

[Er is ook een Bayesologie 1, over basisbegrippen, maar die laat zich niet goed scannen. Het idee van deze stukken is: interne professionalisering. In de categorie 'Goede voornemens'. Afijn, ik ben er zelf wel serieus mee doorgegaan, maar had daar ook de steun van een wiskundige ondergrond bij, na een half jaar econometrie gestudeerd te hebben.]

[NB: figuren ontbreken. Ik heb deze transcriptie van een txt-bestand dat eigenlijk een Word-bestand is; ik moet nog nagaan of ik het met Word kan openen en de figuren op een of andere manier kan overnemen.]

[Lees 'pi' als $p_{sub i}$]

BAYESOLOGIE. 2.: Het leggen van de cesuur bij studietoetsen.

Verondersteld wordt dat Bayesologie 1. grondig bestudeerd is.

In het volgende zullen we een betoog van Hofstee over normhandhaving bij studietoetsen gebruiken als materiaal om onze Bayesiaanse kunsten op te beproeven (Hofstee: Een alternatief voor normhandhaving bij toetsen. N.T.v.d.Ps. 1973, 28 no 34, 215227). Misschien lukt het om het betoog van Hofstee op punten te verbeteren.

Als start nemen we de uitspraak van Hofstee dat alle normen uiteindelijk niet anders dan relatief kunnen zijn.'

"Ze moeten ooit bepaald zijn aan de hand van de prestaties van één of andere 'eerste groep'. Wanneer er géén geoefend typiste meer dan vijf woorden per minuut zou typen, was er geen denken aan dat de norm zou liggen waar hij nu ligt. De tegenwerping, dat docenten in nieuwe situaties wel degelijk bij voorbaat normen kunnen stellen en dat ook doen, kan gemakkelijk worden ontkracht. Wat in zo'n geval plaats vindt is een generalisatie vanuit reeds bekende situaties (op grond van voorkennis over de disposities van de leerlingen en de aard van de stof). Als die generalisatie een acceptabel resultaat oplevert en de tolerantiegrenzen daarvoor plegen zeer ruim te zijn ontstaat al snel de illusie dat van een absoluut oordeel kan worden gesproken. In feite echter is de mens de maat van alle dingen."

Hofstee stelt m.a.w. dat de docent altijd al zijn a priori inzichten over studieprestaties die Verwacht kunnen worden, een rol laat spelen bij het bepalen van de cesuur voor zijn tentamens. Het is uit het artikel van Hofstee niet duidelijk of hij bij het uitwerken van een en ander een Bayesiaanse aanpak heeft gehanteerd (in ieder geval heeft hij geen Bayesiaanse statistische technieken gebruikt). We zullen eens kijken of een Bayesiaanse aanpak tot dezelfde resultaten als van Hofstee leidt.

De docent die de cesuur gaat bepalen heeft daarbij impliciet in zijn hoofd de prestaties van voorgaande groepen studenten in ongeveer dezelfde omstandigheden, en vaak ook nog een idee over de prestaties van de

onderhavige groep in vergelijking met vroegere groepen. Het lijkt beter om deze impliciete noties te expliciteren en in de technische aanpak van het cesuurbepalingsprobleem op te nemen.

De grootte waar het om gaat is de proportie die werkelijk een voldoende verdient: de proportie π . Deze π is een platonische liefde van ons, die een klein beetje aards te maken is door gebruik te maken van de waarnemingen X , de aantallen studenten die een voldoende krijgen. Er van uit gaande dat we individuele studenten in welke groep dan ook niet kennen (geen informatie over individuele studenten hebben), is de waarschijnlijkheid dat een willekeurige student een voldoende behaalt bepaald door π . De waarschijnlijkheidsverdeling van het aantal studenten dat boven de norm scoort is de binomiaalverdeling:

$$(1) \quad p(x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \\ x = 0, 1, 2, \dots, n.$$

De moeilijkheden met formule (1) zijn allereerst dat we π typisch niet kennen (m.b.v. de stelling van Bayes komen we daar wel uit), en bovendien dat x ook nog afhankelijk lijkt te zijn van de willekeur van de docent die een bepaalde cesuur heeft getrokken, op basis waarvan we immers de waarneming X doen. De laatste moeilijkheid zullen we voorlopig laten voor wat ze is, misschien kunnen we er later nog iets zinnigs over zeggen. (Ook nu geldt weer dat we er in eerste instantie op uit zijn iets over Bayesiaanse methoden op te steken, en niet over het leggen van cesuren bij tentamens).

