Was man in SAS Genetics vergeblich sucht: Allelfrequenzschätzungen bei Dominanz von Allelen mittels EM-Algorithmus

Bernd Paul Jäger
Ernst-Moritz-Arndt-Universität
Greifswald, Institut für Biometrie und
Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
bjaeger@biometrie.uni-greifswald.de

Catharina Schüler
Ernst-Moritz-Arndt-Universität
Greifswald, Institut für Biometrie und
Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
Catharina.Schueler@gmx.de

Karl-Ernst Biebler
Ernst-Moritz-Arndt-Universität
Greifswald, Institut für Biometrie und
Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
biebler@biometrie.uni-greifswald.de

Paul Eberhard Rudolph
Leibnizinstitut für Nutztierbiologie, FBN
FB Genetik u. Biometrie
(bis 8/2012)
Wilhelm-Stahl-Allee 2
Dummerstorf
pe.rudolph@kabelmail.de

Zusammenfassung

Es liegen drei Berechnungsmethoden für Allelfrequenzschätzungen von Verer-bungssystemen bei vorliegender Dominanz eines oder mehrerer Allele vor, einschließlich der zugehörigen SAS-Programme: der EM-Algorithmus, die MKQ- und die Minimum-χ^2 – Methode. Sie können leicht auf neu hinzukommende Vererbungssysteme angepasst werden.
Eine Empfehlung wird für den EM-Algorithmus abgegeben, weil diese Schätzung mit der MLH-Schätzung zusammenfällt (siehe Schüler, [5]). Damit erbt die EM-Schätzung alle guten Eigenschaften der MLH-Schätzung, nämlich die asymptotische Erwartungstreue und dass ihre Varianz asymptotisch gegen die Minimalvarianz konvergiert. Darüber hinaus weiß man, dass mit der MLH-Schätzung auch die EM-Schätzung asymptotisch gegen eine Normalverteilung konvergiert, wobei μ der erwartete Parameter und σ^2 die zugehörige Minimalvarianz nach Rao/Cramer/Freschet ist. Damit wird die Konstruktion von Konfidenzbereichen numerisch möglich.
Die Untersuchung der Eigenschaften der EM-Schätzung in Simulationsexperimenten steht für kleine Stichprobenumfänge n noch aus. Man möchte wissen, bei etwa welchem n die leichter zu bestimmende Asymptotik die exakte Verteilung der Schätzung ersetzen kann. Insbesondere interessiert man sich für solche Vererbungssysteme, die zahlreiche Genotypen besitzen und bei denen kleine Allelwahrscheinlichkeiten vorkommen, bei denen die Asymptotik erst für sehr große n näherungsweise erfüllt ist.
Die Implementierung der MLH-Methode wird nicht empfohlen. Die Likelihoodfunktion ist auf großen Bereichen des Definitionsbereichs verschwindend klein, damit auch die Gradienten. Die Abbruchbedingungen sind dann bereits beim Startwert erfüllt, die partiellen Ableitungen sind nahe Null und das Programm bricht ab, ohne zu iterieren. Der Startwert muss sehr dicht am Maximumpunkt liegen! Das lässt sich in Praxi nur schwer realisieren.

Risikomaße in epidemiologischen Fall-Kontroll-Studien, etwa OR, lassen sich wesentlich effektiver auf der Allelebene als auf der Phänotypenebene statistisch beurteilen (siehe PROC CASECONTROL). Auch deshalb benötigt man für Vererbungssysteme mit Dominanz das vorgestellte SAS-Programm 2.

Schlüsselwörter: SAS/Genetics, Allelfrequenz, Schätzung, Dominanz, EM-Algorithmus

1 Einleitung

Die Schätzung von Allelwahrscheinlichkeiten in Mehrallelensystemen ist nur so lange einfach, wie kein Allel dominiert und alle Genotypen beobachtbar sind. Dann lässt sich die Schätzung analog zum Zweiallelenfall ohne Dominanz durchführen, sie ist eine verallgemeinerte „Genzählmethode".
Die Schätzung der Allelwahrscheinlichkeit p = P(A) eines beliebigen Allels A wird in einem solchen Fall als relative Häufigkeit der A-Allele bezüglich aller vorhandenen Allele aufgefasst. Eine Stichprobe vom Umfang n enthält insgesamt 2n Allele und beim Durchzählen der A-Allele der Stichprobe muss man berücksichtigen, dass jeder Homozygote AA zwei Allele und jeder Heterozygote A. nur ein A-Allel enthält.
Liegt aber Dominanz vor, wird die Berechnung aufwändig. Der Maximum-Likelihood-Ansatz führt zu einem nichtlinearen Gleichungssystem, dessen numerische Behandlung in der Regel Schwierigkeiten bereitet. Diese Schwierigkeiten zu umgehen ist das Ziel des EM-Algorithmus, eines Iterationsalgorithmus, bei dessen Konstruktion die „Genzählmethode" Pate stand. Allerdings gehen dabei nicht die beobachteten Genotypen ein, weil sie verborgen hinter den Phänotypen liegen, sondern die erwarteten Häufigkeiten für die Genotypen unter der Bedingung des Phänotyps. Die Anzahlen nicht erkennbarer Genotypen, die einem Phänotypen unterliegen und in die Berechnung der Allerwahrscheinlichkeit eingehen, werden durch die erwarteten Anzahlen bezüglich des Vererbungsmodells ersetzt. So kommt man ausgehend von beliebigen Startwerten für die Allerwahrscheinlichkeiten zu neuen Schätzwerten, die ihrerseits wieder als Startwerte in die folgende Iteration einfließen. Man beendet die Iteration, wenn vorher festgesetzte Genauigkeitsforderungen eingehalten werden.
Der EM-Algorithmus wird als ein ausführlich kommentiertes SAS-Programm mitgeteilt, das sich auf beliebige Vererbungssysteme, bisher bekannte aber auch zukünftig hinzukommende, anwenden lässt, sofern man den Erbgang kennt. Ebenso kann mit dem EM-Algorithmus auch die Berechnung von Haplotypenfrequenzen erfolgen. Der Nachweis der Konvergenz der EM-Iteration gegen die MLH-Lösung stammt von Excoffier u. Slatkin [6]. Ausführlich beschrieben und verallgemeinert findet man den Beweis bei Schüler [5]. Am Beispiel des bekannten AB0-Systems, eines Drei-Allelen-Modells (mit Dominanz von A über 0 und B über 0), wird die Vorgehensweise erläutert und an den jeweiligen Stellen auf die Schwierigkeiten hingewiesen.

