Exakte Konfidenzbereiche für die Parameter von Polynomialverteilungen

Bernd Paul Jäger
Ernst-Moritz-Arndt-
Universität Greifswald,
Institut für Biometrie und
Med. Informatik
Walther-Rathenau-Str. 48
17487 Greifswald
bjaeger@biometrie.uni-greifswald.de

Ole Geldschläger
Ernst-Moritz-Arndt-
Universität Greifswald,
Institut für Biometrie und
Med. Informatik
Walther-Rathenau-Str. 48
17487 Greifswald
ogeld@biometrie.uni-greifswald.de

Paul Eberhard Rudolph
ehemals Leibnizinstitut für
Nutztierbiologie (FBN)
Wilhelm-Stahl-Allee 2
18196 Dummerstorf
pe.rudolph@kabelmail.de

Zusammenfassung

Neben einer asymptotischen Methode für Konfidenzbereiche von Trinomialverteilungen werden eine variierte asymptotische und eine exakte Methode besprochen. Die hier formulierten Überlegungen für Trinomialverteilungen sind auch für beliebige Polynomialverteilungen gültig.

Schlüsselwörter: Konfidenzintervalle, Polynomialverteilung

1 Einleitung

Während es für den Parameter p der Binomialverteilung üblich ist exakte Konfidenzintervalle anzugeben, ist das bei Trinomialverteilungen nicht der Fall. Bestenfalls gibt man die auf der n-dimensionalen asymptotischen Normalverteilung beruhenden Ellipsoide an, obwohl man im eindimensionalen Fall - der Binomialverteilung - weiß, dass diese Methode sehr konservativ ist und das Konfidenzniveau nicht ausschöpft.
Neben dieser asymptotischen Methode für Konfidenzbereiche werden hier eine variierte asymptotische und eine exakte Methode besprochen, insbesondere die Vor- und Nachteile dargelegt und für den Anwender eine Empfehlung ausgesprochen.
Wir beschränken uns aus didaktischen Gründen zwar auf die Trinomialverteilung, die Überlegungen sind aber für beliebige Polynomialverteilungen gültig.

2 Die Trinomialverteilung

Aus einer Urne mit weißen, schwarzen und roten Kugeln wird n-mal mit Zurücklegen gezogen. Die Anteile der einzelnen Farben sind:

$$P(W) = p_1,\ P(S) = p_2 \text{ und } P(R) = 1 - p_1 - p_2.$$

Das dreimalige Ziehen kann durch einen Wahrscheinlichkeitsbaum veranschaulicht werden.

Das zufällige Geschehen wird durch eine zweidimensionale Zufallsgröße (W,S) beschrieben, wobei die Anzahl der weißen Kugeln $W = n_1$ zwischen 0 und n sowie die Anzahl der schwarzen Kugeln $S = n_2$ zwischen 0 und $n - n_1$ variiert. Die Anzahl der roten Kugeln ist nicht zufällig, sondern determiniert, denn $n_3 = n - n_1 - n_2$.

Die Wahrscheinlichkeit eines Astes des Wahrscheinlichkeitsbaumes ist $p^{n_1} q^{n_2} (1 - p - q)^{n_3}$, der Polynomialkoeffizient $\binom{n}{n_1 \quad n_2 \quad n_3}$ ist die Anzahl der Äste mit gleichen Anzahlen $n_{w,}$ n_s und n_r . Nach Definition gilt $\binom{n}{n_1 \quad n_2 \quad n_3} = \frac{n!}{n_1!\, n_2!\, n_3!}$.

Definition (Polynomialverteilung):

Eine Zufallsgröße $X = (X_1, X_2, \ldots, X_r)$ heißt polynomialverteilt zu den Parametern n und $p_1, p_2, \ldots, p_r$ mit $0 \le p_i \le 1$, $\sum_{i=1}^{r} p_i = 1$, wenn ihre Wahrscheinlichkeitsfunktion durch

$$P\big((X_1, X_2, \ldots, X_n) = (n_1, n_2, \ldots, n_r)\big) = \binom{n}{n_1\, n_2 \ldots n_r} p_1^{n_1}\, p_2^{n_2} \ldots p_r^{n_r}$$

beschrieben wird.

Die Abb.1 zeigt die Wahrscheinlichkeitsfunktion für die Trinomialverteilung zu den Parametern n = 50, p_1 = 0.3 und p_2 = 0.5. Alle Schnitte mit dem Graphen der Wahrscheinlichkeitsfunktion parallel zur n_1-n_2-Ebene ähneln in etwa Ellipsen. Dass dies für Konfidenzbereiche nicht zutreffen muss, wird im Folgenden erläutert.

