Tentamen toegepaste statistiek voor TA (WI1275TA)

18 april 2012, 14:00-17:00

Afdeling Toegepaste Wiskunde, Faculteit Elektrotechniek, Wiskunde en Informatica

---

Toelichting: 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. Rekenmachine toegestaan. Formuleblad/boek/aantekeningen niet toegestaan.

---

1. Stel dat een stochastische variabele X kansdichtheid f heeft, waarbij

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 < x \leq 2 \\ 0 & x \notin [0, 2] \end{cases}.$$

   Bereken de verdelingsfunctie $F(x)$ van X voor alle $x \in \mathbb{R}$.

2. We testen het bloed 1000 personen op een zeldzame ziekte. Bekend is dat

$$\mathrm{P}(\text{persoon test positief}) = 0.005.$$

   De standaard procedure is om ieder persoon te testen. We bekijken nu een alternatieve procedure. Bij deze procedure verdelen we de personen over 25 groepen van 40 en testen per groep een mengsel van het bloed van de 40 personen. Als een mengsel positief test, dan worden alsnog alle 40 personen binnen de betreffende groep afzonderlijk getest.

   (a) Noem nu het aantal tests voor de i-de groep X_i. Geef de kansmassafunctie van X_i.

   (b) Bereken het verwachte totaal aantal testen dat nodig is met de alternatieve procedure (om alle 1000 personen getest te hebben).

   (c) Bereken de kans dat je met de alternatieve procedure meer testen nodig hebt dan met de standaard procedure. Je kunt dit direct uitrekenen (benaderen met de centrale limietstelling is niet nodig).

3. Voor een visgroothandel is een geautomatiseerd vissorteringssysteem ontwikkeld. Om het systeem te testen zijn 100 zeebaarsen en 100 zalmen aangeboden ter klassificatie, met het volgende resultaat:

werkelijke soort	geklassificeerd als zeebaars	zalm
zeebaars	80	20
zalm	10	90

   Een willekeurige vis, gevangen in een gebied waar de verhouding zeebaars-zalm ongeveer 70 : 30 is, wordt door het systeem als zeebaars geklassificeerd. Bereken de kans dat

de gevangen vis daadwerkelijk in het echt zeebaars is.
Hint: Definieer de gebeurtenissen

$$T = \{\text{vis is als zeebaars geklassificeerd}\} \quad \text{en} \quad B = \{\text{vis is zeebaars}\}.$$

Schrijf vervolgens de gevraagde kans als conditionele kans met behulp van deze gebeurtnissen en pas de regel van Bayes toe.

4. De Rayleigh verdeling is een kansverdeling met cumulatieve verdelingsfunctie

$$F(x) = \begin{cases} 0 & \text{als } x < 0 \\ 1 - e^{-x^2/\theta^2} & \text{als } x \geq 0 \end{cases}.$$

Deze verdeling wordt vaak gebruikt voor het modelleren van bijvoorbeeld golfhoogten of windsnelheden. We veronderstellen $\theta > 0$. Het is eenvoudig in te zien dat de kansdichtheid van X voor $x \geq 0$ gegeven wordt door

$$f(x) = \frac{2x}{\theta^2} e^{-x^2/\theta^2}.$$

Bereken de meest aannemelijke (maximum likelihood) schatter voor θ op grond van realisaties $x_1, \ldots, x_n$ van onafhankelijke stochastische variabelen $X_1, \ldots, X_n$ met deze verdeling.

5. Stel dat we data $x_1, \ldots, x_n$ hebben die we opvatten als een realisatie van een steekproef $X_1, \ldots, X_n$ uit een normale verdeling met onbekende verwachting μ en bekende variantie σ^2. In dit geval wordt een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor μ gegeven door

$$\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right). \qquad (*)$$

Geef bij opgaven (a) en (b) aan of de bewering goed of fout is. Licht je antwoord toe (een correct antwoord zonder toelichting wordt levert geen punten op).

