80-10-15_Voorspellen_studieresultaten.rtf
80-10-15_Voorspellen_studieresultaten.pdf

COW0 Ben Wilbrink [eigen aantekeningen]

Voorspellen van eigen studieresultaten.

Over het voorspellen van eigen toetsresultaten vallen verschillende dingen te zeggen. Allereerst: welke techniek kan de student gebruiken om zo'n voorspelling te doen? Vervolgens: hoe waarschijnlijk is het dat studenten op intuitieve wijze tot ongeveer dezelfde voorspelling komen als op basis van de gegeven techniek het geval zou zijn? Hoe combineer je voorspellingen van individuele studenten tot een soort groepsvoorspelling van het groepsresultaat op de toets? Hoe kun je vanuit de verkregen toetsresultaten terugredeneren naar de mate waarin individuele studenten hun toetsresultaat hebben kunnen voorspellen? Hoe check je het laatste modelmatige resultaat tegen empirische gegevens over de individuele voorspellingen, gedaan vlak voor het afleggen van de toets? Wat kun je doen met achterafvoorspelling van de toetsscore, wanneer de student extra informatie heeft gekregen over zijn prestaties, maar nog niet precies weet wat de uitslag van zijn toets zal zijn? Hoe gaan dit soort voorspellingen in hun werk wanneer alleen maar de uitslag zakken/slagen voorspeld wordt? En hoe gaat een en ander wanneer een driedeling onvoldoende/ compenseerbaar/voldoende voorspeld wordt, bij een

gedeeltelijk compensatorische examenregeling zoals bij Psychologie aan de UvA gehanteerd? In termen van percentages correcte voorspellingen: wat zijn voor dergelijke voorspellingen onder en bovengrenzen? Hoe verhouden die onder en bovengrenzen zich tot de percentages die Hoogstraten en Vorst rapporteren? Zijn de resultaten van Hoogstraten en Vorst voorspelbaar op basis van het hier te ontwikkelen model, waarin aantal toetsvragen, gemiddeld toetsresultaat, en spreiding van dat toetsresultaat een rol spelen?

In het volgende zal ik proberen deze vragen in deze volgorde te beantwoorden.
(dit is geen concept voor een TOR commentaar; het te geven materiaal kan daartoe eventueel wel omgewerkt worden).

techniek voor het voorspellen van het eigen toetsresultaat.

[2002: introduceert gebruik van de betaverdeling als prior. In het Algemene ToetsModel is die omslachtige procedure vervangen door de aannemelijkheid gegeven een proeftoetsscore, of andere informatie die in termen van een proeftoetsscore is te vertalen, of een combinatie van beide (optellen van aantal goed en van aantal items)]

Voor de student ziet de toets er uit als een random steekproef uit een (denkbaar) domein van vragen

over de opgegeven stof. Dat betekent dat, gegeven zijn ware stofbeheersing, de voorspellende waarschijnlijkheidsverdeling voor de toetsscore of het toetsresultaat de vorm van een binomiaalverdeling heeft met als parameters de ware stofbeheersing p en het aantal vragen in de toets n;

(1) $f(x|p) = (n \text{ boven } x)\ p^{x}\ (1p)^{nx}$

De student kan ook voor zijn ware stofbeheersing een waarschijnlijkheidsverdeling opzetten. Daar zijn verschillende methoden voor te gebruiken, waarvan de meest precieze is de techniek voor het specificeren van priors zoals bijv, gegeven door Noviek & Jackson (1974). Gegeven dat deze verdeling te combineren moet zijn met de binomiaal in vergelijking (1) wordt de waarschijnlijkheidsverdeling voor de eigen stofbeheersing gekozen als een betaverdeling Be(c,d):
deling Be(c,d):

(2) $f(p) = Be(c, d) = B^{-1}\ (c, d)\ p^{c-1}\ (1-p)^{d-1}.$

waarbij

(3) $B(c, d) = (c-1)!\ (d-1)!\ /\ (c+d-1)!$

Gebruik makend van de relatie

(4) $f(x,p) = f(x|p)\ f(p)$

$$= B^{-1}(c,d)\ (n \text{ boven } x)\ p^{c+x-1}\ (1-p)^{d+n-x-1}$$

kan de voorspellende waarschijnlijkheidsverdeling voor het toetsresultaat f(x) verkregen worden door f(x,p) te integreren over alle waarden van p:

(5) $$f(x) = \int_0^1 f(x,p)\ dp$$

$$= B^{-1}(a,d)\ (n \text{ boven } x) \int_0^1 p^{a+x-1}\ (1-p)^{d+n-x-1}\ dp.$$

De integraal in het rechterlid van vergelijking (5) is gelijk aan de incomplete beta:

(6) $$B(c+x, d+n-x) = \int_0^1 p^{c+x-1}\ (1-p)^{d+n-x-1}\ dp.$$

zodat

(7) $$f(x) = (n \text{ boven } x)\ B^{-1}(c,d)\ B(c+x,d+n-x)$$
$$= BeBi(c,d,n)$$

De verdeling in vergelijking (7) is de betabinomiaalverdeling BeBi(c,d,n), ook wel bekend onder de naam negatiefhypergeometrische verdeling, Pólyaverdeling, of PólyaEggenberger.