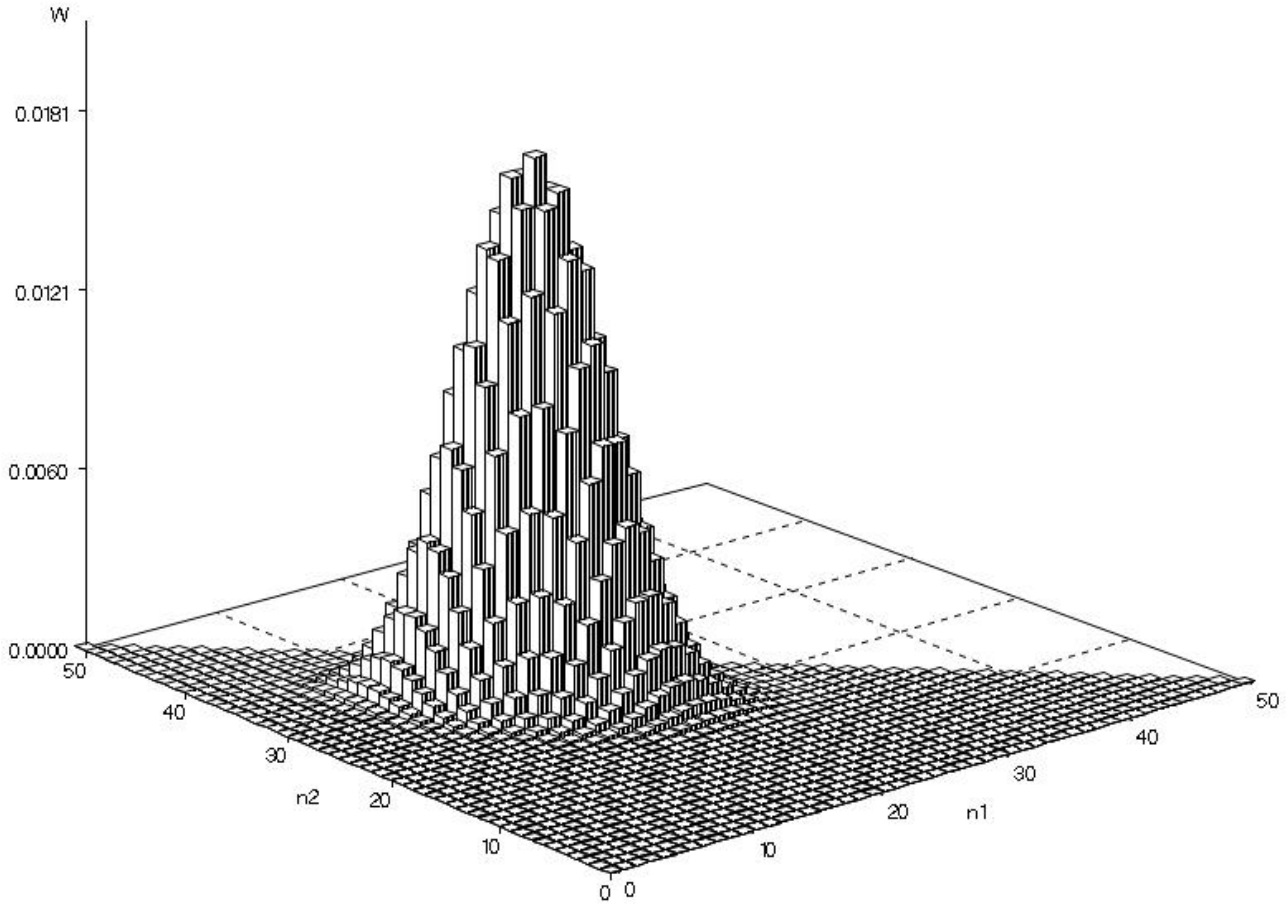

Abbildung 1: Wahrscheinlichkeitsfunktion der Trinomialverteilung zu den Parametern $n = 50$, $p_1 = 0.3$ und $p_2 = 0.5$

3 MLH-Schätzungen der Parameter der Trinomialverteilung

In einer unendlich großen Grundgesamtheit findet man die sich gegenseitig ausschließenden Merkmale *X*, *Y* und *Z* mit den Wahrscheinlichkeiten $p = P(X)$, $q = P(Y)$ und $r = 1 - p - q = P(Z)$. Eine Zufallsstichprobe vom Umfang *n* enthält n_x Elemente aus *X*, , n_y aus *Y* und aus *Z* sind es deterministisch $n_z = n - n_x - n_y$. Wegen $n_x + n_y + n_z = n$ müssen gelten:

$$0 \leq n_x \leq n,\ 0 \leq n_y \leq n - n_x \text{ und } 0 \leq n_z \leq n - n_x - n_y\ .$$

