Tentamen Inleiding Statistiek

december 2013

Bij deze tentamen mogen leerboeken en aantekeningen worden geraadpleegd. Vergeet niet al Uw antwoorden goed te motiveren!

Opgave 1.
Een aselecte steekproef $X_1, \ldots, X_n$ wordt genomen uit een verdeling F met dichtheid $f(x) = (\theta^3/2)x^{-4}e^{-\theta/x}$ voor $x > 0$, waar $\theta \in \Theta = (0, \infty)$ een onbekende parameter is.

$Var(X) = \frac{\theta^2}{4}$

De volgende feiten zijn wellicht in het volgende van nut: als $X \sim F$ een enkele trekking uit de de verdeling F is, geldt $\mathrm{E}(X) = \theta/2$ and $\mathrm{Var}(X) = \theta^2/4$. Bovendien, geldt dat $1/X \sim \mathrm{Gamma}(3, \theta)$: de gamma verdeling met vorm parameter $\alpha = 3$ en (inverse) schaal parameter $\lambda = \theta$, in de terminologie van Rice.

(a) Bepaal de meest aannemelijke schatter (MLE) $\widehat{\theta}$ van θ.

(b) Laat zien zien dat $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i^{-1}$ een voldoende grootheid (Engels: "sufficient statistic") is. (Hint: gebruik het *factorizatie criterium*).

(c) Een alternative schatter van theta is $\widetilde{\theta} = 2\overline{X}$. Bepaal de onzuiverheid, variantie en verwachte kwadratische fout (Engels: "bias", "variance", "mean square error") van deze schatter.

(d) Wat zijn voor grote n, bij benadering, de kansverdelingen van deze twee schatters? Welke is dan te preferen?

(e) Hoe zou U, voor grote n, de variantie van de MLE $\widehat{\theta}$ schatten?

(f) Hoe zou U een benaderend 95% betrouwbaarheids interval voor θ berekenen?

Opgave 2. De leden van een biologische populatie kunnen ingedeeld worden in drie genetische klassen A, B en C. Volgens een bepaald genetisch model komen de twee klassen B en C even vaak voor. Maar in het algemeen zouden de drie klassen volledig willekeurige kansen $\theta_A, \theta_B, \theta_C$ kunnen hebben, waarbij natuurlijk wel moet gelden $\theta_A + \theta_B + \theta_C = 1$. Een aselecte steekproef van 100 individuen uit de populatie levert de volgende aantallen op voor elk van de drie klassen, in dezelfde volgorde A, B, C: 63, 23, 14.

(a) Wat is de meest aannemelijke schatter voor de vector parameter $\theta = (\theta_A, \theta_B, \theta_C)$ onder de *nul hypothese* dat het eerder genoemde genetisch model geldt, en wat is de meest aannemelijke schatter voor θ *in het algemeen*?

(b) Geef een formule voor de toetsingsgrootheid $-2\log\Lambda$ in termen van de data en de eerder afgeleide schattingen. Hier is Λ de *generalized likelihood ratio* voor het toetsen van de eerder genoemde nul hypothese tegen het alternatief (binnen het algemene context van de opgave) dat de nul hypothese niet geldt. U hoeft hierbij de numerieke waarde van $-2\log\Lambda$ niet uit te rekenen als U geen zakrekenmachine bij de hand hebt: de opgave vraagt om een eenvoudige formule waarmee men dat in principe zou kunnen doen.

(c) Ik vond een waarde 2.211 voor $-2\log\Lambda$. Als ik deze toets wil uitvoeren met significantie niveau $\alpha = 10\%$, moet ik wel of niet de nul hypothese verwerpen? (Hint: appendix B van het boek van Rice bevat, o.m., tabellen met quantielen/percentielen van verschillende standaard verdelingen.)

Opgave 3. Definieer $f(x;\theta) = \theta x^{\theta-1}$ voor $0 \le x \le 1$, $f(x) = 0$ voor $x < 0$ en voor $x > 1$, waar $\theta > 0$ een vaste parameter is.

(a) Laat zien dat de bijbehorende cumulatieve verdelingsfunctie gelijk is aan $F(x;\theta) = x^\theta$ voor $0 \le x \le 1$; $F(x;\theta) = 0$ voor $x \le 0$; $F(x;\theta) = 1$ voor $x \ge 1$. Welke bekende kansverdeling stelt dit voor, in het geval dat $\theta = 1$?

(b) Laat zien dat de quantiel functie $Q(p;\theta) = p^{\frac{1}{\theta}}$ is $(0 \le p \le 1)$.

(c) Voor gegeven θ, hoe zou u een trekking uit deze verdeling op een computer kunnen nabootsen? Uw computer kan wel uniform verdeelde getallen op het interval $[0,1]$ nabootsen.

(d) Hoe ziet een QQ plot uit van de logarithmes van de quantielen van één verdeling uit deze klasse van verdelingen tegen de logaritmes van de quantielen van een andere verdeling van dezelfde klasse? Noem de parameters van de twee verdelingen θ_1 en θ_2 en gebruik de x-as voor de log quantielen van de eerste verdeling, de y-as voor de log quantielen van de andere.