Schätzung der Amphibienbestände in einem Teich durch ein Wiederfangverfahren

Von

H. RÜST

Es soll die unbekannte Anzahl N der Individuen einer Amphibienart in einem Teich geschätzt werden. Es werden jedoch lediglich statistische Schätzverfahren ohne Berücksichtigung der spezifischen Fang- und Markiermethoden betrachtet (diese werden in der Arbeit von BLANKENHORN, BURLA, MÜLLER und VILLIGER [7] beschrieben). N wird dabei für die ganze Dauer des Experiments als konstant angenommen. Es werden in einem ersten Arbeitsgang n Individuen gefangen, markiert und wieder in den Teich gesetzt. Später werden in k Stichproben je $g_1, g_2, \ldots, g_k$ Individuen gefangen. Es wird angenommen, dass von den noch nicht gefangenen Individuen jedes dieselbe Chance hat, gefangen zu werden (unabhängig davon, ob markiert oder nicht). In der Stichprobe i werden m_i markierte Individuen gezählt ($i = 1, \ldots, k$).

Wir wollen nun zunächst den Fall $k = 1$ betrachten.

1. Eine Stichprobe

Es werde nur eine Stichprobe vom Umfang g gemacht; darin befinden sich m markierte Individuen.

1.1. Binomialverteilung

Es wird angenommen, dass jedes gefangene Individuum gleich wieder ausgesetzt wird, und dass es beim nächsten Fang die gleiche Wahrscheinlichkeit hat wie die übrigen Individuen, wieder gefangen zu werden. (Diese Voraussetzung ist natürlich unrealistisch. Das Modell ist jedoch wichtig als Approximation des in (1.2.) beschriebenen Verfahrens).

Man erhält als Wahrscheinlichkeit, unter g Individuen gerade m markierte zu fangen:

$$P_B(m \mid N, n, g) = L(N) = \binom{g}{m} \left(\frac{n}{N}\right)^m \left(\frac{N-n}{N}\right)^{g-m}. \tag{1}$$

(1) gibt gleichzeitig die Likelihoodfunktion von N. Wir müssen nun die Schätzung $\hat{N}$ für N so bestimmen, dass

$$L(\hat{N}) = \max_N L(N). \tag{2}$$

Dabei ist $L(N)$ zunächst nur für ganzzahlige N definiert. Doch sieht man leicht, dass man einen Fehler kleiner als 1 macht, wenn man das Maximum der auf der ganzen positiven reellen Zahlenachse definierten Funktion (1) bestimmt. $\hat{N}$ berechnet sich dann aus

$$0 = \frac{d \log L(N)}{dN} = -\frac{m}{N} + \frac{g-m}{N-n} - \frac{g-m}{N};$$

also

$$\hat{N} = \frac{ng}{m}. \tag{3}$$

Der Erwartungswert von $\hat{N}$ existiert natürlich nicht, da $m=0$ mit positiver Wahrscheinlichkeit. BAILEY [1] empfiehlt statt $\hat{N}$

$$\hat{N}_B = \frac{n(g+1)}{m+1} \tag{4}$$

zu verwenden. Man rechnet leicht nach, dass nun

$$E(\hat{N}_B) = N\{1 - P_B(0 \mid N, n, g+1)\}. \tag{5}$$

BAILEY schlägt auch vor, als Schätzung für die Varianz $\sigma^2(\hat{N}_B)$ die folgende Statistik S_B^2 zu verwenden

$$S^2_B = \frac{n^2(g+1)(g-m)}{(m+1)^2(m+2)}. \tag{6}$$

Es ist

$$E(S_B^2) = \sigma^2(\hat{N}_B) + R, \tag{7}$$

$$\begin{aligned} R = {} & N^2\{P_B(0 \mid N, n, g+2) + P_B(1 \mid N, n, g+2) + P_B(0 \mid N, n, g+1)^2 \\ & - 2P_B(0 \mid N, n, g+1)\} \\ = {} & -N^2\left(\frac{N-n}{N}\right)^{g+1}\left\{\frac{N-n}{N} + (g+2)\left(\frac{n}{N}\right) + \left(\frac{N-n}{N}\right)^{g+1} - 2\right\}. \end{aligned} \tag{8}$$

1.2. Hypergeometrische Verteilung

Wir nehmen hier nun an, dass die Individuen erst nach der Stichprobe wieder zurückgelegt werden.

Man erhält als Wahrscheinlichkeit $P_H(m|N,n,g)$ unter den g gefangenen Individuen gerade m markierte zu fangen und gleichzeitig als Likelihoodfunktion $L(N)$:

$$P_H(m \mid N, n, g) = L(N) = \frac{\binom{n}{m}\binom{N-n}{g-m}}{\binom{N}{g}}. \tag{9}$$

Hier soll auch wieder die Maximum-Likelihood-Schätzung $\hat{N}$ für N nach (2) bestimmt werden.

