

Divisie Wiskunde en Informatica	Tentamen Algemene Statistiek
Vrije Universiteit	UITWERKING
	2 april 2002

1. (a) De aannemelijkheidsfunctie op basis van de waarnemingen $x_1, \dots, x_n$ is:

$$\theta \mapsto L(\lambda; x_1, \dots, x_n) = \frac{\lambda^{rn}}{[(r-1)!]^n} e^{-\lambda \sum_{i=1}^n x_i} \prod_{i=1}^n x_i^{r-1}$$

Dit is een differentieerbare functie in λ . De afgeleide van de log-aannemelijkheidsfunctie is:

$$\frac{rn}{\lambda} - \sum_{i=1}^n x_i$$

Afgeleide nul stellen en aantonen dat extreem een maximum is geeft als meest aannemelijke schatter voor λ :

$$\hat{\lambda} = \frac{rn}{\sum_{i=1}^n X_i} = \frac{r}{\bar{X}_n}$$

- (b) De momentenschatter op basis van het eerste moment wordt verkregen door gelijkstellen van $E_\lambda X_i$ aan $\bar{X}_n$. Er geldt

$$E_\lambda X_i = \int_0^\infty \frac{\lambda^r x^r e^{-\lambda x}}{(r-1)!} dx = \frac{1}{\lambda(r-1)!} \int_0^\infty y^r e^{-y} dy = \frac{r}{\lambda},$$

Dus wegens gelijkstellen krijgen we $\hat{\lambda} = \frac{r}{\bar{X}_n}$.

- (c) Stel $\Lambda \sim \pi(\lambda)$. De Bayes schatter wordt verkregen via

$$\begin{aligned} T &= E(\Lambda | (X_1, \dots, X_n) = (x_1, \dots, x_n)) = \frac{\int_0^\infty \lambda \prod_{i=1}^n f_\lambda(x_i) \pi(\lambda) d\lambda}{\int_0^\infty \prod_{i=1}^n f_\lambda(x_i) \pi(\lambda) d\lambda} \\ &= \frac{\int_0^\infty \frac{\lambda^{rn+1}}{[(r-1)!]^n} e^{-\lambda(\sum_{i=1}^n x_i + 1)} \prod_{i=1}^n x_i^{r-1} d\lambda}{\int_0^\infty \frac{\lambda^{rn}}{[(r-1)!]^n} e^{-\lambda(\sum_{i=1}^n x_i + 1)} \prod_{i=1}^n x_i^{r-1} d\lambda} \\ &= \frac{\frac{1}{(\sum x_i + 1)^{nr+2}} \int_0^\infty (\lambda(\sum x_i + 1))^{nr+1} e^{-\lambda(\sum x_i + 1)} d(\lambda(\sum x_i + 1))}{\frac{1}{(\sum x_i + 1)^{nr+1}} \int_0^\infty (\lambda(\sum x_i + 1))^{rn} e^{-\lambda(\sum x_i + 1)} d(\lambda(\sum x_i + 1))} \\ &= \frac{(nr+1)! (\sum x_i + 1)^{nr+1}}{(nr)! (\sum x_i + 1)^{nr+2}} = \frac{nr+1}{\sum_{i=1}^n x_i + 1} \end{aligned}$$

Dus de Bayes schatter is $\hat{\lambda} = \frac{nr+1}{\sum_{i=1}^n X_i + 1}$

2. (a) Er geldt dat Y hypergeometrisch verdeeld is. De populatiegrootte, N , is onbekend. We trekken zonder teruglegging 4400 “elementen” uit de populatie, en in de populatie zit een fractie $p = 0.103$ “elementen” met eigenschap A : doet in de avonduren boodschappen. Omdat het aantal elementen in de trekking klein is t.o.v. de totale Nederlandse populatie gezinnen, mogen we dit benaderen met een binomiale verdeling, dus $Y \sim \text{Bin}(4400, 0.103)$.

We gebruiken de normale benadering van de binomiale verdeling: $Y \sim \mathcal{N}(np, np(1-p))$, met $n = 4400$ en $p = 0.103$, dus $np(1-p) > 5$. Dus we krijgen, met continuïteitscorrectie en $U \sim \mathcal{N}(0, 1)$:

$$\begin{aligned} P_{p=0.103}(Y \leq 427) &\approx P\left(U \leq \frac{427 - 4400 \times 0.103 + 0.5}{\sqrt{4400 \times 0.103 \times 0.897}}\right) \\ &= P(U \leq -1.275) = P(U \geq 1.275) = 1 - 0.8988 = 0.1012 \end{aligned}$$

