_{jj}$ Cauchi-Schwarz Ungleichung

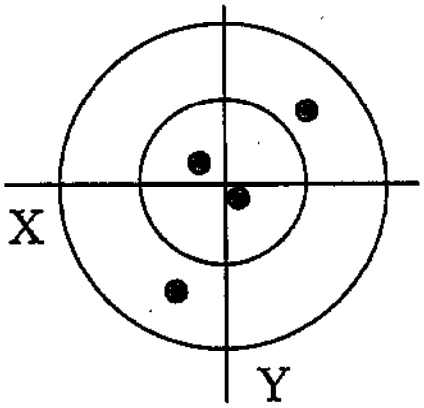
iii) $-1 \leq \rho_{ij} \leq 1,\ \rho_{ij}=\sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ Korrelationskoeffizient

iv) $D>0 \Leftrightarrow$ Es gibt keine lineare Beziehung zwischen den Komponenten von Y, die mit Wahrscheinlichkeit 1 gilt.

v) $r(D)=r<n \Rightarrow$ Dann gibt es einen Zufallsvektor X, eine (n,r) Matrix B und einen konstanten Vektor C mit $Y=BX+C$. $r(B)=r,\ \dim X=r,\ D(X)=I_r$

$$\begin{aligned} cov(a'Y,b'Y) &= E(a'Y-a'EY)(b'Y-b'EY) \\ &= E(a'(Y-EY)(Y-EY)'b) \\ &= a'E(Y-EY)(Y-EY)'b \\ &= a'Db \end{aligned}$$

Normalverteilung



Herschels Hypothese

(a) Randverteilungen $P(x)$ und $P(y)$ haben stetige Dichten $f(x)$ und $f(y)$

(b) Die gemeinsame Dichtefunktion $f(x, y)$ hängt nur von $r = \sqrt{x^2 + y^2}$ ab.

(c) X und Y sind unabhängig verteilt.

$$\Rightarrow f(x,y) = f(x)f(y) = \frac{1}{\sigma^2 2\pi} e^{-(x^2+y^2)/2\sigma^2}$$

Definition: Ein Zufallsvektor $x = (x_1, x_2, ..., x_p)'$ mit der gemeinsamen Dichtefunktion

$$f(x) = \frac{1}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}X'(\sigma^2 I_p)^{-1}X}$$

heißt normalverteilt, genauer $N(0, \sigma^2 I_p)$

$$y = Ax + m \Rightarrow Ey = m,\ Cov(y) = A(Cov(x))A'$$

Resultat: Der Zufallsvektor $y = (y_1, y_2, ..., y_s)'$ hat die gemeinsame Dichtefunktion

$$f(y) = \frac{|AA'|^{-1/2}}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}(X-m)'(\sigma^2 AA')^{-1}(X-m)}$$

Problem: $(\sigma^2 AA')^{-1}$ existiert nur wenn $r(A) = s$

<u>Moderne und einfachste Definition:</u>

$$x \sim N_p(m, D) \Leftrightarrow \forall a : a'x \sim N_1(a'm, a'Da)$$

Weitere Eigenschaften der Normalverteilung

(a) $$X \sim N(m, D), r(D) = r \Leftrightarrow \left\{ \begin{array}{l} X = m + BY, \\ BB' = D, r(B) = r, \\ Y \sim N_r(0, I) \end{array} \right\}$$

(b) $$X \sim N(m, D) \Rightarrow P(X \in \{m + R(D)\}) = 1$$

(c) $$X \sim N(m, D) \Rightarrow CX \sim N(Cm, CDC')$$

(d) $X_i \sim N_p(m_i, D_i)$ unabhängig verteilt
$$\Rightarrow Y = \sum_i a_i X_i \sim N_p(\sum_i a_i m_i, \sum_i a_i^2 D_i)$$

Für Teilvektoren

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \right\}$$ gilt:

(e) X_1 und X_2 sind unabhängig $\Leftrightarrow D_{12} = 0$

(f) bedingte Verteilung:

$$(X_2 \mid X_1) \sim N(d, D_{22} - D_{21} D_{11}^- D_{12})$$
$$\text{mit } d = m_2 + D_{21} D_{11}^- (X_1 - m_1)$$

unabhängig von der Wahl von D_{11}^-.

Cauchy-Schwarz Ungleichung

$$|\Sigma x_i y_i| \leq \sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}$$

Betrachte für $X = (x_1 \ldots x_n)'$ und $Y = (y_1 \ldots y_n)'$

$$(aX+Y)'(aX+Y) = a^2X'X + 2aX'Y + Y'Y \geq 0 \ \forall a$$

Quadratische Ergänzung liefert

$$X'X\left(a+\frac{X'Y}{X'X}\right)^2 + Y'Y - \frac{(X'Y)^2}{X'X} \geq 0 \ \forall a$$

$$\Leftrightarrow \ Y'Y - \frac{(X'Y)^2}{X'X} \geq 0$$

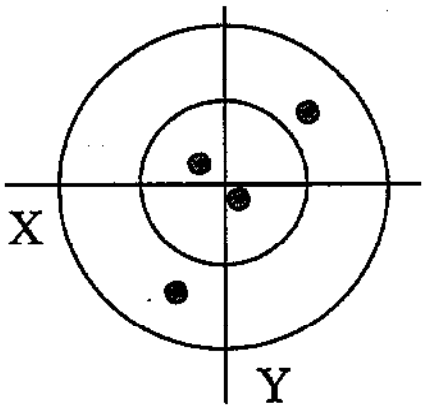
Völlig analog zeigt man

$$E(aX+Y)^2 = a^2EX^2 + 2aEXY + EY^2 \geq 0 \ \forall a$$

$$EX^2\left(a+\frac{EXY}{EX^2}\right)^2 + EY^2 - \frac{(EXY)^2}{EX^2} \geq 0 \ \forall a$$

$$\Leftrightarrow \ EY^2 - \frac{(EXY)^2}{EX^2} \geq 0$$

Normalverteilung



Herschels Hypothese

(a) Randverteilungen $P(x)$ und $P(y)$ haben stetige Dichten $f(x)$ und $f(y)$

(b) Die gemeinsame Dichtefunktion $f(x, y)$ hängt nur von $r = \sqrt{x^2 + y^2}$ ab.