Es ist:
$$\frac{L(N)}{L(N-1)} = \frac{(N-n)(N-g)}{N(N-n-g+m)}. \tag{10}$$

Uns interessieren nur Werte von N, die grösser sind als $n+g-m$. Der Quotient (10) nimmt rechts von $n+g-m$ genau einmal den Wert 1 an, nämlich für

$$\hat{N} = \frac{ng}{m} \tag{11}$$

falls $m>0$. $\hat{N}$ (genauer die nächst kleinere ganze Zahl) ist zugleich die Maximum-Likelihood-Schätzung für N. Für $N<\hat{N}$ bzw. $N>\hat{N}$ ist nämlich der Quotient (10) grösser bzw. kleiner als 1 und somit $L(N)>L(N-1)$ bzw. $L(N)<L(N-1)$. D. h. $L(N)$ nimmt links von $\hat{N}$ monoton zu und rechts monoton ab.

Wir benötigen für spätere Zwecke noch den Wert

$$\overline{N} = \hat{N}\left\{1+\left(1-\frac{n+g-m}{\hat{N}}\right)^{1/2}\right\}, \tag{12}$$

wo der Quotient (10) sein Minimum annimmt (rechts von $n+g-m$). Zwischen $n+g-m$ und $\overline{N}$ ist der Quotient (10) eine monoton fallende Funktion von N.

Auch hier wieder hat man sehr grosse Mühe, Erwartungswert und Streuung zu berechnen. Da zudem $E(\hat{N}|m>0)$ einen beachtlichen Bias besitzt, verwendet man besser die von CHAPMAN [2] vorgeschlagene Schätzung

$$\hat{N}_{II} = \frac{(n+1)(g+1)}{m+1} - 1 \tag{13}$$

mit
$$E(\hat{N}_H) = N-(N+1)\,P_H(0 \mid N+1, n+1, g+1). \tag{14}$$

Als approximativ biasfreie Schätzung für die Varianz $\sigma^2(\hat{N}_H)$ kann die folgende Statistik S_{II}^2 verwendet werden:

$$S_{II}^2 = \frac{(n+1)(g+1)(g-m)(n-m)}{(m+1)^2(m+2)}. \tag{15}$$

Es ist dann
$$E(S_H^2) = \sigma^2(\hat{N}_H) + R_{II}, \tag{16}$$

wobei

$$\begin{aligned} R_H = (N+1)\{&-2(N+1)\,P_H(0 \mid N+1, n+1, g+1) \\ &+(N+1)\,P_H(0 \mid N+1, n+1, g+1)^2+(N+2)\,P_{II}(0 \mid N+2, n+2, g+2) \\ &+(N+2)\,P_H(1 \mid N+2, n+2, g+2)-P_H(0 \mid N+1, n+1, g+1)\}. \end{aligned} \tag{17}$$

1.3. Vergleich der beiden Verfahren

BAILEY [1] empfiehlt, falls g klein ist gegen N, das Modell mit der hypergeometrischen Verteilung (9) durch das mit der Binomialverteilung (1) zu ersetzen.

Da die empfohlenen Schätzgrössen (4) und (13) praktisch gleich sind und man mit der Schätzung (6) statt (15) für die Streuung auf der sicheren Seite bleibt, ist gegen dieses Verfahren nicht allzuviel einzuwenden.

Interessant sind auch noch die in BAILEY [1] ausführlich beschriebenen Schätzgrössen für $1/N$ bzw. für N bei inverser Stichprobe. In beiden Fällen tritt die zufällige Variable im Zähler und nicht mehr im Nenner der Schätzgrösse auf. Damit kann ihre Verteilungsfunktion besser überblickt werden.

Bei der inversen Stichprobe wird der Fangprozess solange fortgesetzt, bis eine vorgegebene Anzahl m markierter Individuen vorhanden ist. Es ist nun also g die zufällige Variable.

Bei der praktischen Durchführung können hier natürlich Schwierigkeiten auftreten, da es oft schwierig sein dürfte, die vorgegebene (allerdings zu Beginn frei wählbare) Anzahl m markierter Individuen wiederzufangen.

Ähnliche Schwierigkeiten können natürlich auch bereits bei den früheren Verfahren auftreten, da ja auch dort die Anzahl g der zu fangenden Individuen vorher festzulegen ist.

DARROCH [3] stellt ein Modell auf, in welchem sowohl m wie auch g als Zufallsvariable betrachtet werden. Er zeigt, dass dann auch $\hat{N}_H$ und S_H^2 als Schätzgrössen für N und $\sigma^2(\hat{N}_H)$ verwendet werden können.

2. Mehrere Stichproben

Es werden also in den Stichproben $i=1,\ldots,k$ je g_i Individuen gefangen, von denen m_i markiert sind.

Wir werden wieder Maximum-Likelihood-Schätzungen für den Fall der Binomial- bzw. hypergeometrischen Verteilung untersuchen.

