

Tentamen Inleiding Statistiek (wi2615)
22 januari 2010, 9.00-12.00u.

Een antwoord alleen is *niet* voldoende: er dient een berekening, toelichting en/of motivatie aanwezig te zijn. Dit alles goed leesbaar en in goed Nederlands. Succes gewenst met dit tentamen!

1. We beschikken over een steekproef $X_1, \dots, X_n$ uit de verdeling met kansdichtheid

$$f(x|\theta) = \begin{cases} \theta(1+x)^{-\theta-1}, & \text{als } x \geq 0 \\ 0, & \text{als } x < 0. \end{cases}$$

Hier is $\theta > 0$ een onbekende parameter.

- a. Laat zien dat de familie $\{f(\cdot|\theta) : \theta > 0\}$ een exponentiële familie van dichtheden is. Concludeer uit de bijbehorende representatie dat $\sum_{i=1}^n \log(1 + X_i)$ een voldoende (sufficient) grootte is voor θ .

Oplossing: Schrijf de dichtheid van X_i als

$$f(x|\theta) = \exp \{ \log \theta - \log(1+x) - \theta \log(1+x) \} = \exp \{ c(\theta)T(x) + d(\theta) + S(x) \}, x \in A$$

Hier geldt $c(\theta) = -\theta$, $T(x) = \log(1+x)$, $S(x) = -\log(1+x)$, $d(\theta) = \log \theta$ en $A = [0, \infty)$. Dit laat zien dat het een exponentiële familie is. Voor de simultane dichtheid van $X_1, \dots, X_n$ geldt dan dat

$$f(x_1, \dots, x_n|\theta) = \exp \left\{ n \log \theta - \sum_{i=1}^n \log(1+x_i) - \theta \sum_{i=1}^n \log(1+x_i) \right\}, x_i \geq 0$$

De factoriseringsstelling van Neyman geeft dat $\sum_{i=1}^n T(X_i) = \sum_{i=1}^n \log(1+X_i)$ voldoende is voor θ .

- b. Leid de maximum likelihood schatter voor θ af, gebaseerd op $X_1, \dots, X_n$.

Oplossing: De log likelihood wordt gegeven door

$$\ell(\theta) = \sum_{i=1}^n \log f_{\theta}(X_i) = \sum_{i=1}^n (\log \theta - (\theta + 1) \log(1 + X_i)).$$

De afgeleide hiervan is

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^n \log(1 + X_i)$$

Nulstellen levert $\hat{\theta}_n = \left(\frac{1}{n} \sum_{i=1}^n \log(1 + X_i) \right)^{-1}$; tekenoverzicht laat zien dat de log likelihood in dit punt maximaal is.

2. We hebben een enkele observatie X uit een uniforme verdeling op $[0, \theta]$. De uniforme verdeling op $[0, \theta]$ heeft kansdichtheid $f(x|\theta) = \theta^{-1} 1_{[0, \theta]}(x)$. We willen de hypothese $H_0 : \theta = 1$ toetsen tegen alternatief $H_1 : \theta = 2$. De toets wordt gebaseerd op X , waarbij H_0 ten gunste van H_1 wordt verworpen voor grote waarden van X .

- a. Bepaal de kritieke waarde horend bij deze toets, bij significantieniveau $\alpha = 0.10$.

Oplossing: De nulhypothese wordt verworpen voor $X > c$. De kritieke waarde c wordt bepaald door de eis $P_{\theta=1}(X > c) = \int_c^1 dx = 0.10$, dus $c = 0.90$.

- b. Bepaal het onderscheidingsvermogen (de power) van deze toets.

Oplossing: Het onderscheidingsvermogen is de kans op verwerpen van H_0 als de alternatieve hypothese waar is. In dit geval dus $P_{\theta=2}(X > 0.90) = \int_{0.90}^2 \frac{1}{2} dx = 0.55$.

- c. In problemen waar we een enkelvoudige nulhypothese tegen een enkelvoudige alternatieve hypothese toetsen, ligt het voor de hand de toets op de likelihood ratio te baseren. Geef de uitdrukking voor de likelihood ratio in de huidige context. Beargumenteer dat de beslisregel om H_0 ten gunste van H_1 te verwerpen voor kleine waarden van de likelihood ratio niet eenduidig leidt tot de toets hierboven die hier is beschreven.

Oplossing: De likelihood ratio is in dit geval

$$LR(x) = \frac{f(x|1)}{f(x|2)} = \begin{cases} 2 & \text{als } 0 \leq x \leq 1 \\ 0 & \text{als } x \in (1, 2] \\ \text{ongedefinieerd} & \text{als } x > 2 \end{cases}$$

Onder beide hypothesen treedt $X > 2$ met kans nul op. Op het relevante gebied $[0, 2]$ is LR het kleinst op $(1, 2]$. De toets die op basis van een realisatie in dit gebied verwerpt, zal echter significantieniveau 0 hebben. Het verwerpingsgebied dient te worden uitgebreid. In deze opgave is gekozen om dat te doen naar het interval $[0.90, 2]$, volgens het LR principe had echter ook kunnen worden gekozen voor een verwerpingsgebied $[0, 0.10] \cup [1, 2]$. Deze toets heeft hetzelfde onderscheidingsvermogen.

