Institut für Angewandte Mathematik
SS 2017
Patrik Ferrari, Eva Kopfer & Angelo Profeta

2. Übungsblatt „Einführung in die Statistik”

Abgabe: Mittwoch 10.5.2017 in der Vorlesung

---

1. (Schätzung des Alters im infinite sites model) [*4 Pkt*]

Um das Alter des letzten gemeinsamen Vorfahrens V von zwei Organismen A und B zu bestimmen, wird angenommen, dass die Mutationen entlang der Stammbaumlinien von A nach V und B nach V zu den Zeitpunkten von unabhängigen Poisson-Prozessen mit bekannter Intensität (Mutationsrate) $\mu > 0$ erfolgt sind. Weiter wird angenommen, dass die Mutationen jeweils ein anderes Nukleotid in der Gensequenz verändern. Sei nun x die beobachtete Anzahl der unterschiedlichen Nukleotide in den Sequenzen von A und B.

Präzisieren Sie das statistische Modell und geben Sie einen Maximum-Likelihood-Schätzer für das Alter on V an.

2. (Randomized Response) [*6 Pkt*]

Um bei Umfragen zu heiklen Themen ("Haben Sie schon mal harte Drogen genommen?") die Privatsphäre der befragen Personen zu schützen und zuverlässige Antworten zu bekommen, kann die folgende Variante des *unrelated question*-Befragungsmodells verwendet werden:
Der Interviewer läßt den Befragten zweimal einem Würfel werfen, ohne dass der Interviewer den Ausgang des Würfelexperiments sieht. Wenn der Befragte die Augensumme 5 oder 6 erzielt, so soll er die ihm vorgelegt heikle Frage bejahen, bei der Augensumme 8 oder 9 soll er sie verneinen. Anderenfalls soll die gestellte Frage wahrheitsgemäß beantworten.

Es werden nun n Personen unabhängig voneinander befragt, wobei $\vartheta \in [0, 1]$ die Wahrscheinlichkeit bezeichnet, dass die heikle Frage wahrheitsgemäß bejaht wird. Präzisieren Sie das statistische Modell, geben Sie einen besten Schätzer für ϑ und bestimmen Sie dessen Varianz.

3. (Schätzung der Mutationsrate im infinite alleles model) [*10 Pkt*]

Im sog. *inifinite alleles model* wird angenommen, dass jede Mutation innerhalb eines Genes zu einem neuen Allel führt. Sei nun ϑ ein fester Parameter, der die Mutationsrate beschreibt.
Betrachten Sie eine Urne, die zum Zeitpunkt 0 nur die "Mutationskugel" mit Gewicht ϑ enthält. In jedem Zug wird nun eine Kugel aus der Urne *mit der Wahrscheinlichkeit proportional zu ihrem Gewicht* gezogen. Handelt es sich bei der gezogenen Kugel um die Mutationskugel, so wird diese zusammen mit einer Kugel mit Gewicht 1, deren Farbe in

der Urne noch nicht vorhanden ist, zurückgelegt. Ist die gezogene Kugel hingegen farbig, so wird diese zusammen mit einer weiteren Kugel mit Gewicht 1 der gleichen Farbe in die Urne zurückgelegt. Sei nun $X = (X_i)_{i\geq 1}$ eine Folge von Zufallsvariablen, wobei $X_i \in \mathbb{N}_0$ die Anzahl der Farben, die i Mal vorkommen, beschreibt. Die Gesamtzahl der farbigen Kugeln nach dem n-ten Zug ist dann $\sum_{i\geq 1} i\, X_i = n$. Weiter bezeichne $\varrho_{n,\vartheta}$ die Verteilung von x nach n Zügen.

a) Zeigen Sie durch Induktion über n, dass

$$\varrho_{n,\vartheta}(x) \;=\; \mathrm{P}_\vartheta[X_1 = x_1, \ldots, X_n = x_n] \;=\; \frac{n!}{\vartheta^{(n)}} \prod_{i\geq 1} \frac{(\vartheta/i)^{x_i}}{x_i!}$$

wobei $\vartheta^{(n)} = \vartheta(\vartheta+1)\ldots(\vartheta+n-1)$ ist.

$\varrho_{n,\vartheta}$ ist die sogenannte *Ewens'sche Stichprobenverteilung.* Sie beschreibt die Verteilung der Typhäufigkeiten in einer zufällen Stichprobe von n Individuen einer Population mit Mutationsrate μ.

b) Zeigen Sie, dass $\{\varrho_{n,\vartheta} : \vartheta \in (0,\infty)\}$ für festes n eine exponentielle Familie (siehe Skript für die Definition) und $K_n(x) = \sum_{i=1}^n x_i$ ein bester erwartungstreuer Schätzer für $\tau_n(\vartheta) := \sum_{i=0}^{n-1} \frac{\vartheta}{\vartheta+i}$ ist. Bestimmen Sie zudem die Varianz von K_n sowie die Fischer-Information $I(\vartheta)$.

c) Zeigen Sie, dass der Maximum-Likelihood-Schätzer für ϑ durch $T_n := \tau_n^{-1} \circ K_n$ gegeben ist. *Hinweis: Betrachten Sie hierzu die Monotonie von τ_n.*

d) Zeigen Sie, dass für festes $\vartheta > 0$ im Limes $n \to \infty$ $\mathrm{Var}_{n,\vartheta}[K_n]/\ln n \to \theta$ strebt. Zeigen Sie weiterhin, dass $\big(K_n/\ln n\big)_{n\geq 1}$ eine asymptotisch erwartungstreue und konsistente Schätzfolge für θ ist.