Oplossing Examen Statistiek I

S. Vansteelandt

Academiejaar 2022-2023

---

1. (15 min, 15p) Antwoord met ja / nee en geef een korte motivatie (u krijgt geen punten indien het ofwel niet duidelijk is of u ja of nee bedoelt, ofwel een geldige motivatie ontbreekt). Stel dat we op basis van gegevens uit hetzij een gerandomiseerde studie, hetzij een observationele studie een betrouwbaarheidsinterval berekenen voor het gemiddelde verschil in uitkomst tussen mensen met versus zonder een gegeven behandeling. Gaat u akkoord dat dit interval onder andere de volgende onnauwkeurigheid/onzekerheid in rekening brengt:

   (a) het feit dat in een gerandomiseerde studie door toeval relatief meer mannen in de ene versus andere groep kunnen belanden, wat de groepen minder vergelijkbaar maakt.
   Ja, een betrouwbaarheidsinterval brengt onzekerheid in kaart ten gevolge van toevallige wijzigingen van studie tot studie.

   (b) het feit dat in een observationele studie het gemiddeld verschil tussen beide groepen ten gevolge van confounding systematisch kan afwijken van het behandelingseffect.
   Nee, confounding is een systematische afwijking die niets met toeval te maken heeft.

2. (45 min) Het doel van een studie is om het gemiddelde inkomen te berekenen in een populatie. Daartoe worden lukraak deelnemers gerecruteerd uit de populatie, maar het blijkt dat sommigen (vooral mensen met een drukke job) geen data voor hun inkomen Y verstrekken. Zij $R = 1$ als Y geobserveerd wordt en $R = 0$ anders. Zij verder X een vector van gemeten kenmerken van het individu (zoals geslacht, leeftijd, beroep, scholingsgraad).

   Stel dat we beschikken over i.i.d. data $(R_i, X_i), i = 1, ..., n$, en dat voor deze individuen met $R_i = 1$ ook i.i.d. metingen van Y_i voorhanden zijn. Stel tot slot dat $R \perp\!\!\!\perp Y|X$.

   (a) (5p) Lijkt $R \perp\!\!\!\perp Y|X$ u meer plausibel dan $R \perp\!\!\!\perp Y$, of eerder niet? Leg uit.
   Ja, $R \perp\!\!\!\perp Y$ lijkt niet plausibel vermits ontbrekende data vaker voorkomen bij mensen met een drukke job die op hun beurt mogelijks een hoger inkomen hebben. We kunnen dit verhelpen door te conditioneren op X vermits we dan mensen met een zelfde beroep, ... evalueren.

   (b) (10p) Zij $P(R_i = 1|X_i)$ gekend. Toon dan aan dat voor elke functie $m(X)$ (vb. $m(X) = X$),

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{P(R_i = 1|X_i)} \{Y_i - m(X_i)\} + m(X_i) \tag{1}$$

   een onvertekende schatter is voor $E(Y)$. (Merk op dat deze schatter enkel metingen Y_i gebruikt voor individuen met $R_i = 1$ en bijgevolg berekend kan worden op de data.)
   Het gemiddelde hiervan is

$$\begin{aligned}
&E\left[\frac{R_i}{P(R_i = 1|X_i)} Y_i\right] + E\left[\left\{1 - \frac{R_i}{P(R_i = 1|X_i)}\right\} m(X_i)\right] \\
&= E\left[\frac{E(R_i Y_i|X_i)}{P(R_i = 1|X_i)}\right] + E\left[E\left\{1 - \frac{R_i}{P(R_i = 1|X_i)}|X_i\right\} m(X_i)\right] \\
&= E\left[\frac{E(R_i|X_i)E(Y_i|X_i)}{P(R_i = 1|X_i)}\right] + E\left[\left\{1 - \frac{E(R_i|X_i)}{P(R_i = 1|X_i)}|X_i\right\} m(X_i)\right] \\
&= E\left[E(Y_i|X_i)\right] + E\left[0 \times m(X_i)\right] = E(Y_i)
\end{aligned}$$

(c) (5p) Stel dat we de functie $m(X) = \hat{\delta}'X$ gebruiken voor een schatter $\hat{\delta}$ die in kans convergeert naar een eindige limiet δ. Stel verder dat we de variantie van de resulterende schatter