Deze verdeling is m.b.v. een recursieformule eenvoudig te berekenen (zie Wilbrink Cesuurbepaling 1980 bijlage D).

Wanneer de parameters c en d niet al te klein zijn, en de verdeling niet al te asymmetrisch is, kan de betabinomiaal benaderd worden met de normaalverdeling.

Zijn de specificaties van de waarschijnlijkheidsverdeling voor de eigen stofbeheersing reëel, dan zou voor deelgroepen studenten die **dezelfde** specificatie geven moeten gelden dat de toetsscoreverdeling voor deze deelgroep gelijk is aan de voorspelde (voorspellende) waarschijnlijkheidsverdeling.

Hoe verhoudt deze techniek zich met meer intuitieve voorspellingen?

[2002: voor intuïtieve voorspellen is er nu het recente boek van Hogarth. Ik ben nu bezig om het Algemene ToetsModel op de dataset van Rechten uit de 8--er jaren te leggen, precies wat in de laatste alinea van deze paragraaf een wens is.]

De bovenbeschreven techniek is een **voorschrijvende** techniek. Het is dan ook nog maar de vraag of intuitief tot stand gekomen voorspellingen in overeenstemming zijn met wat de gegeventen techniek opgeleverd zou hebben. Ofwel: kan de voorschrijvende techniek gebruikt worden als model voor het **beschrijven** van de voorspelling die studenten voor de eigen toetsscore kunnen doen?

Het is op voorhand wel zeker dat er verschillen beide zullen zijn. Die verschillen kun je vervolgens proberen terug te voeren tot verklarende variabelen. Die kunnen liggen in de sfeer van de persoonlijke verschillen (voorzichtigheid, of overmoed, bij het voorspellen,bijvoorbeeld). De verschillen kunnen ook een meer systematisch karakter hebben, bijvoorbeeld voortvloeiend uit het onvermogen om rekening te houden met het steekproefkarakter van de toets, of met de rol van raadkansen wanneer de toets uit meerkeuzevragen bestaat. Hier kan alleen empirisch onderzoek opheldering over verschaffen. Voorshands lijkt deze onderzoeksrichting niet direct van grote relevantie voor het te voeren onderwijsbeleid.

Een interessanter gebruik van de voorschrijvende techniek als model voor de meer intuitieve voorspellingen zoals studenten die in werkelijkheid (impliciet) doen is de volgende. Toetssituaties zullen doorgaans verschillen vertonen in de mate van doorzichtigheid, in het aantal vragen waaruit de toets bestaat, de plaats waar de cesuur gelegd wordt, e.d. Uit daarvoor te construeren modellen zijn predicties af te leiden voor het studentengedrag bij de voorbereiding op de toetsing. En omdat bedoeld gedrag afhankelijk is van de voorspelling die de student doet over zijn toetsscore, gegeven wat hij denkt dat zijn stofbeheersing is, ligt hier een duidelijke relatie met studentvoorspellingen, die hopelijk op tamelijk eenvoudige wijze ook aan (routinematig verzamelbare) gegevens empirisch na

te gaan is.

Dat de gegeven techniek er erg ingewikkeld uitziet hoeft geenszins te betekenen dat de resultaten niet erg in overeenstemming zullen zijn met de meer intuitieve benadering uit de dagelijkse studiepraktijk van de student. Hij heeft geleerd om tamelijk realistische voorspellingen te doen, en dat die voorspellingen op tamelijk globale overwegingen berusten sluit overeenstemming met een complex model niet op voorhand uit.

Hoe combineer je individuele voorspellingen?

