1e deeltentamen Statistiek
18 april 2012

I Schrijf je naam en studentnummer op elk vel dat je inlevert.
II Het gebruik van het boek van J.A. Rice, aantekeningen, handouts en zakrekenmachines is toegestaan.
III U mag in ieder onderdeel de conclusies van voorgaande onderdelen gebruiken, ook als u die (nog) niet bewezen hebt.
IV Motiveer steeds uw antwoord door duidelijk aan te geven welke argumenten en welke resultaten u gebruikt om een bepaalde conclusie te trekken.
V U heeft 3 uur de tijd voor het tentamen.
VI Achter elke vraag staat het aantal punten dat met de vraag te behalen is. De puntenverdeling is: 1 - 20, 2 - 15, 3 - 40, 4 - 25.
VII Veel succes!

Opgave 1 Zij X_1, X_2, ..., X_n onafhankelijke, identiek verdeelde stochasten met kansdichtheidsfunctie: $f(x|\theta) = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2} x^{\theta-1}(1-x)^{\theta-1}$ voor $0 < x < 1$ met $\theta > 0$.
U mag gebruiken dat voor $\alpha > -1$ en $\beta > -1$ geldt dat $\int_0^1 x^\alpha (1-x)^\beta dx = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}$ en dat $\Gamma(x+1) = x\Gamma(x)$.

a 7pt) Bepaal de momentschatter T_M van θ.
ANTWOORD: $\mathbb{E}(X|\theta) = \frac{1}{2}$. Dit volgt uit het feit dat $f(x|\theta)$ symmetrisch is rond $x = \frac{1}{2}$ of door de integraal uit te rekenen. $\mathbb{E}(X|\theta) = \int_0^1 x f(x|\theta)dx = \int_0^1 \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2} x^\theta (1-x)^{\theta-1} dx = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2}\frac{\Gamma(\theta+1)\Gamma(\theta)}{\Gamma(2\theta+1)} = \frac{\Gamma(2\theta)}{\Gamma(2\theta+1)}\frac{\Gamma(\theta+1)}{\Gamma(\theta)} = \frac{\Gamma(2\theta)}{2\theta\Gamma(2\theta)}\frac{\theta\Gamma(\theta)}{\Gamma(\theta)} = \frac{1}{2}$.
Omdat het eerste moment onafhankelijk is van θ, moeten we het tweede moment bekijken.
$\mathbb{E}(X^2|\theta) = \int_0^1 x^2 f(x|\theta)dx = \int_0^1 \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2} x^{\theta+1}(1-x)^{\theta-1}dx = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2}\frac{\Gamma(\theta+2)\Gamma(\theta)}{\Gamma(2\theta+2)} = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2}\frac{(\theta+1)\Gamma(\theta+1)\Gamma(\theta)}{(2\theta+1)\Gamma(2\theta+1)} = \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2}\frac{(\theta+1)\theta\Gamma(\theta)\Gamma(\theta)}{(2\theta+1)(2\theta)\Gamma(2\theta)} = \frac{\theta(\theta+1)}{2\theta(2\theta+1)} = \frac{\theta+1}{2(2\theta+1)}$. Definieer $K = \frac{1}{n}\sum_{i=1}^n x_i^2$, dan voldoet de momentschatter $\hat{\theta}_M$ van θ aan: $K = \frac{\hat{\theta}_M+1}{2(2\hat{\theta}_M+1)}$, oftewel, $\hat{\theta}_M = \frac{1-2K}{4K-1}$.