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P(R_i=1|X_i)}\left\{Y_i-\hat{\delta}'X_i\right\}+\hat{\delta}'X_i \tag{2}$$

berekenen als n^{-1} vermenigvuldigd met de steekproefvariantie van

$$\frac{R_i}{P(R_i=1|X_i)}\left\{Y_i-\hat{\delta}'X_i\right\}+\hat{\delta}'X_i.$$

Verwacht u dat dit een 'correcte' schatter van deze variantie zal opleveren (met name, verwacht u dat de verhouding van deze variantieschatter ten opzichte van de populatievariantie van de schatter bij benadering 1 zal zijn in grote steekproeven)? Leg uit. Hier wordt geen rekenwerk verwacht; een duidelijke argumentatie volstaat.
Nee, vermits dit ervan uitgaat dat $\hat{\delta}$ niet varieert van studie tot studie, en de onzekerheid in $\hat{\delta}$ bijgevolg niet in rekening brengt. Deze vraag werd als bonusvraag gerekend (zefs indien u ze niet correct opgelost had, kon u nog steeds alle punten op het examen halen; deze vraag kon dus punten 'bovenop 20' opleveren).

Open boek:

1. (1u) We gaan even verder met voorgaande opgave, maar vanaf hier wordt het examen open boek.

   (a) (10p) Toon aan dat (2) een consistente schatter is voor $E(Y)$.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{P(R_i=1|X_i)}Y_i+\hat{\delta}'\frac{1}{n}\sum_{i=1}^{n}\left\{1-\frac{R_i}{P(R_i=1|X_i)}\right\}X_i$$

   waarbij de eerste term in kans convergeert naar $E(Y)$, $\hat{\delta}$ naar een zekere δ en

$$\frac{1}{n}\sum_{i=1}^{n}\left\{1-\frac{R_i}{P(R_i=1|X_i)}\right\}X_i$$

   naar 0.

   (b) (10p) Laat ons even terugstappen op het originele voorstel waarbij een gekende functie $m(.)$ wordt gebruikt. Toon aan dat de variantie van de schatter (1) dan gelijk is aan

$$\frac{\mathrm{Var}(Y)}{n}+\frac{P(R=0)}{n}E\left[\frac{E\left[\{Y-m(X)\}^2|X\right]}{P(R=1|X)}|R=0\right]$$

   (en dus niet kleiner dan de variantie van het steekproefgemiddelde van alle data Y_i, zoals gegeven door de eerste term).

We gebruiken de regel van herhaalde variantie.

$$\begin{aligned}
&\mathrm{Var}\left(E\left[\frac{R_i}{P(R_i=1|X_i)}\left\{Y_i-m(X_i)\right\}+m(X_i)|X_i,Y_i\right]\right)\\
&+E\left(\mathrm{Var}\left[\frac{R_i}{P(R_i=1|X_i)}\left\{Y_i-m(X_i)\right\}+m(X_i)|X_i,Y_i\right]\right)\\
&=\mathrm{Var}\left(Y_i\right)+E\left[\frac{P(R_i=1|X_i)P(R_i=0|X_i)}{P(R_i=1|X_i)^2}\left\{Y_i-m(X_i)\right\}^2\right]\\
&=\mathrm{Var}\left(Y_i\right)+E\left[P(R_i=0|X_i)\frac{E\left\{\left\{Y_i-m(X_i)\right\}^2|X_i\right\}}{P(R_i=1|X_i)}\right]\\
&=\mathrm{Var}\left(Y_i\right)+E\left[(1-R_i)\frac{E\left\{\left\{Y_i-m(X_i)\right\}^2|X_i\right\}}{P(R_i=1|X_i)}\right]\\
&=\mathrm{Var}\left(Y_i\right)+P(R_i=0)E\left[\frac{E\left\{\left\{Y_i-m(X_i)\right\}^2|X_i\right\}}{P(R_i=1|X_i)}|R_i=0\right].
\end{aligned}$$

De variantie van het overeenkomstige steekproefgemiddelde bekomt men door dit te delen door n. Deze vraag werd als bonusvraag gerekend (zefs indien u ze niet correct opgelost had, kon u nog steeds alle punten op het examen halen; deze vraag kon dus punten 'bovenop 20' opleveren).

