Examen Kansrekening en Wiskundige Statistiek: oplossingen

S. Vansteelandt

Academiejaar 2006-2007

---

1. Een team van onderzoekers wil nagaan of een bepaald geneesmiddel Triptan meer effectief is dan aspirine tegen migraine-aanvallen. Men heeft daartoe 100 migrainepatiënten aangeschreven en deze werden lukraak toegewezen aan de ene of de andere behandeling gedurende 1 maand (er werd hen opgelegd dagelijks de medicatie te nemen). Er werd hen gevraagd om een dagboek bij te houden van het totaal aantal uren hoofdpijn ze die maand ondervonden. Daarna volgde een periode van 2 maand waarin de patiënten geen behandeling werd toegediend. De daaropvolgende maand werd elke patiënt opnieuw behandeld maar nu met het andere middel (dus diegenen die in de eerste periode aspirine kregen ontvingen nu Triptan en vice versa). Op het einde van de eerste en de vierde maand sinds de start van de studie werd het totaal aantal uren hoofdpijn gerapporteerd aan de onderzoekers. Dit gaf de volgende data die aangegeven wordt als gemiddelde (SD)

n	Behandeling	Periode 1	Periode 2	Aspirine-Triptan
52	Triptan/Aspirine	59.2 (1.2)	68.3 (1.9)	9 (1.7)
48	Aspirine/Triptan	72.1 (1.4)	61.5 (2.1)	11.7 (2.5)

   (a) *Men kan aantonen dat er een periode effect is (d.i. een natuurlijke wijziging in gemiddeld aantal uren hoofdpijn per maand over de tijd, ongeacht de inname van een bepaald geneesmiddel) wanneer het gemiddeld verschil in aantal uren hoofdpijn tussen Aspirine en Triptan anders is in beide behandelingsgroepen. Stel een 95% betrouwbaarheidsinterval op om na te gaan of er in deze studie een periode effect aanwezig is.*

   Vermits de gegevens uit beide groepen ongepaard zijn, vinden we:

$$9 - 11.7 \pm t_{98,0.025}\sqrt{\frac{1.7^2 \times 51 + 2.5^2 \times 47}{98}}\sqrt{\frac{1}{52} + \frac{1}{48}} = [-3.5, -1.9]$$

   (b) *Wat is uw conclusie? Leg uit.*

   Het BI sluit 0 uit. Bijgevolg is er een significante indicatie van een periode-effect op het 5% significantieniveau.

   (c) *Aan welke voorwaarde(n) moet voldaan zijn opdat dit interval het juiste betrouwbaarheidsniveau zou hebben? Leg uit.*

   De verschillen tussen Aspirine en Triptan moeten in elke groep Normaal verdeeld zijn met dezelfde variantie, en zowel onafhankelijk binnen als tussen de groepen.

   (d) *Stel (ongeacht het resultaat dat u hierboven bekwam) dat er geen periode effect aanwezig is, zodat de wijziging in aantal uren hoofdpijn tussen beide studieperiodes kan toegeschreven worden aan de verschillende behandeling in beide periodes, en niet het gevolg is van een natuurlijke tijdsevolutie. Gebruik dan een - bij voorkeur*

zo krachtig mogelijke - toets om op het 5% significantieniveau na te gaan of er een gemiddeld verschil is in aantal uren hoofdpijn per maand na inname van Triptan versus Aspirine. Geef de p-waarde en formuleer een besluit.

In de afwezigheid van een periode-effect duidt het verschil tussen Aspirine en Triptan op het behandelingseffect. Beiden $\bar{X}_n$ en $\bar{Y}_m$, de steekproefgemiddelden van deze verschillen in de eerste en tweede groep, zijn dus onvertekende schattingen voor het behandelingseffect. We kiezen als teststatistiek

$$\frac{n\bar{X}_n + m\bar{Y}_m}{n+m}$$

met n en m de steekproefgroottes van beide groepen. Onder de nulhypothese dat beide geneesmiddelen even goed werken, is deze teststatistiek Normaal verdeeld met gemiddelde 0 en variantie $\sigma^2/(n+m)$, waarbij σ^2 de variantie is op de verschillen tussen Aspirine en Triptan voor 1 enkel individu. De p-waarde is bijgevolg

$$2\Phi\left(-\frac{(9\times 52 + 11.7\times 48)/100}{\sqrt{\frac{1.7^2\times 51 + 2.5^2\times 47}{98}}/\sqrt{100}}\right) = 0$$

hetgeen wijst op een zeer sterk significant behandelingseffect.