Natürlich wäre vor allem auch das Verfahren interessant, bei welchem die Individuen sämtlicher Stichproben markiert werden, bevor sie wieder ausgesetzt werden. Schätzverfahren dafür sind jedoch in der Arbeit von DARROCH [3] ausführlich beschrieben.

2.1. Binomialverteilungen

Jedes gefangene Individuum wird sofort wieder ausgesetzt. Die Einteilung in die einzelnen Stichproben ist – wenigstens von der mathematischen Analyse her betrachtet – rein willkürlich und bringt keine zusätzliche Information. Die Summe der m_i ist suffizient für die Schätzung von N. Wir können diesen Fall somit genau gleich wie in (1.1) behandeln, wobei $g=g_1+g_2+\ldots+g_k$ und $m=m_1+m_2+\ldots+m_k$ zu setzen sind.

2.2. Hypergeometrische Verteilungen

Als für die mathematische Analyse kompliziertester Fall erweist sich nun der in der praktischen Durchführung einfachste. Die Individuen werden lediglich beim ersten Arbeitsgang markiert und bei den nachfolgenden Stichproben nur gezählt und am Ende jeder Stichprobe wieder ausgesetzt.

Wir erhalten als Wahrscheinlichkeit, jeweils $m_1, m_2, \ldots, m_k$ markierte unter $g_1, g_2, \ldots, g_k$ gefangenen Individuen zu finden und zugleich als Likelihoodfunktion:

$$L(N) = \prod_{i=1}^{k} P_H(m_i \mid N, n, g_i) \tag{18}$$

(siehe (9)).

Als Schätzungen interessieren uns nun natürlich nur Werte, die grösser sind als $\max_{i=1,\ldots,k} (n+g_i-m_i)$.

Um das Maximum der Likelihoodfunktion in diesem Intervall zu finden, untersuchen wir wieder den Quotienten

$$Q(N) = \frac{L(N)}{L(N-1)} = \prod_{i=1}^{k} \frac{(N-n)(N-g_i)}{N(N-n-g_i+m_i)}, \tag{19}$$

wobei wir die Voraussetzung treffen, dass nicht alle m_i verschwinden. Wir definieren noch die folgenden vier zufälligen Variablen

$$\begin{aligned} N_1 &= \max_{i=1,\ldots,k} (n+g_i-m_i), \\ N_2 &= \min_{i=1,\ldots,k} \frac{n\,g_i}{m_i}, \\ N_3 &= \max_{i=1,\ldots,k} \frac{n\,g_i}{m_i}, \\ N_4 &= \min_{i=1,\ldots,k} \left\{ \frac{n\,g_i}{m_i} \left[1+\left(1-\frac{(n+g_i-m_i)\,m_i}{n\,g_i} \right)^{1/2} \right] \right\}. \end{aligned} \tag{20}$$

Wir betrachten nun $Q(N)$ gemäss (19) als Funktion auf dem ganzen reellen Intervall $N>N_1$.

Satz 1: Falls

$$N_1 < N_2, \tag{21}$$

$$N_3 \leqq N_4 \tag{22}$$

und

$$\sum_{i=1}^{k} m_i > 0, \tag{23}$$

so besitzt die Gleichung

$$Q(N) = \prod_{i=1}^{k} \frac{(N-n)(N-g_i)}{N(N-n-g_i+m_i)} = 1 \tag{24}$$

genau eine Lösung $\hat{N}>N_1$. $\hat{N}$ bzw. die nächstkleinere ganze Zahl ist dann auch die Maximum-Likelihood-Schätzung für N.

Beweis: $Q(N)$ ist stetig für $N>N_1$. Aus (1.2.) folgt: $Q(N_2)\geqq 1$ und $Q(N_3)\leqq 1$. (Da N_4 wegen (23) endlich ist, ist wegen (22) auch N_3 endlich.) Somit nimmt $Q(N)$ zwischen N_2 und N_3 mindestens einmal den Wert 1 an. Zwischen N_1 und N_4 ist

$Q(N)$ eine monoton fallende Funktion (siehe (12)). Da zudem $Q(N)>1$ für $N_1<N<N_2$ und $Q(N)<1$ für $N>N_3$, gibt es genau einen Wert $\hat{N}>N_1$, so dass $Q(\hat{N})=1$. $L(N)$ ist dann monoton wachsend für $N_1<N<\hat{N}$ und monoton fallend für $N>\hat{N}$. Daraus folgt die Behauptung.

In den meisten praktischen Beispielen werden die Voraussetzungen (21), (22) und (23) wohl erfüllt sein (wie wir beim Beweis von Satz 2 noch sehen werden). Da sie zudem lediglich hinreichend für die Gültigkeit der Aussagen von Satz 1 sind, und bei ihrer Herleitung ziemlich grobe Abschätzungen verwendet wurden, kann wohl in praktisch allen Fällen die Maximum-Likelihood-Schätzung aus (24) gewonnen werden.