   (a) Het betrouwbaarheidsinterval met betrouwbaarheid 0.90 omvat het betrouwbaarheidsinterval met betrouwbaarheid 0.95.

   (b) Als σ onbekend is, dan ziet het betrouwbaarheidsinterval er hetzelfde uit als in formule $(*)$, behalve dat σ wordt vervangen door zijn schatter S_n.

   (c) Stel dat $\sigma = 2$ en $\alpha = 0.05$. Hoe groot moet n zijn opdat de lengte van het betrouwbaarheidsinterval maximaal 1 is?

6. Stel dat $X_1, X_2, , X_3$ een steekproef is uit een normale verdeling met verwachting μ en variantie 1. We toetsen de nulhypothese $\mu = 0$ tegen de alternatieve hypothese $\mu > 0$. We nemen als toetsingsgrootheid $T = \bar{X}_3$ en als kritiek gebied $[1, \infty)$.

   (a) Bereken de kans op een type 1 fout.

   (b) Bereken de kans op een type 2 fout als $\mu = 1$.

7. Je onderzoekt de response time voor een bepaald type database query. Je bent van plan een aantal vergelijkbare queries uit te voeren waarvoor het volgende model geldt:

$$R_i = c + U_i, \quad i = 1, \ldots, n,$$

waarbij R_i de response time is voor query i en c de maatgevende response time; U_i staat voor toevallige variatie, er geldt: $\mathrm{E}[U_i] = 0$, $\mathrm{Var}(U_i) = 4$, en $U_1, \ldots, U_n$ zijn onafhankelijk. De tijdseenheid is milliseconde. Je bent van plan het gemiddelde $\bar{R}_n$ als schatting voor c te gebruiken. We veronderstellen verder dat $n = 30$.

   (a) Bepaal $\mathrm{E}\left[\bar{R}_n\right]$ en $\mathrm{Var}\left(\bar{R}_n\right)$.

   (b) Geef met behulp van de centrale limietstelling een schatting van de kans $\mathrm{P}\left(|\bar{R}_n - c| < 1\right)$.

8. Beschouw het 2 steekproeven probleem (dwz. het toetsen op verschil tussen de verwachtingen in 2 populaties); MIPS hoofdstuk 28). Geef van de volgende beweringen aan of ze waar zijn of niet. Licht je antwoord toe (een correct antwoord zonder toelichting wordt levert geen punten op).

   (a) Het verdient altijd de voorkeur om de gepoolde variantieschatter te gebruiken in plaats van de ongepoolde variantieschatter.

   (b) Bij gebruik van de ongepoolde variantieschatter heeft de toetsingsgrootheid een t-verdeling, waarbij het aantal vrijheidsgraden gelijk is aan het totaal aantal waarnemingen minus 2.

Puntenverdeling bij open vragen:

Opgave:	1	2a	2b	2c	3	4	5a	5b	5c	6a	6b	7a	7b	8a	8b
Punten:	3	2	2	2	3	4	2	2	2	2	2	2	3	2	2

Beknopte uitwerkingen

1. Er geldt $F(x) = \int_{-\infty}^{x} f(a)da$. Het is eenvoudig om te zien dat $F(x) = 0$ als $x \leq 0$. Net zo, $F(x) = 1$ als $x \geq 2$. Voor $x \in [0,1]$ geldt $F(x) = \frac{1}{2}x^2$. Voor $x \in [1,2]$ geldt:

$$\begin{aligned} F(x) &= F(1) + \int_1^x (2-a)da = \frac{1}{2} + \left[2a - \frac{1}{2}a^2\right]_1^x \\ &= 2x - \frac{1}{2}x^2 - 1 \end{aligned}$$

2. (a) $X_i = 1$ met kans $p = 0.995^{40} \approx 0,82$; $X_i = 41$ met kans $1-p$.
   (b) $\mathrm{E}[X_i] = 1 \cdot p + 41 \cdot (1-p) \approx 8,27$. Het gevraagde antwoord is $40\mathrm{E}[X_1] = 331$.
   (c) Dit gebeurt alleen als $X_i = 41$ voor alle $1 \leq i \leq 25$. Aangezien alle X_i onafhankelijk en identiek verdeeld zijn, geldt dat dit gebeurt met kans $\mathrm{P}(X_1 = 41)^{25} = (1-p)^{25} \approx 3 \cdot 10^{-19}$.