(e) *Aan welke voorwaarde(n) moet voldaan zijn opdat deze toets geldig zou zijn? Leg uit.*

Dezelfde als in vraag (c) (en het feit dat er geen periode-effect is).

Op basis van het aantal uren hoofdpijn in maand 1 heeft men de patiënten in verschillende hoofdpijnklassen (Klasse 1 heeft minder erge symptomen dan Klasse 2) opgedeeld. De data ziet er dan als volgt uit:

Behandeling	Klasse 1	Klasse 2	Totaal
Aspirine	$a = 19$	$b = 29$	48
Triptan	$c = 34$	$d = 18$	52
Totaal	$n_1 = 53$	$n_2 = 47$	100

(a) *Het relatief risico op een gebeurtenis A voor een behandeling 1 versus 2 is gedefinieerd als $P(A|$ behandeling $1)/P(A|$ behandeling $2)$. Schat het relatief risico om in klasse 1 terecht te komen voor patiënten met Aspirine versus Triptan.*

$$(19/48)/(34/52) = 0.61$$

(b) *De natuurlijke logaritme van het relatief risico is asymptotisch standaard normaal verdeeld met standaard error die kan geschat worden als*

$$\sqrt{\frac{c}{an_1} + \frac{d}{bn_2}}$$

Bereken een 95% betrouwbaarheidsinterval voor het relatief risico om in klasse 1 terecht te komen voor patiënten met Aspirine versus Triptan.

$$\exp(\ln(0.61) \pm 1.96\sqrt{34/(19\times 53) + 18/(29\times 47)} = [0.40, 0.93]$$

(c) *Interpreteer dit interval zorgvuldig. Kan u op basis hiervan een besluit formuleren of behandeling met Triptan meer effectief lijkt te zijn dan behandeling met Aspirine (m.b.t. de kans om minder erge symptomen van hoofdpijn te vertonen)? Leg uit.*

Het relatief risico om in klasse 1 terecht te komen, ligt met 95% kans, 7% tot 60% lager in de Aspirine groep dan in de Triptan groep.

2. Veronderstel dat we over een reeks onafhankelijke metingen $X_1, ..., X_n$ beschikken, die uniform verdeeld zijn over het interval $[0, \theta]$, waarbij θ een ongekende parameter is. Stel dat we willen toetsen of $H_0 : \theta = 1/2$ versus $H_A : \theta > 1/2$ en beslissen om de nulhypothese te verwerpen zodra het maximum $Y = \max(X_1, ..., X_n)$ van alle observaties een verder te bepalen constante c overschrijdt; d.i. we verwerpen H_0 zodra $Y > c$.

(a) *Bepaal dan welke keuze van c men moet maken om de test op het 5% significantieniveau uit te voeren.*

We weten dat

$$\begin{aligned} 0.05 &= \mathcal{P}(Y > c|H_0) \\ \Leftrightarrow 0.95 &= \mathcal{P}(Y \leq c|H_0) \\ &= \mathcal{P}(\max(X_1, ..., X_n) \leq c|H_0) \\ &= \mathcal{P}(X_i \leq c|H_0)^n \\ &= \frac{c^n}{\theta^n} \end{aligned}$$

waaruit volgt dat $c = 0.95^{1/n}\theta = 0.95^{1/n}/2$.

(b) *Stel dat Y de waarde 0.48 aanneemt in een steekproef van $n = 20$ metingen. Wat is dan de p-waarde bij bovenstaande toets? En wat is het bijhorende besluit van de toets?*

De p-waarde is

$$\mathcal{P}(Y > 0.48|\theta = 1/2) = 1 - \frac{0.48^{20}}{0.5^{20}} = 0.56$$

waaruit we besluiten dat we de nulhypothese niet kunnen verwerpen.

(c) *Stel dat Y de waarde 0.52 aanneemt in een steekproef van $n = 20$ metingen. Wat is dan de p-waarde bij bovenstaande toets? En wat is het bijhorende besluit van de toets?*

Als $Y = 0.52$, dan kan θ onmogelijk gelijk zijn aan 0.5 en is de p-waarde bijgevolg 0.

