Hertentamen IST (Inleiding Statistiek)

Docent: Richard Gill

29 januari 2015

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

Opgave 1

De beta verdeling is een continue verdeling op het interval $[0,1]$ met kansdichtheid $x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ waarbij $\alpha > 0$ en $\beta > 0$ twee parameters zijn, en de beta-functie $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

U kunt in de appendix van het boek van Rice de volgende uitdrukkingen vinden voor de verwachtingswaarde van X en de variantie van X:

$$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Onderdeel 1 A

Ga na dat

$$\mathrm{E}(1-X) = \frac{\beta}{\alpha+\beta}, \quad \frac{\mathrm{E}(X)\mathrm{E}(1-X)}{\mathrm{var}(X)} = \alpha+\beta+1$$

Onderdeel 1 B

Stel dat we beschikken over een aselecte steekproef $X_1, \ldots, X_n$ uit deze verdeling, en dat de parameters α en β onbekend zijn. Leid de momenten schatters voor α en β af.

Onderdeel 1 C

Laat zien dat $\mathbf{T} = (T_1, T_2) = (\sum_{i=1}^n \log X_i, \sum_{i=1}^n \log(1-X_i))$ een twee-dimensionele voldoende (Engels: *sufficient*) statistiek is voor de twee-dimensionele parameter $\theta = (\alpha, \beta)$.

Onderdeel 1 D

Leidt twee vergelijkingen in twee onbekenden af, wiens oplossing de meest aannemelijke schatters voor α en β is. Laat zien dat de meest aannemelijke schatters alleen van de data afhangen via de voldoende grootheid $\mathbf{T}$.

Het is misschien handig om de volgende notatie te gebruiken: $\frac{\mathrm{d}}{\mathrm{d}\beta}(\log\Gamma(\beta)) = \Psi_0(\beta)$. De functie Ψ_0 heet de *digamma functie.*

Onderdeel 1 E

Stel dat u de meest-aannemelijke schatters in de praktijk kan uitrekenen. Hoe zou u (bij benadering) de standaard deviaties van deze schattingen bepalen?

Opgave 2

Deze opgave is een voortzetting van Opgave 1.

In de populatie genetica, blijkt dat de beta verdeling een goede model is voor allele frequentie in een grote populatie die uit twee sub-populaties bestaat. Hierbij wordt een andere parametrisatie gebruikt dan in Opgave 1: $p = \alpha/(\alpha+\beta)$ en $F = 1/(\alpha+\beta+1)$; oftewel, $\alpha = p(1-F)/F$, $\beta = (1-p)(1-F)/F$. Het model heet dan de *Balding-Nichols model*, en heeft belangrijke toepassingen bij genetische epidemiologie en forensisch DNA.

Onderdeel 2 A

Stel dat een onderzoeker over een aselecte steekproef ter groote n uit de Balding-Nichols verdeling beschikt en de nul hypothese wilt toetsen dat F een bepaalde waarde, F_0, heeft; onder deze hypothese is dus p nog steeds onbekend. De alternatief hypothese is natuurlijk $F \neq F_0$.

Leg uit hoe je een likelihood ratio toets (meer precies: "Generalized Likelihood Ratio Toets") zou uitvoeren in deze situatie. Als we een toets willen hebben met (bij benadering) significantie niveau 5%, wat is de kritieke waarde van onze toetsingsgrootheid?

Onderdeel 2 B

Stel nu dat de onderzoeker een hypothese heeft omtrent bepaalde waardes p_0 en F_0 van beide parameters. De nul-hypothese is dus enkelvoudig ("simple null hypothesis") Leg uit hoe je een nagenoeg *exacte* niveau 5% toets kan uitvoeren m.b.v. een computer simulatie (Monte-Carlo, bootstrap, ...)

Opgave 3

In deze opgave hebben we weer te maken met een aselecte steekproef $X_1, \ldots, X_n$ uit een of andere verdeling. Schrijf ook X voor een generieke trekking uit dezelfde verdeling. Zij F en Q de bijbehorende verdelingsfunctie en quantiel functie: dus $F(x) = \mathrm{P}(X \leq x)$, $x \in (-\infty, \infty)$, en $Q(p) = F^{-1}(p)$, $p \in (0,1)$. Voor het gemak gaan we vanuit dat de verdelingsfunctie F continu en strict monotoon stijgend is op de hele lijn $(-\infty, \infty)$.

Onderdeel 3 A

De empirische verdelingsfunctie F_n wordt gedefinieerd door: $F_n(x) = \frac{1}{n}\#\{i : X_i \leq x\}$ waar $\#$ staat voor "aantal". Definieer $c = 1.9599...$, de 97.5% quantiel van de standaard normale verdeling. Laat zien dat, voor gegeven x en n groot, $F_n(x) \pm c\sqrt{(F_n(x)(1-F_n(x))/n)}$ een benaderend 95% betrouwbaarheids interval is voor $F(x)$.