We zijn op zoek naar de a posteriori verdeling voor de werkelijke proportie voldoende gegeven het aantal studenten dat een voldoende beoordeling heeft gekregen:

$$(2) \quad p(\pi|x) \propto p(x|\pi) p(\pi).$$

Opmerking tussendoor: we gaan er dus geenszins vanuit dat x/n gelijk zou behoren te zijn aan π (beter: de modus van $p(\pi|x)$). Niet alleen houdt de feilbaarheid van de toetsscores in dat er ook bij $\pi=x/n$ altijd nog fouten in de beoordeling gemaakt zullen worden, maar ook is vrijwel altijd het 'verlies' van een ten onrechte gegeven voldoende niet gelijk aan het 'verlies' van een ten onrechte gegeven voldoende, wat bij een besliskundige aanpak van het cesuurprobleem leidt tot $\pi \neq x/n$

De a priori verdeling $p(\pi)$ bevat alle informatie over π die we op grond van ervaring met vroegere groepen, en kennismaking met de huidige groep, menen te hebben. Zoals we weten is het vrijwel altijd mogelijk om die a priori informatie ongeveer te representeren in een beta verdeling $B(a,b)$. Voor het

kiezen van een geschikte beta verdeling staan verschillende wegen open. Het is altijd handig om deze verschillende wegen altijd alle te begaan: op die wijze komen we er achter of de gekozen $B(a,b)$ wel in alle opzichten in overeenstemming is met onze a priori opvattingen. Het is maar al te eenvoudig een Beta verdeling te kiezen die bijv. wel een modus en variantie heeft die in overeenstemming is met onze a priori inzichten, maar die een dermate groot (of juist klein) aantal waarnemingen veronderstelt dat ze op dat punt niet in overeenstemming is met wat we redelijkerwijs aan aantal impliciete waarnemingen willen veronderstellen. Zouden we onwetend van dit alles toch een dergelijke Beta verdeling hebben gekozen, dan zouden we beschuldigd worden niet coherent te zijn in onze a priori opvattingen. En bijna niets is zo erg voor een Bayesiaan als incoherent te zijn in zijn opvattingen.

Voor het specificeren van onze a priori beta verdeling zouden we tabellen en nomogrammen kunnen gebruiken, maar bij voorkeur laten we ons in deze taak bijstaan door een interactief data analyse programma, CADA, in elkaar gezet door Novick en zijn medewerkers. (in de bijlage is een beschrijving te vinden van de manier om dat interactieve programma op het beeldscherm te krijgen). Ons probleem is aan te pakken met programma 2 uit het CADA pakket (zit in file CADA1).