3. T: "vis als zeebaars geklassificeerd"; B: "vis *is* zeebaars". Uit de tabel: $\mathrm{P}(T \mid B) = 0.8$, $\mathrm{P}(T \mid B^c) = 0.1$; voor de klassificatie zouden we schatten $\mathrm{P}(B) = 0.7$. Uit de regel van Bayes volgt: $\mathrm{P}(T) = 0.7 \cdot 0.8 + 0.3 \cdot 0.1 = 0.56 + 0.03 = 0.59$ en dus $\mathrm{P}(B \mid T) = 56/59 = 0.949$.

4. De loglikelihood wordt gegeven door

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i) = \sum_{i=1}^{n} \left(\ln(2x_i) - 2\ln\theta - \frac{x_i^2}{\theta^2}\right)$$

   Differentiëren naar θ geeft:

$$\ell'(\theta) = -\frac{2n}{\theta} + 2\frac{\sum_{i=1}^n x_i^2}{\theta^3}.$$

   Oplossen van $\ell'(\theta) = 0$ geeft

$$\theta = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}.$$

   Verifiëren van het tekenoverzicht van $\ell'(\theta)$ toont aan dat dit een maximum is.

5. (a) $z_{0.025} > z_{0.05}$, en dus is het 95%-BI breder dan het 90% BI. De bewering is dus onjuist.
   (b) Bewering is onjuist, de kritieke waarde $z_{\alpha/2}$ moet ook nog vervangen worden door $t_{n-1,\alpha/2}$.
   (c) De lengte is $2z_{\alpha/2}\sigma/\sqrt{n} = 7.84/\sqrt{n}$. Dit moet ≤ 1 zijn. Dus $n \geq (7.84)^2 \approx 61.5$. Neem dus $n \geq 62$.

6. (a)

$$\begin{aligned} P_{\mu=0}(T \in K) &= P_{\mu=0}(\bar{X}_3 > 1) = P_{\mu=0}(\sqrt{3}\bar{X}_n > \sqrt{3}) \\ &= P(Z > \sqrt{3}) \approx 0.042 \end{aligned}$$

(b) $P_{\mu=1}(T \notin K) = P_{\mu=1}(\bar{X}_3 < 1) = 0.5$

7. (a) Zie Hoofdstuk 13: $\mathrm{E}\left[\bar{R}_n\right] = c$ en $\mathrm{Var}\left(\bar{R}_n\right) = \mathrm{Var}(R_1)/n = 4/n$.

   (b) Volgens de CLS geldt $\bar{R}_n \sim N(c, 4/n)$; en $4/n = 4/30$, dus $\bar{R}_n - c \sim N(0, 4/30)$. Er volgt:

$$\begin{aligned}\mathrm{P}\left(|\bar{R}_n - c| < 1\right) &= \mathrm{P}\left(\frac{|\bar{R}_n - c|}{\sqrt{4/30}} < \frac{1}{\sqrt{4/30}}\right) \\ &\approx \mathrm{P}\left(|Z| < \sqrt{7.5}\right) = 1 - 2\mathrm{P}\left(Z > \sqrt{7.5}\right) \approx 0.9938.\end{aligned}$$

8. (a) Onjuist: dit hangt ervan af of de variantie in de 2 groepen gelijk is of niet.

   (b) Onjuist: als we de ongepoolde variantieschatter gebruiken, dan is de verdeling van de toetsingsgrootheid onbekend.