3. Zij $X_1, X_2, \dots, X_n$ onafhankelijke gelijkverdeelde stochastische variabelen uit een verdeling met kansdichtheid f_{μ, σ^2} ($\mu \in \mathbb{R}, \sigma^2 > 0$) met verwachting μ en variantie σ^2 . Beschouw, voor $c_1, \dots, c_n \in \mathbb{R}$, de volgende schatter voor μ : $\hat{\mu} = \sum_{i=1}^n c_i X_i$.

- a. Leid de voorwaarde af waar $c_1, \dots, c_n$ aan moeten voldoen om $\hat{\mu}$ een zuivere schatter voor μ te maken.

Oplossing: De eis voor zuiverheid is $E\hat{\mu} = \mu$. Lineariteit van de verwachting levert dan dat voor alle $\mu \in \mathbb{R}$, $\sum_{i=1}^n c_i \mu = \mu$. Dus: $\sum_{i=1}^n c_i = 1$.

- b. Bepaal $c_1, \dots, c_n$ zodanig dat $\hat{\mu}$ de zuivere schatter voor μ is met minimale variantie. Bereken hiertoe eerst de variantie van $\hat{\mu}$ als functie van $c_1, \dots, c_n$ en minimaliseer die over de toegestane waarden van $c_1, \dots, c_n$.

Oplossing: De directe manier gaat via een Lagrange multiplier. De variantie van $\hat{\mu}$ wordt gegeven door

$$\phi(c_1, \dots, c_n) = \sigma^2 \sum_{i=1}^n c_i^2$$

Minimaliseren van deze functie over die c_i 'tjes waarvoor geldt $\sum_{i=1}^n c_i = 1$, is hetzelfde als minimaliseren van de functie

$$\phi_\lambda(c_1, \dots, c_n) = \phi(c_1, \dots, c_n) + \lambda \left(\sum_{i=1}^n c_i - 1 \right)$$

over die c_i 'tjes. Minimaliseren we ϕ_λ over $\mathbb{R}^n$, kan door differentiëren worden gedaan:

$$\frac{\partial}{\partial c_i} \phi_\lambda(c_1, \dots, c_n) = 2\sigma^2 c_i + \lambda = 0 \iff c_i = c_i^{(\lambda)} = -\frac{\lambda}{2\sigma^2}.$$

Voor negatieve λ is het stationaire punt inderdaad een minimum. Merk op dat $\sum_1^n c_i^\lambda = -\frac{n\lambda}{2\sigma^2}$. Als $\lambda = -2\sigma^2/n$ wordt genomen (dit correspondeert met $c_i = 1/n$ voor alle i), wordt aan de 'zuiverheidsvoorwaarde' voldaan. Dit geeft voor iedere $(c_1, \dots, c_n) \in \mathbb{R}^n$ met $\sum_1^n c_i = 1$,

$$\phi(c_1, \dots, c_n) = \phi_\lambda(c_1, \dots, c_n) \geq \phi_\lambda(c_1^{(-2\sigma^2/n)}, \dots, c_n^{(-2\sigma^2/n)}) = \phi(c_1^{(-2\sigma^2/n)}, \dots, c_n^{(-2\sigma^2/n)}).$$

4. Zij $X_1, X_2, \dots, X_n$ onafhankelijke stochastische variabelen, allemaal uniform verdeeld op het interval $[0, 1]$. Zij $\mathbb{F}_n$ de empirische verdelingsfunctie gebaseerd op $X_1, X_2, \dots, X_n$. Bepaal voor $x \in (0, 1)$ de verwachting en de variantie van $\mathbb{F}_n(x)$.

Oplossing: De empirische verdelingsfunctie in punt x is in feite een gemiddelde van n onafhankelijke identiek verdeelde Bernoulli variabelen. Er geldt

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, x]}(X_i).$$

Dit betekent voor $x \in (0, 1)$ dat $E\mathbb{F}_n(x) = E1_{(-\infty, x]}(X_i) = P(X_i \leq x) = x$. Bovendien geldt $\text{Var}(\mathbb{F}_n(x)) = x(1-x)/n$.

5. Beschouw een kansdichtheid f op $(0, \infty)$ met $f(x) > 0$ voor alle $x > 0$. Schrijf F voor de bijbehorende verdelingsfunctie, h voor de bijbehorende hazard functie en H voor de cumulatieve hazard functie. Ten overvloede,

$$h(x) = -\frac{d}{dx} \log(1 - F(x)) \text{ en } H(x) = \int_0^x h(y) dy.$$

Zij Y een standaard exponentieel verdeelde stochastische variabele, dus met kansdichtheid $g(y) = e^{-y} 1_{[0, \infty)}(y)$. Laat zien dat de stochastische variabele X , gedefinieerd als $X = H^{-1}(Y)$ verdeeld is volgens de verdeling met kansdichtheid f .