Onderdeel 3 B

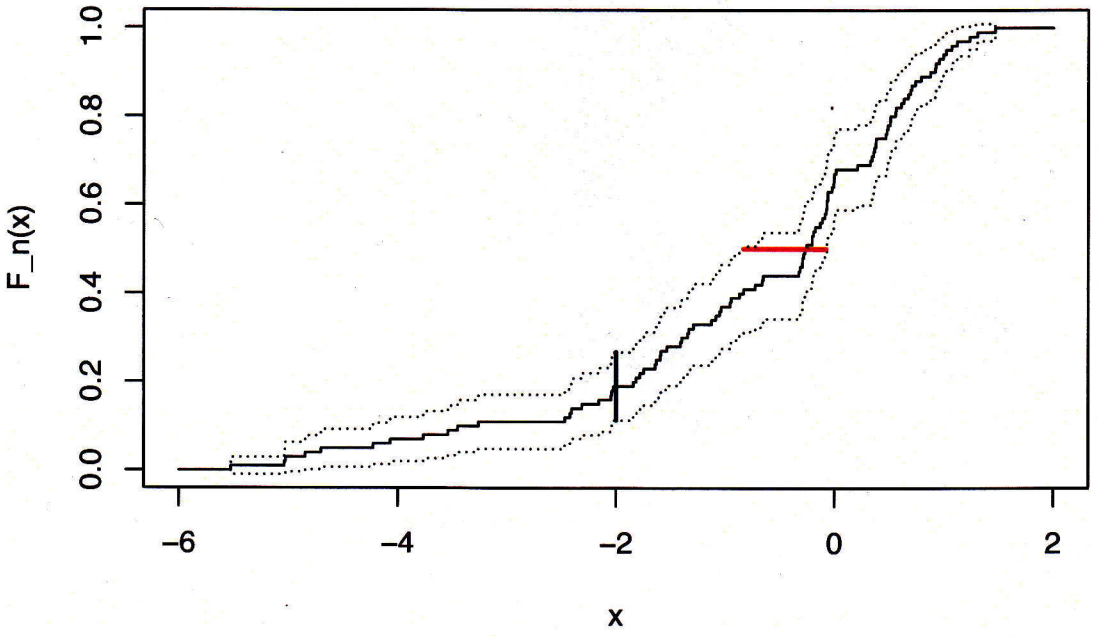
Kunt u ook een exacte 95% betrouwbaarheids interval construeren?

Onderdeel 3 C

Onderstaand grafiek toont de empirische verdelingsfunctie gebaseerd op een steekproef ter grootte 100 uit zekere verdeling. De twee stippel-lijnen geven de grenzen aan van de betrouwbaarheidsintervallen van onderdeel 2A. Dus, voor elke x apart, kan men uit de grafiek de ondergrens en de bovengrens aflezen voor een (benaderend) 95% betrouwbaarheid interval voor $F(x)$.

Bijvoorbeeld, voor $F(-2)$: verticale blauwe streep in figuur.

```
set.seed(1234); N <- 100; c <- qnorm(0.975)
X <- sort(log(rexp(N))); x <- c(-6, X, 2); y <- c(0, (1:N)/N, 1)
plot(x, y, type = "s", xlim = c(-6, 2), ylim = c(0, 1),
     xlab = "x", ylab = "F_n(x)")
lines(x, y + c * sqrt(y * (1-y)/N), type = "s", lty = 3)
lines(x, y - c * sqrt(y * (1-y)/N), type = "s", lty = 3)
medLow <- max(x[y - c * sqrt(y * (1-y)/N) < 0.5])
medHigh <- min(x[y + c * sqrt(y * (1-y)/N) > 0.5])
lines(c(-2, -2),
  0.19 + c*c(-sqrt(0.19 * 0.81/N), sqrt(0.19 * 0.81/N)),
  col = "blue", lwd = 2)
lines(c(medLow, medHigh), c(0.5, 0.5), col = "red", lwd = 2)
```



Laten we voor het gemak het feit dat we hier alleen over een "benadering" spreken buiten beschouwing laten. Noem de ondergrens $O_n(x)$ en de bovengrens $B_n(x)$. We veronderstellen dus dat voor elke x geldt $\mathrm{P}(O_n(x) \leq F(x) \leq B_n(x)) = 0.95$.

Stel, ik wil een betrouwbaarheidsinterval voor de mediaan $Q(0.5)$ van de verdeling waaruit mijn data komt. Bewijs dat $(\max\{x : O_n(x) < 0.5\}, \min\{x : B_n(x) > 0.5\})$ zo'n betrouwbaarheidsinterval is: zie horizontale rode streep in de figuur.

Hint: je kunt toetsen dat de mediaan een bepaalde waarde, zeg maar x_0 heeft, door te kijken of de waarde $p = 1/2$ binnen de betrouwbaarheids interval voor $F(x_0)$ ligt. Gebruik nu de dualiteit tussen toetsen en betrouwbaarheidsintervallen.

Footnote

Text from wikipedia:

Florence Nightingale (1820 – 1910) was a celebrated English social reformer and statistician, and the founder of modern nursing. Florence Nightingale exhibited a gift for mathematics from an early age. Later, Nightingale became a pioneer in the visual presentation of information and statistical graphics. In 1859, Nightingale was elected the first female member of the Royal Statistical Society.