(c) X und Y sind unabhängig verteilt.

$$\Rightarrow f(x,y) = f(x)f(y) = \frac{1}{\sigma^2 2\pi} e^{-(x^2+y^2)/2\sigma^2}$$

Definition: Ein Zufallsvektor $x = (x_1, x_2, ..., x_p)'$ mit der gemeinsamen Dichtefunktion

$$f(x) = \frac{1}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}x'(\sigma^2 I_p)^{-1}x}$$

heißt normalverteilt, genauer $N(0, \sigma^2 I_p)$

$$y = Ax + m \Rightarrow Ey = m,\ Cov(y) = A(Cov(x))A'$$

Resultat: Der Zufallsvektor $y = (y_1, y_2, ..., y_p)'$ hat die gemeinsame Dichtefunktion

$$f(y) = \frac{|AA'|^{-1/2}}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}(y-m)'(\sigma^2 AA')^{-1}(y-m)}$$

Problem: $(\sigma^2 AA')^{-1}$ existiert nur wenn $r(A) = p$

<u>Moderne und einfachste Definition:</u>

$$x \sim N_p(m, D) \Leftrightarrow \forall a : a'x \sim N_1(a'm, a'Da)$$

Orthogonale Transformationen

Eine Matrix C heißt orthogonal wenn $C'C = I$

$$y = C'x + m \Rightarrow Ey = m,\ Cov(y) = C'(Cov(x))C = \sigma^2 I_p$$

Es folgt

$$\begin{aligned} f(y) &= \frac{|C'C|^{-1/2}}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}(y-m)'(\sigma^2 C'C)^{-1}(y-m)} \\ &= \frac{1}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}(y-m)'(\sigma^2 I_p)^{-1}(y-m)} \\ &= \frac{1}{(\sigma^2 2\pi)^{p/2}} e^{-\frac{1}{2}\sum_i^p (y_i-m_i)^2/\sigma^2} \\ &\quad \prod_i^p \frac{1}{(\sigma^2 2\pi)^{1/2}} e^{-\frac{1}{2}(y_i-m_i)^2/\sigma^2} \end{aligned}$$

Unabhängige normalverteilte ZG gehen durch orthogonale Transformationen in solche über.

Linerare Funktionen unabhängige normalverteilter ZG sind normalverteilt.

Die ch. Funkt. von $x \sim N(m, \sigma^2)$ ist mit $t = i\theta$

$$\begin{aligned} \phi(t) = \phi(i\theta) &= E(e^{-\theta x}) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int e^{-\theta x} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \\ &= e^{\frac{\sigma^4\theta^2 + 2\sigma^2 m\theta}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}} \int e^{-\frac{(x-m-\sigma^2\theta)^2}{2\sigma^2}} dx \\ &= e^{\frac{\sigma^2\theta^2 + 2m\theta}{2}} = e^{-\frac{\sigma^2 t^2}{2} + itm} \end{aligned}$$

Die ch. Funkt. von $x \sim N_p(m, D)$ ist $\phi(t) = e^{-\frac{t'Dt}{2} + it'm}$

Grenzverteilung von Häufigkeiten

$E_i\ i = 1, ..., k$ disjunkte Ereignisse $P(E_i) = p_i$

n_i Häufigkeit von E_i in n Versuchen

Wir definieren:
$$V' = \left(\frac{n_1 - np_1}{\sqrt{np_1}}, ..., \frac{n_k - np_k}{\sqrt{np_k}} \right)$$
$$\phi' = \left(\sqrt{p_1}, ..., \sqrt{p_p} \right)$$
$$b' = (b_1, ..., b_k)$$

Satz: Die lineare Funktion $b'V$ besitzt eine asymptotische Normalverteilung $N(0, b'(I - \phi\phi')b)$.
Es gilt $(I - \phi\phi')\phi = 0$.

Beweis: Sei X die ZG mit Werten $b_i/\sqrt{p_i}$ im Fall von Ereignis E_i mit WS p_i.

$$E(X) = \sum \frac{b_i}{\sqrt{p_i}} p_i = \sum b_i \sqrt{p_i} = b'\phi$$

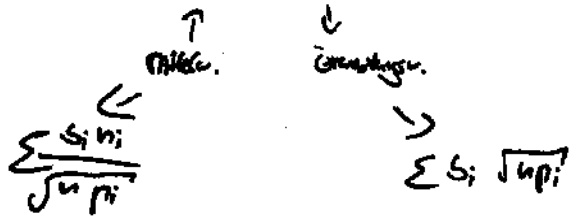
$$V(X) = \sum \frac{b_i^2}{p_i} p_i - (b'\phi)^2 = b'b - (b'\phi)^2 = b'(I - \phi\phi')b$$

Der Mittelwert der $X_j\ j = 1, ..., n$ ist

$$\bar{X}_n = \frac{1}{n} \sum \frac{b_i n_i}{\sqrt{p_i}} = \sum \frac{b_i n_i}{n\sqrt{p_i}}$$

Nach dem zentralen Grenzwertsatz gilt

$$\sqrt{n}(\bar{X}_n - b'\phi) = b'V \rightarrow N(0, b'(I - \phi\phi')b)$$



Für eine vektorwertige lineare Funktion $B'V$ gilt

$$B'V \rightarrow N_p(0, B'(I-\phi\phi')B)$$

da $$\lambda'B'V \rightarrow N_p(0, \lambda'B'(I-\phi\phi')B\lambda) \; \forall\lambda$$

Satz von Wald & Wolfowitz (1944):

Es sei F_n die Folge der Verteilungsfunktionen einer Folge von k-dimensionalen ZG $X_n = (x_1^{(n)}, ..., x_k^{(n)})'$ und $F_{n\lambda}$ die Folge der Verteilungsfuktionen von $\sum \lambda_i x_i^{(n)}$.
Dann konvergiert F_n genau dann gegen eine k-dimensionale Veteilungsfunktion F wenn für alle λ $F_{n\lambda}$ gegen eine Verteilungsfunktion F_λ konvergiert.

Spezialfall: Es sei A eine kx(k-1) Matrix, so daß $(\phi : A)$ eine orthogonale Matrix ist. (per orthogonale Ergänzung)

$$A'V = G \rightarrow N_{k-1}(0, \underbrace{A'(I-\phi\phi')A}) = N_{k-1}(0, I)$$

↳ Gleiche Matrix

Die Grenzverteilung von G ist die eines Vektors von k-1 unabhängigen und standard - normalverteilten ZG.

Dann hat $G'G$ per Definition eine Chi-quadrat -Verteilung mit k-1 Freiheitsgraden $G'G \sim \chi^2_{k-1}$.

Satz: Es gilt $V'V = G'G$ und $V'V \sim \chi^2_{k-1}$

Betrachte die Gleichung

$$\phi' V = \sum \frac{n_i - np_i}{\sqrt{np_i}} \sqrt{p_i} = \sum \frac{n_i - np_i}{\sqrt{n}} = 0$$

Zusammen mit $G = A'V$ folgt

$$\begin{matrix} \phi' V = 0 \\ A' V = G \end{matrix} \Leftrightarrow \begin{pmatrix} \phi' \\ A' \end{pmatrix} V = \begin{pmatrix} 0 \\ G \end{pmatrix}$$

Die Matrix $(\phi : A)$ ist orthogonal, d.h

$$(\phi : A) \begin{pmatrix} \phi' \\ A' \end{pmatrix} = \phi\phi' + AA' = I$$

Dann folgt

$$V = (\phi : A) \begin{pmatrix} 0 \\ G \end{pmatrix} = AG$$

Schließlich: $V'V = G'A'AG = G'G$

Zusammenfassung: Chi-quadrat Test von Pearson (1900)

$$\chi^2 = \sum \frac{(n_i - np_i)^2}{np_i} = \sum \frac{(beobachtet - erwartet)^2}{erwartet} \sim \chi^2_{k-1}$$

Hat viele Anwendungen in der Bioinformatik:

Kann eine Sequenz mit n_A, n_C, n_G, n_T Nukleotiden A,C,G,T eine zufällige Sequenz mit gleichverteilten Nukleotiden sein? Unterscheiden sich die Zusammensetzungen zweier Sequenzen?