Daar wordt gevraagd aan te geven waar het 25e, 50e en 75e percentiel van je beta verdeling liggen. Het is voldoende om in eerste termijn de meest grove schatting in te tikken, omdat het programma je langs interactieve weg tot een scherpere specificatie zal brengen. Aangegeven moeten dan worden de waarden (op schaal 0 tot 1) links waarvan je schat dat een kwart van het oppervlak onder de beta verdeling zich bevindt, rechts waarvan je denkt dat een kwart van de opp. onder de verdeling zit, en de modus (de meest waarschijnlijke waarde). Bedenk dat in gevallen waarin de modus aan een uiteinde van de schaal ligt, het niet kan zijn dat zowel het 25e als het 75e percentiel op gelijke afstand van de modus liggen, de verdeling is dan nl. noodzakelijkerwijs scheef. Het programma waarschuwt je wel. Het is duidelijk dat de afstand tussen opgegeven 25e en 75e percentiel in relatie staat tot de standaard deviatie van de aldus gespecificeerde beta verdeling, m.a.w. een smaller of een breder interval betekent grotere of kleinere zekerheid van je a priori gedachten. Maar ook de beide parameters van de beta verdeling hebben daar mee te maken, immer $a + b$ is het aantal waarnemingen dat je in gedachten hebt, de sterkte van je a priori verwachting dus. Nu geeft CADA op basis van de opgegeven percentielen een beta verdeling die daar dicht bij in de buurt komt (je kunt je percentielen alsnog anders kiezen om de 'fit' te verbeteren), en die beta verdeling is gekenmerkt door de parameters a en b . Blijkt dus dat je met het opgeven van deze drie percentielen impliciet ook een aantal veronderstelde

waarnemingen opgegeven hebt. En dat aantal waarnemingen kan wel eens een stuk hoger zijn dan je voor je rekening wilt nemen (of lager). Het programma geeft je alle gelegenheid om de noodzakelijke correcties aan te brengen. Om m.a.w. de opgegeven beta verdeling geheel in overeenstemming te brengen met je a priori opvattingen. Of nog anders gezegd: om eventueel de nodige correcties in die a priori opvattingen aan te brengen, voorzover ze niet coherent blijken te zijn,

In het cesuurprobleem zou bijv. wat ervaringen met vroegere groepen betreft, gedacht kunnen worden aan een modus voor de a priori verdeling van 0,8: door de bank genomen menen we dat de proportie kandidaten die voor dit tentamen opkwamen en daar werkelijk een voldoende voor verdienden 0,8 was. Van de waarden van π boven 0,8 zou de helft daarvan boven de 8,25 kunnen liggen (m.a.w.: de kans dat π groter dan 8,25 zou zijn schatten we op 1/4). Idem voor het 25e percentiel: beneden welke waarde schatten we dat de kans dat π daar ligt 1/4 is? Zeg 7,25.

We gebruiken CADA om hiermee aan de slag te gaan, om uit te zoeken welke beta verdeling hier het best bij past. We zouden ook, nog voordat we met CADA gaan werken, kunnen bedenken hoeveel waarnemingen onze a priori inzichten 'waard' zijn. Dat kan het beste gebeuren in vergelijking met het aantal waarnemingen dat we nu gaan verzamelen (het aantal kandidaten dat nu tentamen moet doen of juist gedaan heeft). Zeg $n=100$.

En stel dat we onze a priori ideeën toch zeker ook wel 100 waarnemingen waard achten, zodat we een beta verdeling voor onze prior moeten hebben waar $a + b$ ongeveer gelijk 100 is.

Omdat we de modus van $B(a,b)$ a priori op 0,8 willen stellen, moet a ongeveer 80 zijn (immers, a is het aantal 'successen', personen die een voldoende verdienen, en b het aantal personen die geen voldoende verdienen; en $a+b=100$).

Attach, CADA1, ID=COWO, en kies het subprogramma dat je helpt bij het opstellen van de a priori beta verdeling (programma 2).

We specificeren de 25e, 50e en 75e percentielen als
? .725, .8, .825

en krijgen als antwoord dat de beta verdeling die daar het best bij past $B(a=49,74, b=12,68)$ is, met 25e, 50e, en 75e percentielen .764, .8, .833. Voorlopig accepteren we dit als redelijk en gaan door (type 0), waarna CADA de gelegenheid geeft om voor $a+b$ (het aantal hypothetische waarnemingen) een andere waarde te kiezen. De opgegeven verdeling bleek $a+b = 62,42$ te hebben, wat we veranderen in 100 (type 100). CADA wijzigt in overeenstemming hiermee de beta verdeling, die er nu uit ziet: $B(80,05, 19,95)$, 25e, 50e en 75e percentiel : .775, .802, .828.