Oplossing: Er geldt $P(Y \leq y) = 1 - e^{-y}$ voor $y \geq 0$. Tevens geldt direct op basis van de definitie van H dat $1 - F(x) = \exp(-H(x))$ en dat H een strikt stijgende continue functie is. We kunnen schrijven voor $x > 0$

$$P(X \leq x) = P(H^{-1}(Y) \leq x) = P(Y \leq H(x)) = 1 - e^{-H(x)} = F(x).$$

6. Er wordt een experiment uitgevoerd om het effect van een bepaalde behandeling na te gaan. Vijf experimentele eenheden krijgen de behandeling, vijf andere niet. De resulterende metingen worden als volgt gemodelleerd.

$$X_1, X_2, \dots, X_5 \stackrel{iid}{\sim} N(\mu_X, 25), \quad Y_1, Y_2, \dots, Y_5 \stackrel{iid}{\sim} N(\mu_Y, 25),$$

hier is bedoeld dat de varianties gelijk zijn aan 25, niet de standaardafwijkingen. De beide steekproeven worden ook onafhankelijk verondersteld. De onderzoeksvraag vertaalt zich in het toetsen van de hypothese $H_0 : \mu_X = \mu_Y$ tegen alternatief $H_1 : \mu_X > \mu_Y$; de situatie van het zogenaamde ‘twee steekproevenprobleem’. De toets wordt gebaseerd op het verschil in gemiddelden, $T = \bar{X} - \bar{Y}$.

- a. Geef de standaardafwijking van toetsingsgrootheid T .

Oplossing: Wegens onafhankelijkheid van de steekproeven geldt

$$\text{Var}(T) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{1}{5} \text{Var}(X_1) + \frac{1}{5} \text{Var}(Y_1) = 5 + 5 = 10,$$

zodat de standaardafwijking gelijk is aan $\sqrt{10}$.

- b. Geef aan voor welke waarden (groot, klein, beide) van de toetsingsgrootheid H_0 ten gunste van H_1 wordt verworpen. Met andere woorden, geeft de kwalitatieve vorm van het verwerpsgebied.

Oplossing: Verwerpen doe je als de waarnemingen meer overeenkomen met H_1 dan met H_0 . Als H_1 geldt, zal T de neiging hebben hoger uit te vallen dan als H_0 waar is. De nulhypothese wordt dus verworpen voor grote waarden van T .

- c. Het experiment wordt uitgevoerd en de uitkomst voor T blijkt $t = 3.05$ te zijn. Geef de p -waarde behorende bij deze observatie. Om hierbij te helpen, worden de volgende regels in het statistisch pakket **R** uitgevoerd (met bijbehorende resultaten):

```
> dnorm(3.05/sqrt(10))
[1] 0.2505583
> dnorm(3.05/10)
[1] 0.3808114
> pnorm(3.05/sqrt(10))
[1] 0.832601
> pnorm(3.05/10)
[1] 0.6198169
> qnorm(3.05/sqrt(10))
[1] 1.805409
> qnorm(3.05/10)
[1] -0.5100735
```

Oplossing: H_0 wordt verworpen voor grote waarden van T . De p -waarde is daarom de kans op $T > 3.05$, onder de nulhypothese. Onder de nulhypothese is T normaal verdeeld met verwachting nul en variantie 10. De p -waarde wordt dus $1 - \Phi(3.05/\sqrt{10})$, waar Φ de verdelingsfunctie van een standaardnormale stochast is. In **R** is dit `1-pnorm(3.05/sqrt(10))`. De p -waarde is dan $1 - 0.833 = 0.167$.

7. Zij $0 < x_1 < x_2 < \dots < x_n < 1$ en de stochastische variabelen Y_i , $1 \leq i \leq n$, gedefinieerd door het enkelvoudige regressiemodel zonder intercept:

$$Y_i = \beta x_i + \epsilon_i, \text{ waar } \epsilon_1, \epsilon_2, \dots, \epsilon_n \sim^{iid} N(0, 1).$$

- a. Laat zien dat de kleinste kwadratenschatter $\hat{\beta}$ voor β wordt gegeven door

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Oplossing: De KK schatter is gedefinieerd als die β waarvoor de som van kwadraten

$$\phi(\beta) = \sum_{i=1}^n (Y_i - \beta x_i)^2$$

minimaal is. Differentiëren, nulstellen en een tekenoverzicht leiden tot het resultaat

- b. Bereken op basis van deze uitdrukking $E\hat{\beta}$ en $\text{Var}(\hat{\beta})$.

Oplossing: Er geldt, waarbij we voor de variantie de onafhankelijkheid gebruiken,

$$E\hat{\beta} = \frac{\sum_{i=1}^n x_i EY_i}{\sum_{i=1}^n x_i^2} = \beta \text{ en } \text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^n x_i^2 \text{Var}(Y_i)}{(\sum_{i=1}^n x_i^2)^2} = \left(\sum_{i=1}^n x_i^2 \right)^{-1}.$$

Puntenverdeling:

1a:3	2a:2	2c:2	3b:4	5:3	6b:2	7a:2
1b:3	2b:2	3a:2	4:3	6a:2	6c:2	7b:3

Het tentamencijfer wordt als volgt bepaald: $\frac{\text{aantal punten} + 4}{3.9}$.