Multipliziert man die Gleichung (24) mit dem Nenner von $Q(N)$, so muss $\hat{N}$ als Nullstelle eines Polynoms vom Grad $2k-1$ bestimmt werden. Für $k>2$ lässt sich somit die Lösung, falls man sie nicht errät, nicht angeben. Dies ist mir leider nicht gelungen; und auch für $k=2$ musste ich das Verfahren wegen des grossen Rechenaufwandes vorzeitig abbrechen.

Hingegen liesse sich die Lösung durch ein Iterationsverfahren mit genügender Genauigkeit approximieren.

2.3. Asymptotische Eigenschaften von $\hat{N}$

Im Abschnitt 3 werden wir eine für praktische Zwecke besser geeignete Schätzgrösse untersuchen. Wir benötigen jedoch für Vergleichszwecke noch einige asymptotische Eigenschaften (für $N\to\infty$) der Maximum-Likelihood-Schätzung $\hat{N}$ (siehe Satz 1).

Für die Beschreibung des asymptotischen Verhaltens einer (in der Regel komplizierten) Funktion $\varphi(N)$ von N werden wir folgende Abkürzung verwenden:

$$\varphi(N)=0\,(f(N)). \tag{25}$$

Dabei wird $f(N)$ meist eine wesentlich einfachere Funktion sein (etwa $f(N)=N^p$). (25) bedeutet, dass eine von N unabhängige Konstante $c>0$ existiert, so dass

$$\left|\frac{\varphi(N)}{f(N)}\right|<c, \text{ für alle } N>0. \tag{26}$$

(Falls $f(N)=0$ für endlich viele Werte von N, kann man auch noch zulassen, dass (26) dort nicht erfüllt ist.)

Durch (18) ist im Raum der Gitterpunkte $(m_1,m_2,\ldots,m_k)$ eine Wahrscheinlichkeitsverteilung definiert, die wir mit P bezeichnen werden. Wir betrachten $\hat{N}$ als Funktion

$$\hat{N}=\hat{N}(m_1,m_2,\ldots,m_k\mid n,g_1,g_2,\ldots,g_k) \tag{27}$$

(oder abgekürzt $\hat{N}(m_1,m_2,\ldots,m_k)$) von $m_1,m_2,\ldots,m_k$ mit den Parametern $n,g_1,g_2,\ldots,g_k$. $\hat{N}$ ist eine zufällige Variable, deren Verteilungsfunktion von $N,g_1,\ldots,g_k$ und n abhängt.

Satz 2: Geht mit $N \to \infty$ auch $n \to \infty$ und $g_i \to \infty$ derart, dass ausser für endlich viele N die Bedingungen

$$\begin{aligned} 0 < \alpha_1 \leqq \frac{n}{N} \leqq \alpha_2 < 1, \\ 0 < \beta_1 \leqq \frac{g_i}{N} \leqq \beta_2 < 1, \quad \text{für } i = 1, \ldots, k \end{aligned} \tag{28}$$

erfüllt sind, so gibt es zu jedem δ mit $0 \leqq \delta \leqq 1$ eine Folge von Gebieten $G_{N,\delta}$ im Raum der Gitterpunkte $(m_1, m_2, \ldots, m_k)$, für die gilt:

$$P(G_{N,\delta}) = 1 - 0(N^{-\delta}), \tag{29}$$

$$\hat{N}(m_1, \ldots, m_k \mid n, g_1, \ldots, g_k) = N - \sum_{i=1}^{k} \frac{N^2}{n(N-g_i)} \left(\sum_{j=1}^{k} \frac{g_j}{N-g_j} \right)^{-1} (m_i - \mu_i) + 0(N^{\delta}) \tag{30}$$

für $(m_1, \ldots, m_k) \in G_{N,\delta}$, wobei

$$\mu_i = E(m_i) = \frac{n g_i}{N}. \tag{31}$$

Beweis: Sei G das Gebiet der reellen Punkte $(m_1, \ldots, m_k)$, die die Voraussetzungen (21), (22) und (23) erfüllen. $\hat{N}(m_1, \ldots, m_k)$ lässt sich dann mit Hilfe von (24) unmittelbar auf ganz G definieren. $\hat{N}$ ist auf G stetig und beliebig oft differenzierbar.

$(\mu_1, \ldots, \mu_k)$ liegt in G. Da $Q(N) = 1$ für $m_i = \mu_i$ ist $\hat{N}(\mu_1, \ldots, \mu_k) = N$.