[2002: een interessant thema. Maar misschien moet je dat niet letterlijk uitwerken, en is de Coleman-benadering beter, zoals door mij in 1992 op de conferentie in Twente gerapporteerd. Ik vat de techniek uit deze paragraaf nu als volgt samen: de groep studenten heeft een scoreverdeling waarop een betabinomiaal waarschijnlijk goed zal passen, in welk geval de ware beheersing in de groep verdeeld is te denken als beta. De truc is nu de groep te behandelen als zou deze een individu zijn. (zo heb ik het in deze paragraaf niet verwoord, maar zo is het wel het meest helder.) Voor de groep is de voorspellingssituatie dan volkomen analoog aan die van de individuele student, en het verschil is natuurlijk dat de groepsresultaten per definitie meer gespreid zijn dan bij herhaalde toetsing (bij simulatie bijv.) van een individu het geval zou zijn. Maar in de

techniek die ik hier gebruik doe ik iets moeilijks: ik ga een binomiaal fitten op de beta voor de groep. Conceptueel maak ik dat niet meteen erg helder, maar in de bijlage van Cesuurbepaling 1980 doe ik dat samen met figuren, als ik me goed herinner. Tenzij die bijlage identiek is aan dit stuk! Wat mist is de volkomen heldere constructie van een aannemelijkheid, dat is voor het Algemen Toetsmodel inderdaad ook een van de doorbraken geweest. In oktober 1980 ontbrak dat dus nog ten enenmale. Nu is het voor een groepsresultata ook niet meteen dudielijk hoe je daarvoor een aannemelijheid zou kunnen construeren of simuleren; neem aan dat er een beta voor ware beheersing aan ten grondslag ligt, dan kun je althans in beginsel iets construeren van een aannemelijkheid voor de basis-beta. Probleem is dat je twee parameters hebt, dus iets driedimensionaals moet doen, wat natuurlijk niet onmogelijk is: zowel de som a+b variëren, als de verhouding a/b. Aardig idee, eigenlijk. Moet als simulatie heel goed te doen zijn, toch?]

Veronderstel eens de ideale situatie dat álle studenten via de beschreven voorspellingstechniek de toetsscore voorspellen (of beter: de voorspellende waarschijnlijkheidsverdeling voor de toetsscore specificeren). Hoe combineer je al die individuele voorspellingen tot een enkele voorspelling voor de verdeling van de toetsscores voor deze groep studenten? Zou je die groepsvoorspelling doen, dan kun je de voorspelling vergelijken met het verkregen

resultaat, de empirische toetsscoreverdeling.

Een mogelijkheid zou kunnen zijn om alle individuele betaverdelingen te combineren tot één enkele verdeling, op de wijze zoals Molenaar (1980) dat voor de combinatie van twee betapriors heeft gedaan. Dit is een mogelijkheid die ik no niet heb onderzocht, waarvan ik gen dan ook niet weet of ze leidt tot wiskundig hanteerbare oplossing.

Een andere aanpak is wél uitgewerkt, en daarin wordt wederom het betabinomiale model uitgebuit. In het betabinomiale toetsmodel wordt het binomiale model voor de individuele toetsscore gecombineerd met de betaverdeling voor de ware stofbeheersing in de **groep** studenten. Dat is geheel analoog aan het hierboven besproken geval voor de **individuele** student die een waarschijnlijkheidsverdeling voor zijn ware stofbeheersing specificeert.

Zoals het binomiale model voor de toetsscore via combinatie met de waarschijnlijkheidsverdeling over ware stofbeheersing voor de groep studenten de empirische toetsscoreverdeling genereert, zo kun je deze empirische toetsscoreverdeling ook als gegenereerd beschouwen door de individuele voorspellende toetsscoreverdelingen gecombineerd met een geschikte mengende verdeling die op een of andere wijze de verschillen tussen studenten weergeeft.

Ik heb laten zien dat de voorspellende

waarschijnlijkheidsverdeling voor de toetsscore (voor de individuele student) een betabinomiaal verdeling BeBi(c,d,n) is. Ook is bekend dat empirische toetsscoreverdelingen zich doorgaans goed tot uitstekend laten benaderen door eveneens een beta~binomiaal verdeling, laten we zeggen BeBi(a,b,n). Houden we vast aan de betabinomiaal 'fit' op de empirische toetsscoreverdeling, dan is het probleem dat de individuele betabinomiaal verdelingen zich niet eenvoudig laten mengen tot wederom een betabinomiaal verdeling. Aan dat probleem is een mouw te passen door de betabinomiaalverdeling voor de voorspellende waarschijnlijkheidsverdeling van de individuele student te vervangen door de binomiaalverdeling die hetzelfde gemiddelde en dezelfde stnadaarddeviatie SD heeft, en dat is een benadering die doorgaans wel redelijk goed uit zal vallen. Dan past daar als mengende verdeling wederom een betaverdeling Be(v,w) bij, en dat blijkt de verdeling te zijn **over studenten** van de door hen aangegeven meest waarschijnlijk te behalen toetsscore (of, wat op hetzelfde neerkomt, van de door hen aangegeven meest waarschijnlijke eigen stofbeheersing; 'meest waarschijnlijk' is daarbij te verstaan als de verwachte waarde of het gemiddelde voor iedere individuele student). Daarbij is dan wel de aanname gemaakt dat de individuele parameters c i en d de eigenschap hebben dat voor alle i c_i+d_i = constant. D.w.z.: aangenomen wordt dat alle studenten even nauwkeurig schatten.