2 Herleitung des EM-Algorithmus für das AB0 Blutgruppensystem

Es wird das allgemein bekannte Blutgruppensystem AB0 ausgewählt, bei dem die beiden Allele A und B über das Allel 0 dominieren. Das AB0-System und seine Bedeutung bei der Bluttransfusion wurden 1901 von Karl Landsteiner [1] beschrieben. 1930 erhielt er dafür den Medizin-Nobelpreis. Die ersten, allerdings falschen, Vererbungsregeln im AB0-System wurden von Ludwik Hirszfeld und Emil von Dungern 1910-11 aufgestellt [2]. Der Mathematiker F. Bernstein brachte 1924 diese falsche Hypothese zu Fall [3] und gab 1930 sowohl die bis heute gültigen Vererbungsregeln als auch eine näherungsweise Berechnungsmethode für die Allelwahrscheinlichkeiten an (siehe unten Bernstein-Näherung, [4]).

Als Bezeichnungen für die Allelwahrscheinlichkeiten werden gewählt:

$p = P(A)$, $q = P(B)$ und $r = 1-p-q = P(0)$.

Es entstehen die folgenden vier Phänotypen mit den zugehörigen Phänotypen-Wahrscheinlichkeiten:

- A mit den unterliegenden Genotypen AA und A0 und der entsprechenden Wahrscheinlichkeit $P(A) = p^2+2p(1-p-q)$,
- B mit den Genotypen BB und B0 und der Wahrscheinlichkeit $P(B) = q^2 + 2q(1-p-q)$,
- sowie die beiden Phänotypen(= Genotypen) AB mit $P(AB) = 2pq$ und
- 0 mit $P(0) = (1-p-q)^2$.

Wenn man in einer Stichprobe vom Umfang n die Genotypenhäufigkeiten mit n_A, n_B, n_{AB} und n_0 bezeichnet, ergibt sich der Likelihood-Ansatz:

$$L(p,q) = \binom{n}{n_A n_B n_{AB} n_0} (p^2 + 2p(1-p-q))^{n_A} (q^2 + 2q(1-p-q))^{n_B}$$
$$(q^2 + 2q(1-p-q))^{n_B} (2pq)^{n_{AB}} ((1-p-q)^2)^{n_0}.$$

Von dieser Funktion L der zwei Veränderlichen p und q gilt es, das Maximum zu berechnen. Die Logarithmus-Transformation als eine die Ordnung erhaltende Abbildung verändert nicht die Maximalstelle, so dass ersatzweise von der leichter zu behandelnden Funktion $\log(L(p,q))$ ausgegangen werden kann. Notwendige Bedingung für ein Extremum ist das Verschwinden der partiellen Ableitungen nach p bzw. nach q.

$$\frac{\partial \log(L(p,q))}{\partial p} = \frac{\begin{pmatrix} 2n_A(1-p-q)^2(2p+q-2) + (p+2q-2) \\ \cdot(2p(n_B(p+q-1) + n_0(2p+q-2)) \\ +n_{AB}(2+2p^2-3q+q^2+p(3q-4))) \end{pmatrix}}{p(1-p-q)(2-2p-q)(p+2q-2)}$$

$$\frac{\partial \log(L(p,q))}{\partial q} = \frac{\begin{pmatrix} n_{AB}(2p^3 + 2(q-2)(1-q)^2 + p^2(7q-8) + p(10-17q+7q^2)) \\ +2(n_B(1-p-q)^2(p+2q-2) + q(2p+q-2)(-n_A(1-p-q) \\ +n_0(p+2q-2))) \end{pmatrix}}{q(1-p-q)(2-2p-q)(p+2q-2)}$$

Nach Nullsetzen ergibt sich ein nichtlineares Gleichungssystem, von dem man keine expliziten Wurzeln angegeben kann. Mit geeigneten Startwerten, beispielsweise den Bernstein-Lösungen, erhält man numerische Näherungslösungen für dieses System. Man kann leicht nachweisen, dass es im Definitionsbereich von L(p,q), einem Dreieck mit den Punkten (0,0), (1,0) und (0,1), auf dessen Rand genau eine Lösung (p_0, q_0) gibt. Die Bernstein-Lösungen erhält man durch folgende Überlegungen:

$$P(0) + P(A) = (1-p-q)^2 + (p^2 + 2p(1-p-q)) = (1-q)^2,$$
$$P(0) + P(B) = (1-p-q)^2 + (q^2 + 2q(1-p-q)) = (1-p)^2 \text{ und}$$
$$P(0) = (1-p-q)^2$$

Die zugehörige Momentenschätzungen - Wenn man nämlich diese drei Gleichungen mit dem Stichprobenumfang multipliziert, entstehen auf der linken Seite die Erwartungswerte der Phänotypen, die man ihren Häufigkeiten näherungsweise gleich setzt. - ergeben:

$$n_0 + n_A \approx E(0) + E(A) = n(1-q)^2,$$
$$n_0 + n_B \approx E(0) + E(B) = n(1-p)^2 \text{ und}$$
$$n_0 \approx E(0) = n(1-p-q)^2.$$

Diese Momentenschätzung ist die Bernstein-Lösung, nämlich

$$q = 1 - \sqrt{\frac{n_0 + n_A}{n}},$$
$$p = 1 - \sqrt{\frac{n_0 + n_B}{n}} \text{ und}$$
$$r = 1 - p - q = \sqrt{\frac{n_0}{n}}.$$

Diese Schätzung hat natürlich nicht die guten Eigenschaften einer MLH-Lösung. Für große Stichprobenumfänge n liegt sie aber befriedigend genau an der MLH-Lösung.