Chi-quadrat Anpassungstest

Betrachte disjunkte Ereignisse E_i $i = 1, ..., k$ mit Wahrscheinlichkeiten $p_i(\theta)$, wobei $\theta' = (\theta_1, ..., \theta q)$ ein unbekannter Parametervektor mit $q < k - 1$ ist. Es seinen n_i die Häufigkeiten der Ereignisse E_i in n Versuchen.

Die Funktionen $p_i(\theta)$ haben stetige partielle Ableitungen und es ist $\hat{\theta}$ eine Lösung der Maximum-Likelihood-Gleichungen

$$\frac{\partial}{\partial \theta_j} \sum n_i \ln p_i(\theta) = 0 \ \forall j$$

Die Matrix der partiellen Ableitungen $\partial p_i / \partial \theta_j$ an der Stelle der richtigen Parameter habe vollen Rang. Dann gilt

$$\chi^2 = \sum \frac{\left(n_i - n\hat{p}_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})} \rightarrow \chi^2_{k-1-q}$$

Hauptanwendung: Vergleich und Prüfung von Modellen.

Beispiel: Vergleich zweier Sequenzzusammensetzungen

Hypothese:

$$\begin{aligned} p_1 &= p_5 = \theta_1 \\ p_2 &= p_6 = \theta_2 \\ p_3 &= p_7 = \theta_3 \\ p_4 &= p_8 = \theta_4 \end{aligned} \qquad \chi^2_{8-1-4}$$

Aufgabenblatt 6

1. Es sei A eine nxn Matrix und B eine nxn Diagonalmatrix. Beweise $|AB| = |A||B|$

2. Ergänze den Vektor (1,1,1)' zu einer 3x3 orthognalen Matix.

3. Berechne die Determinante von

$$A = \begin{pmatrix} c_{11} & c_{12} & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ 0 & 0 & b_{11} & b_{12} \\ 0 & 0 & b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & B \end{pmatrix}$$

   Gilt $|A| = |C||B|$?

4. Es sei $X \sim N(0, \sigma^2)$ und $Y = X^2$. Berechne die Dichtefunktion von Y.
   Hinweis: $P(Y < K) = 2P(0 \leq X < \sqrt{K})$

Matrix Operationen

Addition von Matrizen gleichen Typs

$$A+B=B+A, \quad A+0=A$$
$$A+(B+C)=(A+B)+C=A+B+C$$
$$(A+B)'=A'+B'$$

Multiplikation mit einem Faktor

$$(c+d)A=cA+dA$$
$$c(A+B)=cA+cB$$
$$(cA)'=cA'$$

Multiplikation von Matrizen passender Typen

$$AB=(c_{ij}),\ c_{ij}=\sum_r a_{ir}b_{rj} \qquad AB \neq BA!$$
$$A(BC)=(AB)C=ABC$$
$$(AB)'=B'A',\ (ABC)'=C'B'A'$$

Einheitsmatrix

$$AI=A,\ IA=A$$

Inverse einer regulären Matrix

$$AA^{-1}=A^{-1}A=I$$
$$(AB)^{-1}=B^{-1}A^{-1},\ (ABC)^{-1}=C^{-1}B^{-1}A^{-1}$$

Spur einer quadratischen Matrix

$$tr(A)=\sum_i a_{ii}, \quad tr(A)=tr(A'), \quad tr(AB)=tr(BA)$$

Lineare Gleichungssysteme

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m = y_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m = y_2$$
$$\vdots$$
$$a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m = y_n$$

Vektor Form: $a_1x_1 + a_2x_2 + \ldots + a_mx_m = y$
Matrix Form: $Ax = y, \quad R(A) = \{Ax,\ x \in R^m\}$

$Ax = y$ ist lösbar $\Leftrightarrow\ y \in R(A)$
$Ax = 0$ ist immer lösbar !
$N(A) = \{x \in R^m,\ Ax = 0\}$

Grundkenntnisse: i) $\dim R(A) = r(A) \leq \min(n, m)$
ii) $\dim N(A) = m - r(A)$

Bew: $r = r(A), \quad a_1, a_2, \ldots, a_r$ unabhängig
$a_j = a_1x_{1j} + a_2x_{2j} + \ldots + a_rx_{rj} \quad j = r+1, \ldots m$
$x_j' = (x_{1j}, x_{2j}, \ldots, x_{rj}, 0, \ldots, 0, 1, 0 \ldots 0) \in N(A)$
↗ Stelle j

iii) Es existieren reguläre (n,n) bzw.
(q,q) Matrizen A und B mit $X = A\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}B$

Lineare Vektorräume

Definition: Eine Menge $\Re$ von Vektoren heißt linearer Raum wenn

$$x_i \in \Re \; i = 1, \ldots, n \;\Rightarrow\; \sum_i \alpha_i x_i \in \Re \; \forall \alpha_i$$

Definition: Vektoren $x_i \; i = 1, \ldots, n$ sind linear unabhängig wenn

$$\sum_i \alpha_i x_i = 0 \;\Rightarrow\; \alpha_i = 0 \; \forall \; i = 1, \ldots, n$$

Dimension: Maximale Anzahl linear unabhängiger Vektoren in $\Re$.

Der **Spaltenrang** von $A = (a_1 \; a_2 \; \ldots \; a_k)$ ist die Dimension des Spaltenraums $\Re_s$. Der **Zeilenrang** ist die Dimension des Zeilenraums $\Re_z$

$$\dim \Re_s = \dim \Re_z = Rang(A)$$

Transponierte Matrix: $A = (a_{ij}) \;\Rightarrow\; A' = (a_{ji})$

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \quad A' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 \end{pmatrix}$$

$$\Re_s(A) = \Re_z(A')$$

Summen linearer Räume

Wir schreiben $R = R_1 + R_2$ wenn

$$R_1 \cap R_2 = \{0\}$$
$$\dim(R) = \dim(R_1) + \dim(R_2)$$

Jeder Vektor $a \in R$ besitzt eine eindeutige Zerlegung
$a \in R \Rightarrow a = a_1 + a_2,\ a_1 \in R_1,\ a_2 \in R_2$

<u>Basisergänzungssatz:</u>
Zu jedem Teilraum $R_1 \subseteq R$ gibt es einen Teilraum R_2 mit $R = R_1 + R_2$ und genau ein R_2^* mit zusätzlich $R_1 \perp R_2^*$