(c) (5p) Gebruik dit resultaat om aan te tonen dat de beste keuze van $m(X)$, die een schatter met de laagste variantie oplevert, gelijk is aan $m(X)=E(Y|X)$. *Uit*

$$\begin{aligned}
E\left\{\left\{Y_i-m(X_i)\right\}^2|X_i\right\} &= E\left\{\left\{Y_i-E(Y_i|X_i)+E(Y_i|X_i)-m(X_i)\right\}^2|X_i\right\}\\
&= E\left\{\left\{Y_i-E(Y_i|X_i)\right\}^2|X_i\right\}+E\left\{\left\{E(Y_i|X_i)-m(X_i)\right\}^2|X_i\right\}\\
&\quad +2E\left[\left\{Y_i-E(Y_i|X_i)\right\}\left\{E(Y_i|X_i)-m(X_i)\right\}|X_i\right]\\
&= E\left\{\left\{Y_i-E(Y_i|X_i)\right\}^2|X_i\right\}+E\left\{\left\{E(Y_i|X_i)-m(X_i)\right\}^2|X_i\right\},
\end{aligned}$$

hetgeen minimaal is in $m(X_i)=E(Y_i|X_i)$.

(d) (15p) Stel dat $P(R_i=1|X_i)$ niet gekend is, maar dat we weten dat $P(R_i=1|X_i)=\Phi(\alpha' X_i)$ voor een ongekende parameter α, met Φ de cumulatieve distributiefunctie van een standaardnormale meting. Stel dan een vergelijking op waaruit de maximum kans schatter van α kan opgelost worden (d.i., waar ze een nulpunt van is).
De likelihood is

$$\prod_{i=1}^{n}\Phi(\alpha' X_i)^{R_i}\left\{1-\Phi(\alpha' X_i)\right\}^{1-R_i}$$

Vervolgens de logaritme nemen

$$\sum_{i=1}^{n}R_i\log\Phi(\alpha' X_i)+(1-R_i)\log\left\{1-\Phi(\alpha' X_i)\right\}$$

en afleiden naar α:

$$0=\sum_{i=1}^{n}X_i\varphi(\alpha' X_i)\left\{\frac{R_i}{\Phi(\alpha' X_i)}-\frac{1-R_i}{1-\Phi(\alpha' X_i)}\right\}$$

met $\varphi(.)$ de standaardnormale dichtheid.

2. (30 min) Dertig sollicitanten moeten een mondelinge en schriftelijke proef afleggen, waarvan de resultaten onderaan samengevat zijn. De scores van de sollicitanten op een gegeven examen zijn onderling onafhankelijk, en zowel de 2 examenscores als de verschillen ertussen zijn Normaal verdeeld.

	Mondeling	Schriftelijk	Mondeling − Schriftelijk
Gemiddelde	66.563	70.471	−3.908
Standaard error	2.193	2.137	1.194

   (a) (10p) Schat de correlatie tussen de metingen van het mondelinge en schriftelijke examen. *Gebruik hiervoor de formule voor de variantie van een verschil van 2 toevalsveranderlijken, waaruit de covariantie kan berekend worden.*

   (b) (10p) Bepaal een 95% betrouwbaarheidsinterval voor het gemiddeld verschil in testscores tussen beide examens.

$$-3.908 \pm t_{29,0.025} 1.194$$

   (c) (10p) Stel dat men op basis van deze gegevens een *algemene uitspraak* wil doen dat het mondelinge examen dat werd opgesteld gemiddeld lagere scores oplevert dan het schriftelijke examen dat werd opgesteld. Gaat u akkoord met deze bewering? Waarom wel/niet?
   Ja, vermits 0 niet tot het 95% betrouwbaarheidsinterval behoort en bijgevolg tegengesproken wordt door de gegevens.

3. (1u 30 min) In een studie naar het effect van genotypes X en Z op de levensduur T onderstelt men dat de kans om een vast tijdstip t te overleven, gelijk is aan

$$P(T > t|X, Z) = \exp(-\theta t X - \beta t Z)$$

   met θ, β ongekende parameters. Op basis van biologische kennis weet men dat $X \perp\!\!\!\perp Z$.