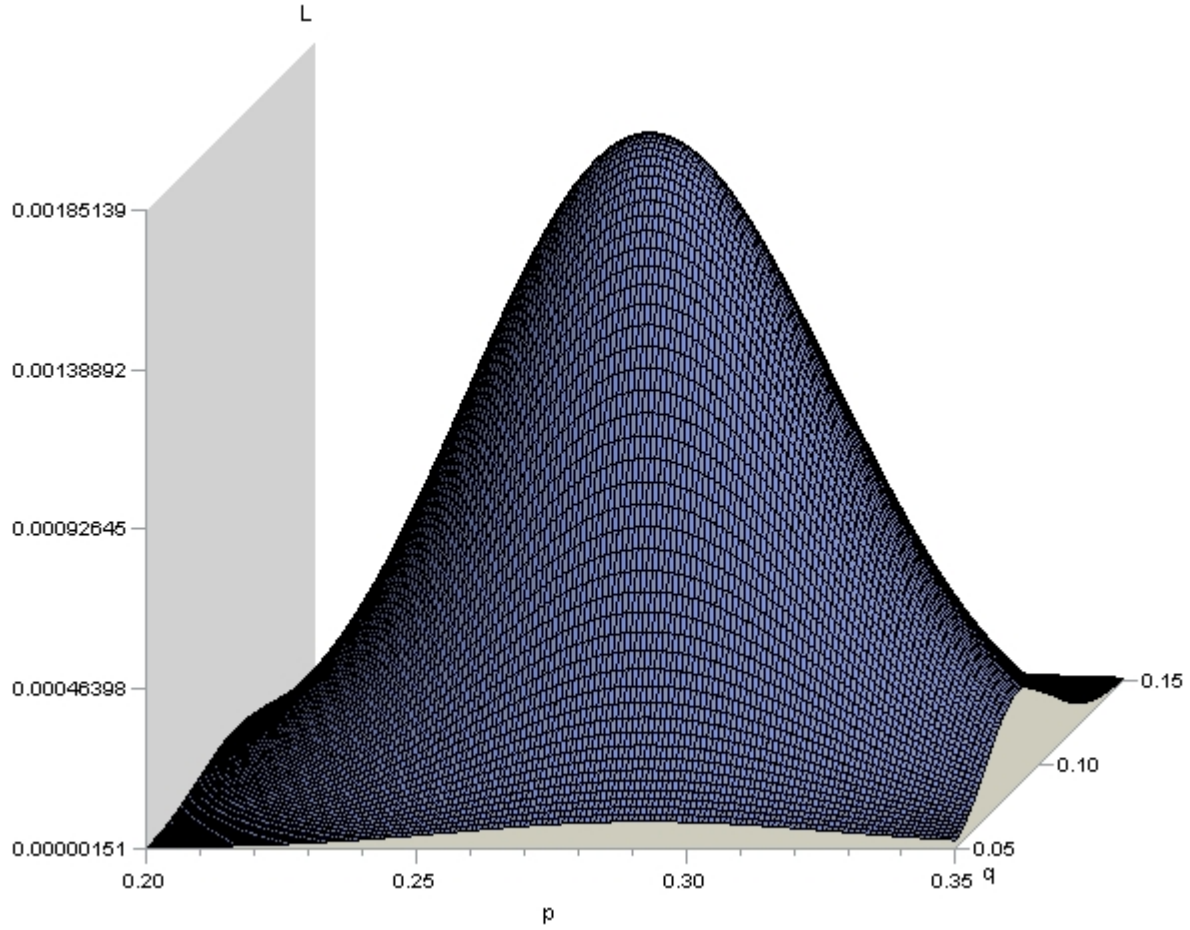

Abbildung 1: Likelihoodfunktion L(p,q) für das AB0-System mit $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$ in einer Umgebung des Maximums

Das Maximum von L(p,q) wird an der Stelle $(p_0, q_0) = (0.2791, 0.0945)$ angenommen, wie man mit der PROC MEANS leicht nachrechnen kann (vergleiche auch Abb.1). Die Bernsteinsche Momentenschätzung (0.2789, 0.0945) ist dicht an dieser Lösung.

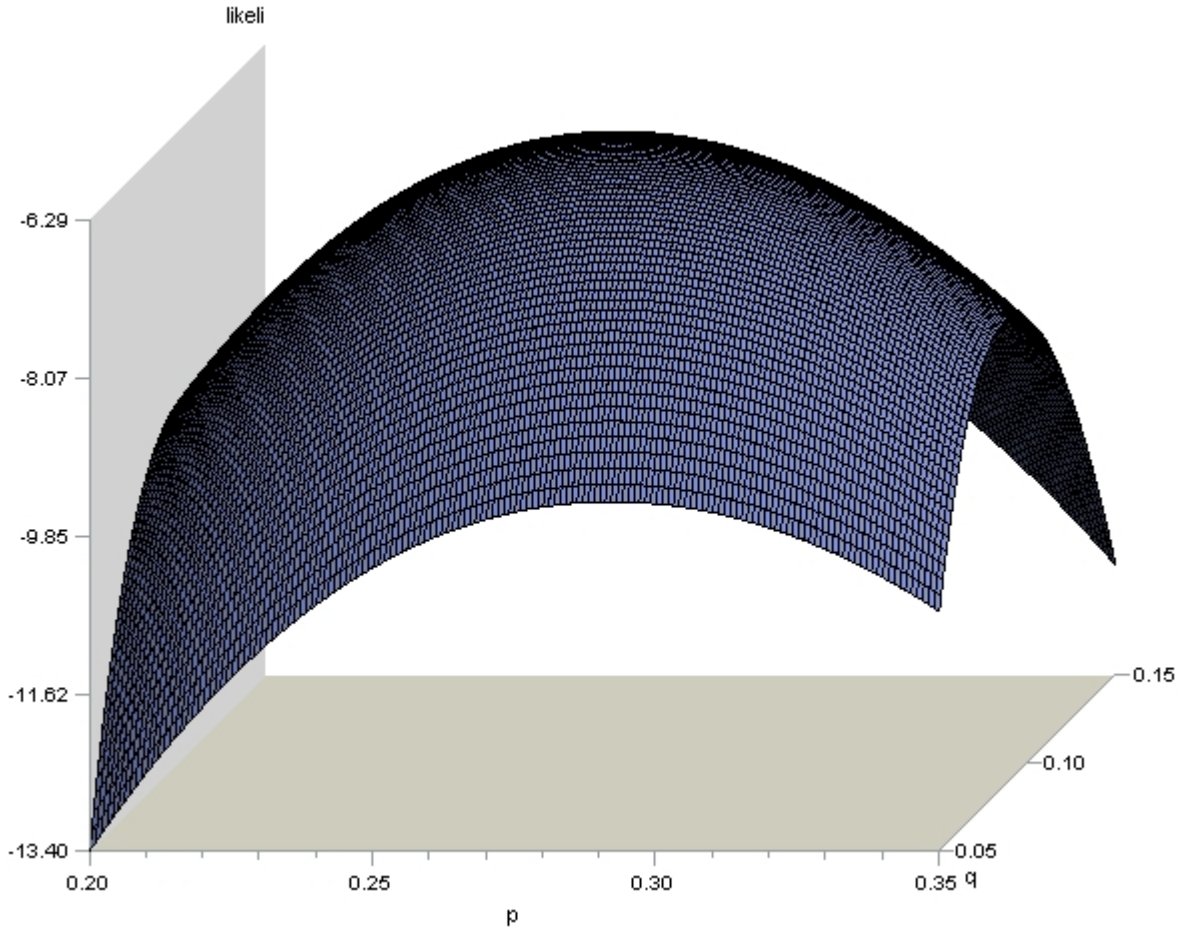

Abbildung 2: Log-Likelihoodfunktion Log(L(p,q) für das AB0-System mit $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$ in einer Umgebung des Maximums