Es sei A eine nxm Matrix

	A nxm	A' mxn	
$R(A) \subseteq R_n$			$R(A') \subseteq R_m$
$\dim R(A) = r(A)$			$\dim R(A') = r(A)$
$R_n = R(A) \oplus N_1$			$R_m = R(A') \oplus N_2$
$\dim N_1 = n - r(A)$			$\dim N_2 = m - r(A)$
$N_1 = N(A')$			$N_2 = N(A)$

$\oplus$: orthogonale Summe

N : Nullraum
$N(A) = \{x : Ax = 0\}$

Beweis: $x \in R(A'), x \in N(A) \Rightarrow x = A'z, AA'z = 0, A'z = 0, x = 0$
$R(A') \cap N(A) = \{0\}$

$x \in R(A'), y \in N(A) \Rightarrow x = A'z, Ay = 0, x'y = z'Ay = 0$

A ist nxm Matrix $\Rightarrow R_n = R(A) \oplus N(A'), \quad R_m = R(A') \oplus N(A)$

Weiter gilt: $N(A) = N(A'A)$

$$a \in N(A) \Rightarrow Aa = 0 \Rightarrow A'A = 0 \Rightarrow a \in N(A'A)$$
$$a \in N(A'A) \Rightarrow A'Aa = 0 \Rightarrow a'A'Aa = 0 \Rightarrow Aa = 0$$

Durch orthogonale Ergänzung folgt $R(A) = R(AA')$

Das Bild der Matrix A wird bereits dann vollständig erzeugt, wenn wir A auf alle Vektoren aus $R(A')$ anwenden!!

Das Gleichungssystem $(I + uv')x = w$

$$\begin{aligned} (I+uv')x &= w \\ x + (v'x)u &= w \\ (v'x) + (v'x)(v'u) &= (v'w) \\ (v'x)(1 + (v'u)) &= (v'w) \end{aligned}$$

Ist $1 + (v'u) \neq 0$ so folgt $(v'x) = \dfrac{(v'w)}{1 + (v'u)}$

Eindeutige Lösung: $x = w - \dfrac{(v'w)}{1 + (v'u)}u$

Ist $1 + (v'u) = 0 \Rightarrow (v'w) = 0$, d.h Gleichung ist nur Lösbar für solche rechte Seiten w. Dann ist $x = w$ eine Lösung!

Faktorisationssatz:

Sei A nxm Matrix. Dann gibt es eine reguläre nxn Matrix B und eine reguläre mxm Matrix C mit $A = B^{-1}\Delta C^{-1}$, wobei

$$\Delta = \begin{pmatrix} d_1 & & & 0 \ldots 0 \\ & \ddots & & \vdots \\ & & d_n & 0 \ldots 0 \end{pmatrix} = (D\ 0) \quad \text{für } m \geq n$$

$$\Delta = \begin{pmatrix} d_1 & & & \\ & \ddots & & \\ & & d_m & \\ 0 & \ldots & 0 & \\ \vdots & \ldots & \vdots & \\ 0 & \ldots & 0 & \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix} \text{für } n \geq m .$$

Wir definieren: $d_i^- = \begin{cases} d_i^{-1} & \text{für } d_i \neq 0 \\ 0 & \text{für } d_i = 0 \end{cases}$

verallgemeinerte Inverse

D^- durch Ersetzen von d_i mit d_i^- in D
Δ^- durch Ersetzen von D mit D^- in Δ'

Es folgt: $d_i d_i^- d_i = d_i$ $DD^-D = D$ $\Delta\Delta^-\Delta = \Delta$

und mit $A^- = C\Delta^- B$ $AA^-A = A$

Ist das Gleichungssystem $Ax = y$ lösbar, so ist A^-y eine Lösung.

Probe: $A(A^-y) = AA^-Ax = Ax$

Verallgemeinerte Inverse

Definition:

A^- heißt G-Inverse von A falls für jedes $b \in R(A)$ der Vektor A^-b eine Lösung von $Ax = b$ ist.

Das gilt genau dann wenn $\boxed{AA^-A = A}$

Beweis:

$A = (a_1, a_2, ..., a_m) \Rightarrow Ax = a_i$ ist lösbar
$AA^-a_i = a_i \quad \forall i \quad \Rightarrow \quad AA^-A = A$

$AA^-A = A, \quad Ax = b$ lösbar $\Rightarrow \exists x : b = Ax$
$AA^-b = AA^-Ax = Ax = b \Rightarrow A^-b$ ist Lösung.

Folgende Bedingungen sind äquivalent.

i) $G \in \{A^-\}$
ii) $(GA)^2 = GA, \quad r(GA) = r(A)$
iii) $(AG)^2 = AG, \quad r(AG) = r(A)$
iv) $r(I - GA) = r(I) - r(A)$

r : Rang

Allgemeine Form der G-Inversen

$\{A^-\} = A^- + U - A^-AUAA^-, \quad \forall U$
$\{A^-\} = A^- + V(I - AA^-) + (I - A^-A)U, \quad \forall V, U$

Moore Penrose G-Inverse

Die KQS G-Inverse liefert irgend ein x mit $\|Ax-y\| = \min$.
Wir suchen nun das kleinste x mit dieser Eigenschaft.

Definition:

> A^+ heißt Moore-Penrose G-Inverse falls jedes y, $x = A^+y$ eine Lösung von $\min_x \|Ax-y\|$ mit minimaler Norm ist.

Resultat:

i) A^+ ist eindeutig wie A^{-1}

ii) $AA^+A = A, \quad A^+AA^+ = A^+,$
$(A^+A)' = A^+A; \quad (AA^+)' = AA^+$

iii) $A^+ = A'A(A'AA'A)^-A'$

Ganz wichtiger Satz für beliebige G-Inverse!

i) $A(A'A)^-A'A = A,\ A'A(A'A)^-A' = A'$

ii) $BA^-A = B \Leftrightarrow R(B') \subseteq R(A')$

iii) $AA^-C = C \Leftrightarrow R(C) \subseteq R(A)$

iv) $B, C \neq 0$. BA^-C unabhängig von der Wahl von $A^- \Leftrightarrow R(B') \subseteq R(A'),\ R(C) \subseteq R(A)$

$(I - 2uu')$

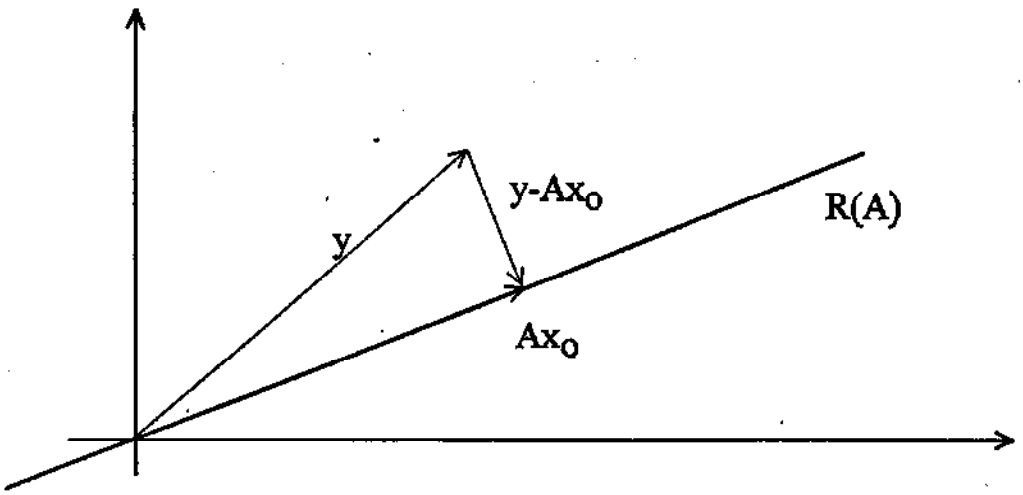
Näherungslösungen für lineare Gleichungssysteme

