3.2.1. Beispiel: Taxiproblem. In einer Stadt gebe es N Taxis mit den vom Straßenrand aus lesbaren Nummern $1, \dots, N$. Ein Passant stehe eine gewisse Zeit lang an einer viel befahrenen Straße und notiere sich die Nummern $x_1, \dots, x_k$ der vorbeifahrenden Taxis. Es sei angenommen, daß [1] $x_1 < \cdots < x_k$ und daß der Passant ein mehrmals vorbeifahrendes Taxi nur einmal zählt. Unter der Annahme, daß im Beobachtungszeitraum alle Taxis in Betrieb sind, ist die Anzahl N aller Taxis zu schätzen.

Als *statistisches Modell* [2] kann [3] $(\mathfrak{X}_k, \mathfrak{S}_k; \mathbf{P}_{k;N}, N \in \mathbb{N},\ N \geq k)$ mit

$$\begin{aligned} \mathfrak{X}_k &= \text{Menge der } k\text{-elementigen Teilmengen von } \mathbb{N} \text{ [4]}, \\ \mathfrak{S}_k &= \mathrm{Pot}(\mathfrak{X}_k), \\ \mathbf{P}_{k;N} &= \text{Gleichverteilung auf der Menge der } k\text{-elementigen} \\ &\qquad \text{Teilmengen von } \{1, \dots, N\}, \quad N \in \mathbb{N},\ N \geq k \text{ [5]}, \end{aligned}$$

gewählt werden. Dieser Ansatz führt zur Likelihood-Funktion

$$L_{(k;x_1,\dots,x_k)}(N) = \begin{cases} \text{[6]}\ \binom{N}{k}^{-1}, & \text{falls } x_k \leq N, \\ \text{[7]}\ 0, & \text{falls } x_k > N, \end{cases}$$

zur Beobachtung von k Taxis mit den Nummern $x_1 < x_2 < \cdots < x_k$. Da für jedes k die Funktion $\{k, k+1, \dots\} \ni N \to \binom{N}{k}^{-1}$ monoton fällt, ist [8]

$$S_1 = x_k$$

der *Maximum-Likelihood-Schätzer* für die Gesamtzahl N der Taxis.

Der Maximum-Likelihood-Schätzer S_1 ist in der vorliegenden Situation unbefriedigend, da offensichtlich immer $S_1 \leq N$ gilt, d.h., die „wahre“ Anzahl aller Taxis wird systematisch unterschätzt.

Alternative Schätzer. Mit heuristischen Argumenten können zwei weitere, evtl. [9] plausiblere Schätzer vorgeschlagen werden.

- Aus „Symmetriegründen“ sollte [10] $x_1 - 1 \approx N - x_k$ gelten [11]. Als Schätzer für N ergibt sich dann:

$$S_2 = x_k + x_1 - 1.$$

- Es wäre auch sinnvoll, den Ansatz [10] [12]

$$N - x_k \approx \frac{1}{k} \sum_{r=1}^{k} (x_r - x_{r-1} - 1) = \frac{1}{k}(x_k - k),$$

[1] Die Nummern der vorbeifahrenden Taxis werden in aufsteigender Reihenfolge notiert.

[2] Vgl. Abschnitt 3.1.

[3] Die Anzahl k der beobachteten Taxis wird nicht als eine Beobachtungsgröße, die zu den statistischen Schlußfolgerungen herangezogen wird, betrachtet. Nach dem Ende der Beobachtungen steht k fest und wird dann vor dem eigentlichen Beginn der statistischen Überlegungen als eine deterministische, d.h. nicht zufällige Zahl festgehalten.

[4] Beachte, daß $\mathfrak{X}_k$ abzählbar ist.

[5] Hier geht die Annahme ein, daß alle Taxis gleichmäßig im Stadtgebiet im Einsatz sind.

[6] In der Menge $\{1, \dots, N\}$ existieren $\binom{N}{k}$ Teilmengen mit k Elementen.

[7] Offensichtlich muß die Anzahl N aller Taxis $\geq$ der höchsten beobachteten Nummer x_k sein.

[8] Der Maximum-Likelihood-Schätzer für die Gesamtzahl aller Taxis ist somit die größte der beobachteten Nummern.

[9] Dies ist natürlich Ansichtssache.

[10] Diese Vermutung sollte zumindest „im Mittel bei vielen Beobachtungsreihen“ gelten.

[11] Die Lücke bis zur kleinsten beobachteten Nummer x_1, bzw. die Lücke nach der größten beobachteten Nummer x_k sollten in etwa gleich sein.

[12] Es wird hier $x_0 = 0$ gesetzt.

zu wählen [13]. Diese Überlegung führt nun zu [14]

$$S_3 = x_k + \lceil (x_k - k)/k \rceil$$

als Schätzer für N [15].

Literatur

[1] U. Krengel. Einführung in die Wahrscheinlichkeitstheorie und Statistik, 7. Auflage. Vieweg, 2003.

[13] Hierbei wird die Größe der Lücke nach der größten beobachteten Nummer x_k durch die „mittlere Größe aller Lücken“ geschätzt.

[14] Es sollte sinnvollerweise N durch eine ganze Zahl geschätzt werden.

[15] Die drei Schätzer S_1, S_2 und S_3 für die Gesamtzahl N der Taxis besitzen unterschiedliche Eigenschaften, vgl. [1], Abschnitte 4.2 - 4.4. Zunächst kann nachgewiesen werden, daß S_2 und S_3 *erwartungstreue Schätzer* sind, d.h., für $i = 2, 3$ gilt:

$$\mathbf{E}_{k;N}[S_i] := \sum_{l=k}^{\infty} l \cdot \mathbf{P}_{k;N}[S_i = l] = N, \quad N \in \mathbb{N},\ N \geq k. \tag{$*$}$$

Andererseits ist S_1 nicht erwartungstreu, d.h., S_1 erfüllt $(*)$ nicht. „Im Mittel“ wird daher durch die Schätzer S_2 und S_3 der wahre Wert von N gefunden. Hingegen wird durch S_1 „im Mittel“ ein falscher Wert geschätzt.

Beim Vergleich von S_2 und S_3 zeigt sich, daß der *mittlere quadratische Fehler* für S_3 kleiner als für S_2 ist, d.h.,

$$\begin{aligned}\mathbf{E}_{k;N}\big[(S_3 - \mathbf{E}_{k;N}[S_3])^2\big] &= \sum_{l=k}^{\infty} (l - \mathbf{E}_{k;N}[S_3])^2 \cdot \mathbf{P}_{k;N}[S_3 = l] \\ &< \mathbf{E}_{k;N}\big[(S_2 - \mathbf{E}_{k;N}[S_2])^2\big], \quad N \in \mathbb{N},\ N \geq k.\end{aligned}$$

Der Schätzer S_3 schwankt daher „im quadratischen Mittel“ weniger als S_2 um den wahren Wert von N.

Zusammenfassend ist also der Schätzer S_3 gegenüber den beiden anderen Schätzern zu bevorzugen.