Beim EM-Algorithmus wird ein Ansatz gewählt, der an die „Genzählmethode“ erinnert. Die Allelwahrscheinlichkeit p beispielsweise erhält man aus der Gleichung

$$p = \frac{1}{2n}(2E(AA \mid A) + E(A0 \mid A) + n_{AB}) = \frac{1}{2n}(2n_A P(AA \mid A) + n_A P(A0 \mid A) + n_{AB}),$$

wobei $E(AA|A) = n_A \cdot P(AA|A)$ der bedingte Erwartungswert ist und P(AA|A) die bedingte Wahrscheinlichkeit, dass der Genotyp AA vorliegt, wenn man bereits den Phänotyp A erkannt hat. Entsprechendes gilt für die bedingten Erwartungswerte E(A0|A), E(BB|B) und E(B0|B), sowie die bedingten Wahrscheinlichkeiten P(A0|A), P(BB|B) und P(B0|B). Leider sind diese bedingten Wahrscheinlichkeiten Funktionen der unbekannten Allelfrequenzen:

$$P(AA \mid A) = \frac{p^2}{p^2 + 2p(1-p-q)} \text{ und } P(A0 \mid A) = \frac{2p(1-p-q)}{p^2 + 2p(1-p-q)}, \text{ sowie}$$

$$P(BB \mid B) = \frac{q^2}{q^2 + 2q(1-p-q)} \text{ und } P(B0 \mid B) = \frac{2q(1-p-q)}{q^2 + 2q(1-p-q)}.$$

Startet man nun von einer beliebigen Näherung p_n, q_n und $r_n = 1 - p_n - q_n$ und wendet die gleichen Überlegungen auf eine Schätzfunktion für das Allel q an, so ergeben sich folgende Iterationsgleichungen:

$$p_{n+1} = \frac{1}{2n}\left(2n_A \frac{p_n^2}{p_n^2 + 2p_n(1-p_n-q_n)} + n_A \frac{2p_n(1-p_n-q_n)}{p_n^2 + 2p_n(1-p_n-q_n)} + n_{AB} \right),$$

$$q_{n+1} = \frac{1}{2n}\left(2n_B \frac{q_n^2}{q_n^2 + 2q_n(1-p_n-q_n)} + n_B \frac{2q_n(1-p_n-q_n)}{q_n^2 + 2q_n(1-p_n-q_n)} + n_{AB} \right)$$

und schließlich $r_{n+1} = (1 - p_{n+1} - q_{n+1})$.

Nach wenigen Iterationsschritten ist man selbst bei schlecht gewählten Startwerten dicht an der MLH-Lösung. Den Nachweis, dass der Iterationsalgorithmus konvergiert, findet man unter allgemeineren Bedingungen bewiesen bei Excoffier und Slatkin [6].

Beispiel:
Für $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$, das sind etwa die Phänotypenanzahlen für A, B, AB und 0 von Deutschland für einen Stichprobenumfang von n = 100, erhält man bei den ersten sechs Iterationsschritten die in Tabelle 1 enthaltenen Ergebnisse, wobei von p = 0.5, q = 0.5 und r = 0 gestartet wurde.

Tabelle 1: Iterationsschritte des EM-Algorithmus beim Startwert $(p_0, q_0, r_0) = (0.5, 0.5, 0)$

Iteration	p = P(A)	q = P(B)	r = P(0)
0	0.5	0.5	0.0
1	0.455	0.155	0.39
2	0.319211	0.100775	0.580014
3	0.286396	0.095195	0.618409
4	0.280424	0.094654	0.624493
5	0.279399	0.094575	0.626026
6	0.279225	0.094565	0.626210
…	…	…	…
∞	0.279189	0.094563	0.626248

Bemerkungen:

- Die Buchstaben EM im Algorithmus stehen für die beiden Schritte der Iteration. Im ersten wird die erwartete Häufigkeit der Genotypen beim Startwert (p, q) berechnet (E von Erwartung oder expectation). Im zweiten Schritt wird die „Genzählmethode“ auf die Erwartungswerte angewandt. Diese Lösung ist das Maximum der Likelihood-Funktion (M von Maximum oder maximisation).
- Der EM-Algorithmus lässt sich problemlos auch auf die Schätzung von Haplotypenfrequenzen anwenden.
- Nach wenigen Iterationsschritten ist der Algorithmus am Fixpunkt angelangt, wenn man von einem beliebigen Startwert $(p_0, q_0, r_0) \neq (0,0,0)$ ausgeht. In der folgenden Abbildung ist diese Konvergenz dargestellt. Die Realisierung erfolgt mit SAS-Programm 1. Ausgehend von 20 zufällig gewählten Startwerten erhält man Abb.3.