$y \notin R(A) \Rightarrow Ax = y$ ist nicht lösbar
Ersatz: $x : \|Ax - y\|^2 = \min$

Definition:

A_l^- heißt KQS (least squares) G-Inverse von A falls für alle y, $x = A_l^- y$ den Ausdruck $\|Ax - y\|^2$ minimiert.

Ist A_l^- überhaupt eine G-Inverse?

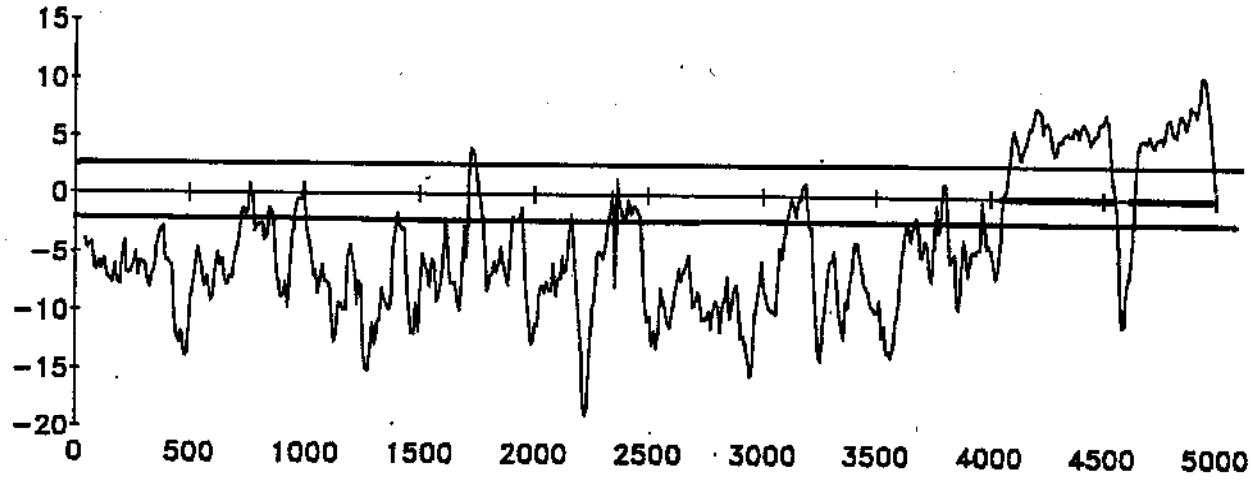


Optimalitätsbedingung: $(AA_l^- y - y) \perp R(A)$

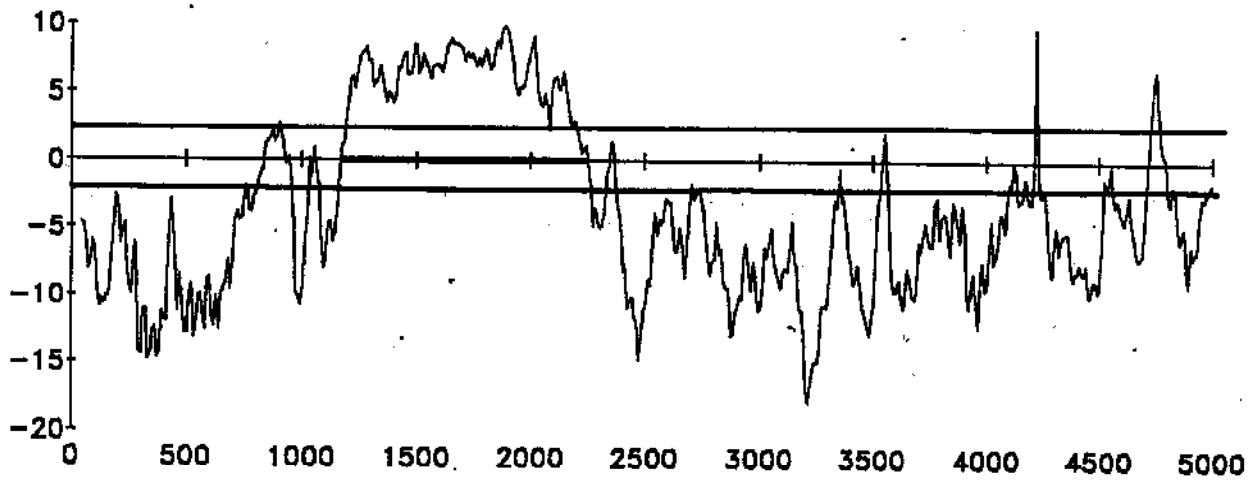
Resultat:

i) A_l^- ist genau dann KQS G-Inverse von A wenn $AA_l^- A = A$, $(AA_l^-)' = AA_l^-$
ii) $(A'A)^- A'$ ist eine Wahl von A_l^-.
iii) $x = A_l^- y$ ist Lösung der Gleichung $A'Ax = A'y$