b 7pt) Bepaal een voldoende statistiek voor θ.
ANTWOORD: $f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) = \prod_{i=1}^n \frac{\Gamma(2\theta)}{(\Gamma(\theta))^2} x_i^{\theta-1}(1-x_i)^{\theta-1} = \left(\frac{\Gamma(2\theta)}{(\Gamma(\theta))^2}\right)^n \prod_{i=1}^n (x_i(1-x_i))^{\theta-1} = \frac{\Gamma(2\theta)^n}{\Gamma(\theta)^{2n}} e^{(\theta-1)\sum_{i=1}^n \log(x_i(1-x_i))}$
Definieer $T = \sum_{i=1}^n \log(x_i(1-x_i))$, $h(x_1, x_2, \ldots, x_n) = 1$ en $g(t,\theta) = \frac{\Gamma(2\theta)^n}{\Gamma(\theta)^{2n}} e^{(\theta-1)t}$.
Met deze definities geldt: $f(x_1, x_2, \ldots, x_n|\theta) = g(T(x_1, x_2, \ldots, x_n), \theta)h(x_1, x_2, \ldots, x_n)$.
Dan volgt uit de factorisatiestelling dat T een voldoende statistiek is voor θ.

c 6pt) Bepaal een minimaal voldoende statistiek voor θ (mag dezelfde zijn als in onderdeel b)) en bewijs dat deze minimaal is.
ANTWOORD: Gebruik de volgende stelling: Zij $T(x_1, x_2, \ldots, x_n) := T(\mathbf{x})$ een functie zodanig dat $\forall \mathbf{x}, \mathbf{y}$ geldt: $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ onafhankelijk van $\theta \Leftrightarrow T(\mathbf{x}) = T(\mathbf{y})$. Dan is T een minimaal voldoende statistische grootheid voor θ.
BEWIJS: Stel $T(\mathbf{x}) = T(\mathbf{y})$. Dan geldt dat $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g(T(\mathbf{x}),\theta)}{g(T(\mathbf{y}),\theta)} = 1$, onafhankelijk van θ.

Neem vervolgens aan dat $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ onafhankelijk is van θ.
$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g(T(\mathbf{x}),\theta)h(x_1,x_2,\ldots,x_n)}{g(T(\mathbf{y}),\theta)h(y_1,y_2,\ldots,y_n)} = \frac{g(T(\mathbf{x}),\theta)}{g(T(\mathbf{y}),\theta)} = \frac{e^{(\theta-1)T(\mathbf{x})}}{e^{(\theta-1)T(\mathbf{y})}} = e^{(\theta-1)(T(\mathbf{x})-T(\mathbf{y}))}$.
Deze uitdrukking is alleen onafhankelijk van θ als $T(\mathbf{x}) = T(\mathbf{y})$. $\square$

Opgave 2 Stel dat een bestaand geneesmiddel (middel A) een ziekte in 80% van de gevallen geneest. Een bedrijf heeft een nieuw geneesmiddel (middel B) voor dezelfde ziekte ontwikkeld. Het bedrijf beweert dat het middel significant meer patiënten geneest dan middel A. Om dit te testen wordt een onderzoek gedaan waarbij 200 patiënten die de ziekte hebben middel B toegediend krijgen. Het resultaat van het onderzoek is dat 167 van deze patiënten genezen. De onbetrouwbaarheidsdrempel (α) is 0.05.

a 6pt) Formuleer een kansmodel en beschrijf het toetsingsprobleem.
ANTWOORD: Een patiënt heeft een Bernoulli-verdeelde kans om te genezen na toediening van medicijn B en zij p de parameter van de Bernoulli-verdeling.
H_0: $p = 0.8$
H_A: $p > 0.8$

b 9pt) Bepaal door een geschikte toets of de bewering van het bedrijf waar is.
Kies $p > 0.8$ vast. De likelihood ratio test is dan de meest onderscheidende test. Zij k het aantal patiënten dat geneest. $\Lambda = \frac{0.8^k 0.2^{200-k}}{p^k(1-p)^{200-k}} = \left(\frac{0.8}{p}\right)^k \left(\frac{0.2}{1-p}\right)^{200-k}$. Λ is klein als k groot is. Omdat dit geldt voor elke $p > 0.8$ is de test die H_0 verwerpt als $\frac{k}{200} > c$ uniform het meest onderscheidend. De standaardfout in het steekproefgemiddelde onder H_0 is $\sqrt{\frac{0.8 \cdot 0.2}{200}} \approx 0.028$. $P(\frac{k}{200} > c|H_0) = P(\frac{k}{200} - 0.8 > c - 0.8) = P(\frac{\frac{k}{200} - 0.8}{0.028} > \frac{c-0.8}{0.028})$. Gebruik de normale benadering. Hieruit volgt dat $\frac{c-0.8}{0.028} = 1.64$, oftewel $c = 0.847$. Verwerp H_0 als het steekproefgemiddelde groter is dan 0.847. In de steekproef was het gemiddelde 167/200=0.835. De nulhypothese wordt dus niet verworpen en de bewering van het bedrijf volgt niet uit het onderzoek.