Bij de BeBi(c,d,n) is de binomiaalverdeling Bi(q,m) te fitten door gemiddelde en variantie voor beide verdelingen aan elkaar gelijk te stellen.

Gemiddelde en variantie van BeBi(c, d, n) zijn:

(8) $nc / (c+d), \quad ncd (c+d+n) / (c+d)^2 (c+d+1).$

Gemiddelde en variantie van Bi(q,m) zijn:

(9) $mq, mq (1-q)$

Algebraïsche uitwerking resulteert dan in:

(10) $q = c / (c+d),$

(11) $m = n (c+d+1) / (c+d+n),$

waarbij eraan gedacht moet worden beide verdelingen eerst op dezelfde schaal te brengen, bijvoorbeeld de schaal 0 tot 1, en pas daarna gemiddelden en varianties gelijk te stellen.

Merk op dat de nu verkregen binomiaalverdeling Bi(q, m) niet meer overeenstemt met de empirische situatie, waarin immers de toets nog steeds uit n vragen bestaat. Omdat in het volgende met de Bi(q, m) verder gewerkt zal worden, worden ook de empirische toetsresultaten getransformeerd naar de schaal van 0 tot m, door het gemiddelde te delen door n en te vermenigvuldigen met m, en hetzelfde te doen met de SD.

Neem aan dat de nauwkeurigheid waarmee voorspeld wordt voor alle studenten i gelijk is, d.w.z. dat de parameters e en d voor alle studenten tot dezelfde constante sommeren: c_i+d_i = constant voor alle i.

Dan is de parameter q = c / (c+d) gelijk aan de verwachte waarde van de (proportionele) toetsscore, of de verwachte waarde voor de eigen stofbeheersing. Laat de verdeling over studenten van deze parameter q de betaverdeling Be(v,w) zijn. In beginsel is deze betaverdeling ook op te vatten als benadering of fit voor de empirische verdeling die je zou krijgen wanneer alle studenten gevraagd wordt de verwachte waarde voor hun toetsscore op te geven, d.i. de meest waarschijnlijke waarde (maar niet de modus).

Dan is het complete model, analoog aan het in vergelijkingen (1) tot en met (7) behandelde model:

(12) f(xlq) = (m boven x) q^x $(1q)^{m-x}$ = Bi(q, m),

(13) g(q) = Be(v, w) = B^{-1} (v, w) q^{v-1} $(1q)^{w-1}$,

(14) g(x) = (m boven x) B^{-1} (v,w) B(v+x, w+m-x)
= BeBi(v, w, m).

Vergelijking (14) is dan de voorspelling van de empirische toetsscoreverdeling (waarbij deze laatste overgezet is van de schaal van 0 tot n naar de schaal van 0 tot m).

De voorspelling is niet helemaal bepaald, omdat voor c+d een constante waarde gekozen moet worden. In beginsel kan deze constante waarde natuurlijk ook empirisch bepaald worden, maar dan zijn puntschattingen niet voldoende.

Wat zeggen de toetsresultaten over de voorspelbaarheid?

In de voorgaande paragraaf werd een model opgezet om van individuele voorspellingen te komen tot de voorspelling van het groepsresultaat, d.i. de empirische toetsscoreverdeling.

In deze paragraaf wordt het omgekeerde gedaan: vanuit de empirische toetsscoreverdeling terugredeneren naar de individuele voorspellingen die daar waarschijnlijk mee in overeenstemming zijn.

Om te beginnen is daarvoor nodig dat de empirische toetsscoreverdeling in wiskundige termen vertaald wordt. Met andere woorden: er wordt een waarschijnlijkheidsverdeling bij gezocht. Het ligt voor de hand, gezien de gunstige ervaringen met deze verdeling opgedaan, om een beta binomiaalverdeling te fitten. De daarbij te gebruiken methode is de eenvoudige momentenmethode, waarbij gemiddelden en varianties gelijk gesteld worden.