De percentielen van deze verdeling zijn niet helemaal zoals we ze wilden hebben, echter op dit moment kunnen we daar niets in wijzigen. We zijn wel tevreden met $m=100$, en gaan door naar het volgende programmaonderdeel waar we het gemiddelde van deze verdeling kunnen wijzigen.

Onze meest waarschijnlijke a priori waarde voor π was .8, d.w.z. de modus van de beta prior willen we stellen op .8. De modus is i.h.a. niet gelijk aan het 50e percentiel (= de mediaan), en voor de onderhavige beta verdeling is de modus .807. Die waarde veranderen we door .8 op te geven. Daarmee wordt de betaverdeling $B(79.40, 20.60)$, en 25e, 50e, 75e percentielen .768, .796, .822, gemiddelde: .794. CADA geeft dan nog een keer gelegenheid om het aantal hypothetische waarnemingen te wijzigen, en de modus te wijzigen.

In het programma zijn we dan al een aantal malen HDR tegengekomen, afkorting van Highest Density Region, het kortste interval waarbinnen zich een opgegeven percentage van de verdeling bevindt. Zie Novick & Jackson blz 119-120 voor een al te summiere uitleg. Bijv.: het 50-90, HDR interval (géén betrouwbaarheidsinterval; dat zijn geheel verschillende zaken) is een interval waarbinnen π ligt met waarschijnlijkheid .5, en wel in het bijzonder van al dergelijke intervallen de kortst mogelijke.

In de figuur is een HDR interval geschetst, om ongeveer een idee te geven. Stel het is een 50% HDR, dan zou, voor een coherente persoon, zijn schatting van de waarschijnlijkheid dat π binnen dat interval ligt, .5 zijn. Zou nu het 50% HDR interval van de gekozen beta a priori verdeling naar ons idee te breed zijn, dan hebben we kennelijk toch te weinig hypothetische waarnemingen verondersteld, en moeten we om coherent te blijven de beta verdeling in die zin wijzigen.

Vervolgens kan ook nakijken van andere HDR intervallen (90% bijv.) geen kwaad.

Dit was het dan voor wat betreft het opstellen van een a priori beta verdeling, de explicitering van informatie over prestaties van groepen studenten die we in ons hoofd menen te hebben, informatie die we vaak impliciet gebruiken als referentiepunt om er de beoordeling van nieuwe groepen studenten op te baseren.

Terug naar Hofstee.

P.S. Het is een goed idee om nu al vast wat te experimenteren met CADA.

Je weet wel, login, identificatie, rekennummer, sup
connect, input, output
attach, cadal, id=cowo
cadal

In de aanpak van Hofstee wordt geen procedure gegeven om bijv. de eerste keer dat een bepaald tentamen wordt afgenomen een a priori schatting op te stellen over de proportie personen die een voldoende beheersing van de stof heeft. Misschien dat hij er wel een oplossing voor heeft, maar die wordt in het artikel niet gegeven.

Bij een bayesiaanse aanpak worden we wél in staat gesteld om ook wanneer een tentamen voor de allereerste keer wordt afgenomen onze a priori verwachting behoorlijk te specificeren, en wel in de vorm van een beta verdeling voor π , de proportie personen die een voldoende beheersing over de stof heeft (in termen van 'ware scores' zou je kunnen zeggen). Er is geen objectieve weg om tot een dergelijke verdeling te komen, een bezwaar dat te ondervangen is door de erbij gehanteerde overwegingen te expliciteren. Dergelijke overwegingen zouden kunnen zijn: ik richt mijn onderwijssituatie zo in dat in principe iedereen daarvoor moet kunnen slagen bij behoorlijke studieinspanning; wanneer studieprestaties beneden verwachting zouden zijn, en toch de zekerheid bestaat dat een behoorlijke studieinspanning inderdaad geleverd is, dan zal in verreweg de meeste onderwijssituaties de schuld niet op de student afgeschoven moeten worden (alleen daar waar objectieve absolute criteria voor de beheersing van inzichten en vaardigheden gesteld kunnen worden en gehandhaafd moeten worden in het belang van een veilige en verantwoorde beroepsuitoefening zou je daar van af moeten wijken); e.d. Subjectieve elementen in de gegeven beta verdeling kunnen onderwerp van overleg zijn tussen betrokkenen, waarbij andere partijen met hun eigen a priori verwachtingen (beta verdeling) op de proppen kunnen komen.