Als Wahrscheinlichkeit für das Tripel (n_x, n_y, n_z) erhält man

$$P(T = (n_x, n_y, n_z)) = \binom{n}{n_x \quad n_y \quad n_z} p^{n_x} q^{n_y} r^{n_z},$$

oder besser, da nur n_x und n_y frei wählbar sind,

$$P_{p,q}(T=(n_x,n_y)) = \binom{n}{n_x \quad n_y \quad n-n_x-n_y} p^{n_x} q^{n_y} (1-p-q)^{n_z}.$$

Als erwartungstreue Maximum-Likelihood-Schätzung (MLH) für die Parameter *p* und *q* erhält man:

$$(\tilde{p},\ \tilde{q}) = \left(n_x/n\,,\ n_y/n\ \right).$$

Die asymptotische Minimalvarianz ist $1/(n \cdot I(p,q))$, wobei *I*(*p*,*q*) als Fisher Information bezeichnet wird. Aus $f = (p, q, 1-p-q)$ erhält man:

$$f_1 = \frac{\partial \ln(f)}{\partial p} = \left(\frac{1}{p}, 0, -\frac{1}{1-p-q}\right),$$

$$f_2 = \frac{\partial \ln(f)}{\partial q} = \left(\ 0, \frac{1}{q}, -\frac{1}{1-p-q}\right)$$

und daraus

$$I(p,q) = \begin{pmatrix} E(f_1^2) & E(f_1 \cdot f_2) \\ E(f_1 \cdot f_2) & E(f_2^2) \end{pmatrix} = \begin{pmatrix} \frac{1}{p} + \frac{1}{1-p-q} & \frac{1}{1-p-q} \\ \frac{1}{1-p-q} & \frac{1}{q} + \frac{1}{1-p-q} \end{pmatrix}.$$

Da die Trinomialverteilung „regulär“ ist, dies wird hier nicht demonstriert, ergibt sich nach der klassischen Schätztheorie die asymptotische Minimalvarianz des Schätzers als

$$V(\tilde{p},\tilde{q}) = 1/(I(p,q) \cdot n).$$

Außerdem ist $(\tilde{p}-p, \tilde{q}-q)\sqrt{I(p,q)\cdot n}$ asymptotisch standardnormalverteilt, folglich das Quadrat

$$n \cdot (\tilde{p}-p, \tilde{q}-q) \cdot I(p,q) \cdot \begin{pmatrix} \tilde{p}-p \\ \tilde{q}-q \end{pmatrix}$$

asymptotisch χ^2-verteilt mit zwei Freiheitsgraden.

4 Drei Methoden zur Konstruktion von Konfidenzbereichen für die Parameter (p, q) der Trinomialverteilung

4.1 Die asymptotische Methode

Die letzte Formel kann verwandt werden, um beispielsweise für $\alpha = 0.05$ einen Konfidenzbereich zu konstruieren. Diejenigen (*p, q*) mit

$$n \cdot (\tilde{p}-p, \tilde{q}-q) \cdot I(p,q) \cdot \begin{pmatrix} \tilde{p}-p \\ \tilde{q}-q \end{pmatrix} = 5.991 = \chi^2_{2,0.95}\ ,$$

begrenzen den asymptotischen Konfidenzbereich. Die Schätzung $(\tilde{p}, \tilde{q})$ und der Stichprobenumfang n sind bekannt. Die Matrizengleichung ergibt eine quadratische Gleichung in p, wenn man q festhält:

$$Ap^2 + B\,p + C = 0,$$

mit

$$A = -q - 5991 * q/n + \tilde{q}^2\ ,$$

$$B = 5.991q/n - 5.991q^2/n - (q - \tilde{q})^2 - 2\tilde{p}q(-1 + \tilde{q})) \text{ und}$$

$$C = \tilde{p}^2(-1 + q)q\ .$$

4.2 Die variierte asymptotische Methode

Ähnlich der Vorgehensweise bei der asymptotischen Methode werden lediglich in der Formel der Fisher-Information an Stelle der Parameter p und q Schätzungen $\tilde{p}$ und $\tilde{q}$ eingesetzt:

$$I(\tilde{p}, \tilde{q}) = \begin{pmatrix} \frac{1}{\tilde{p}} + \frac{1}{1-\tilde{p}-\tilde{q}} & \frac{1}{1-\tilde{p}-\tilde{q}} \\ \frac{1}{1-\tilde{p}-\tilde{q}} & \frac{1}{\tilde{q}} + \frac{1}{1-\tilde{p}-\tilde{q}} \end{pmatrix}.$$