Voordat dat op empirische data toegepast gaat

worden, moeten we eraan denken de schaal waarop die data staan (de schaal van 0 tot n, waarbij n het aantal vragen in de toets is) te transformeren naar 0 tot m, waar m gelijk is aan:

(15) $$m = n (c + d + 1) / (c + d + n)$$

Daarvoor is nodig om voor c+d een constante waarde te kiezen, waarvoor een bepaald veelvoud van n gekozen kan worden. Er zijn hier twee verschillende benaderingen mogelijk. De ene benadering maakt gebruik van informatie die op andere wijze al verkregen is over de grootte van c+d. De andere benadering gaat uit van de onbekendheid van c+d, en probeert juist door in het model met trial waarden voor c+d te werken er achter te komen welke waarde voor c+d het meest waarschijnlijk de empirische scoreverdeling heeft gegenereerd. Voor de laatste benadering zou je om te beginnen c+d=n, c+d=2n, en c+d=3n kunnen kiezen (ieder afzonderlijk uit te werken, natuurlijk).

Laat BeBi(v, w, m) de te fitten betabinomiaalverdeling zijn. Dan zijn de parameters v en w als volgt uit te drukken in gemiddelde M en standaarddeviatie SD van de empirische scoreverdeling (op schaal 0 tot m gezet):

(16) $$w = \frac{SD^2 - M (m - M)}{M - m\, SD^2 / (m - M)} ,$$

(17) $$v = wm /(m - M)$$

Met deze fit is ook de betaverdeling Be(v, w) bepaald, de verdeling over studenten van individueel verwachte toetsresultaten.

Het is mogelijk dat (16) en (17) niet tot een oplossing met positieve waarden voor v en w leiden. In dat geval is c+d te klein gekozen, en moet deze constante opgehoogd worden.

Uit een en ander blijkt dat de empirische toetsscoreverdeling in ieder geval een ondergrens stelt aan de nauwkeurigheid waarmee studenten hun toetsresultaat voorspellen. De empirische resultaten laten niet een willekeurige mate van onnauwkeurigheid toe. Immers, de parameters v en w moeten positieve waarden aannemen.

Wat de bovengrens aan de nauwkeurigheid is, is nog niet direct duidelijk, hoewel uiteindelijk het steekproefkarakter van de toets een absoluut plafond bepaalt.

voorbeeld

Een toets van 80 vragen levert de empirische scoreverdeling op met gemiddelde M=59 en standaarddeviatie SD=9. (vgl. Meerum Terwogt-Kouwenhoven 1980 tabel 2, inleiding in de psychologie toets C).

Kies om te beginnen eens c+d=17. Transformeer gemiddelde en standaarddeviatie naar de nieuwe schaal van 0 tot m, waar m=14,85 (uit vergelijking

15). Duid de getransformeerde waarden met een accent aan. Dan is M'= 10,95 en SD'=1,67. Dan zijn de gezochte parameters w en v te bepalen uit vergelijking (16) en (17):

$$(18)\quad w = \frac{1{,}67^2 - 10{,}95 \,.\, 3{,}90}{10{,}95 - 14{,}85\ 1{,}67^2 /\ 3{,}90}$$

$$= -\ 39{,}92/0{,}33 = -120{,}97.$$

Omdat w positief moet zijn, blijkt c+d te klein gekozen. In dit geval wordt wél een goed resultaat verkregen wanneer c+d=18 gekozen wordt. Dan is m=15,35, waarna M'=11,32 en SD'=1,73, zodat nu:

$$(19)\quad w = \frac{1{,}73^2 - 11{,}32 \,.\, 4{,}03}{11{,}32 - 15{,}35 \,.\, 1{,}73^2 /\ 4{,}03}$$

$$= -\ 42{,}63 / 0{,}080 = 532{,}88.$$

$$(20)\quad v = 532{,}88 \,.\, 15{,}35 / 4{,}03 = 2029{,}70.$$

Deze resultaten betekenen dat voor c+d=18 alle studenten dezelfde verwachte waarde voor hun toetsseore (of hun stofbeheersing) hebben, dus dezelfde voorspelling doen. De variantie van Be(v, w) is namelijk bijzonder klein. De waarde c+d=18 zit heel dicht bij de absolute ondergrens voor de nauwkeurigheid waarmee studenten de eigen stofbeheersing of het eigen toetsresultaat voorspellen. In de praktijk kun je eigenlijk wel zeggen dat c+d=18 die ondergrens is. Dit alles aangenomen

dat studenten onderling niet verschillen in de mate van nauwkeurigheid.

Vertaald zou je kunnen zeggen dat een ondergrens aan de voorspellingsnauwkeurigheid gesteld is die overeenkomt met de informatie zoals een proeftoets bestaande uit c+d-2=16 items die aan de student levert (na afloop, en wanneer die proeftoets correct gescoord is).