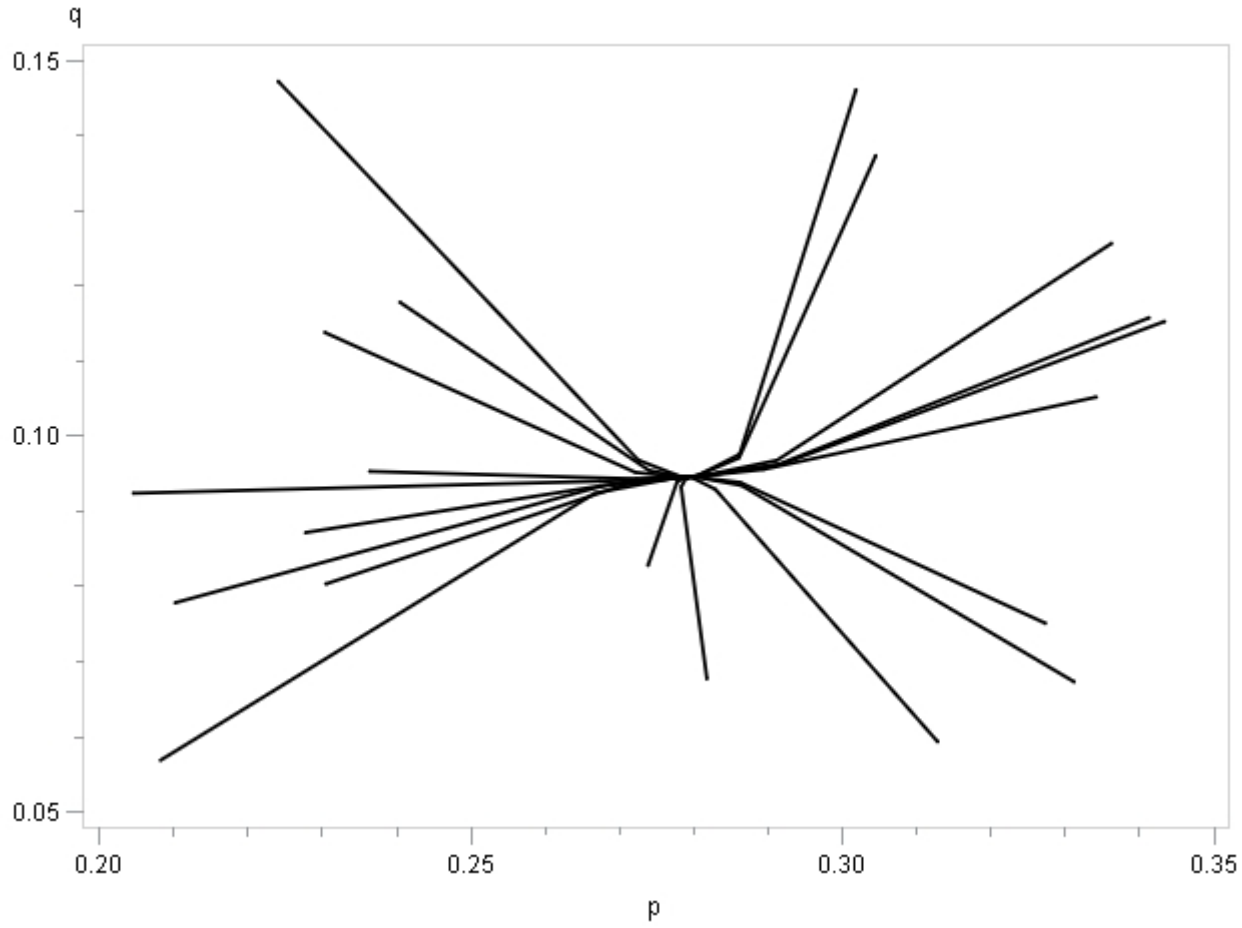

Abbildung 3: Konvergenzverhalten des EM-Algorithmus beim AB0-System gegen die Lösung (p_0, q_0) = (0.27919, 0.09456) von 20 zufällig gewählten Startpunkten aus bei $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$

SAS-Programm 1: Konvergenz des EM-Algorithmus für das AB0-System von beliebigen Startwerten aus

```
data ab0;
na=43;nb=13;nab=5;n0=39;n=na+nb+nab+n0;
Do j=1 to 20;
     marke: ps=UNIFORM(j);
     qs=UNIFORM(j);
     if (ps<0.24 or ps>0.32 or qs<0.08 or qs>0.11) then goto marke;
     p=ps;q=qs;
     do i=1 to 20;
          EAA=na*p**2/        (p**2+2*p*(1-p-q));
          EA0=na*2*p*(1-p-q)/(p**2+2*p*(1-p-q));
          pn=(2*EAA+EA0+nab)/(2*n);

          EBB=nb*q**2/        (q**2+2*q*(1-p-q));
          EB0=nb*2*q*(1-p-q)/(q**2+2*q*(1-p-q));
          qn=(2*EBB+EB0+nab)/(2*n);
          output;
          p=pn;q=qn;r=1-p-q;
     end;
end;run;
symbol i=join l=1 w=2 c=black;
axis1 order=(0.08 to 0.11 by 0.01);
axis2 order=(0.24 to 0.32 by 0.02);
proc gplot;
plot q*p=ps/vaxis=axis1 haxis=axis2 nolegend;
run;quit;
```

3 Das SAS-Programm zum EM-Algorithmus

Das folgende SAS-Programm ist mit vielen Kommentaren versehen, um das Verständnis für die Programmschritte zu erhöhen.

- Das jeweilige Vererbungssystem ist in den Funktionsaufruf Likeli(x) einzubinden. Im Feld g werden die Phänotypenwahrscheinlichkeiten eingetragen, die durch additive Verknüpfungen der unterliegenden Genotypenwahrscheinlichkeiten entstehen. Die beobachteten Phänotypenhäufigkeiten werden als globales Feld *n* und die Phänotypenanzahl als globale Variable *pt* festgelegt.
- Herzstück des Programms ist der Aufruf `CALL NLPFDD,` in dem beide Schritte des EM-Algorithmus zusammen gefasst sind. Der Aufruf erfolgt in einer DO-Schleife, die solange durchlaufen wird, bis der Absolutbetrag zwischen altem und neuen Iterationswert unter *eps* = `10**(-5)` fällt.
- Die folgende Tab. 2 enthält alle weiteren im SAS verfügbaren nichtlinearen Optimierungsverfahren und ist dem SAS-Hilfesystem entnommen. Dabei sind einheitlich bei allen Subroutinen (außer `CALL NLPFEA`) *fun* die zu optimierende Funktion und *x0* der Startvektor, die beide zu den notwendigen Eingabeoptionen zählen. Im Abschnitt 4 sind bei den alternativen Schätzverfahren weitere Optimierungsbeispiele angewandt.

- Für den EM-Algorithmus sind einige Programmschritte hinzugefügt, mit denen man die erwarteten Genotypwahrscheinlichkeiten berechnen und einen Test durchführen kann, ob die beobachteten und die nach Hardy-Weinberg erwarteten Phänotypenhäufigkeiten in Übereinstimmung sind. Das Programm stellt damit die Grundlagen zur Berechnung von Vaterschaftswahrscheinlichkeiten dar, für bekannte aber auch zukünftige Vererbungssysteme.