Bedacht moet worden dat de a priori verdeling betrekking heeft op een soort 'ware proportie' π , en daaruit volgt geenszins direct een aftestgrens: daarvoor zijn nog geheel andere overwegingen eveneens van belang. Met andere woorden: onze a priori verdeling specificceert nog geen a priori aftestgrens. Hofstee zit met hetzelfde probleem, zonder daar overigens een punt van te maken. De vraag die er uit volgt is dan ook waar je bij dit eerste tentamen de cesuur in termen van waargenomen score moet gaan leggen. Niet verrassend, omdat we vroeg of laat deze vraag toch moeten beantwoorden; wél verrassend in die zin dat een antwoord op die vraag niet rechtlijnig maar misschien via een soort iteratieve procedure verkregen moet worden.

Laten we voorlopig eens de volgende werkwijze uitproberen: bereken met behulp van CADA1 (programma 3) een groot aantal a posteriori verdelingen voor diverse mogelijke cesuren die in aanmerking zouden kunnen komen. Veronderstel dat we een a priori verdeling opstellen $B(88.83, 11.17)$, 25e, 50e en 75e percentiel resp. .869, .891, .911, aantal hypothetische waarnemingen 100, modus .896, gem. .888. Veronderstel dat we een tentamen afgenomen hebben waarvoor in de onderste regionen de volgende scores waargenomen werden: (n=100)

score	10	..	..	..	14	15
16	..	18	19			
frequentie	2				1	3
5	17					
aantal daarboven	98				97	
94	87		82	65		

We zouden de cesuur kunnen leggen ofwel bij 10 (2 onvoldoendes), bij 14, 15, 16, 18 of 19. Om de keuze te vereenvoudigen voeren we voor alle zes mogelijke cesuren een a posteriori analyse uit, m.b.v. CADA 1 programma 3, waaruit de resultaten zijn:

aantal successen	a	b	modus	gem	s.d.
50% HDR					
waargenomen					

98	187	13.94	.93	.02	.926	.949
97	186	14.93	.93	.02	.921	.945
94	183	17.92	.91	.02	.905	.931
87	176	24.88	.88	.02	.867	.898
82	171	29.86	.85	.02	.840	.874
65	154	46.72	.77	.03	.751	.792

TABEL 21. A posteriori verdelingen voor 6 mogelijke cesuren.

Wat doen we met deze resultaten? Op welke wijze valt daar uit af te leiden welke cesuur gekozen kan worden? Een mogelijke redenering is bij voorkeur die cesuur te kiezen die modus en 50% HDR interval in de a posteriori verdeling heeft die het dichtst liggen bij de waarden in de gekozen a priori verdeling. Laten we er van uit gaan dat de a priori verdeling is opgesteld voordat het onderwijs gegeven werd en voordat de docent kennis maakte met deze groep studenten. Als dat het geval is, dan zou de docent zich kunnen afvragen of er uit de geregistreerde gebeurtenissen tijdens het voorafgaande onderwijs aanwijzingen geput kunnen worden die er op wijzen dat de a priori verdeling in opwaartse of neerwaartse richting gewijzigd zou

moeten worden. Is dat het geval, dan zou een hogere of lagere aftestgrens gekozen moeten worden. (Een en ander natuurlijk onverlet het overleg dat betrokkenen over een en ander met elkaar zouden moeten voeren).