Een geringere nauwkeurigheid dan deze ondergrens is eenvoudig niet in overeenstemming te brengen met de empirische gegevens, tenzij er aanwijzingen zijn dat de modelaannamen in dit speciale geval niet opgaan (maar dan moet de analyse overgedaan worden nadat de data gecorrigeerd zijn door bijv. de gegevens van een afwijkende deelgroep studenten uit de berekening te verwijderen).

Maximale nauwkeurigheid is bereikt wanneer iedere student zijn ware beheersing precies kent. In dat geval wordt de betaverdeling Be(v,w) gelijk aan de ware score verdeling Be(a,b). In het voorbeeld is deze ware score verdeling, formules (16) en (17) gebruikend:

(21) $b = (9^2 - 59 \cdot 21) / (59 - 80 \cdot 81 / 21) = 4{,}64$

(22) $a = 4{,}64 \cdot 59 / 21 = 13{,}04.$

De standaard deviatie van Be(4,64, 13,04) is 0,102. Op de schaal van de toets van 80 vragen is dat **8,14**

Voor enkele tussenliggende waarden voor c+d zijn de resultaten:

c+d=42 : m=28,20, M'=20,80, SD'=3,168
w=8,24 en v=23,16
standaard deviatie op toetsscoreschaal: **6,18**.

c+d=82 : m=40,99, M'=30,23, SD'=4,608
w=6,00 en v=16,86
standaard deviatie op toetssooreschaal: **7,20**.

c+d=162: m=53,88, M'=39,74, SD'=6,062
w=5,24 en v=14,73
standaard deviatie op toetsscoreschaal: **7,69**.

De standaarddeviaties op toetsseoreschaal zijn dan te vergelijken met de standaard deviatie van de verdeling over studenten van de voorspelde toetssoore. Zonodig wordt de berekening voor nog weer andere waarden voor c+d gedaan, bijv:

c+d=32 m=23,57, M'=17,38, SD'=2,655
w=10,63 en v=29,85
standaard deviatie op toetsscoreschaal: **5,47**.

Een toets van 30 vragen levert de empirische scoreverdeling op met gemiddelde M=18 en SD=4,7. (vgl. Meerum. TerwogtKouwenhoven 1980 tabel 2, testleer) . Berekeningen:

c+d=13: resulteert in negatieve w.

c+d=14 : m=10,23, M'=6,14, SD'=1,603
w=70,06 en v=105,69

standaard deviatie op toetsscoreschaal: **1,10**.

c+d=22 : m=13,27, M'=7,96, SD'=2,077,
w=13,51 en v=20,25
standaard deviatie op.toetsscoreschaal: **2,49**.

c+d=32 : m= 15,97, M'=9,58, SD'=2,500,
w=9,10 en v=13,64
standaard deviatie op toetsscoreschaal: **3,02.**

c+d=62 : m=20,54, M'=12,33, SD'=3,220,
w=6,65 en v=9,99
standaard deviatie op toetsscoreschaal: **3,50**

c+d=92 : m=22,87, M'=13,72, SD'=3,581,
w=6,14 en v= 9,21
standaard deviatie op toetsscoreschaal: **3,63**.

ware score n=30, M=18, SD=4,7, b=5,21 en a=7,82
verdeling: standaard deviatie op toetsscoreschaal: **3,92**.

Voorspellenachteraf

Met een geheel andere situatie krijg je te maken wanneer de student nádat de toets is afgelegd, maar nog voordat de uitslag bekend kan zijn, gevraagd wordt om de behaalde score te schatten.

Zou de student na afloop van de toets gevraagd worden om zijn **stofbeheersing** te schatten, dan kan

een model opgesteld worden met behulp van technieken uit de Bayesiaanse statistiek. De student stelt voorafgaand aan het tentamen zijn prior op voor wat hij denkt dat zijn stofbeheersing is, en hij herziet die prior na afloop van het tentamen op grond van de extra informatie die hij heeft gekregen. Ook dan is nog een probleem dat hij erg onzeker zal zijn over het aantal op de toets goed gemaakte vragen, zodat het misschien niet mogelijk of tenminste erg omslachtig is om hier een wiskundig model te hanteren. Eenvoudiger is het om de student na afloop opnieuw de prior voor wat hij denkt dat zijn stofbeheersing is, te laten bepalen.

De student de vraag voorleggen welke score hij denkt behaald te hebben lijkt niet erg zinvol. Welke informatie kan dat opleveren? Er is geen enkele reden om op voorhand te veronderstellen dat de voorspelling achteraf een bepaald verband met de voorspelling vooraf zal hebben (het is niet duidelijk welke relatie tussen beide gelegd kan worden). Het lijkt er nog het meest op dat je de student op deze wijze vraagt naar het aantal vragen waarop hij met zekerheid het juiste antwoord gegeven denkt te hebben, waarop hij geraden heeft, of waarop hij geraden heeft na één of meer alternatieven (bij meerkeuzevragen) als in ieder geval onjuist, te hebben afgestreept.