(d) *Leid een uitdrukking af voor de grootte n van de steekproef die nodig is om bij een gegeven waarde van θ een power gelijk aan $1 - \beta$ te bekomen, indien men de toets op het 5% significantieniveau uitvoert*[1]*.*

De power is gelijk aan

$$\begin{aligned} 1 - \beta &= \mathcal{P}(Y > 0.95^{1/n}/2|\theta) \\ &= \frac{0.95}{(2\theta)^n} \end{aligned}$$

[1]Indien u de waarde van c niet gevonden heeft uit deel (a) van deze oefening, laat ze dan gewoon als onbekende in de berekening staan.

waaruit

$$n = \frac{\ln(0.95/(1-\beta)}{\ln(2\theta)}$$

3. Zij $(X_1, Y_1), ..., (X_n, Y_n)$ onafhankelijke en identiek verdeelde toevalsvectoren van binaire observaties X_i en $Y_i, i = 1, ..., n$ waarvoor $P(X_i = 1) = 1/2, P(Y_i = 1|X_i = 0) = e^{-a\theta}$ en $P(Y_i = 1|X_i = 1) = e^{-b\theta}, i = 1, ..., n$ voor een ongekende parameter $\theta > 0$ en gekende constanten $a > 0$ en $b > 0$.

   (a) *Veronderstel dat de toevalsvectoren $(X_1, Y_1), ..., (X_n, Y_n)$ geobserveerd worden. Bepaal dan een vergelijking waaruit men de maximum-waarschijnlijkheidsschatter voor θ op basis van deze gegevens kan oplossen (het is **niet** nodig om deze vergelijking op te lossen).*

   De likelihood is de gezamenlijke verdeling van de toevalsvectoren $(X_1, Y_1), ..., (X_n, Y_n)$:

$$\prod_{i=1}^{n} \frac{1}{2} \left\{ e^{-a\theta y_i} (1 - e^{-a\theta})^{1-y_i} \right\}^{1-x_i} \left\{ e^{-b\theta y_i} (1 - e^{-b\theta})^{1-y_i} \right\}^{x_i}$$

   De logaritme nemen, afleiden naar θ en gelijkstellen aan 0 levert de gewenste vergelijking.

   (b) *Veronderstel dat enkel $Y_1, ..., Y_n$ geobserveerd worden (en $X_1, ..., X_n$ m.a.w. ongekend zijn). Bepaal dan een vergelijking waaruit men de maximum-waarschijnlijkheidsschatter voor θ op basis van deze gegevens[2] kan oplossen (het is **niet** nodig om deze vergelijking op te lossen).*

   De likelihood is nu de gezamenlijke verdeling van de toevalsveranderlijken $Y_1, ..., Y_n$ (omdat $X_1, ..., X_n$ niet geobserveerd zijn). De verdeling van Y_i wordt gegeven door:

$$\mathcal{P}(Y_i = 1) = \frac{e^{-a\theta} + e^{-b\theta}}{2}$$

   De likelihood is bijgevolg

$$\prod_{i=1}^{n} \left\{ \frac{e^{-a\theta} + e^{-b\theta}}{2} \right\}^{y_i} \left\{ 1 - \frac{e^{-a\theta} + e^{-b\theta}}{2} \right\}^{1-y_i}$$

   De logaritme nemen, afleiden naar θ en gelijkstellen aan 0 levert de gewenste vergelijking.

4. Zij $X_1, ..., X_n$ onafhankelijke en identiek verdeelde, binaire toevalsveranderlijken met $P(X_i = 1) = p$ voor $i = 1, ..., n$ met p een ongekende parameter in $]0, 1[$. Zij $\hat{\theta}_n = \bar{X}_n(1 - \bar{X}_n)$ een schatter voor $\theta = p(1 - p)$.

   (a) *Toon dan aan dat $\hat{\theta}_n$ een $\sqrt{n}$-consistente, asymptotisch normaal verdeelde schatter is van θ wanneer $p \neq 1/2$.*

   Toepassing van de Centrale Limietstelling toont aan dat

$$\sqrt{n}(\bar{X}_n - p) \to N\left(0, p(1-p)\right)$$

   Vervolgens, toepassing van de Delta methode toont aan dat

$$\sqrt{n}(\bar{X}_n(1 - \bar{X}_n) - p(1-p)) \to N\left(0, (1-2p)^2 p(1-p)\right)$$

[2] Zorg er m.a.w. voor dat deze schatter enkel van de geobserveerde gegevens afhangt.

(b) *Leid vervolgens de asymptotische verdeling*[3] *van deze schatter af in het bijzonder geval dat* $p = 1/2$.
Merk op dat

$$n(\hat{\theta}_n - \theta) = -n(\bar{X}_n - 1/2)^2 = -\left\{\sqrt{n}(\bar{X}_n - 1/2)\right\}^2 \to -0.25\chi_1^2$$

omdat

$$\sqrt{n}(\bar{X}_n - 1/2) \to 0.5N(0,1)$$

(c) *Convergeert* $\hat{\theta}_n$ *sneller, trager of even snel naar* θ *wanneer* $p = 1/2$ *dan wanneer* $p \neq 1/2$*? Leg uit.*
Sneller omdat het met snelheid n i.p.v. $\sqrt{n}$ convergeert. De variantie van de schatter is dus van de orde n^{-2} i.p.v. n^{-1}.