Differenziert man beide Seiten von (24) partiell nach m_j und löst nach $\frac{\partial \hat{N}}{\partial m_j}$ auf, so erhält man für Punkte in G:

$$\frac{\partial \hat{N}}{\partial m_j} = \frac{1}{\hat{N} - n - g_j + m_j} \left\{ \sum_{i=1}^{k} \left(\frac{1}{\hat{N}-n} + \frac{1}{\hat{N}-g_i} - \frac{1}{\hat{N}} - \frac{1}{\hat{N}-n-g_i+m_i} \right) \right\}^{-1}. \tag{32}$$

Setzt man in (32) $m_i = \mu_i$ und verwendet, dass dann $\hat{N} = N$, so ergibt sich der Koeffizient des linearen Terms in (30).

Bildet man mit Hilfe von (32) die zweite Ableitung von $\hat{N}$ in G, so sieht man, dass

$$\frac{\partial^2 N}{\partial m_i \, \partial m_j} = 0\left(\frac{1}{N}\right) \text{ solange } \frac{1}{\hat{N} - n - g_h + m_h} = 0\left(\frac{1}{N}\right) \tag{33}$$

für $h = 1, \ldots, k$.

Wir konstruieren nun $G_{N,\delta}$ so, dass die Bedingungen (21), (22), (23), (33) und

$$m_i - \mu_i = 0(N^{\frac{1}{2}+\frac{\delta}{2}}), \quad i = 1, \ldots, k, \tag{34}$$

erfüllt sind. Aus der Verallgemeinerung des Taylorschen Satzes auf mehrere Dimensionen und der Lagrangeschen Restgliedformel folgt dann (30).

Sei $$G_N^1 = \left\{(m_1, \ldots, m_k) \,\middle|\, \frac{n\,g_i}{m_i} - n - g_j + m_j > c_1 N \text{ für alle } i,j\right\}. \tag{35}$$

In G_N^1 sind (21) und (33) erfüllt. $\frac{1}{N}\left(\frac{n\,g_i}{m_i} - n - g_j + m_j\right)$ ist asymptotisch äquivalent einer zufälligen Variablen mit Erwartungswert $\mu_{ij} = \frac{1}{N}\left(N - n - g_j + \frac{n\,g_j}{N}\right) = \frac{1}{N^2}(N-n)(N-g_j)$ und Varianz $\sigma_{ij}^2 = 0\left(\frac{1}{N}\right)$. Daraus folgt mit Hilfe der Tschebyscheffschen Ungleichung

$$P(G_N^1)^c \leqq \sum_{i,j}' P\left\{\left|\frac{1}{N}\left(\frac{n\,g_i}{m_i} - n - g_j + m_j\right) - \mu_{ij}\right| > \mu_{ij} - c_1\right\}$$

$$\leqq \frac{\sigma_{ij}^2}{(\mu_{ij} - c_1)^2} = 0\left(\frac{1}{N}\right), \text{ sofern } \mu_{ij} > c_1.$$

D. h. $$P(G_N^1) = 1 - 0\left(\frac{1}{N}\right), \text{ wenn } c_1 < (1-\alpha_2)(1-\beta_2).$$

Weiter sei $$G_{N,\delta}^2 = \{(m_1, \ldots, m_k) \mid |m_i - \mu_i| < c_2 N^{1/2(1+\delta)} \text{ für alle } i\}. \tag{36}$$

Ebenfalls aus der Tschebyscheffschen Ungleichung folgt $P(G_{N,\delta}^2) = 1 - 0(N^{-\delta})$. In $G_{N,\delta}^2$ ist (34) und bei geeigneter Wahl von c_2 auch (23) erfüllt. Wir wählen nun

$$G_{N,\delta} = G_N^1 \cap G_{N,\delta}^2. \tag{37}$$

Aus (20) und (35) folgt dann für $(m_1, \ldots, m_k) \in G_{N,\delta}$:

$$N_4 \geqq \min_{i=1,\ldots,k}\left\{\frac{n\,g_i}{m_i} + \left(\frac{n\,g_i}{m_i} - n - g_i + m_i\right)\right\} \geqq N_2 + c_1 N.$$

Da jedoch wegen (36) auch $N_3 - N_2 = 0(N^{\frac{1}{2}+\frac{\delta}{2}})$ in $G_{N,\delta}$, so lassen sich für genügend grosses N, c_1 und c_2 so bestimmen, dass $N_3 = N_2 + 0(N^{\frac{1}{2}+\frac{\delta}{2}}) \leqq N_2 + c_1 N \leqq N_4$. Natürlich ist auch $P(G_{N,\delta}) = P(G_N^1 \cap G_{N,\delta}^2) = 1 - 0(N^{-\delta})$. (29) ist somit auch erfüllt.