De data die op deze wijze verzameld worden vertonen dan naar verwachting enige verwantschap aan de empirische resultaten die **zekerheidsscoring**

oplevert. In verband met de vraag naar de voorspelbaarheid van studieresultaten hebben ze geen enkele betekenis.

Voorspellen van de uitslag gezakt / geslaagd

Het voorspellen van de uitslag gezakt/geslaagd hoeft weinig méér problemen op te leveren dan de voorspelling van de toetsscore. Wanneer de student zijn voorspellende waarschijnlijkheidsverdeling voor de toetsscore heeft opgesteld, dan is de waarschijnlijkheid dat de score lager is dan de gestelde cesuur daaruit op eenvoudige wijze af te leiden (uit de cumulatieve BetaBinomiaalverdeling). Zo geeft Wilbrink (1978, bijlage F) enkele tabellen voor het zakrisico, wanneer de voorspellings 'nauwkeurigheid' c+d=25.

Voor de betaverdelingen in de voorgaande paragraaf berekend, kan uit Pearson's 'Tables of the incomplete betafunction' afgeleid worden welk percentage studenten een verwachte (=voorspelde) toetsscore heeft beneden de cesuur, of tenminste gelijk aan de cesuur. Deze percentages, opgesteld voor verschillende waarden van de 'nauwkeurigheid' van voorspellen c+d, kunnen vergeleken worden met de empirische percentages.

Uiteraard volgt uit het opgestelde model eveneens hoeveel van deze voorspellingen correct zullen blijken te zijn, en ook die voorspellingen kunnen met

de empirische resultaten vergeleken worden. Het probleem is dat een en ander niet exact uitvoerbaar is, omdat het daarvoor benodigde bivariate model niet wiskundig traceerbaar lijkt te zijn. Wanneer verdelingen niet symmetrisch zijn, en de parameterwaarden klein zijn, is ook een normaalverdeling als benadering (voor de voorspellende evenals voor de empirische verdeling) niet goed bruikbaar, hoewel je langs deze weg snel enige voorlopige resultaten binnen zou kunnen halen. Alléén in de situatie van maximale nauwkeurigheid, waarin studenten de eigen ware stofbeheersing exact kennen, vereenvoudigt het bivariate model tot een bivariaat betabinomiaal model met dezelfde parameterwaarden voor beide marginale verdelingen, en daar kan waarschijnlijk mee gerekend worden (Lord en Novick 1968 P. 519520), (Ishii en Hayakawa 1960).

Een andere mogelijkheid om tenminste een indruk te krijgen van de theoretisch te verwachten succespereentages bij het voorspellen van zakken of slagen is deze: bereken voor een aantal 'representatieve' waarden voor de verwachte (=voorspelde) toetsscore wat de waarschijnlijkheid is dat de voorspelling ook tot een juiste zakslaag voorspelling leidt, en doe dat voor ófwel de waarschijnlijk geachte waarde voor de bekende parameters c+d, ófwel voor verschillende gekozen waarden voor c+d. Dit komt er op neer dat enkele cumulatieve betabinomiaal verdelingen geëvalueerd moeten worden, en dat kan hopelijk in veel gevallen

zonder veel bezwaar via de normaalverdeling als benadering (zoals gedaan in Wilbrink 'Kansberekeningen bij Pais' voorontwerp Machtigingswet', juli 1980).

Enkele berekeningen die aansluiten bij de voorbeelden gegeven op blz. 7 en 8, leveren de resultaten zoals vermeld op de volgende blz. Ieder item is als volgt berekend. Gegeven wordt verondersteld dat de voorspellingsnauwkeurigheid c+d=82, de toets bevat 80 vragen de cesuur is 56 (voor de berekeningen wordt dan 55,5 gehanteerd), en veronderstel dat de voorspelde toetsscore (=verwachte toetsscore) 65 is. De voorspellende waarschijnlijkheidsverdeling voor de toetsscore is in dit geval BeBi(66,62, 15,38, 80). De variantie van BeBi(c, d, n) is

(23) $n\ c\ d\ (\ c + d + n\ ))\ /\ (\ c + d\ )^2\ (\ c + d + 1\ )$

wat in dit geval de SD 4,88 oplevert. De verwachte toetsscore ligt 91 punt boven de cesuur, ofwel 1,95 standaard deviaties. De normaal2 verdeling als benaderin gebruikend levert de tabel voor de standaard normaal verdeling dat de waarschijnlijkheid beneden 55,5 te scoren 0,026 is, of 3%