   (a) (10p) Toon aan dat X en Z ook onafhankelijk zijn bij mensen die tijdstip t overleven.

$$\begin{aligned} f(X, Z|T > t) &= \frac{P(T > t|X, Z) f(X, Z)}{P(T > t)} \\ &= \frac{\exp(-\theta t X - \beta t Z) f(X) f(Z)}{P(T > t)} \end{aligned}$$

   Dit factoriseert in een functie van X en een functie van Z, waaruit de onafhankelijkheid volgt.

   (b) (10p) Toon aan dat

$$E(X|T > t) = \frac{E\{X \exp(-\theta X t)\}}{E\{\exp(-\theta X t)\}}$$

$$
\begin{aligned}
E(X|T>t,Z) &= \int x f(x|T>t,Z)dx \\
&= \int x \frac{P(T>t|x,Z)}{P(T>t|Z)} f(x|Z)dx \\
&= \int x \frac{P(T>t|x,Z)}{E\left\{P(T>t|X,Z)|Z\right\}} f(x|Z)dx \\
&= \frac{E\left\{XP(T>t|x,Z)|Z\right\}}{E\left\{P(T>t|X,Z)|Z\right\}} \\
&= \frac{E\left\{X\exp\left(-\theta tX-\beta tZ\right)|Z\right\}}{E\left\{\exp\left(-\theta tX-\beta tZ\right)|Z\right\}} \\
&= \frac{E\left\{X\exp\left(-\theta tX\right)|Z\right\}}{E\left\{\exp\left(-\theta tX\right)|Z\right\}} \\
&= \frac{E\left\{X\exp\left(-\theta tX\right)\right\}}{E\left\{\exp\left(-\theta tX\right)\right\}},
\end{aligned}
$$

waarbij we gebruiken dat $X \perp\!\!\!\perp Z$. Merk tot slot op wegens vraag 3a dat $E(X|T>t,Z) = E(X|T>t)$.

(c) (15p) Onderstel dat we over onafhankelijke, lukrake metingen $X_1, ..., X_n$ beschikken. Het rechterlid kan dan geschat worden als

$$
\frac{\sum_{i=1}^n X_i \exp\left(-\theta X_i t\right)}{\sum_{i=1}^n \exp\left(-\theta X_i t\right)} \tag{3}
$$

Bepaal de asymptotische verdeling[1] van deze schatter (voor gekende θ). Hint: het kan helpen om eerst aan te tonen dat

$$
\sqrt{n}\left(\frac{\sum_{i=1}^n X_i \exp\left(-\theta X_i t\right)}{\sum_{i=1}^n \exp\left(-\theta X_i t\right)} - \frac{E\left\{X\exp\left(-\theta Xt\right)\right\}}{E\left\{\exp\left(-\theta Xt\right)\right\}}\right)
$$

kan geschreven worden als

$$
\frac{1}{\sqrt{n}E\left\{\exp\left(-\theta Xt\right)\right\}} \sum_{i=1}^n \exp\left(-\theta X_i t\right)\left[X_i - \frac{E\left\{X\exp\left(-\theta Xt\right)\right\}}{E\left\{\exp\left(-\theta Xt\right)\right\}}\right] + o_p(1).
$$

Indien u meteen van deze laatste uitdrukking vertrekt zonder ze eerst aan te tonen, dan krijgt u maximaal 5 punten.

$$
\begin{aligned}
&\sqrt{n}\left(\frac{\sum_{i=1}^n X_i \exp\left(-\theta X_i t\right)}{\sum_{i=1}^n \exp\left(-\theta X_i t\right)} - \frac{E\left\{X\exp\left(-\theta Xt\right)\right\}}{E\left\{\exp\left(-\theta Xt\right)\right\}}\right) \\
&= \frac{1}{\sqrt{n}n^{-1}\sum_{i=1}^n \exp\left(-\theta X_i t\right)}\left(\sum_{i=1}^n X_i \exp\left(-\theta X_i t\right) - \sum_{i=1}^n \exp\left(-\theta X_i t\right)\frac{E\left\{X\exp\left(-\theta Xt\right)\right\}}{E\left\{\exp\left(-\theta Xt\right)\right\}}\right) \\
&= \frac{1}{\left[E\left\{\exp\left(-\theta Xt\right)\right\} + o_p(1)\right]}\frac{1}{\sqrt{n}}\sum_{i=1}^n \exp\left(-\theta X_i t\right)\left[X_i - \frac{E\left\{X\exp\left(-\theta Xt\right)\right\}}{E\left\{\exp\left(-\theta Xt\right)\right\}}\right]
\end{aligned}
$$