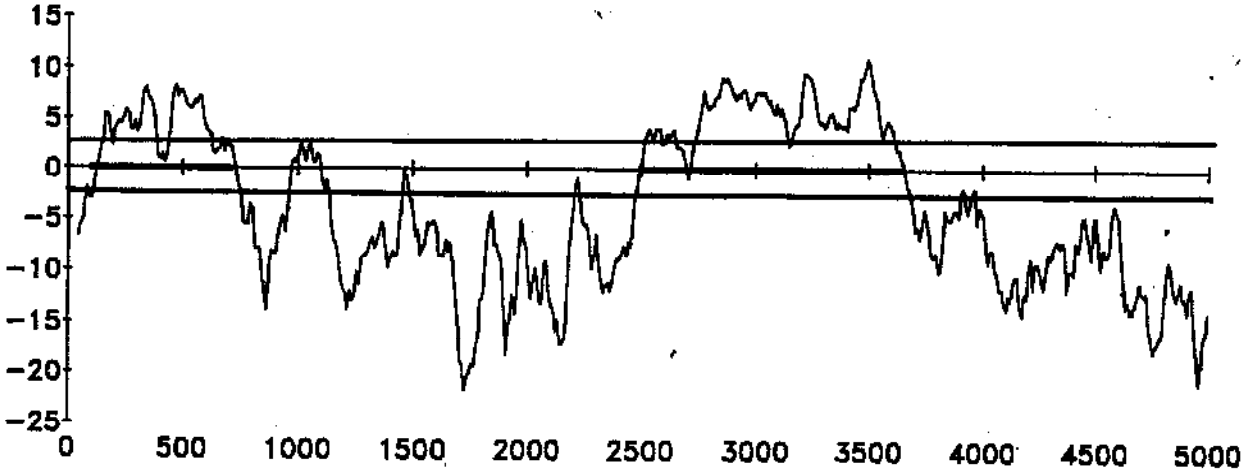
coding signal frame 1



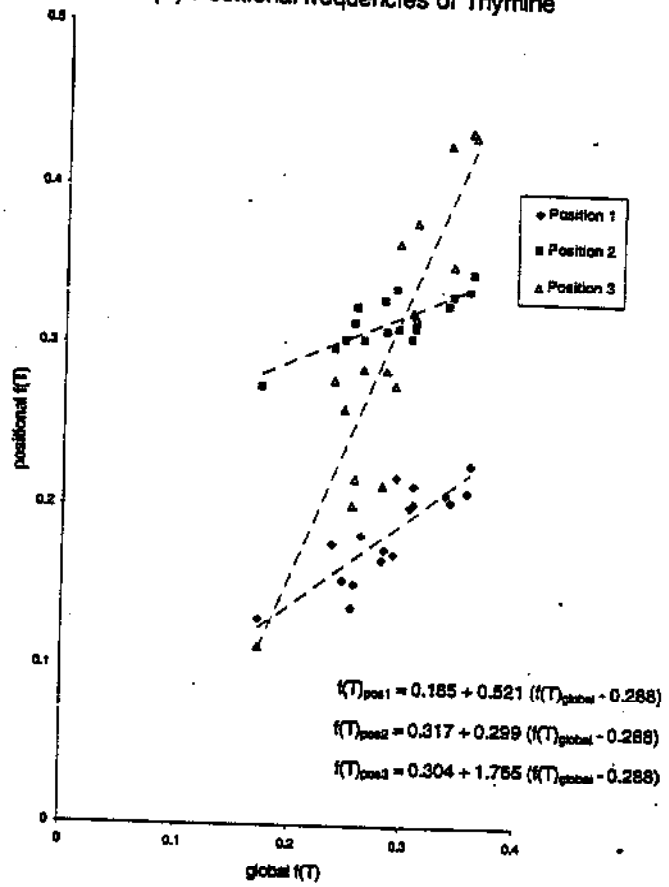
coding signal frame 2



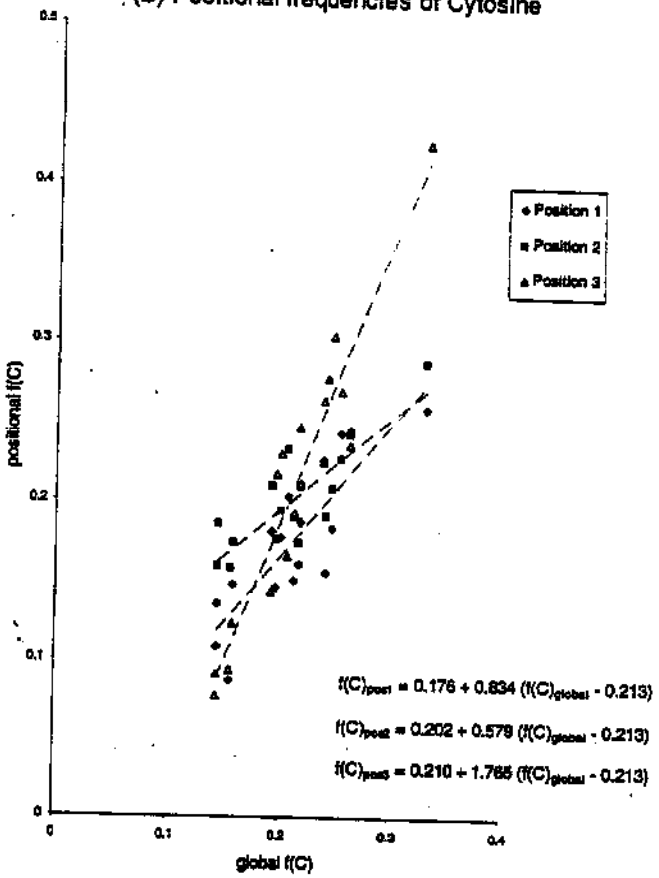
coding signal frame 3



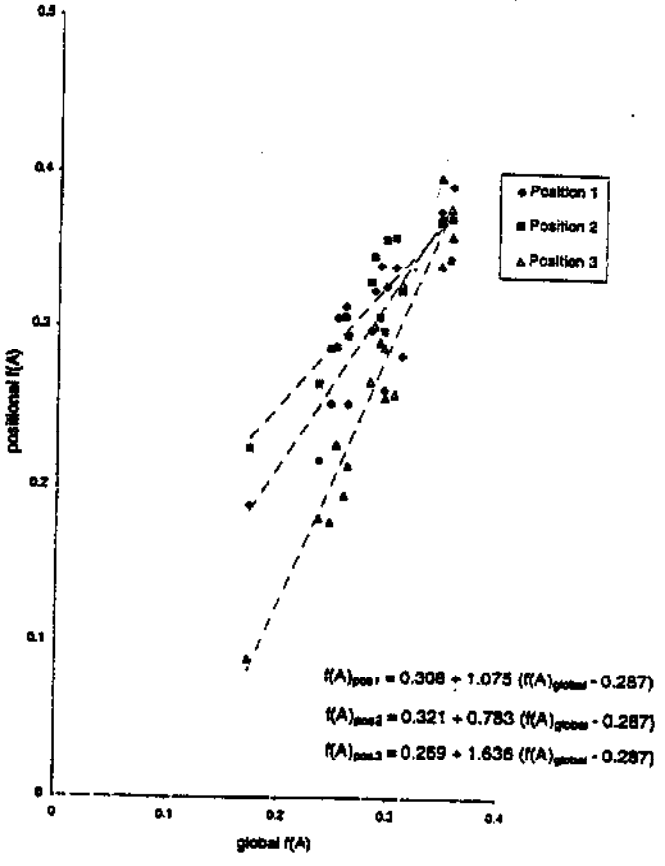
(A) Positional frequencies of Thymine



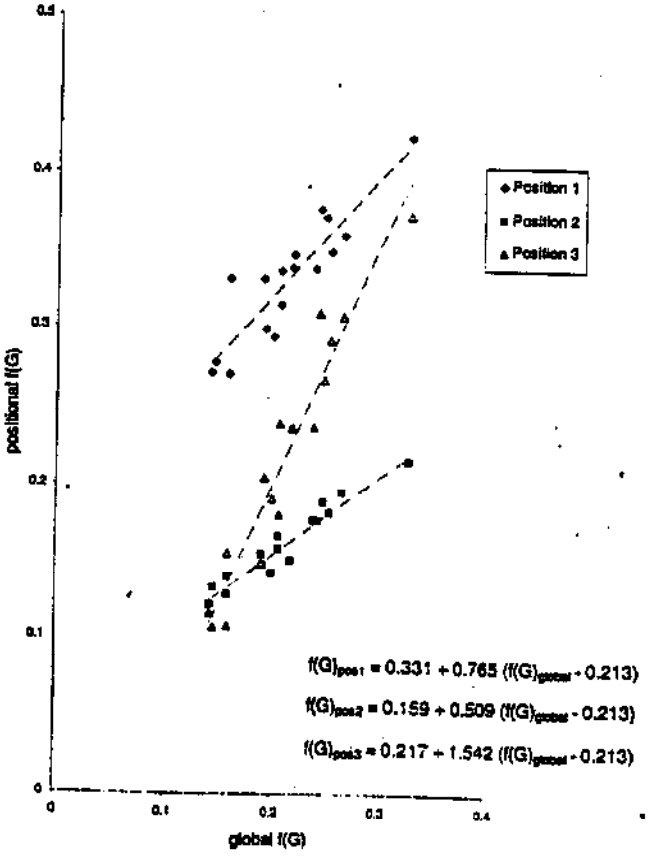
(B) Positional frequencies of Cytosine



(C) Positional frequencies of Adenine



(D) Positional frequencies of Guanine



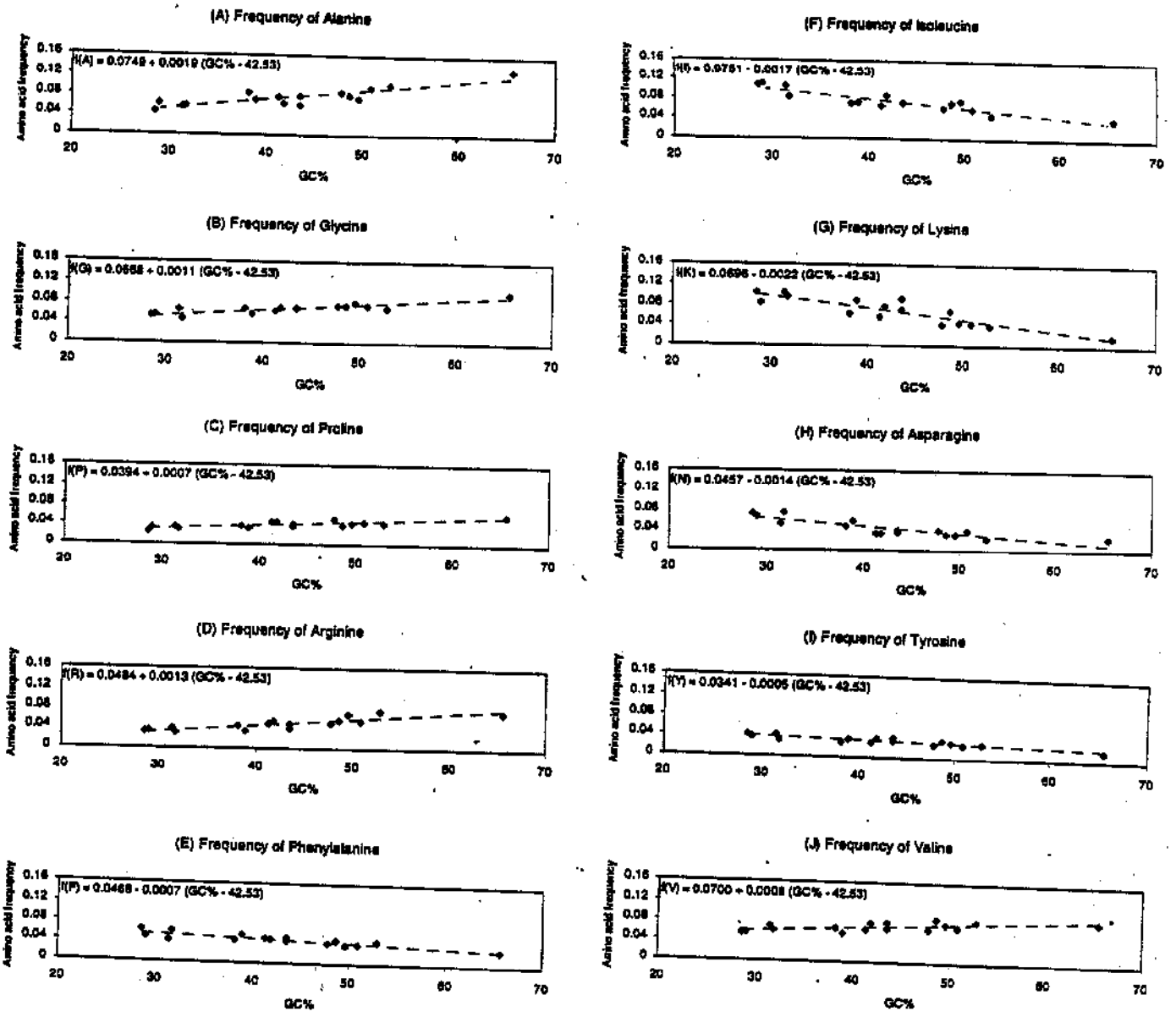


Figure 2. (A) Frequency of amino acid alanine shown as a function of GC content of the bacterial genome for 17 genomes. The equation of the line approximating the observed data was obtained by linear regression analysis. (B–J) As (A) for glycine, proline, arginine, phenylalanine, isoleucine, lysine, asparagine, tyrosine and valine, respectively.

however, functioned properly independently of the number of synonymous codons, so all amino acids were handled in the same manner. Obviously, this method guarantees that the sum of the refined frequencies of codons is equal to the sum of the frequencies of the amino acids. By completing this computation for all 61 codons, we produced the heuristically built codon usage table for the input genomic sequence.

To construct the three-periodic zero order Markov model of a protein coding region the codon usage table is all that is needed. For example, to determine the probability of A in the first position of a codon, the probabilities of all codons that start with A were added together. In the zero order model of non-coding sequence the global frequencies of the respective nucleotides were used.

For the first order three-periodic Markov model, the codon usage table provides enough data to calculate only two matrices of transition probabilities out of three. To define the values of transition probabilities related to nucleotides occupying the third position of one codon and the first position of the next codon it was assumed that occurrences of adjacent codons are independent events. Indeed, a rather weak correlation exists between nucleotides of adjacent codons. Thus the probability of nucleotide Y in the first position of a codon following a nucleotide X in the third position of the previous codon, $P(X \rightarrow Y)$ for the (..X||Y..) configuration, is equal to the probability of nucleotide Y in the first position of a codon defined previously for the zero order Markov model.

For the second order Markov model, only the transition probabilities for the nucleotide in the third codon position could be produced from the codon usage table.