Satz 3: Geht mit $N \to \infty$ auch $n \to \infty$ und $g_i \to \infty$ derart, dass

$$\lim_{N\to\infty} \frac{n}{N} = \eta, \qquad \lim_{N\to\infty} \frac{g_i}{N} = \gamma_i \tag{38}$$

mit $0 < \eta < 1$ und $0 < \gamma_i < 1$ für $i = 1, \ldots, k$, so konvergiert die Folge der zufälligen Variablen

$$Y_N = \frac{\hat{N} - N}{\sqrt{N}} \tag{39}$$

stochastisch gegen eine normal verteilte Variable Y mit Mittelwert Null und Varianz

$$\text{var}(Y) = \frac{1-\eta}{\eta}\left(\sum_{i=1}^{k} \frac{\gamma_i}{1-\gamma_i}\right)^{-1}. \tag{40}$$

D. h. zu jedem $\epsilon_1>0$ und $\epsilon_2>0$ existiert ein N_0, so dass für jedes $N>N_0$

$$P\{|Y_N-Y|>\epsilon_1\}<\epsilon_2. \tag{41}$$

Korollar: Unter den Voraussetzungen von Satz 3 gilt auch, dass die Verteilungsfunktion von Y_N punktweise gegen die Verteilungsfunktion von Y, also die Normalverteilung mit Mittelwert Null und Varianz (40), konvergiert. (Dies folgt aus Satz 2.10. Seite 99 in KRICKEBERG [5].)

Beweis: Wir setzen $0<\delta<\frac{1}{2}$. Dann folgt (41) aus Satz 2, wenn wir zunächst statt Y_N die Variable

$$\tilde{Y}_N=-\sum_{i=1}^{k}\frac{N^{3/2}}{n\,(N-g_i)}\left(\sum_{j=1}^{k}\frac{g_j}{N-g_j}\right)^{-1}(m_i-\mu_i) \tag{42}$$

einsetzen. Die Folge $\tilde{Y}_N$ ist eine Cauchy-Folge im Sinne der stochastischen Konvergenz, d. h.:

$$\lim_{N,M\to\infty}P\{|\tilde{Y}_N-\tilde{Y}_M|\geqq\epsilon\}=0 \text{ für alle } \epsilon>0.$$

Daraus folgt die Existenz einer Variablen Y als Grenzwert von $\tilde{Y}_N$ und somit von Y_N. Die Verteilungsfunktion der Variablen $\frac{m_i-\mu_i}{\sqrt{N}}$ strebt gegen eine Normalverteilung (siehe VAN DER WAERDEN [6] § 8, Seite 36ff.). Der Mittelwert von $\tilde{Y}_N$ ist Null und die Varianz

$$\operatorname{var}(\tilde{Y}_N)=\sum_{i=1}^{k}\left\{\frac{N^{3/2}}{n\,(N-g_i)}\left(\sum_{j=1}^{k}\frac{g_j}{N-g_j}\right)^{-1}\right\}^2\operatorname{var}(m_i)=\frac{N\,(N-n)}{n\,(N-1)}\left(\sum_{j=1}^{k}\frac{g_j}{N-g_j}\right)^{-1}$$

strebt gegen den Ausdruck (40). Daraus folgt die Behauptung.

3. Praktische Schätzgrösse für N

Die Sätze 2 und 3 sind nun leider bei der praktischen Berechnung der Maximum-Likelihood-Schätzung $\hat{N}$ nicht verwendbar. Sie ergeben aber, sobald eine Schätzung für N vorliegt, Approximationen für die Verteilungsfunktion von $\hat{N}$. Wir müssten also zusätzlich noch ein Iterationsverfahren zur Auflösung von (24) aufstellen. Es soll jedoch zunächst in diesem Abschnitt ein wesentlich einfacheres Verfahren untersucht und mit der Maximum-Likelihood-Schätzung verglichen werden.

3.1. Die Schätzgrösse von Schumacher und Eschmeyer

SCHUMACHER und ESCHMEYER schlagen in ihrer Arbeit [4] für den in (2.2.) betrachteten Fall die Schätzgrösse

$$\hat{N}_P=\frac{n\,g}{m}\quad\text{mit}\quad g=\frac{1}{k}\sum_{i=1}^{k}g_i,\quad m=\frac{1}{k}\sum_{i=1}^{k}m_i \tag{43}$$

vor. Sie geben auch eine Schätzgrösse S_P^2 für die Varianz von $\hat{N}_P$ an, nämlich:

$$S_P^2 = \frac{\hat{N}_P^4}{n^2} \frac{1}{k-1} \sum_{i=1}^{k} \frac{g_i}{k\,g} \left(\frac{m_i}{g_i} - \frac{n}{\hat{N}_P} \right)^2. \tag{44}$$

(Die Formeln wurden allerdings für einen allgemeineren Fall hergeleitet. Für unsern Spezialfall ergeben sich jedoch nach einigen Umformungen die Ausdrücke (43) und (44).)

