

Tentamen Inleiding Statistiek (TW2080)
10 november 2017, 13:30-16:30u

Formulebladen, rekenmachines, boeken, aantekeningen etc. zijn bij dit tentamen niet toegestaan.

1. Beschouw een populatie $\{x_1, x_2, \dots, x_N\}$ bestaande uit nullen en enen. Zij p de fractie enen in de populatie, dus het populatiegemiddelde $p = \frac{1}{N} \sum_{i=1}^N x_i$. Zij σ^2 de populatievariantie, dus $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - p)^2$.

(a) [1pt] Laat zien dat $\sigma^2 = p(1 - p)$.

(b) [2pt] Beschouw een aselechte steekproef $X_1, \dots, X_n$ met teruglegging, uit de populatie. In het licht van a) is $T_n = \bar{X}_n(1 - \bar{X}_n)$ een voor de hand liggende schatter voor σ^2 . Bereken de verwachting van T_n en concludeer of dit een zuivere schatter voor σ^2 is of niet.

2. Beschouw de volgende klasse van Weibull kansdichtheden:

$$f(x|\theta) = \frac{3}{\theta^3} x^2 e^{-(x/\theta)^3} 1_{[0,\infty)}(x) \quad (1)$$

waar $\theta \in \Theta = (0, \infty)$.

- (a) [2pt] Laat zien dat de bovenstaande klasse van dichtheden een exponentiële familie vormen.
- (b) [2pt] Stel U is een uniform op $[0, 1]$ verdeelde stochastische variabele. Geef aan hoe met behulp hiervan een stochastische variabele met kansdichtheid $f(\cdot|1)$ (dus met $\theta = 1$) kan worden gegenereerd. Als u de correcte R-code geeft, levert dit een halve bonuspunt op.
- (c) [3pt] We beschikken over een steekproef $X_1, \dots, X_n$ gegenereerd door een dichtheid als weergegeven in (1) met $\theta > 0$ onbekend. Laat zien dat de maximum likelihood schatter voor θ wordt gegeven door

$$\hat{\theta}_n = \sqrt[3]{T_n}, \quad \text{met} \quad T_n = \frac{1}{n} \sum_{i=1}^n X_i^3.$$

3. We beschikken over de realisatie x van een stochastische variabele X met een normale verdeling met onbekende verwachting $\mu \in \mathbb{R}$ en variantie $\sigma^2 = 1$. Dit betekent dat deze stochastische variabele dichtheid

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}, \quad x \in \mathbb{R}$$

heeft. We willen de nulhypothese $H_0 : \mu = 3$ toetsen tegen de alternatieve hypothese $H_1 : \mu > 3$.

- (a) [2pt] Laat zien dat de gegeneraliseerde likelihood ratio toets neerkomt op het verwerpen van H_0 ten gunste van H_1 voor grote waarden van X .
- (b) [2pt] Ga na of de resulterende toets uniform meest onderscheidend is.
4. We beschikken over een dataset $\{x_1, \dots, x_{25}\}$. Deze vatten we op als realisatie van een steekproef $X_1, \dots, X_{25}$ uit een onbekende kansdichtheid f . Het vermoeden bestaat dat f behoort tot de klasse van normale dichtheden, dus dat

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x - \mu)^2/2\sigma^2) \text{ voor } x \in \mathbb{R} \text{ en } (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty).$$

- (a) [1pt] We willen grafisch nagaan of het vermoeden redelijk is. Geef aan hoe u dit kunt doen; geef eventueel ook aan welke functie in R u hiervoor kunt gebruiken.
- (b) [2pt] Iemand claimt dat de verdeling waaruit de data zijn gegenereerd de verdeling met verdelingsfunctie

vergelijk F met eedf $F(x) = \frac{\exp(x)}{1 + \exp(x)}, x \in \mathbb{R}$

is. Beschrijf een procedure volgens welke deze claim formeel kan worden getoetst. Ga hierbij in op de keuze van een mogelijke toetsingsgrootheid en het bepalen (of benaderen) van een bijbehorende kritieke waarde.