Diejenigen (p, q), für die die Gleichung

$$n \cdot (\tilde{p} - p, \tilde{q} - q) \cdot I(\tilde{p}, \tilde{q}) \cdot \begin{pmatrix} \tilde{p} - p \\ \tilde{q} - q \end{pmatrix} = 5.991 = \chi^2_{2,0.95}$$

gelten, bilden die Grenzen des (1-α)-Konfidenzbereichs. Die Begrenzung ist durch eine ***Ellipse*** gegeben:

$$Mp^2 + N\,p + K = 0,$$

mit

$$M = n(1 - \tilde{q})\tilde{q} + n\tilde{p}\tilde{q}^2 + n\tilde{p}^2(-q^2 + \tilde{q}),$$

$$N = 2n\tilde{p}\tilde{q}(-1 + q) \text{ und}$$

$$K = n\tilde{p}q^2 - n\tilde{p}\tilde{q}q - 5.991(1 - \tilde{p} - \tilde{q})\tilde{q}.$$

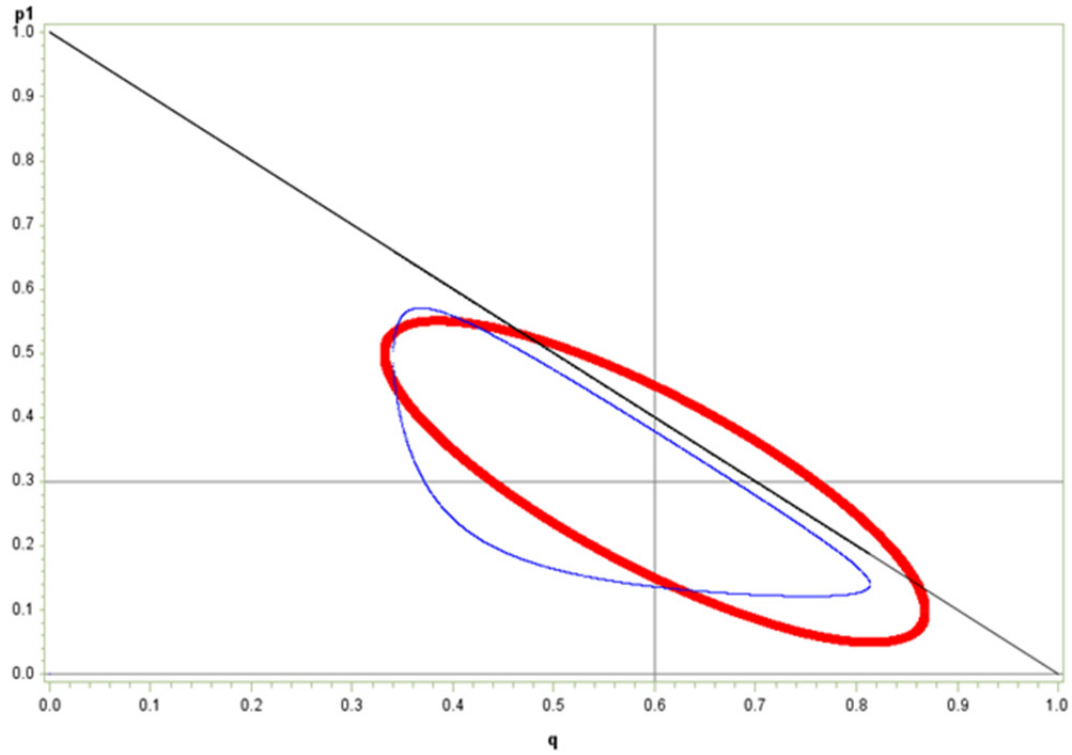

Abbildung 2: Asymptotischer Konfidenzbereich („Hinkelstein“, zarte Linie) und variierter asymptotischer Konfidenzbereich (Ellipse, dicke Linie) für die Parameter p und q der Trinomialverteilung (gesetzte Parameter $\tilde{p} = 0.3$ und $\tilde{q} = 0.6$, $n = 20$)

In den Abbildungen 2 und 3 erkennt man die Unterschiede zwischen dem asymptotischen und dem variierten asymptotischen Konfidenzbereich. Bei kleinen Stichprobenumfängen und Parametern am Rande des Definitionsbereichs verlässt der variierte asymptotische Konfidenzbereich den definierenden Bereich der Zufallsgröße, ein Dreieck, das durch die Punkte (0, 0), (1, 0) und (0, 1) begrenzt wird. Simulationsexperimente haben darüber hinaus gezeigt, dass der asymptotische Konfidenzbereich das Konfidenzniveau näherungsweise einhält, der variierte asymptotische Bereich (die Ellipse) aber nicht (Beitrag Wodny in [1])
Mit größer werdendem Stichprobenumfang nähert sich die Gestalt des asymptotischen Konfidenzbereichs der Ellipse an (siehe Abb. 4 für $n = 100$).