Dabei steht der erste Faktor offensichtlich als Schätzung für

$$\left(\frac{d\hat{N}_P}{d\left(\frac{m}{g}\right)} \right)^2_{\frac{m}{g} = E\left(\frac{m}{g}\right) = \frac{n}{N}} = \frac{N^4}{n^2} \tag{45}$$

und das gewichtete quadratische Mittel als Schätzung für die Varianz von m/g. Dafür kann man aber unter Ausnutzung des bekannten Ausdrucks für die Varianzen der hypergeometrisch verteilten zufälligen Variablen m_i bessere Schätzungen verwenden (siehe Satz 5). (44) wird insbesondere für kleine k schlechte Ergebnisse liefern. Für $k=1$ ist die Formel (44) gar nicht anwendbar.

Satz 4: Wenn die Voraussetzungen von Satz 2 erfüllt sind, so gibt es ebenfalls zu jedem δ mit $0 \leqq \delta \leqq 1$ eine Folge von Gebieten $G_{N,\delta}$ im Raum der Gitterpunkte, für die gilt:

$$P(G_{N,\delta}) = 1 - 0\,(N^{-\delta}) \tag{46}$$

und

$$\hat{N}_P(m_1, \ldots, m_k \mid n, g_1, \ldots, g_k) = \frac{n\,g}{m} = N - \frac{N^2}{n\,g}(m-\mu) + \frac{N^3}{n^2\,g^2}(m-\mu)^2 + 0\,(N^{1/2(3\delta-1)}) \tag{47}$$

für $(m_1, \ldots, m_k) \in G_{N,\delta}$, wobei $\mu = E(m) = \frac{n\,g}{N}$. ($m, g$ siehe (43)).

Beweis: Man wählt $G_{N,\delta}$ so, dass für ein $c>0$

$$G_{N,\delta} = \{(m_1, \ldots, m_k) \mid |m-\mu| \leqq c\,N^{1/2(1+\delta)}\}. \tag{48}$$

Der Beweis verläuft dann ähnlich wie bei Satz 2.

Satz 5: Unter den Voraussetzungen von Satz 3 konvergiert die Folge der zufälligen Variabeln

$$Y_N^P = \frac{\hat{N}_P - N}{\sqrt{N}} \tag{49}$$

stochastisch gegen eine normal verteilte Variable Y^P mit Mittelwert Null und Varianz

$$\operatorname{var}(Y^P) = \frac{1-\eta}{\eta} \left(\frac{1}{k\,\gamma} - \frac{1}{k^2\,\gamma^2} \sum_{i=1}^{k} \gamma_j^2 \right). \tag{50}$$

Gleichzeitig konvergieren die Variablen

$$Z_N = \hat{N}_P - N + \frac{N^2}{n\,g}(m-\mu) \tag{51}$$

stochastisch gegen eine Variable Z mit Mittelwert

$$\beta = E(Z) = \frac{1-\eta}{\eta}\left\{\frac{1}{k\,\gamma} - \frac{1}{k^2\,\gamma^2}\sum_{i=1}^{k}\gamma_i^2\right\} \tag{52}$$

und var $(Z) = 0(1)$. Dabei ist γ in (50) und (52) eine Abkürzung für das arithmetische Mittel der γ_i.

Beweis: Der Beweis wird analog wie für Satz 3 geführt unter Benützung des Ausdrucks:

$$E(m-\mu)^2 = \text{var}\,(m) = \frac{1}{k^2}\sum_{i=1}^{k}\text{var}\,(m_i) = \frac{1}{k^2}\sum_{i=1}^{k}\frac{n\,(N-n)\,g_i\,(N-g_i)}{N^2\,(N-1)}. \tag{53}$$

Natürlich gilt auch hier analog dem Korollar zu Satz 3, dass die Verteilungsfunktionen der Y_N^P punktweise gegen die Verteilungsfunktion von Y konvergieren. β in (52) ist der asymptotische Wert (für grosse N) für den Bias, der bei Verwendung der Schätzgrösse $\hat{N}_P$ zu erwarten ist. Man kann deshalb die folgende verbesserte Schätzung für N verwenden:

$$\hat{N}_u = \hat{N}_P - \frac{\hat{N}_P - n}{n}\left\{\frac{\hat{N}_P}{k\,g} - \frac{1}{k^2\,g^2}\sum_{i=1}^{k}g_i^2\right\}. \tag{54}$$

Diesen Wert wird man dann in der Schätzgrösse für die asymptotische Varianz von $\hat{N}_u$ bzw. $\hat{N}_P$ einsetzen. Man erhält dann aus Satz 5 und (54)

$$\text{var}\,(\hat{N}_u) \sim \text{var}\,(\hat{N}_P) \sim \frac{\hat{N}_u\,(\hat{N}_u - n)}{n}\left\{\frac{\hat{N}_u}{k\,g} - \frac{1}{k^2\,g^2}\sum_{i=1}^{k}g_i^2\right\}. \tag{55}$$

Damit hätten wir ein wesentlich einfacheres Schätzverfahren gefunden, das in der Praxis der Maximum-Likelihood-Schätzung $\hat{N}$ sicher vorgezogen wird, falls es nicht wesentlich unzuverlässiger ist. Diese Frage soll jedoch im nächsten Abschnitt noch untersucht werden.