- (b) Omdat de p-waarde gelijk is aan 2×0.1012 kunnen we op grond van opgave (a) al concluderen dat we de nulhypothese niet verwerpen. Er is dus niet voldoende aanwijzing om de uitspraak uit het krantenartikel te rechtvaardigen. Als we officieel de toets uitvoeren krijgen we dat voor het kritieke gebied geldt $K = \{0, \dots, k_1\} \cup \{k_2, \dots, 4400\}$ met k_1 en k_2 gegeven door $P(Y \leq k_1) \leq 0.025$ en $P(Y \geq k_2) \leq 0.025$, oftewel

$$\frac{k_1 - 4400 \times 0.103 + 0.5}{\sqrt{4400 \times 0.103 \times 0.897}} \leq -1.96 \quad \text{en} \quad \frac{k_2 - 4400 \times 0.103 - 0.5}{\sqrt{4400 \times 0.103 \times 0.897}} \geq 1.96$$

Dit geeft $k_1 \leq 413.2$ en $k_2 \geq 493.2$, oftewel $k_1 = 413$ en $k_2 = 494$.

- (c) We moeten uitrekenen $P_{p=0.097}(Y \leq 413) + P_{p=0.097}(Y \geq 494)$. Dit is gelijk aan $P(U \leq -0.677) + P(U \geq 3.398) = 1 - 0.7508 + 0.0003 = 0.2495$.
- (d) We benaderen de verdeling van $Y \sim \text{Bin}(4400, p)$ met $Y \sim \mathcal{N}(4400p, 4400p(1-p))$. Vervolgens benaderen we de variantie van deze normale verdeling met $4400 \frac{y}{4400} (1 - \frac{y}{4400})$. Dit geeft als 95% betrouwbaarheidsinterval voor p :

$$\hat{p} = \frac{y}{4400} \pm 1.96 \frac{1}{\sqrt{4400}} \sqrt{\frac{y}{4400} (1 - \frac{y}{4400})}$$

Invullen van $y = 427$ levert op $\hat{p} = 0.0970 \pm 0.0087 = [0.088, 0.106]$.

3. (a) Er geldt $\prod_{i=1}^n f_{\theta}(x_i) = (1 - \theta)^{\sum x_i - n} \theta^n$. De likelihood ratio statistiek (LRS) wordt gegeven door

$$\begin{aligned} \text{LRS} &= \frac{\sup_{\theta \in (0,1)} \prod_{i=1}^n f_{\theta}(X_i)}{\prod_{i=1}^n f_{\theta=\theta_0}(X_i)} \\ &= \frac{\prod_{i=1}^n f_{\theta=\hat{\theta}}(X_i)}{\prod_{i=1}^n f_{\theta=\theta_0}(X_i)} \\ &= \frac{(1 - \frac{1}{\bar{X}})^{n(\bar{X}-1)} \frac{1}{\bar{X}^n}}{(1 - \theta_0)^{n(\bar{X}-1)} \theta_0^n} \end{aligned}$$

- (b) De score functie op basis van één waarneming is gelijk aan:

$$\dot{\ell}_{\theta}(x) = \frac{-(x-1)}{1-\theta} + \frac{1}{\theta}.$$

De Fisher informatie op basis van één waarneming is derhalve

$$\text{var}_{\theta} \dot{\ell}_{\theta}(X) = \frac{1}{(1-\theta)^2} \text{var}_{\theta}(X) = \frac{1}{(1-\theta)\theta^2}$$

Dus de plug-in schatter wordt gegeven door $\frac{\bar{X}^2}{(1-\frac{1}{\bar{X}})} = \frac{\bar{X}^3}{\bar{X}-1}$.

De waargenomen Fisher informatie wordt gegeven door $-\frac{1}{n} \sum_{i=1}^n \ddot{\ell}_{\hat{\theta}}(X_i)$.

Vanwege $\ddot{\ell}_{\theta}(x) = \frac{-(x-1)}{(1-\theta)^2} - \frac{1}{\theta^2}$, krijgen we

$$-\frac{1}{n} \sum_{i=1}^n \ddot{\ell}_{\hat{\theta}}(X_i) = -\frac{-\frac{1}{n} \sum_{i=1}^n (X_i - 1)}{(1 - \hat{\theta})^2} + \frac{1}{\hat{\theta}^2} = \frac{-(1 - \bar{X})/\bar{X}^2 + (1 - \frac{1}{\bar{X}})^2}{(1 - \frac{1}{\bar{X}})^2 \frac{1}{\bar{X}^2}}$$