Opgave 3 Zij X_1, X_2, ..., X_n onafhankelijke, identiek verdeelde stochasten met kansdichtheidsfunctie: $f(k|\theta) = \begin{cases} \theta & \text{als } k = 0 \\ \theta(1-\theta) & \text{als } k = 1 \\ (1-\theta)^2 & \text{als } k = 2 \end{cases}$ met $0 < \theta < 1$
Zij n_k het aantal observaties met waarde k ($k \in \{0, 1, 2\}$).
We beschouwen twee schatters, T_1 en T_2 voor θ die als volgt zijn gedefinieerd:
$T_1 = \frac{n_0}{n}$. T_2 is de meest waarschijnlijke schatter (MLE).

a 7pt) Bereken de meest waarschijnlijke schatter T_2 van θ.
ANTWOORD: $L = f(0|\theta)^{n_0} f(1|\theta)^{n_1} f(2|\theta)^{n_2}$, oftewel,
$l = n_0 \log f(0|\theta) + n_1 \log f(1|\theta) + n_2 \log f(2|\theta) = n_0 \log(\theta) + n_1 \log(\theta) + n_1 \log(1-\theta) + 2n_2 \log(1-\theta) = (n_0 + n_1)\log(\theta) + (n_1 + 2n_2)\log(1-\theta)$
$\frac{dl}{d\theta} = \frac{n_0+n_1}{\theta} - \frac{n_1+2n_2}{1-\theta}$
$\frac{d}{d\theta} l(\hat{\theta}) = 0 \Rightarrow \hat{\theta} = \frac{n_0+n_1}{n_0+2(n_1+n_2)}$.

b 8pt) Bepaal of de volgende uitspraken waar zijn:

1) T_1 is zuiver voor elke steekproefgrootte.
2) T_1 is consistent.
3) T_2 is zuiver voor elke steekproefgrootte.
4) T_2 is consistent.

ANTWOORD: $\mathbb{E}(T_1) = \mathbb{E}(\frac{n_0}{n}) = \frac{n\theta}{n} = \theta$, dus T_1 is zuiver.
T_1 is ook consistent omdat $Var(T_1) = \frac{\theta(1-\theta)}{n}$ en dit gaat naar 0 als $n \to \infty$

Stel $n = 1$. $\mathbb{E}(T_2) = \mathbb{E}(\frac{n_0+n_1}{n_0+2(n_1+n_2)}) = \theta\frac{1}{1} + \theta(1-\theta)\frac{1}{2} + (1-\theta)^2\frac{0}{2} = \theta + \theta\frac{(1-\theta)}{2}$. De meest waarschijnlijke schatter is dus niet zuiver voor elke steekproefgrootte.
De meest waarschijnlijke schatter is wel consistent want we hebben een gladde kansdichtheidsfunctie (als functie van θ) en dus kunnen we stelling A op pagina 275 van Rice gebruiken.