[1] De variantie van deze verdeling kan uitgedrukt worden als de variantie van een transformatie van de metingen van een lukraak individu. Het is niet nodig om dergelijke variantie uit te werken of te vereenvoudigen.

wat het gestelde resultaat levert vermits

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\exp\left(-\theta X_i t\right)\left[X_i - \frac{E\left\{X\exp\left(-\theta X t\right)\right\}}{E\left\{\exp\left(-\theta X t\right)\right\}}\right] = O_p(1)$$

wegens de Centrale Limietstelling en het feit dat de termen binnen de sommatie gemiddeld 0 zijn, en omdat $o_p(1)O_p(1) = o_p(1)$*. De asymptotische verdeling is dus Normaal met gemiddelde 0 en variantie gegeven door de variantie van*

$$\frac{1}{E\left\{\exp\left(-\theta X t\right)\right\}}\exp\left(-\theta X_i t\right)\left[X_i - \frac{E\left\{X\exp\left(-\theta X t\right)\right\}}{E\left\{\exp\left(-\theta X t\right)\right\}}\right].$$

(d) Stel dat we gegevens $X_1, ..., X_n$ *en* $Z_1, ..., Z_n$ *verzamelen voor lukrake, ongerelateerde hoogbejaarden. We stellen vast dat hun leeftijd* $T_1, ..., T_n$ *bedraagt.*

i. (10p) Als de schatter (3) berekend wordt op basis van deze gegevens, is ze volgens u dan onvertekend? Hier wordt geen rekenwerk verwacht; een duidelijke argumentatie volstaat.
Neen, vermits we in dat geval niet over een lukrake steekproef beschikken.

ii. (15p) Bepaal een maximum kans schatter voor θ *in de onderstelling dat* X *Normaal verdeeld is met een gekend gemiddelde en variantie (bvb. op basis van kennis uit externe populatieregisters).*
Omdat we enkel metingen verzamelen voor mensen die nog in leven zijn, is de likelihood

$$\prod_{i=1}^{n} f(X_i, Z_i | T > T_i) = \prod_{i=1}^{n} P(T > T_i | X_i, Z_i) f(X_i) f(Z_i) / P(T > T_i)$$
$$= \prod_{i=1}^{n} \frac{\exp\left(-\theta T_i X_i - \beta T_i Z_i\right)}{E\left\{\exp\left(-\theta T_i X_i - \beta T_i Z_i\right)\right\}} f(X_i) f(Z_i)$$

waarbij het gemiddelde enkel m.b.t. de verdeling van X_i *en* Z_i *(niet* T_i*) genomen wordt. De loglikelihood bedraagt op een constante na*

$$\sum_{i=1}^{n} -\theta T_i X_i - \beta T_i Z_i - \log E\left\{\exp\left(-\theta T_i X_i - \beta T_i Z_i\right)\right\}$$

Afleiden naar θ *levert*

$$\begin{aligned} 0 &= \sum_{i=1}^{n} -T_i X_i + \frac{E\left\{T_i X_i \exp\left(-\theta T_i X_i - \beta T_i Z_i\right)\right\}}{E\left\{\exp\left(-\theta T_i X_i - \beta T_i Z_i\right)\right\}} \\ &= \sum_{i=1}^{n} -T_i X_i + \frac{E\left\{T_i X_i \exp\left(-\theta T_i X_i\right)\right\}}{E\left\{\exp\left(-\theta T_i X_i\right)\right\}} \end{aligned}$$

Deze gemiddelden kunnen op basis van de momentgenererende functie worden bepaald, al werd niet verwacht dat u dit verder zou uitwerken. Deze vraag werd als bonusvraag gerekend (zefs indien u ze niet correct opgelost had, kon u nog steeds alle punten op het examen halen; deze vraag kon dus punten 'bovenop 20' opleveren).