Examen Kansrekening en Wiskundige Statistiek

S. Vansteelandt

Academiejaar 2005-2006

Examenvragen

1. Tijdens de productie van een bepaald geneesmiddel wordt een gemiddelde potentie van 50 mg beoogd en weet men op basis van vroegere ervaring dat de potenties Normaal verdeeld zijn met een gekende standaarddeviatie van 3.5 mg.

   (a) Tijdens de kwaliteitscontrole worden lukraak 20 tabletten uit Batch #1234 getrokken en wordt een gemiddelde assay waarde voor de potentie van 48.3 mg vastgesteld. Maak een beslissing of de Batch moet afgekeurd worden als zijnde subpotent als men 1% risico toelaat om verkeerdelijk de Batch af te keuren (en men zowel Batches met te hoge als te lage potentie wenst te detecteren).

   (b) Bij veel kwaliteitscontroles gebruikt men statistische control charts waarbij men beslist om een Batch af te keuren zodra haar gemiddelde assay waarde (bvb. gebaseerd op 20 lukrake controlestalen uit de Batch) minstens 3 standaarddeviaties afwijkt van het gekende populatiegemiddelde (d.i. hier 50 mg). Hierbij gebruikt men de standaarddeviatie van die gemiddelde assay waarde.

      i. Bereken die standaarddeviatie.

      ii. Welk risico loopt men hier om verkeerdelijk de Batch te verwerpen?

2. Als $X_i, i = 1, ..., n$ onafhankelijk en Cauchy verdeeld zijn, wat is dan de verdeling van hun steekproefgemiddelde

$$Y = \frac{X_1 + ... + X_n}{n}?$$

   Toon aan.

3. Ecologische theorie suggereert dat het aantal soorten organismen binnen een natuurreservaat afhankelijk is van de oppervlakte van dat reservaat. Onderzoekers verzamelen gegevens voor $n = 6$ reservaten om dit te verifiëren. Op basis hiervan schatten ze dat bij een toename van de oppervlakte met 1000 km^2, het aantal soorten gemiddeld toeneemt met 3 en dit met een standaard error van 1.

   (a) Bereken een 95% betrouwbaarheidsinterval voor de verwachte stijging in aantal soorten per 1000 km^2 toename van de oppervlakte, als u weet dat de gebruikte schatter $\hat{\beta}$ hiervoor de volgende verdeling bezit

$$\frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t_{n-2}$$

   waarbij β de werkelijke waarde is voor deze verwachte stijging (op populatieniveau) en $se(\hat{\beta})$ de standaard error voorstelt.

(b) Interpreteer het bekomen interval (merk op dat het daartoe niet volstaat om te kijken of het interval bijvoorbeeld 0 bevat). Indien u dit interval niet gevonden heeft, ga er dan van uit dat het gezochte interval [1,5] is.

4. Veronderstel dat $(Y_1, Z_1), ..., (Y_n, Z_n)$ n onafhankelijke en identiek verdeelde toevalsvectoren zijn, dat $Y_i \perp\!\!\!\perp Z_i$ (d.w.z. Y_i onafhankelijk van Z_i),$Y_i \stackrel{d}{=} \text{Exp}(\lambda), Z_i \stackrel{d}{=} \text{Exp}(\mu)$ voor $i = 1, ..., n$.

   (a) bepaal dan de maximum-waarschijnlijkheidsschatter voor (λ, μ).

   (b) stel dat we enkel $X_i = \min(Y_i, Z_i)$ observeren, samen met een indicator Δ_i die aangeeft welk van beide uitkomsten geobserveerd wordt (m.a.w. Δ_i is gelijk aan 1 als $X_i = Y_i$ en 0 anders).

      i. bepaal de gezamenlijke dichtheidsfunctie van de geobserveerde data (X_i, Δ_i). *Hint: het kan helpen om dit afzonderlijk te doen voor een punt* $(X_i, \Delta_i) = (x, 1)$ *en een punt* $(X_i, \Delta_i) = (x, 0)$.

      ii. bepaal de maximum-waarschijnlijkheidsschatter voor (λ, μ) op basis van de geobserveerde data.

