Hertentamen Inleiding Statistiek

januari 2013

Bij het tentamen mogen boek en aantekeningen worden gebruikt.
Opgaven 1 en 2 tellen elk voor 2/5e van het totaal, opgave 3 voor 1/5e.
U hoeft niet alles goed te beantwoorden om een 10 te halen!
Schrijf duidelijk en motiveer Uw antwoorden.

Opgave 1. De Pareto verdeling is de verdeling op $(-\infty, \infty)$ met kansdichtheid $f(x|\theta, \alpha) = \theta\alpha^\theta x^{-\theta-1}$ voor $x \geq \alpha$, $f(x|\theta, \alpha) = 0$ voor $x < \alpha$. Hierbij zijn θ en α twee onbekende parameters, $\theta \in (1, \infty)$, $\alpha \in (-\infty, \infty)$.

Gegeven een steekproef ter grootte n uit deze verdeling, leid de meest aannemelijke schatters voor de onbekende parameters θ en α af. Pas op: de verzameling van waardes x waar de dichtheid positief is hangt van een van deze parameters af. De *likelihood functie* kan dus nul zijn als de waarnemingen $x_1, \ldots, x_n$ niet compatibel zijn met bepaalde parameter waardes θ, α! Voor gegeven data $x_1, \ldots, x_n$, waar is de likelihood positief en waar is hij nul?

Leg uit hoe U de parametrische bootstrap zou kunnen gebruiken om een betrouwbaarheidsinterval van (benaderend) niveau 95% te construeren voor θ.

Opgave 2. Stel dat $X_1, \ldots, X_m$ en $Y_1, \ldots, Y_n$ twee onafhankelijke steekproeven zijn uit twee normale verdelingen met parameters μ en σ^2 en ν en τ^2 respectievelijk; dus vier onbekende parameters. Ik wil de nul-hypothese toetsen dat de twee verdelingen gelijk aan elkaar zijn: dus $\mu = \nu$ en $\sigma = \tau$. Onder de nul-hypothese, heb ik dus één enkel steekproef ter grootte $m + n$ uit een normale verdeling met onbekende verwachting en variantie.

Leid de "generalized likelihood ratio toets" af (d.w.z., bepaal de toetsingsgrootheid Λ), voor deze nul-hypothese tegen het alternatief dat de twee normale verdelingen niet hetzelfde zijn. Wat is het verwerpingskriterium van deze toets voor een (bij benadering) niveau α toets, als U de asymptotische chi-kwadraat verdeling van $-2 \log \Lambda$ gebruikt? (ihb, om welke chi-kwadraat verdeling gaat het?)

Druk uw antwoorden uit in termen van de twee steekproef omvangen, gemiddeldes en standaard deviaties. De gemiddelde en standaard deviatie van de gecombineerd steekproef kunnen uiteraard uit de voorgaande grootheden worden bepaald.

Leg uit waarom de theorie van voldoende grootheden (*sufficient statistics*) ons al kon vertellen dat Λ alleen van de data afhangt via de grootheden $\overline{X}$, $\overline{Y}$, S_X en S_Y.

Opgave 3. Stel dat X en Y twee random variabelen zijn met verdelingsfuncties F en G respectievelijk. De QQ-plot van de verdeling van Y tegen dat van X (ook wel een "probability plot" genoemd, ook al worden geen kansen tegen elkaar uitgezet, maar quantielen) is de grafiek van de quantielen van de verdeling van Y op de y-as tegen die van de verdeling van X op de x-as: dus een grafiek van de kromme $(F^{-1}(p), G^{-1}(p))$, $p \in (0,1)$.

Veronderstel in het vervolg van deze opgave dat de verdelingen in kwestie continue zijn met strikt stijgende verdelingsfuncties (behalve uiteraard voorbij hun eindpunten, als die eindig zijn).

(a) Laat zien dat de QQ plot de grafiek is van de monotoon stijgende functie $\Psi : x \mapsto G^{-1}(F(x))$.

(b) Laat zien dat $\Psi(X)$ dezelfde kansverdeling heeft als Y.

(c) Stel we hebben een familie van kansverdelingen met twee onbekende parameters die zodanig zijn dat elke QQ plot van de quantielen van twee van deze verdelingen tegen elkaar een rechte lijn is $y = a + bx$ (waarbij a en b natuurlijk van de parameters van de twee verdelingen in kwestie afhangen. Dan noemen we de familie een locatie-schaal familie. Zijn de volgende twee-parameter families locatie-schaal families?

(i) $N(\mu, \sigma^2)$

(ii) Gamma(λ, α)

(iii) Uniform op het interval (α, β)