Het is duidelijk dat er wel heel sterke aanwijzingen uit de onderwijspraktijk moeten zijn om de cesuur te leggen op 10, 14 Of 19. De cesuur gelegd bij 15 zou met de a priori verdeling het meest overeenstemmen. Op basis van andere aanwijzingen zou deze cesuur wat verschoven kunnen worden.

Is dat gebeurd, dan moet nog de vraag beantwoord worden of deze gekozen cesuur ook wel geschikt is om als praktische cesuur te hanteren. Ik druk me ongelukkig uit, waar het om gaat is dat de cesuur die op geschetste wijze afgeleid en beredeneerd zou kunnen worden, een cesuur is in termen van 'ware scores' a.h.w., een schatting van de ware proportie voldoende in de onderhavige populatie. Voor concrete beslissingen hebben we echter te maken met feilbare gegevens, zodat op de ideale cesuur, misschien op te vatten als 'referentiepunt', nog een correctie naar boven of beneden (that depends) moet worden gegeven. Van belang daarbij is of men een selectiemodel hanteert, een plaatsingsmodel, een individueel keuze model, of dat men bepaalde andere overwegingen (aanvaardbaarheidsoverwegingen bijv.) hanteert. Het gestelde probleem is van ander karakter dan de in dit stuk bayesologie behandelde problematiek, en zullen we dan ook niet verder bespreken.

Zou de gegeven aanpak in essentie bruikbaar blijken, dan is duidelijk dat ze aangevuld moet worden met technieken of handleidingen die de docent of andere in staat stellen tijdens de onderwijsvoortgang gegevens te verzamelen die bij de eindevaluatie gebruikt kunnen worden om in de keuze van een 'ware' cesuur gebruikt te kunnen worden.

Ik moet bekennen dat ik een en ander een verrassende wending aan het cesuurprobleem vind:opleveren.

De hier gegeven benadering van de cesuurproblematiek is niet vergelijkbaar met die van Hofstee, omdat onze analyse gaat in termen van een verdelingsfunctie voor π , de ware proportie voldoende in de populatie, terwijl Hofstee uitsluitend over waargenomen scores blijft spreken. In deze zin lijkt zijn uitspraak dat dezelfde resultaten als door hem via zijn idiosyncratische aanpak verkregen, ook bij een Bayesiaanse analyse te voorschijn zouden komen. Een Bayesiaanse aanpak gaat langs essentieel andere lijnen. Maar goed, het is niet direct onze bedoeling om met Hofstee in discussie te gaan, maar om bayesiaanse methoden onder de knie te leren krijgen.

Een vervolg op onze succes story zou kunnen zijn dat de gekozen twaref cesuur en de daarbij behorende a posteriori verdeling bij een volgende tentamen afname gaan functioneren als a priori verdeling, waarna dan dezelfde analyse als in de vorige bladzijden beschreven gevolgd kan worden. Etcetera ook voor volgende tentamenafnamen. Waarschuwing daarbij: je kunt niet zo maar doorgaan met vroegere tentamenresultaten te cumuleren in je a priori verdeling; immers, ook het verleden moet een kans krijgen weg te slijten. Onderwijssituaties veranderen, leerstof wordt gewijzigd, vooropleiding van studenten verandert etcetera, zodat de a priori verdeling in het algemeen op minder aantal waarnemingen gebaseerd moet worden dan in feite beschikbaar zijn. Er moet ruimte blijven voor verschuivingen als gevolg van onverwachte gebeurtenissen die zich voorafgaand aan het laatste tentamen kunnen voordoen om invloed uit te oefenen op de a posteriori verdeling van π , en daarmee op de te hanteren aftestgrens. Eenzelfde soort waarschuwing geldt ook voor de aanpak van Hofstee, en zou in zijn artikel niet misstaan hebben.

Hopelijk kan over een en ander een vruchtbare discussie gevoerd worden. In ieder geval is duidelijk dat we hier op een belangrijk spoor zitten voor wat betreft het zoeken naar procedures die de docent kunnen helpen bij het langs rationele weg besluiten tot een bepaalde cesuur op zijn tentamen.