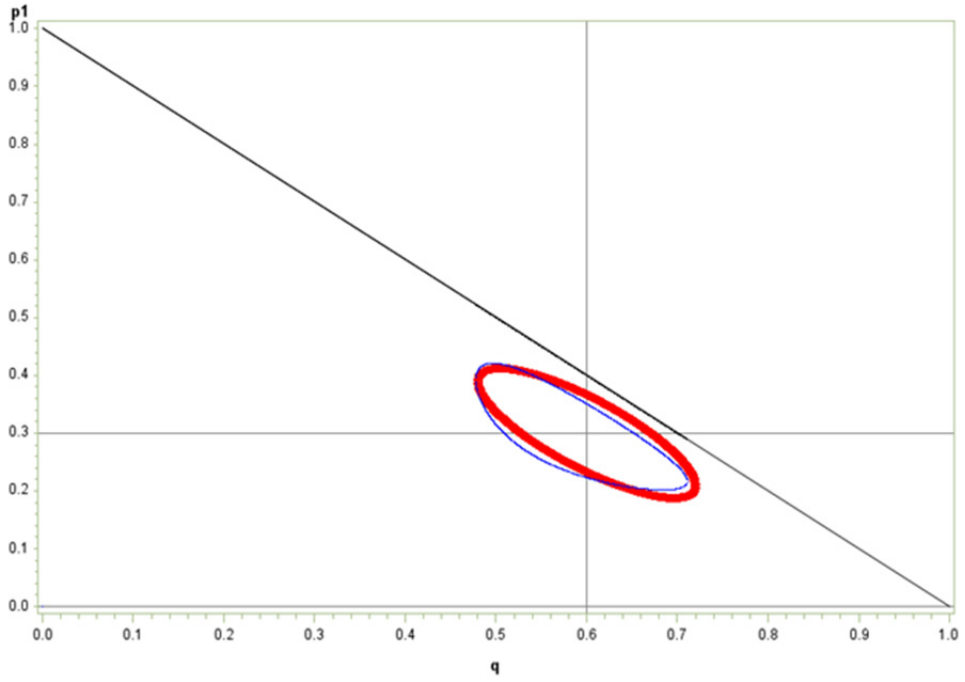

Abbildung 3: Asymptotischer Konfidenzbereich („Hinkelstein“, zarte Linie) und variierter asymptotischer Konfidenzbereich (Ellipse, dicke Linie) für die Parameter p und q der Trinomialverteilung (gesetzte Parameter $\tilde{p} = 0.3$ und $\tilde{q} = 0.6$, $n = 50$)

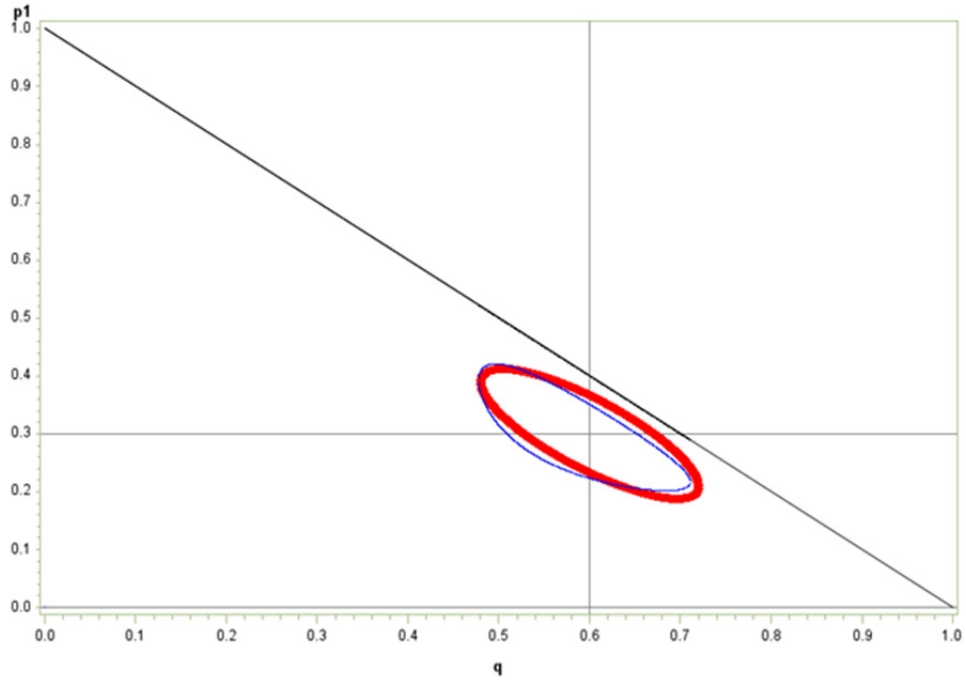

Abbildung 4: Asymptotischer Konfidenzbereich („Hinkelstein“, zarte Linie) und variierter asymptotischer Konfidenzbereich (Ellipse, dicke Linie) für die Parameter p und q der Trinomialverteilung für große Stichprobenumfänge n nahezu gleich
(gesetzte Parameter $\tilde{p} = 0.3$ und $\tilde{q} = 0.6$, $n = 100$)