5. Stel dat we eenzelfde type meting doen aan 3 verschillende machines. We veronderstellen het 1-factor ANOVA-model. De j -de meting op de i -de machine noteren we met Y_{ij} . Per machine verrichten we 2 metingen. Het 1-factor model is te schrijven als

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i \in \{1, 2, 3\} \quad j \in \{1, 2\}$$

waarbij $\{\epsilon_{ij}\}$ onafhankelijke stochastische variabelen zijn met een $N(0, \sigma^2)$ -verdeling. Noteer de vector met metingen als

$$\mathbf{Y} = (Y_{1,1}, Y_{1,2}, Y_{2,1}, Y_{2,2}, Y_{3,1}, Y_{3,2})^T.$$

- (a) [2pt] Schrijf het bovenstaande model als een lineair model. Dwz. in de vorm $\mathbf{Y} = X\beta + \varepsilon$.
- (b) [1pt] Is er een unieke kleinste-kwadraten oplossing? Motiveer uw antwoord.

Het cijfer C voor dit tentamen wordt als volgt bepaald:

$$C = \min \left(1 + 9 \times \frac{\text{aantal punten}}{20}, 10 \right)$$

Eindcijfer TW2080: $(8C + 2\text{TussenToets})/10$

Veel succes!

Antwoorden:

1a Omdat $x_i \in \{0, 1\}$, geldt $x_i^2 = x_i$. Dit geeft

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - p)^2 = \frac{1}{N} \sum_{i=1}^n (x_i^2 - 2px_i + p^2) = p - 2p^2 + p^2 = p(1 - p).$$

1b Omdat voor iedere stochastische variabele Y geldt dat $\text{Var}(Y) = E[Y^2] - (E[Y])^2$, volgt, gebruikmakend van lineariteit van de verwachting dat

$$E[T_n] = E[\bar{X}_n - \bar{X}_n^2] = E[\bar{X}_n] - \text{Var}(\bar{X}_n) - (E[\bar{X}_n])^2.$$

Weer wegens lineariteit, geldt $E[\bar{X}_n] = E[X_1] = p$. Onafhankelijkheid van de X_i 'tjes levert $\text{Var}(\bar{X}_n) = \text{Var}(X_1)/n = p(1 - p)/n$. Hieruit volgt

$$E[T_n] = p - \frac{p(1 - p)}{n} - p^2 = p(1 - p) \left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right) \sigma^2.$$

Dit betekent dat T_n geen zuivere schatter voor σ^2 is. Dan had de verwachting gelijk aan σ^2 moeten zijn.

2a Er geldt

$$\begin{aligned} f(x|\theta) &= \exp \{ \log 3 - 3 \log \theta + 2 \log x - (x/\theta)^3 \} 1_{[0, \infty)}(x) \\ &= \exp(c(\theta)T(x) + d(\theta) + S(x)) \end{aligned}$$

met $S(x) = 2 \log x$, $d(\theta) = \log 3 - 3 \log \theta$, $c(\theta) = -1/\theta^3$, $T(x) = x^3$.

2b Dit kan via de inverse verdelingsfunctie methode. De verdelingsfunctie behorend bij $f(\cdot|1)$ wordt gegeven door

$$F(x|1) = \int_0^x f(y|1) dy = 1 - \exp\{-x^3\}$$

Oplossen van $u = F(x|1)$ naar x geeft $x = (-\log(1 - u))^{1/3}$, waaruit volgt dat $F^{-1}(u|1) = (-\log(1 - u))^{1/3}$. Dit geeft het volgende voorschrift:

$$X = F^{-1}(U|1) = (-\log(1 - U))^{1/3}.$$

Omdat U en $1 - U$ dezelfde verdeling hebben, kan $1 - U$ in het voorschrift ook worden vervangen door U . De R-code is:

```
x<-(-log u)^(1/3).
```

2c De likelihood wordt gegeven door

$$L(\theta) = \frac{3}{\theta^{3n}} \prod_{i=1}^n X_i^2 \exp\left(-\sum_{i=1}^n \frac{X_i^3}{\theta^3}\right).$$