5. Geef uw oordeel omtrent volgende uitspraken. Geef aan of ze waar of vals zijn en **motiveer uw antwoord**.

   (a) Als we de nulhypothese $\mu = \mu_0$ tweezijdig toetsen op basis van een bepaalde steekproef en de p-waarde bedraagt 0.02, dan omvat het 99% betrouwbaarheidsinterval voor μ op basis van deze steekproef het getal μ_0 **niet**.

   (b) Stel dat we de nulhypothese $\mu = \mu_0$ tweezijdig toetsen en dat de p-waarde 0.04 bedraagt. Stel dat we nu op basis van diezelfde steekproef dezelfde nulhypothese toetsen tegenover de alternatieve hypothese $\mu > \mu_0$, dan is de p-waarde gelijk aan 0.02.

   (c) Stel dat 2 klinische studies naar het effect van een bepaald medicijn op een zekere uitkomst respectievelijk p-waarden geven van 0.001 en 0.02. Dan heeft men in de eerste studie een groter effect van het medicijn gevonden dan in de tweede studie.

6. Om een parameter θ^2 te schatten, krijgt u de keuze tussen 2 mogelijkheden: (1) u berekent $(\bar{X}_n)^2$ voor een reeks exponentieel verdeelde gegevens X_i met parameter θ; of (2) u berekent $\bar{Y}_n$ voor een reeks exponentieel verdeelde gegevens Y_i met parameter θ^2. Als n groot is, welke optie verkiest u dan? Leg uit waarom.

7. Zij X, Y en Z toevalsveranderlijken op een zelfde kansruimte, met $E|X| < \infty$ en $Y = h(Z)$ voor een continue functie h. Toon dan aan dat

   (a) als $X \perp\!\!\!\perp Z$ en $E|Z| < \infty$, dan is (bijna zeker)

   $$E(XZ|Y) = E(X)E(Z|Y)$$

   (b) *bonusvraag:* als $E\{f(X)|Z\} = f(Y)$ voor alle continue functies f, dan is $X = Y$ bijna zeker.

Veel succes!

Achtergrondinformatie

1. De Cauchy verdeling met parameters μ en σ heeft dichtheidsfunctie

$$\frac{1}{\sigma\pi}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$$

en karakteristieke functie

$$e^{i\mu t-\sigma|t|}$$

Haar verwachtingswaarde en variantie zijn niet gedefinieerd.

2. De exponentiële verdeling met parameter θ heeft dichtheidsfunctie

$$\theta e^{-x\theta}I(x\geq 0)$$

en momentgenererende functie

$$e^t(1-t/\theta)^{-1}$$

voor $t<\theta$. Haar verwachtingswaarde is $1/\theta$ en haar variantie bedraagt $1/\theta^2$.

3. De regel van de herhaalde verwachtingswaarde uit de cursus, kan triviaal uitgebreid worden naar conditionele verwachtingswaarden. In dat geval geeft ze aan dat voor s.v. X, Y en Z, in de onderstelling dat de verwachtingswaarden bestaan,

$$E\left[E\left\{r\left(X,Y\right)|X,Z\right\}|Z\right]=E\left\{r\left(X,Y\right)|Z\right\},$$

voor een willekeurige functie $r\left(x,y\right)$.

Statistiek I: Examen 1e zit 2005-2006

Peter Vandendriessche

15 januari 2007

1. (a) De nulhypothese is $H_0 : \mu = 50$, de alternatieve $H_A : \mu \neq 50$. We hebben $Z = \frac{50-48.3}{3.5/\sqrt{20}} = 2.172 = z_{0.985}$. Aangezien $0.985 < 0.995$ kunnen we de batch niet afkeuren (=nulhypothese niet verwerpen op het 99% significantieniveau).

   (b) i. $SE(\bar{X}_{20}) = \frac{\sigma}{\sqrt{20}} = 0.783$

   ii. Precies de kans om meer dan drie standaardafwijkingen van de verwachtingswaarde te zitten bij een normale verdeling, dus 0.0026.

2. $\phi_{\frac{X_1+\cdots+X_n}{n}}(t) = E\left(e^{i\left(\frac{t}{n}\right)\sum_{i=1}^n X_i}\right) = \prod_{i=1}^n E\left(e^{i\left(\frac{t}{n}\right)X_i}\right) = \phi_X\left(\frac{t}{n}\right)^n = e^{n\left(i\mu\frac{t}{n}-\sigma\left|\frac{t}{n}\right|\right)} = e^{i\mu t-\sigma|t|}$, dus opnieuw Cauchy verdeeld met dezelfde parameters.