Tabelle 1: SAS-Programm zur Konstruktion des zweidimensionalen asymptotischen Konfidenzbereichs

```
data KB;
n=60;                /* Stichprobenumfang */
chi2=5.99;           /* krit. Wert der Chi-Quadratverteilung mit f=2 */
pdach=.3;qdach=.3; /* Punktschätzungen */

Do q=0.1 to 0.9 by 0.00005;
A=-q-chi2*q/n+qdach**2;
B=chi2*q/n-chi2*q**2/n-(q-qdach)**2-2*pdach*q*(-1+qdach);
C=pdach**2*(-1+q)*q;
/* Ap²+Bp+C=0 */

if B**2-4*A*C>=0 then do;
p1=(-B-SQRT(B**2-4*A*C))/(2*A);
p2=(-B+SQRT(B**2-4*A*C))/(2*A);end;
else do;p1=.; p2=.;end;
output;
end;
run;

Axis1 order=(0 to 1 by 0.1) label=( "p") WIDTH=2;
Axis2 order=(0 to 1 by 0.1) label=( "q") WIDTH=2;
symbol1 i=join c=green v=point l=1;
symbol2 i=join c=green v=point l=1;
proc gplot data=KB;
plot (p2 p1 )*q/ overlay haxis=axis1 vaxis=axis2 ;
run;quit;
```

Das SAS-Programm zur variierten asymptotischen Methode unterscheidet sich nur durch die jeweilige quadratische Bestimmungsgleichung.

4.3 Die exakte Methode

4.3.1 Der Algorithmus

Die exakte Methode zur Bestimmung des Konfidenzbereichs wird als Algorithmus mitgeteilt. Beobachtet seien in einer Stichprobe vom Umfang n das das Merkmalstripel (x_o, y_o, z_o) mit $z_o = n - x_o - y_o$. Von einem vorgegebenen Punkt (p_0, q_0) soll entschieden werden, ob er zum Konfidenzintervall des Niveaus $1 - \alpha$ gehört.

1. Es werden alle Polynomialwahrscheinlichkeiten $P(x,y)$ berechnet.
2. Wahrscheinlichkeiten werden der Größe nach sortiert.
3. Von der größten Wahrscheinlichkeit an werden alle weiteren Wahrscheinlichkeiten kumuliert, bis die Summe $1 - \alpha$ erstmals übersteigt.
4. Alle in die Summe eingehenden Paare (x, y) erhalten eine Markierung.
5. Trägt (x_o, y_o) diese Markierung, so gehört der Punkt (p_0, q_0) zum Konfidenzbereich.

6. Bei genügend dichter Rasterung des Definitionsbereichs (x, y) erhält man einen zusammenhängenden Bereich, den exakten Konfidenzbereich.

4.3.2 Rechentechnische Realisierung der exakten Methode

Numerische Schwierigkeiten bei der Berechnung der Polynomialwahrscheinlichkeiten ergeben sich zum einen bei der Berechnung der Fakultäten, die rasch sehr groß werden, und die Darstellungsmöglichkeit für eine natürliche Zahl übersteigen, und zum anderen bei der Berechnung der Potenzen der Wahrscheinlichkeiten, die mit wachsendem Exponenten sehr klein werden.
Im SAS-Programm werden die Trinomialwahrscheinlichkeiten als Produkt zweier Binomialwahrscheinlichkeiten dargestellt, die als Standardfunktionen PDF('BINOMIAL',*m*,*p*,*n*) für beliebige *m* und *n* verfügbar sind:

$$P(n_x, n_y) = \binom{n}{n_x \quad n_y \quad n-n_x-n_y} p^{n_x} q^{n_y} (1-p-q)^{n_z}$$

$$= \left(\binom{n}{n_x} p^{n_x} (1-p)^{n-n_x}\right) \cdot \left(\binom{n-n_x}{n_y} q^{n_y} (1-q)^{n-n_x-n_y}\right) \cdot \left(\frac{r^{n_z}}{(1-p)^{n-n_x} \cdot (1-q)^{n_z}}\right)$$

$$= \left(\mathrm{PDF}('\mathrm{BINOMIAL}', n_x, p, n)\right) \cdot \left(\mathrm{PDF}('\mathrm{BINOMIAL}', n_y, q, n-n_x)\right) \cdot \left(\frac{r^{n_z}}{(1-p)^{n-n_x} \cdot (1-q)^{n_z}}\right).$$