Dus is de loglikelihood is

$$\ell(\theta) = \log 3 - 3n \log \theta + \log \prod_{i=1}^n X_i^2 - \theta^{-3} \sum_{i=1}^n X_i^3.$$

Afgeleide nemen naar θ geeft

$$\ell'(\theta) = \frac{-3n}{\theta} + \frac{3 \sum_{i=1}^n X_i^3}{\theta^4}.$$

Oplossen van $\ell'(\theta) = 0$ en checken van het tekenoverzicht geeft dat de waarde van θ die L maximaliseert gelijk is aan $\hat{\theta}_n = \sqrt[3]{T_n}$, met $T_n = \frac{1}{n} \sum_{i=1}^n X_i^3$.

3a De GLR grootheid wordt gegeven door

$$\frac{1}{2} \frac{\sup_{\mu \in H_0} f(x|\mu)}{\sup_{\mu \geq 3} f(x|\mu)} = \begin{cases} f(x|3)/f(x|x) = \exp(-(x-3)^2/2) & \text{als } x \geq 3 \\ f(x|3)/f(x|3) = 1 & \text{als } x < 3 \end{cases}$$

Inderdaad, het supremum in de teller wordt aangenomen in $\mu = x$. In de noemer wordt het supremum aangenomen in $\mu = x$ als $x \geq 3$. als $x < 3$, dan wordt het supremum aangenomen in $\mu = 3$.

De nulhypothese wordt verworpen voor kleine waarden van deze grootheid, wat overeenkomt met grote waarden van $x - 3$. $\frac{1}{2}$

3b Voor enkelvoudige $H_1 : \mu = \mu_1$, waar $\mu_1 > 3$, wordt de LR grootheid gegeven door

$$\begin{aligned} \frac{1}{2} \frac{f(x|3)}{f(x|\mu_1)} &= \exp((x - \mu_1)^2/2 - (x - 3)^2/2) = \exp(-\mu_1 x + \mu_1^2/2 + 3x - 9/2) \\ &= \exp((3 - \mu_1)x + \mu_1^2/2 - 9/2). \end{aligned}$$

De nulhypothese wordt verworpen voor kleine waarden van deze LR. Omdat $\mu_1 > 3$, komt de LR toets neer op het verwerpen van H_1 voor grote waarden van X . Dit komt overeen met de GLR toets die onder a) is afgeleid. Wegens het Neyman-Pearson lemma is de LR toets meest onderscheidend. Omdat de beslisregel niet afhangt van de precieze waarde van μ_1 , zolang μ_1 maar groter is dan 3, is de (G)LR toets uniform meest onderscheidend.

4a Hier kan aan een histogram worden gedacht, $\text{hist}(x)$ of aan de empirische verdelingsfunctie, $\text{plot(ecdf}(x))$. Ook een qq-plot ligt voor de hand.

ip noemen
hist, ecdf, qqplot

4b Hier kan de Kolmogorov Smirnov toets worden gebruikt. De toetsingsgrootheid is de supremumafstand tussen de empirische verdelingsfunctie gebaseerd op de data en de geclaimde verdelingsfunctie. De verdeling van de toetsingsgrootheid onder H_0 kan worden gesimuleerd door herhaaldelijk een steekproef van omvang 25 uit de geclaimde verdeling te trekken en vervolgens de KS grootheid uit te rekenen [als opgemerkt wordt dat zelfs uit de uniforme verdeling kan worden getrokken, met bijbehorende variant op de KS-grootheid, is dit een plusje]. Bij significantieniveau $\alpha = 5\%$ wordt de kritieke waarde dan gegeven door het 95% empirische kwantiel van de gegenereerde waarden.

prob plot $\left(\frac{i}{n+1}, F(x_i)\right)$

$x=y$

qq plot $\left(F^{-1}\left(\frac{i}{n+1}\right), x_i\right)$

$x=y$

5a Neem $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)^T$ en

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

5b Nee, kolommen van X zijn lineair afhankelijk, dus $X^T X$ is singulier. Alternatief kan worden opgemerkt dat het model zonder verdere restricties overgeparametriseerd is.