SAS-Programm 2: EM-Algorithmus für das AB0-System

```
proc iml;
%let n={43 13 5 39}; /*beobachtete Phänotypen-Hf. A, B, AB und 0*/
     s=Sum(&n);/* Stichprobenumfang */
%let pt=4; /*Anzahl an Phänotypen */
     start Likeli(x) global(g,n,pt,s); /* Def. Vererbungsmodell */
          g=j(1,&pt,0);
          g[1] = x[1]**2+2*x[1]*x[3]; /* Phänotyp A = {AA, A0} */
            g[2] = x[2]**2+2*x[2]*x[3]; /* Phänotyp B = {BB, B0} */
          g[3] = 2*x[1]*x[2];          /* Phänotyp AB = Genotyp AB*/
            g[4] = x[3]**2;            /* Phänotyp 0 = Genotyp 00 */
          L=sum(&n#log(g));            /*Berechnung Log-Likelihood*/

         return(L);
     finish Likeli;
a=4;                /* Anzahl an Parametern, die der Vektor x
enthält*/
eps=10**(-5);  /* festgelegt Genauigkeit*/
iter=0;        /* Iterationszähler */
x=j(1,a,0);    /* Startvektor x(alt) Allelwkt. mit 1/a belegt */
y=j(1,3,1/3);  /* Startvektor y(neu) mit anderen Werten belegt,
                    damit mindestens ein Iterationsschritt */
do until (ABS(x-y)<eps);
                  /*Do-Schleife durchlaufen, bis jede Komponente <
eps */
iter=iter+1;   /* Iterationszähler erhöhen */
x=y;           /* "neuen" werden "alte" Iterierten für nächsten
Lauf */

CALL NLPFDD(crit,grad,hess,"Likeli",x);
/* approx. Bestimmung der partiellen Abl. grad durch finite Diff.*/
print iter x  crit; /* Iterationsfolge bis Abbr. wird ausgegeben */
y=(grad#x)/sum(grad#x); /*komponentenweise Div. = neuer Startw. */
end;
x=y;
/*********** zusätzliche Ausgabe ***************************/
gtw=J(a-1,a-1,0);/* Matrix der Genotypenwahrscheinlichkeiten*/
do i=1 to a-1;
   do j=i to a-1;
      if i^=j then gtw[i,j]=2*x[1,i]*x[1,j];else
gtw[i,j]=x[1,i]**2;
   end;
end;
```

```
Print 'Genonotypenwahrscheinlichkeiten'; print gtw;
Print 'Phänotypenwahrscheinlichkeiten'; print g;
nn=Sum(&n);nneu = &n;e=nn#g;
print 'beobachtete und erwartete Häufigkeiten'; print nneu, e;
fg = &pt -1 - (a-2);
chi=sum(((&n-e)#(&n-e))/e);p=1-probchi(chi, fg);
print 'Freiheitsgrad',fg;
print chi;print 'P( CHI >' chi') =' p;
run;quit;
```

Tabelle 2: Nonlinear Optimization and Related Subroutines

Conjugate Gradient Optimization Method	Optimization Subroutines
CALL NLPCG(*rc, xr, fun,"x0 <, opt, blc, tc, par, ptit,"grd">);*	
Double Dogleg Optimization Method	
CALL NLPDD(*rc, xr, fun,"x0 <,opt, blc, tc, par, ptit,"grd">);*	
Nelder-Mead Simplex Optimization Method	
CALL NLPNMS(*rc, xr, fun,"x0 <,opt, blc, tc, par, ptit,"hlc">);*	
Newton-Raphson Optimization Method	
CALL NLPNRA(*rc, xr, fun,"x0<,opt, blc, tc, par, ptit,"grd,"hes">);*	
Newton-Raphson Ridge Optimization Method	
CALL NLPNRR(*rc, xr, fun,"x0<,opt, blc, tc, par, ptit,"grd,"hes">);*	
(Dual) Quasi-Newton Optimization Method	
CALL NLPQN(*rc, xr, fun,"x0<,opt, blc, tc, par, ptit,"grd,"hlc,"jacnlc">);*	
Quadratic Optimization Method	
CALL NLPQUA(*rc, xr, quad, x0 <,opt, blc, tc, par, ptit,"lin>);*	
Trust-Region Optimization Method	
CALL NLPTR(*rc, xr, fun,"x0 <,opt, blc, tc, par, ptit,"grd,"hes">);*	
Hybrid Quasi-Newton Least Squares Methods	Least Squares Subroutines
CALL NLPHQN(*rc, xr, fun,"x0, opt <,blc, tc, par, ptit,"jac">);*	
Levenberg-Marquardt Least Squares Method	
CALL NLPLM(*rc, xr, fun,"x0, opt <,blc, tc, par, ptit,"jac">);*	
Approximate Derivatives by Finite Differences	Supplementary Subroutines
CALL NLPFDD(*f, g, h, fun,"x0 <,par, grd">);*	
Feasible Point Subject to Constraints	
CALL NLPFEA(*xr, x0, blc <,par>);*	

4 Alternative Schätzmethoden zum EM-Algorithmus

Neben der EM-Methode, die äquivalent zur MLH-Methode ist, sollen noch zwei weitere Methoden erwähnt werden, die in der Statistik häufig zur Anwendung kommen. Das sind zum einen die χ^2-Methode und zum anderen die Methode der kleinsten Quadrate, von denen man weiß, dass sie zur MLH-Methode asymptotisch äquivalent sind.

4.1 Die Minimum-χ^2-Methode

Nach der Berechnung der Allelfrequenzen aus einer Stichprobe schließt sich in der Regel ein χ^2-Anpassungstest an, der überprüft, ob das Hardy-Weinberg-Gleichgewicht als Hypothese angenommen werden kann. Da liegt es nahe, die Parameter so zu schätzen, dass die Prüfgröße des Anpassungstests minimal wird:

$$\chi^2(p,q) = \frac{\left(n_A - E(N_A)\right)^2}{E(N_A)} + \frac{\left(n_B - E(N_B)\right)^2}{E(N_B)} + \frac{\left(n_{AB} - E(N_{AB})\right)^2}{E(N_{AB})} + \frac{\left(n_0 - E(N_0)\right)^2}{E(N_0)}.$$

Dabei bedeuten E(.) die Erwartungswerte der Zufallsgrößen

$E(N_A) = n \cdot P(A) = n(p^2 + 2p(1-p-q))$,
$E(N_B) = n \cdot P(B) = n(q^2 + 2q(1-p-q))$,
$E(N_{AB}) = n \cdot P(AB) = n(2pq)$ und
$E(N_0) = n \cdot P(0) = n(1-p-q)^2$.