To find the transition probabilities related to the first and second codon positions, we used the same assumption of

Regressionsmodelle

$$y = f(X) + \varepsilon,\ E\varepsilon = 0,\ E\varepsilon^2 = \sigma^2 \Rightarrow$$

$$E(y \mid X) = f(X)$$

$$f(X) \in \left\{ f_\beta(X)\,, \beta = (\beta_1, \beta_2, ..., \beta_k)' \in R^k \right\}$$

Aufgabe: $$\min_\beta E\left(y - f_\beta(X) \right)^2$$

Schätzung von β aus einer Stichprobe:

Schätzung: $$\min_\beta \sum_i \left(y_i - f_\beta(X_i) \right)^2$$ KQS Methode

Lösung: $f_{\hat{\beta}}(X)$ heißt nicht lineare KQS-Schätzfunktion

Funktionensystem:

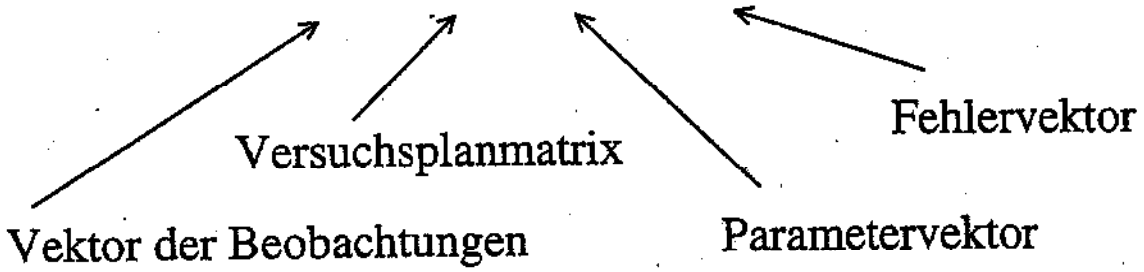
$$f(X) \in \left\{ \sum_i \beta_i f_i(X)\,, \beta = (\beta_1, \beta_2, ..., \beta_k)' \in R^k \right\}$$

↑ Neue unabhängige Variable

$$E\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} f_1(X_1)\, f_2(X_1) & \cdot & \cdot & \cdot & f_k(X_1) \\ f_1(X_2)\, f_2(X_2) & & & & f_k(X_2) \\ \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ f_1(X_n)\, f_2(X_n) & \cdot & \cdot & \cdot & f_k(X_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{pmatrix}$$

Beste lineare erwartungstreue Schätzfunktion

Lineares Modell: $Y = X\beta + \varepsilon, E\varepsilon = 0, E\varepsilon\varepsilon' = \sigma^2 I$



Lineare Schätzfunktionen: $a'Y = a'X\beta + a'\varepsilon$

Erwartungswert: $a'X\beta$

Erwartungstreue: $a'X\beta = c'\beta, \quad \forall\beta$

$\Leftrightarrow \; a'X = c' \Leftrightarrow X'a = c$

Varianz: $var(a'Y) = \sigma^2 a'a$

<u>Definition:</u> $a_0'Y$ heißt lineare erwartungstreue Schätzung für $c'\beta$ mit kleinster Varianz wenn $a_0'a_0' = \min_{a:X'a=c} a'a$

Lagrange-Funktion: $a'a - 2\lambda'(X'a - c)$

Partielle Ableitungen:

$$2a - 2X\lambda = 0 \;\Rightarrow (X'X)\lambda = X'a = c$$

$$\Rightarrow \lambda = (X'X)^{-}c \;\Rightarrow a = X(X'X)^{-}c$$

Ergebnis:

$$a'Y = c'(X'X)^{-}X'Y = c'\hat{\beta}, \hat{\beta} = (X'X)^{-}X'Y$$

Einfache Stichprobe: $y_i \quad i = 1, ..., m$ (Meßwerte)

$$Ey_i = \beta \quad var(y_i) = \sigma^2$$

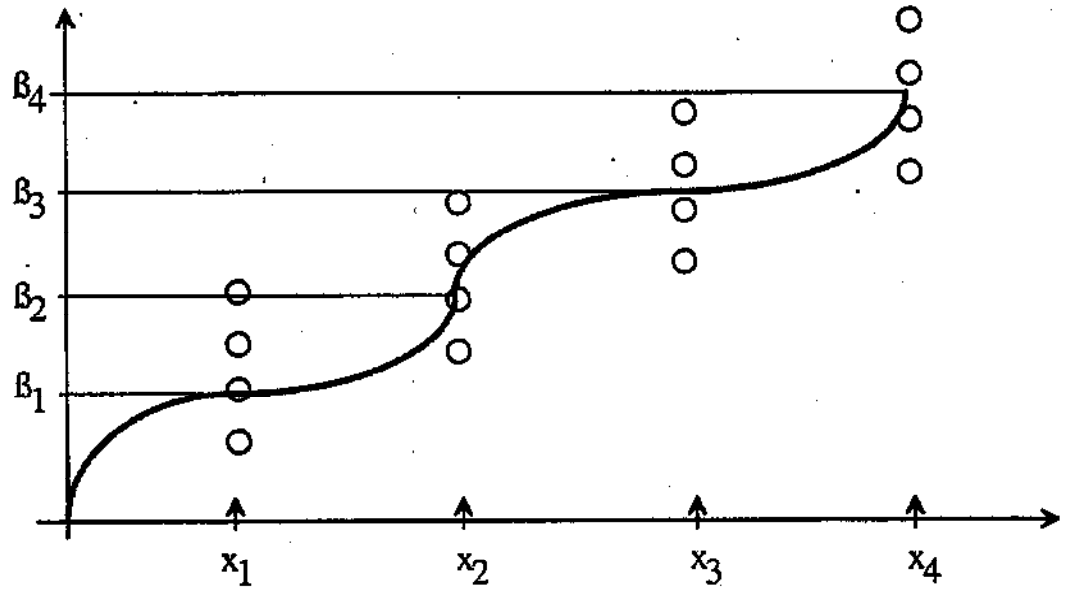
Einfach
klassifizierte Stichprobe $y_{ij} \quad i = 1, ..., m \quad j = 1, ..., n_i$

$$Ey_{ij} = \beta_i \quad var(y_{ij}) = \sigma^2$$

$$z_i = \begin{pmatrix} y_{i1} \\ \cdot \\ \cdot \\ y_{in_i} \end{pmatrix}$$ Gruppe von Meßwerten $Ez_i = \beta_i 1_{n_i}$

$$Y = \begin{pmatrix} z_1 \\ \cdot \\ \cdot \\ z_m \end{pmatrix} \qquad EY = \begin{pmatrix} \beta_1 1_{n_1} \\ \cdot \\ \cdot \\ \beta_m 1_{nm} \end{pmatrix} = \begin{pmatrix} 1_{n_1} & & \\ & \cdot & \\ & & 1_{n_m} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix}$$

$EY = X\beta + \in$ $\quad X$: heißt Versuchsplanmatrix



$$Ey_{ij} = x_i\beta \qquad var(y_{ij}) = \sigma^2$$