3.2. Vergleich mit der Maximum-Likelihood-Schätzung N

Zum Vergleich der Güte der beiden Schätzverfahren verwenden wir die asymptotischen Varianzen von

$$Y_N = \frac{\hat{N}-N}{\sqrt{N}} \quad \text{und} \quad Y_N^P = \frac{\hat{N}_P - N}{\sqrt{N}} \quad \text{(siehe Satz 3 und 5).}$$

Satz 6: Unter den Voraussetzungen von Satz 2 gilt: Falls alle γ_i gleich sind, so haben Y_N und Y_N^P dieselbe asymptotische Varianz, d. h.:

$$\text{var}(Y) = \text{var}(Y^P) = \frac{1}{k}\,\frac{1-\eta}{\eta}\,\frac{1-\gamma}{\gamma}, \quad \text{für } \gamma_i = \gamma. \tag{56}$$

Sind die relativen Abweichungen $\frac{\gamma_i-\gamma}{\gamma}$ vom arithmetischen Mittel γ der γ_i so klein, dass ihre dritten und höheren Potenzen vernachlässigt werden können, so gilt

$$\frac{\text{var}(Y^P)-\text{var}(Y)}{\text{var}(Y)} \approx \left(\frac{\gamma}{1-\gamma}\right)^2 \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\gamma_i-\gamma}{\gamma}\right)^2. \tag{57}$$

Beweis: (56) folgt unmittelbar durch Einsetzen in (40) bzw. (50). Wir benützen die folgenden Abkürzungen:

$$\Delta_i = \frac{\gamma_i-\gamma}{\gamma}, \qquad Q = \sum_{i=1}^{k}\Delta_i^2. \tag{58}$$

Es ist dann:

$$\gamma_i = (1+\Delta_i)\,\gamma \quad \text{und} \quad \sum_{i=1}^{k}\Delta_i = 0. \tag{59}$$

Setzt man (59) in (40) und (50) ein, entwickelt nach Potenzen von Δ_i und vernachlässigt alle höheren als die zweiten Potenzen, so erhält man:

$$\begin{aligned} \text{var}(Y) &\approx \frac{1-\eta}{\eta}\left\{\frac{1-\gamma}{k\,\gamma} - \frac{Q}{k^2(1-\gamma)}\right\}, \\ \text{var}(Y^P) &\approx \frac{1-\eta}{\eta}\left\{\frac{1-\gamma}{k\,\gamma} - \frac{Q}{k^2}\right\}. \end{aligned} \tag{60}$$

Setzt man die Ausdrücke (60) auf der linken Seite der Formel (57) ein, entwickelt nach Potenzen von Q, wobei jedoch nur die erste Potenz berücksichtigt wird, so folgt (57) nach Einsetzen von (58).

Man wird deshalb wohl erst bei grösseren Unterschieden in den Stichprobenumfängen die Likelihood-Schätzung $\hat{N}$ in Betracht ziehen und versuchen, die Gleichung (24) zu lösen. Im Normalfall wird man jedoch, nachdem man sich anhand von (57) überzeugt hat, dass man keine wesentlich grössere Varianz zu erwarten hat, $\hat{N}_u$ (siehe (54)) als Schätzgrösse für N verwenden. Ein approximatives Vertrauensintervall kann dann mit der Approximation (55) für die Varianz und den Schranken der Normalverteilung gewonnen werden.

Bemerkung

Diese Studie ist auf Anregung von Herrn Prof. H. Burla entstanden. Ich möchte ihm für sein Interesse und seine Unterstützung bei der Ausführung der Arbeit danken. Die Karl-Hescheler-Stiftung leistete an die Druckkosten einen Beitrag.

Literatur

[1] NORMAN T. J. BAILEY (1951): On estimating the size of mobile populations from recapture data. Biometrika 38, 293–306.
[2] DOUGLAS G. CHAPMAN (1951): Some properties of the hypergeometric distribution with applications to zoological sample censuses. Univ. Calif. Publ. Statist. 1, 131–160.
[3] J. N. DARROCH (1958): The multiple recapture census. I. Estimation of a closed population. Biometrika 54, 343–359.
[4] F. X. SCHUMACHER, R. W. ESCHMEYER (1943): The estimate of fish population in lakes or ponds. Journal of the Tennessee Academy of Science 18, 228–249.
[5] KLAUS KRICKEBERG (1963): Wahrscheinlichkeitstheorie. Teubner, Stuttgart.
[6] B. L. VAN DER WAERDEN (1957): Mathematische Statistik. Springer, Berlin (1. Aufl.).
[7] H. BLANKENHORN, H. BURLA, P. MÜLLER, M. VILLIGER: Die Bestände an Amphibien zur Laichzeit in drei Gewässern des Kantons Zürich. Erscheint im selben Heft.