Damit ist $\chi^2(p,q)$ eine Funktion der beiden Veränderlichen p und q:

$$\chi^2(p,q) = \frac{\left(n_A - n\left(p^2 + 2p(1-p-q)\right)\right)^2}{n\left(p^2 + 2p(1-p-q)\right)} + \frac{\left(n_B - n\left(q^2 + 2q(1-p-q)\right)\right)^2}{n\left(q^2 + 2q(1-p-q)\right)}$$

$$+ \frac{\left(n_{AB} - n(2pq)\right)^2}{n(2pq)} + \frac{(n_0 - n(1-p-q)^2)^2}{n(1-p-q)^2} \quad .$$

Ein notwendiges Kriterium für das Minimum von $\chi^2(p,q)$ ist das gleichzeitige Verschwinden der ersten partiellen Ableitungen: $\frac{\partial}{\partial p}\chi^2(p,q) = 0$ und $\frac{\partial}{\partial q}\chi^2(p,q) = 0$.

Dieses Bestimmungssystem ist ebenfalls nichtlinear und besitzt ebenso wie das MLH-Bestimmungssystem nur numerische Näherungslösungen. Für den interessierenden Bereich um die Lösung herum ist die Funktion $\chi^2(p,q)$ in Abb. 4 illustriert.

Die hinreichenden Kriterien für ein Minimum müssen ebenfalls noch überprüft werden. Wenn (p_0, q_0) die Lösung dieses Bestimmungssystems ist, so müssen gleichzeitig die hinreichenden Bedingungen

$$\frac{\partial^2}{\partial p^2}\chi^2(p_0, q_0) > 0\ , \frac{\partial^2}{\partial q^2}\chi^2(p_0, q_0) > 0 \text{ und}$$

$$\frac{\partial^2}{\partial p \partial q}\chi^2(p_0, q_0) = \frac{\partial^2}{\partial q \partial p}\chi^2(p_0, q_0) > 0 \text{ erfüllt sein.}$$

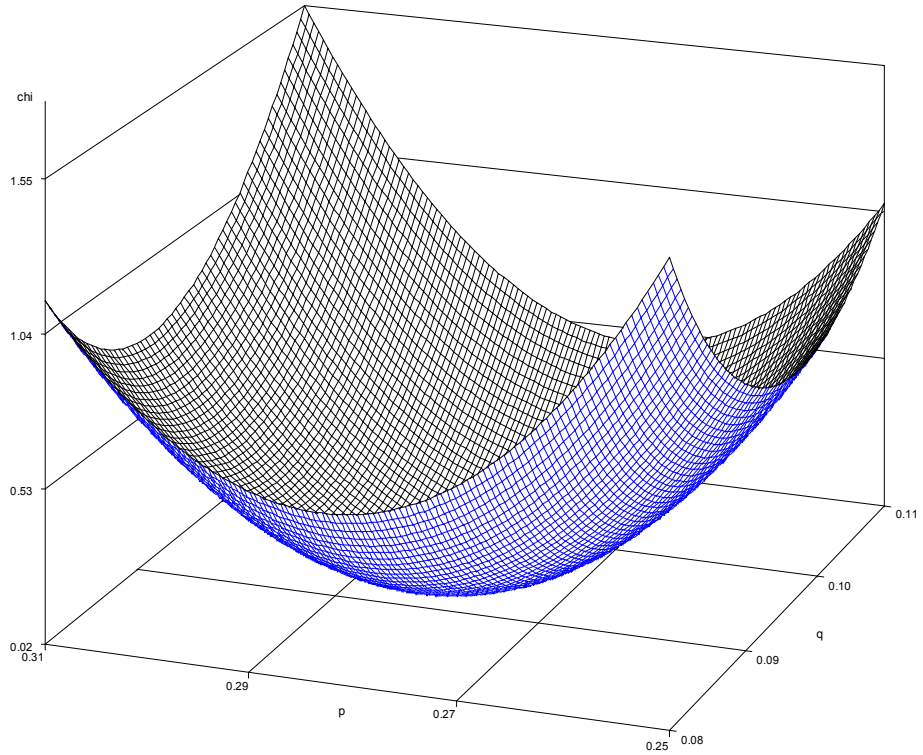

Abbildung 4: $\chi^2(p, q)$ für das AB0-System in der Nähe des Minimums (0.233831, 0.122765) für $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$

SAS-Programm 3: Allelfrequenzschätzung im AB0-System mittels χ^2-Methode und Nutzung von CALL NLPTR

```
title Parameterberechnung mittels Chi-Quadrat-Methode;
proc iml;

   start chi(x) global(n,pt,s);
      /* Definition des Vererbungsmodells */
     n={43 13 5 39};/*beobachtete Phänotypen-Hf. A, B, AB und 0*/
     s=sum(n);/* Stichprobenumfang */
     pt=4; /*Anzahl an Phänotypen */
     g=j(1,pt,0); /* Phänotypenwahrscheinlichkeiten */
     g[1] = x[1]**2+2*x[1]*(1-x[1]-x[2]);
     g[2] = x[2]**2+2*x[1]*(1-x[1]-x[2]);
     g[3] = 2*x[1]*x[2]; g[4] = (1-x[1]-x[2])**2;

     c=j(1,pt,0);
```

```
      c[1]=(n[1]-s*g[1])**2/(s*g[1]); /* Chi²-Anteile der Phänotypen
*/
       c[2]=(n[2]-s*g[2])**2/(s*g[2]);
      c[3]=(n[3]-s*g[3])**2/(s*g[3]);
       c[4]=(n[4]-s*g[4])**2/(s*g[4]);
      L=sum(c); /*Berechnung Chi²-Funktion L*/
      return(L);
     finish chi;
x = {0.28 0.095};/* Startwert */
optn = {0 2}; /* optn[1]=0, dann Minimumsuche */
call nlptr(rc,xres,"chi",x,optn); /* trust-region optimization
method */
run;quit;
```

Hinweise:

1. Da das Gleichungssystem äußerst kompliziert wird - bei den partiellen Ableitungen ist bei jedem Summanden die Quotientenregel anzuwenden - gibt es dafür Vereinfachungen. Es werden solche Terme weggelassen, die Quadrate von Wahrscheinlichkeiten enthalten. Die dürftige Begründung dafür ist lediglich, dass diese Quadrate von Wahrscheinlichkeiten sehr klein werden und dadurch beim Weglassen keine großen Kalkulationsfehler entstehen. Die Methode nennt sich **variierte Minimum-χ^2-Methode.** Sie wird hier nicht behandelt.
2. Mit dem SAS-Programm 3 kann unter Verwendung der Optimierungsroutine CALL NLPTR die Berechnung der Allelfrequenzen mittels χ^2-Methode durchgeführt werden.