Die numerischen Schwierigkeiten sind mit dieser Formeldarstellung nicht vollständig behoben, denn der dritte Faktor wird mit wachsendem Stichprobenumfang rasch klein. Doch während bei der Berechnung der Fakultäten die Probleme etwa ab $n = 180$ anfangen, treten sie bei der Berechnung des Ausgleichsfaktors bei etwa $n = 600$ auf. Für größere n berechne man die Wahrscheinlichkeiten asymptotisch.
Das folgende SAS-Programm prüft, ob der Punkt (p_0, q_0) zum Konfidenzbereich gehört. Die Rechenzeit hängt im Weiteren von der Rasterung der *p*-*q*-Ebene ab. Es ist sinnvoll, nicht den gesamten Definitionsbereich zu untersuchen, sondern beispielsweise die eindimensionalen symmetrischen Konfidenzintervalle für *p* und *q* zu Grunde zu legen. Die Randverteilungen sind bekanntlich Binomialverteilungen: $X \sim \mathrm{B}(n,p)$, $Y \sim \mathrm{B}(n,q)$, sodass deren Konfidenzintervalle leicht bestimmbar sind. Auf ihr Kreuzprodukt kann das Suchgebiet einschränken. Bei größerem Stichprobenumfang *n* ist die Zeitersparnis beträchtlich.

Tabelle 2: Programm prüft, ob ein vorgegebener Punkt zum exakten Konfidenzbereich für die Parameter der Trinomialverteilung gehört

```
%let p0=.3;%let q0=.4;
/* Zu überprüfender Pkt. (p0,q0) */
%let n=20;
/* Stichprobenumfang             */
%let nx=17;%let ny=2;
/* Realisierung durch Stichprobe */
/************** Eingabe abgeschlossen *******/

data aaa;
r0=1-&p0-&q0;
nz=&n-&nx-&ny;
do k=0 to &n;
  do m=0 to &n-k;
    w=PDF('BINOMIAL^',k,&p0,&n)
     *PDF('BINOMIAL^',m,&q0,&n-k)
     *r0**(&n-k-m)/((1-&p0)**(&n-k)*(1-&q0)**(&n-k-m));
/*Polynomial-Wkt. w ist Produkt zweier Binomialwkt. mit
Ausgleichsfaktor*/
    output;
  end;
end;
run;
/* Schritt 1 */

proc sort data=aaa ;
by descending w ;
run;
/* Schritt 2 */

data aaa;
set aaa;
s+w;
/* Schritt 3 */
if s<=0.95 then flag=1;
if s>0.95 and LAG(s)<0.95 then flag=1;/* Schritt 4 */
if k=&nx and m=&ny and flag=1 then Entscheid='ja'; else
Entscheid='ne';
/* Schritt 5 */
run;

proc print;
where k=&nx and m=&ny;run;
```