c 8pt) Bereken de Fisher informatie voor de parameter θ en de Cramér-Rao ondergrens voor de variantie in een zuivere schatter van θ. Is T_1 efficiënt, m.a.w., voldoet T_1 aan de Cramér-Rao ondergrens?
ANTWOORD:
$\frac{\partial^2}{\partial\theta^2}\log(f(0|\theta)) = \frac{\partial^2}{\partial\theta^2}\log(\theta)) = -\frac{1}{\theta^2}$
$\frac{\partial^2}{\partial\theta^2}\log(f(1|\theta)) = \frac{\partial^2}{\partial\theta^2}\log(\theta(1-\theta)) = \frac{-1+2\theta-2\theta^2}{\theta^2(1-\theta)^2}$
$\frac{\partial^2}{\partial\theta^2}\log(f(2|\theta)) = \frac{\partial^2}{\partial\theta^2}\log((1-\theta)^2) = -\frac{2}{(1-\theta)^2}$
Hieruit volgt dat $I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log(f(k|\theta))\right) = -\sum_{k=0}^{2} f(k|\theta)\frac{\partial^2}{\partial\theta^2}\log(f(k|\theta)) =$
$-\left(\theta\left(-\frac{1}{\theta^2}\right) + \theta(1-\theta)\frac{-1+2\theta-2\theta^2}{\theta^2(1-\theta)^2} + (1-\theta)^2\left(-\frac{2}{1-\theta)^2}\right)\right) = \frac{2-\theta}{\theta(1-\theta)}$
De Cramér-Rao ondergrens voor de variantie in een zuivere schatter van θ is gelijk aan $\frac{1}{nI(\theta)} = \frac{\theta(1-\theta)}{n(2-\theta)}$.
$Var(T_1) = \frac{\theta(1-\theta)}{n} > \frac{\theta(1-\theta)}{n(2-\theta)}$. T_1 is dus niet efficiënt.

d 3pt) Beargumenteer of u beter T_1 of T_2 kunt gebruiken als de steekproef groot is.
ANTWOORD: De meest waarschijnlijke schatter is asymptotisch efficiënt en consistent en heeft een kleinere variantie dan T_1. Voor grote steekproeven is T_2 te prefereren boven T_1.

e 7pt) Bepaal een benaderd 99% betrouwbaarheidsinterval voor θ als de steekproef als resultaat gaf: $n_0 = 25$, $n_1 = 24$, $n_2 = 151$.
ANTWOORD: $T_2 = \frac{n_0+n_1}{n_0+2(n_1+n_2)} = \frac{25+24}{25+2(24+151)} = \frac{49}{375} \approx 0.13$.
$I(\hat{\theta}) = \frac{2-\hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = 16.46$. Een benaderd 99% betrouwbaarheidsinterval voor θ wordt gegeven door: $(\hat{\theta} - \frac{z(0.01/2)}{\sqrt{nI(\hat{\theta})}}, \hat{\theta} + \frac{z(0.01/2)}{\sqrt{nI(\hat{\theta})}})$, oftewel $(\frac{49}{375} - \frac{2.58}{\sqrt{200\cdot 16.46}}, \frac{49}{375} + \frac{2.58}{\sqrt{200\cdot 16.46}}) =$ $(0.086, 0.176)$.

f 7pt) Bepaal de "Goodness of fit" van het model met een significantieniveau van 0.05 op basis van de steekproef uit onderdeel e).
ANTWOORD: Gebruik de Pearson χ^2-test. Het verwachte aantal observaties met waarde k is $200f(k|\hat{\theta})$.

k	0	1	2
Observed	25	24	151
Expected	26.13	22.72	151.15
Bijdrage aan X^2	0.05	0.07	0.0001

$X^2 = 0.05+0.07+0.0001 = 0.12$. X^2 is bij benadering χ^2 verdeeld met 1 vrijheidsgraad. 0.12 ligt tussen het 5% en het 10% percentueel van de χ^2 verdeling met 1 vrijheidsgraad, dus er is geen reden om het model te verwerpen.