4.2 Die MKQ-Methode (Methode der kleinsten Fehlerquadrate)

Die Parameter p und q werden so bestimmt, dass die Summe der quadratischen Abstände zwischen den beobachteten Phänotypenhäufigkeiten und den nach Vererbungsmodell erwarteten Häufigkeiten minimal wird:

$$\begin{aligned} \mathrm{mkq}(p,q) &= \left(n_A - E(N_A)\right)^2 + \left(n_B - E(N_B)\right)^2 + \left(n_{AB} - E(N_{AB})\right)^2 + \left(n_0 - E(N_0)\right)^2 \\ &= \left(n_A - n\left(p^2 + 2p(1-p-q)\right)\right)^2 + \left(n_B - n\left(q^2 + 2q(1-p-q)\right)\right)^2 \\ &\quad + \left(n_{AB} - n(2pq)\right)^2 + \left(n_0 - n(1-p-q)^2\right)^2 . \end{aligned}$$

Die Erwartungswerte $E(N_A)$, $E(N_B)$, $E(N_{AB})$ und $E(N_0)$ entsprechen denen der Minimum-χ^2-Methode.

Eine Bestimmungsgleichung für die Parameter p und q erhält man auch hier durch das Nullsetzen der beiden partiellen Ableitungen, das notwendige Kriterium für ein Extremum. Sie bietet ebenfalls keine numerischen Vorteile zur MLH-Methode. Für den interessierenden Bereich um die Lösung herum ist die Funktion $mkq(p,q)$ in Abb. 5

dargestellt. Analog müssen auch hier noch die hinreichenden Bedingungen für ein Minimum überprüft werden.

Mit dem folgenden SAS-Programm 4 kann die Berechnung der Allelfrequenzen mit der MKQ-Methode erfolgen.

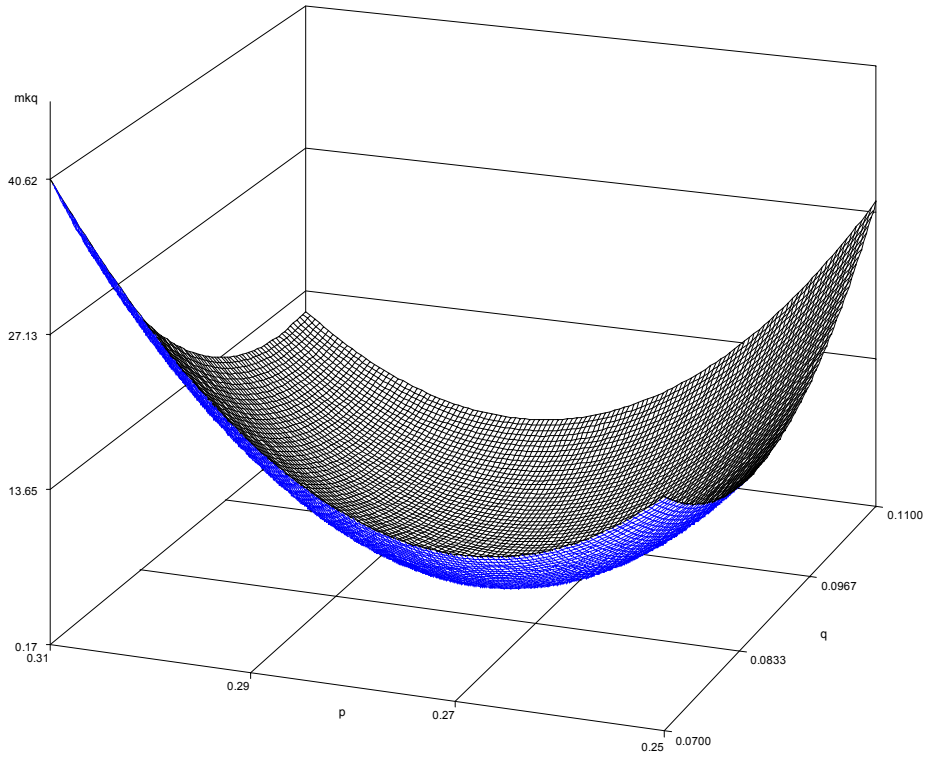

Abbildung 5: mkq(p,q) für das AB0-System mit $n_A = 43$, $n_B = 13$, $n_{AB} = 5$ und $n_0 = 39$

SAS-Programm 4: Allelfrequenzschätzung im AB0-System mittels MKQ-Methode unter Nutzung von CALL NLPTR

```
proc iml;
   start mkq(x) global(n,pn,an,s) ;
      /* Definition des Vererbungsmodells */
     n={43 13 5 39};/* beob. Phänotypenhäufigkeiten A, B, AB und 0 */
     pn=4; /* Anzahl an Phänotypen */
     an=3; /* Allelanzahl  */
     g=j(1,pn,0);
     g[1] = x[1]**2+2*x[1]*(1-x[1]-x[2]);
      g[2] = x[2]**2+2*x[1]*(1-x[1]-x[2]);
     g[3] = 2*x[1]*x[2];g[4] = (1-x[1]-x[2])**2;
     c=j(1,pn,0);
     s=Sum(n);/* Stichprobenumfang */
     c[1]=(n[1]-s*g[1])**2; c[2]=(n[2]-s*g[2])**2;
     c[3]=(n[3]-s*g[3])**2; c[4]=(n[4]-s*g[4])**2;
     L=Sum(c); /*Berechnung der Fehlerquadratsumme */
     return(L);
   finish mkq;
x=J(1,2,1/3); /* Startwert */
```

```
optn = {0 2}; /* optn[1]=0, dann Minimumsuche */
call nlptr(rc,xres,"mkq",x,optn); /* trust-region optimization method
*/ run;quit;
```

Literatur

[1] Landsteiner K.: Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe, Zbl. Bakt. 27, 357-362, 1900

[2] Dunger, E. v.; Hirszfeld, L.: Über Vererbung gruppenspezifischer Strukturen des Blutes II., Z. Immun.-Forsch. 6, 284-292 , 1910

[3] Bernstein, F.: Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstrukturen des Menschen, Klin. Wschr. 3,1495-1497, 1924

[4] Bernstein, F.: Über die Erblichkeit der Blutgruppen, Z. indukt. Abstamm.- und Vererb.-Lehre 54, 400, 1930

[5] Schüler, C.: Über den EM-Algorithmus, Diplomarbeit, Institut für Mathematik der Ernst-Moritz-Arndt-Universität Greifswald, 2012

[6] Excoffier L, Slatkin M.: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. Sep;12(5): 921-7 1995.