Verder uitwerken levert eveneens $\frac{\bar{X}^3}{\bar{X}-1}$

- (c) Een benaderend 95% betrouwbaarheidsinterval voor θ gebaseerd op de meest aannemelijke schatter is

$$\theta = \hat{\theta} \pm \frac{1}{\sqrt{n} \sqrt{\hat{i}_{\theta}}} \xi_{0.025} = \frac{1}{\bar{X}_n} \pm 1.96 \frac{\sqrt{\bar{X}_n - 1}}{\sqrt{n \bar{X}_n^3}} = 0.333 \pm 0.11928 = [0.214, 0.453]$$

Een benaderend 95% betrouwbaarheidsinterval voor θ gebaseerd op de LRS is gebaseerd op de eigenschap dat $2 \log LRS$ in verdeling convergeert naar een χ_1^2 -verdeling als $n \rightarrow \infty$. Met andere woorden, het 95% betrouwbaarheidsinterval wordt gegeven door

$$\begin{aligned} & \{ \theta : 2n(\bar{X} - 1) \log(1 - \frac{1}{\bar{X}}) - 2n \log \bar{X} - 2n(\bar{X} - 1) \log(1 - \theta) - 2n \log \theta \leq \chi_{1,0.95}^2 \} \\ & = \{ \theta : 2n \log \theta + 2n(\bar{X} - 1) \log(1 - \theta) \geq 2n(\bar{X} - 1) \log(1 - \frac{1}{\bar{X}}) - 2n \log \bar{X} - 3.84 \} \\ & = \{ \theta : 40 \log \theta + 80 \log(1 - \theta) \geq -80.22 \} \end{aligned}$$

4. (a) Wegens de invariantie-eigenschap van meest aannemelijke schatters geldt dat de meest aannemelijke schatter voor $(\theta + \frac{1}{2})^2$ wordt gegeven door

$$(\hat{\theta} + \frac{1}{2})^2 = \frac{1}{4} + \frac{1}{n} \sum_{i=1}^n X_i^2$$

- (b) De kansdichtheid van de steekproef wordt gegeven door

$$\begin{aligned} p_{\theta}(x_1, \dots, x_n) &= \left(\frac{1}{\sqrt{2\pi\theta}} \right)^n \exp \left\{ -\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\theta} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi\theta}} \right)^n \exp\{-n\theta/2\} \exp\left\{ \sum_{i=1}^n x_i \right\} \exp \left\{ -\frac{\sum_{i=1}^n x_i^2}{2\theta} \right\} \end{aligned}$$

Dit levert een 1-parameter exponentiële familie, waarbij $\{-\frac{1}{2\theta} \mid \theta > 0\} = (-\infty, 0)$ een niet-leeg inwendig heeft. Dus is $\sum_{i=1}^n X_i^2$ een voldoende en volledige statistiek. De meest aannemelijke schatter voor $(\theta + \frac{1}{2})^2$ is een functie van deze statistiek. Bovendien is hij een zuivere schatter voor $(\theta + \frac{1}{2})^2$:

$$\begin{aligned} E_{\theta} \left(\frac{1}{4} + \frac{1}{n} \sum_{i=1}^n X_i^2 \right) &= \frac{1}{4} + E_{\theta} X_i^2 \\ &= \frac{1}{4} + \text{var}_{\theta}(X_i) + (E_{\theta}(X_i))^2 \\ &= \frac{1}{4} + \theta + \theta^2 = (\theta + \frac{1}{2})^2 \end{aligned}$$

(c) Wegens Cramér-Rao geldt voor iedere zuivere schatter van $g(\theta) = (\theta + \frac{1}{2})^2$:

$$\text{var}_\theta T \geq \frac{g'(\theta)^2}{I_\theta} = \frac{(1 + 2\theta)^2}{n i_\theta}.$$

Omdat $\dot{\ell}_\theta(x) = -\frac{1}{2\theta} - \frac{1}{2} + \frac{1}{2\theta^2}x^2$, geldt

$$i_\theta = \frac{1}{4\theta^4} \text{var}_\theta(X_i^2) = \frac{1 + 2\theta}{2\theta^2},$$

en krijgen we als ondergrens

$$\frac{g'(\theta)^2}{n i_\theta} = \frac{2\theta^2(1 + 2\theta)}{n} = \frac{2\theta^2 + 4\theta^3}{n}.$$

Wegens

$$\text{var}_\theta\left(\frac{1}{4} + \frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n} \text{var}_\theta(X_i^2) = \frac{2\theta^2 + 4\theta^3}{n}$$

wordt de ondergrens bereikt door de UMVZ schatter voor $(\theta + \frac{1}{2})^2$.

Normering:

1(a): 7	2(a): 5	3(a): 6	4(a): 4
1(b): 6	2(b): 8	3(b): 9	4(b): 10
1(c): 8	2(c): 6	3(c): 5	4(c): 8
	2(d): 8		

Eindcijfer = (totaal + 10)/10