5. Definieer

$$U = [d(A,X) - E\{d(A,X)|A_1,X\} - E\{d(A,X)|A_2,X\} + E\{d(A,X)|X\}]\,\epsilon$$

met $d(A,X)$ een willekeurige functie van (A,X), $A = (A_1,A_2)$, $\epsilon = Y - E(Y|A,X) + q(A_1,X)$ voor een willekeurige functie $q(A_1,X)$ van (A_1,X).

(a) *Toon dan aan dat* U *gemiddeld 0 is wanneer* $q(A_1,X) = 0$.
We vinden dat

$$\begin{aligned} E(U) &= E\{E(U|A,X)\} \\ &= E\{[d(A,X) - E\{d(A,X)|A_1,X\} - E\{d(A,X)|A_2,X\} + E\{d(A,X)|X\}] \\ &\quad \times E(\epsilon|A,X)\} \\ &= 0 \end{aligned}$$

omdat $E(\epsilon|A,X) = 0$.

(b) *Toon dan aan dat* U *gemiddeld 0 is wanneer* $A_1 \perp\!\!\!\perp A_2|X$.
Uit voorgaande volgt dat

$$\begin{aligned} E(U) &= E\{[d(A,X) - E\{d(A,X)|A_1,X\} - E\{d(A,X)|A_2,X\} + E\{d(A,X)|X\}] \\ &\quad \times q_1(A_1,X)\} \\ &= E[E\{[d(A,X) - E\{d(A,X)|A_1,X\} - E\{d(A,X)|A_2,X\} + E\{d(A,X)|X\}] \\ &\quad \times q_1(A_1,X)\}|A_1,X] \\ &= 0 \end{aligned}$$

omdat

$$E\{[d(A,X) - E\{d(A,X)|A_1,X\} - E\{d(A,X)|A_2,X\} + E\{d(A,X)|X\}]|A_1,X\} = 0$$

als A_1 II $A_2|X$.

6. (Theorievraag) Op pagina 76 in de tweede paragraaf van sectie 2.4.3 wordt aangetoond dat er voor willekeurige gebeurtenissen A en B een functie $P(A|B)$ bestaat zodat $P(A \cap B) = P(A|B)P(B)$.

[3] Maak hier gebruik van het feit dat $\hat{\theta}_n - \theta = \hat{\theta}_n - 1/4 = -(\bar{X}_n - 1/2)^2$ wanneer $p = 1/2$.

(a) *In de redenering gaat men ervan uit dat B een elementaire gebeurtenis is in een σ-algebra $\mathcal{G}$. Is dit een beperking? Zo ja, vervolledig de redenering voor algemene gebeurtenissen. Zo nee, waarom is dit geen beperking?*

Nee, men kan $\mathcal{G}$ altijd zo kiezen dat de beschouwde gebeurtenis B elementair is.

(b) *Toon aan dat de aldus gevonden functie $P(A|B)$ een kansmaat is.*

Zie bewijs van stelling 1.7. Werk uit als oefening.

(c) *De redenering in sectie 2.4.3 illustreert dat voor willekeurige toevalsveranderlijken X_1, X_2 en Borelverzamelingen A, B de conditionele kans $P(X_1 \in A|X_2 \in B)$ goed gedefinieerd is, zelfs als $P(X_2 \in B) = 0$. Is deze conditionele kans ook goed gedefinieerd als $f_{X_2}(x) = 0, \forall x \in B$? Waarom wel/niet? Leg tevens intuïtief uit of er een onderscheid is/wat het onderscheid is tussen het feit $P(X_2 \in B) = 0$ en het feit $f_{X_2}(x) = 0, \forall x \in B$ voor een gegeven verzameling B.*

Nee, als $f_{X_2}(x) = 0, \forall x \in B$, dan is het onmogelijk om observaties in B te bekomen en is de conditionele kans $P(X_1 \in A|X_2 \in B)$ bijgevolg niet goed gedefinieerd. Als $P(X_2 \in B) = 0$ kunnen observaties in B bekomen worden, maar de kans hierop is oneindig klein. Als $f_{X_2}(x) = 0, \forall x \in B$, dan is het onmogelijk om observaties in B te bekomen.