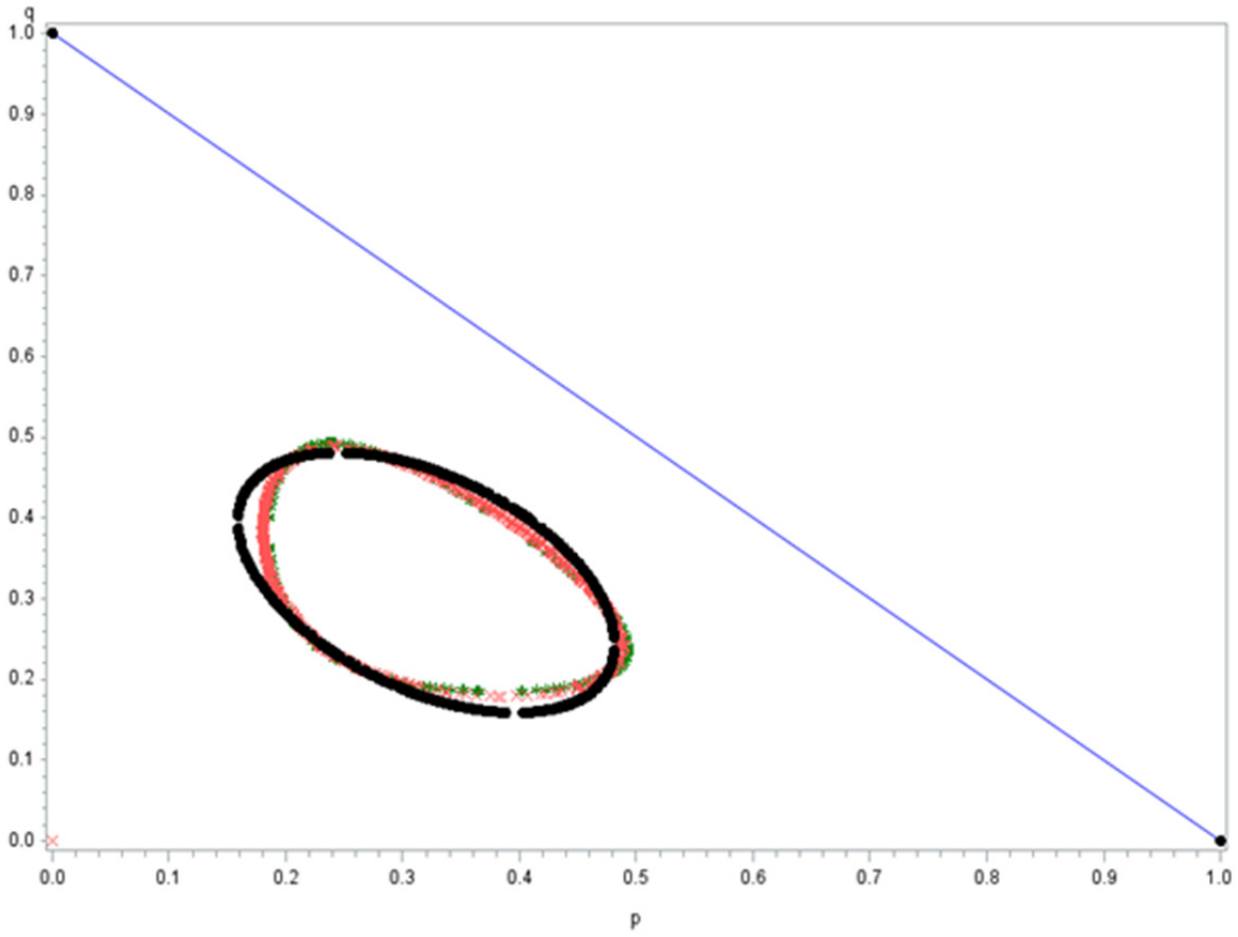

Abbildung 5: Exakter (Sternmarkierung) und asymptotischer Konfidenzbereich („Hinkelstein“, volle Linie) stimmen weitestgehend überein, sowie der variierte asymptotische Bereich (Ellipse) beim Stichprobenumfang $n = 50$ und $n_x = n_y = 15$

In Abb. 5 sind beide asymptotische Bereiche und der exakte Konfidenzbereich eingezeichnet. Man erkennt deutlich, dass bereits ab Stichprobenumfang $n = 50$ der asymptotische Konfidenzbereich gut mit dem exakten übereinstimmt. Man kann nachrechnen, dass der variierte asymptotische Konfidenzbereich im Allgemeinen kein $(1-\alpha)$-Konfidenzbereich ist.

5 Schlußfolgerungen

In Abb. 5 sind die beiden asymptotischen und die exakten Konfidenzbereiche eingezeichnet. Man erkennt deutlich, dass der asymptotische Konfidenzbereich gut mit dem exakten übereinstimmt und das bereits ab Stichprobenumfang $n = 16$. Nur der variierte asymptotische Konfidenzbereich trifft in keiner Weise den tatsächlichen Bereich

Empfehlungen:

- Man wähle niemals den variierten asymptotischen Bereich, weil er zu konservativ ist und Reserven bezüglich der Nichtüberdeckung des wahren Parameters hat.
- Da einerseits die exakte Methode großen Rechenzeitaufwand bedeutet, andererseits exakte und asymptotische Methode gut übereinstimmen, wähle man auf Grund der geringen Rechenzeiten die asymptotische Methode.
- Bei sehr kleinem Stichprobenumfang sollte man den Rechenaufwand der exakten Methode zur Konstruktion des Konfidenzbereichs nicht scheuen, um einen verlässlichen Bereich zu bestimmen.

Literatur

[1] Biebler, K.-E; Jäger, B.; Wodny. M.: Biometrische Methoden der Genomanalyse. Shaker Verlag Aachen, 2013.

[2] Sachs, L.: Angewandte Statistik. Springer Verlag Berlin Heidelberg, 1992.

[3] Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. VEB Deutscher Verlag der Wissenschaft, 1980.

[4] Mattheus, M.: Berechnungen exakter Kondenzbereiche für Polynomialverteilungen, Masterarbeit Universität Greifswald, Institut für Mathematik, 2007.

[5] Beyer, O.; Hackel, H.; Pieper, V.; Jürgen Tiedge, J.: Wahrscheinlichkeitsrechnung und Mathematische Statistik. B. G. Teubner Verlagsgesellschaft Leipzig, 1995.

[6] SAS Institute Inc. (2004) SAS 9.1 Macro Language Reference. Cary, NC: SAS Institute Inc.