Opgave 4 Stel dat een bloedbank bloeddonoren test of ze HIV-positief zijn. De test bepaalt de concentratie (X) van een bepaald eiwit in het bloed. Neem aan dat de gemeten concentratie (in mg/l) van het eiwit bij HIV-negatieve individuen normaal verdeeld is met gemiddelde 11 mg/l en standaarddeviatie 1 mg/l en dat de gemeten concentratie bij HIV-positieve individuen normaal verdeeld is met gemiddelde 18 mg/l en standaarddeviatie 2 mg/l. De test gebruikt een concentratie van 13 mg/l als drempelwaarde, d.w.z., als de gemeten concentratie groter is dan 13 mg/l, dan is het testresultaat positief, als de gemeten concentratie kleiner is dan 13 mg/l, dan is de test negatief. De nulhypothese is dat een bloeddonor HIV-negatief is. De alternatieve hypothese is dat een bloeddonor HIV-positief is.

a 5pt) Wat is het significantieniveau α van de test?
ANTWOORD: $\alpha = P(X > 13 \text{ mg/l}|\text{HIV-negatief})=P(\frac{X-11}{1} > 2|\text{HIV-negatief})=$ 1-0.9772=0.0228 volgens de tabel van de cumulatieve normale verdeling.

b 5pt) Wat is het onderscheidend vermogen (de power) van de test?
ANTWOORD: $1-\beta = P(X > 13 \text{ mg/l}|\text{HIV-positief})=P(\frac{X-18}{2} > -2.5|\text{HIV-positief})=0.9938$.

c 7pt) Neem als prior informatie aan dat 0.15% van de individuen die bloed aanbieden bij de bloedbank HIV positief is. Wat is de posterior kans dat een bloeddonor HIV-positief is gegeven dat de test een positief resultaat gaf.
ANTWOORD: $P(\text{HIV}^+|\text{test}^+)=\frac{P(\text{test}^+|\text{HIV}^+)P(\text{HIV}^+)}{P(\text{test}^+)} = \frac{0.9938 \cdot 0.0015}{0.0015 \cdot 0.9938+0.9985 \cdot 0.0028} = 0.35$

d 8pt) Is er een test met hetzelfde significantieniveau die een groter onderscheidend vermogen heeft. Zo ja, construeer de test met het grootste onderscheidend vermogen, zo nee, bewijs dat de gegeven test het grootste onderscheidend vermogen heeft.
ANTWOORD: H_0 en H_A zijn simpele hypotheses. Het Neyman-Pearson lemma zegt dat de likelihood ratio test het grootste onderscheidend vermogen heeft bij een gegeven significantieniveau.

$$\Lambda = \frac{f_0(x)}{f_A(x)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-11)^2}{2}}}{\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-18)^2}{8}}} = 2e^{-\frac{(x-11)^2}{2}+\frac{(x-18)^2}{8}} = 2e^{-20+\frac{13x}{2}-\frac{3x^2}{8}}$$

Je verwerpt H_0 als Λ klein is, dus als $\frac{13x}{2} - \frac{3x^2}{8} = \frac{3}{8}x(\frac{52}{3} - x)$ klein is.
Verwerp H_0 als $x(\frac{52}{3} - x) < c$ waarbij c zo is gekozen dat het significantieniveau van de test gelijk is aan 0.0228. $x(\frac{52}{3} - x) < c \Leftrightarrow x < \frac{\frac{52}{3}-\sqrt{(\frac{52}{3})^2-4c}}{2}$ of $x > \frac{\frac{52}{3}+\sqrt{(\frac{52}{3})^2-4c}}{2}$.
Definieer $d = \frac{\sqrt{(\frac{52}{3})^2-4c}}{2}$. Verwerp H_0 als $x < \frac{26}{3} - d$ of als $x > \frac{26}{3} + d$. Kies d zodanig dat $\int_{\frac{26}{3}-d}^{\frac{26}{3}+d} \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-11)^2}{2}}dx = 0.9772 \Rightarrow \int_{-\frac{7}{3}-d}^{-\frac{7}{3}+d} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = 0.9772$. Volgens de tabel van de cumulatieve dichtheid van de standaard normale verdeling, is d gelijk aan $\frac{13}{3}$. Dit betekent dat de likelihood ratio test H_0 verwerpt als $x > 13$ of als $x < \frac{13}{3}$.