3. (a) Ons interval is $[\hat{\beta} - SE \cdot t_{4,0.975}, \hat{\beta} + SE \cdot t_{4,0.975}] = [0.224, 5.776]$.

   (b) We zijn 95% zeker dat dit interval het werkelijke gemiddelde β bevat.

4. (a) Aangezien Y_i, Z_i onafhankelijk zijn mogen we de twee componenten apart bepalen. We hebben $L(\theta) = f_X^n(x) = \frac{e^{-\sum x_i/\theta}}{\theta^n} I_{]0,+\infty[}$, zodat voor $\theta > 0$ geldt $\ell := \ln(L) = -n\ln\theta - \frac{1}{\theta}\sum x_i$, $\ell' = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum x_i = 0$ ofte $\theta = \frac{1}{n}\sum x_i$. Of met de definitie uit de bijlage: $\theta = \frac{n}{\sum x_i}$.

   (b) i. We hebben altijd $f(y) = \int_0^\infty f(y)f(z)dz$. Als $\Delta_i = 1$ wordt dit $f(x|\Delta_i = 1) = \int_y^\infty f(y)f(z)dz = \lambda\mu e^{-\lambda y}\int_y^\infty e^{-\mu z}dz = \lambda e^{-(\lambda+\mu)y}$, analoog voor $\Delta_i = 0$ krijgen we $f(x|\Delta_i = 0) = \mu e^{-(\lambda+\mu)y}$). We kunnen dit dus samen schrijven als $f(X_i, \Delta_i) = \lambda^{\Delta_i}\mu^{1-\Delta_i}e^{-(\lambda+\mu)x}$.

   ii. $L(\lambda, \mu) = \lambda^{\sum\Delta_i}\mu^{n-\sum\Delta_i}e^{-(\lambda+\mu)\sum X_i}$ zodat $\ell = \ln(\lambda/\mu)\sum\Delta_i + n\ln(\mu) - (\lambda+\mu)\sum X_i$ en dus $\frac{\partial\ell}{\partial\lambda} = \frac{\partial\ell}{\partial\mu} = 0$ geeft dat $\lambda = \frac{\sum X_i}{\sum\Delta_i}$ en $\mu = \frac{\sum X_i}{n-\sum\Delta_i}$.

5. (a) Fout, alle α100%-BI's met $\alpha \geq 98$ omvatten μ_0.

   (b) Fout, je weet à priori enkel dat die p-waarde in $[0, 0.04]$ ligt.

   (c) Fout, de p-waarde spreekt enkel over statistische significantie, niet over de medische significantie.

6. Enerzijds is $\mathrm{Var}(\bar{Y}_n) = \frac{\theta^4}{n}$, anderzijds is $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} N(0,1)$ en wegens de delta-methode dus $\sqrt{n}((\bar{X}_n)^2 - \theta^2) \xrightarrow{d} 2\theta N\left(0, \theta^2\right)$ ofte $(\bar{X}_n)^2 \xrightarrow{d} N\left(\theta^2, \frac{4\theta^4}{n}\right)$, zodat $\mathrm{Var}((\bar{X}_n)^2) = \frac{4\theta^4}{n}$. Dus $\bar{Y}_n$ heeft (asymptotisch) een kleinere variantie en is dus beter.

7. (a) Daar X, Z onafhankelijk zijn, zijn X, Y ook onafhankelijk. Conditioneren we $E(XZ) = E(X)E(Z)$ op Y dan komt er $E(XZ|Y) = E(X|Y)E(Z|Y) = E(X)E(Z|Y)$.

   (b) Achtereenvolgens $f(t) = t$ en $f(t) = t^2$ kiezen geeft $E(X|Z) = Y, E(X^2|Z) = Y^2$ zodat $Var(X|Z) = 0$ en dus $X = g(Z)$ voor zekere functie g. Er komt te staan dat $E(f(g(Z))|Z) = f(h(Z))$, maar het argument van E hangt enkel af van Z, dus conditioneel op Z doet die E (met kans 1) niets en mogen we ze weglaten. We krijgen $f(g(Z)) = f(h(Z))$ voor alle continue functies f. Kiezen we $f(t) = t$ dan komt er $g(Z) = h(Z)$, wat te bewijzen was.