TABEL
Waarschijnlijkheden voor juiste voorspelling van de uitslag

---

toets 80 vragen, cesuur c+d=42	55,5, c+d=82	i d e m
verwachte score 55 50	65 60 55 50	65 6 0
% juist 53 77	97 80 53 82	95 7 5
toets 80 vragen, cesuur c+d = 42	43,5, c+d=82	i d e m
verwachte score 45 40	55 50 45 40	55 5 0
% juist 58 68	98 86 59 71	95 8 1
toets 30 vragen, cesuur	20,5, c+d = 62	
verwachte score 15	23 22 21 20 19 18 17	1 6
% juist 95	81 69 56 56 68 78 86	9 1
toets 30 vragen, cesuur	20,5, c+d=32	
verwachte score	23 21 19 17 15	
% juist	79 51 66 83 93	
toets 30 vragen, cesuur	16,5, c+d = 62	
verwachte score	23 21 19 17 15 13	
% juist	99 93 78 56 67 86	
toets 30 vragen, cesuur	16,5, c+d=32	
verwachte score	23 21 19 17 15 13 11	
% juist	98 90 75 55 66 83 94	

Uit deze gegevens valt wel een globaal overall percentage juiste uitslagvoorspellingen af te leiden. Voor de toets van 80 vragen, met c+d=82 verondersteld, gebruiken we het resultaat van blzw 8 dat de verwachte (=voorspelde) toetsscores het gemiddelde 59 hebben, en de standaard deviatie 7,20. Leg in gedachten deze verdeling over de zojuist gevonden percentages behorend bij een aantal punten uit die de schaal waar die verdeling op staat. Ongeveer 67 % van de studenten heeft een verwachte toetsscore binnen plus of min één SD van 59, terwijl 95 % binnen plus of min twee SD van 59 de verwachte toetsscore aanwijst. Ongeveer 75 tot 80 % juiste voorspellingen levert dat op. Het is duidelijk dat een preciesere schatting verkregen kan worden door de berekening voor méér punten te doen, en de resultaten te wegen met in dit voorbeeld Be(6,00, 16,86), of in plaats daarvan de geschikte normaalverdeling hanteren en refereren aan de tabel voor de standaard normaalverdeling. Wanneer studenten minder nauwkeurig schatten, c+d = 42, dan komt een en ander uit op 70 tot 75 % juiste uitslagvoorspellingen (bij cesuur 55,5).

De toets van 30 vragen uit het voorbeeld, cesuur 20,5, c+d=62, gemiddeld voorspelde score 18, SD 3,5 in deze voorspellingen, levert als resultaat een percentage correcte voorspellingen van 80 tot 85 %.

Merk op dat rond de cesuur de waarschijnlijkheid van juiste voorspellingen vlak boven de 50 % ligt, en dat 50% de ondergrens is.

Merk ook op dat daaruit volgt dat bij het groter worden van het verschil tussen de cesuur en het gemiddeld (voorspeld) toetsresultaat het overall percentage correcte voorspellingen stijgt.

Het percentage correcte voorspellingen van de uitslag zal ook stijgen wanneer de standaard deviatie van de door studenten voorspelde toetsscores groter wordt, omdat daardoor scores zich ook meer van de cesuur áf zullen spreiden; dit verband geldt alleen wanneer cesuur en gemiddeld voorspelde score niet te ver uit elkaar liggen, terwijl het omgekeerde geldt wanneer er juist wél een forse afstand tussen cesuur en gemiddeld voorspelde score bestaat. Dit laatste effect is vooral van belang gezien het effect dat de nauwkeurigheid van voorspellen heeft op deze spreiding; in bepaalde situaties zou dat tot tegenintuitieve resultaten kunnen leiden. (Bijv.: in een situatie waarin studenten in hun tentamenvoorbereiding mikken op een score in de buurt van de cesuur, zal verhogen van de doorzichtigheid van de toetsing kunnen leiden tot een moeilijker te voorspellen uitslag, niettegenstaande het feit dat de toetsscore in de doorzichtiger situatie wél beter voorspelbaar is)

Voorspellen van de uitslag in een compensatorische variant

De studierichting Psychologie aan de Universiteit van Amsterdam hanteert al enige jaren voor de propedeuse een examenregeling die een tussenvorm behelst van conjunctieg en compensatorisch toetsen. Voor ieder propedeuse tentamen geldt een absolute minimumgrens van 55 % correct; resultaten daar beneden zijn onvoldoende, en kunnen in een herkansing verbéterd worden (de herkansingsregeling is door de jaren heen nogal eens veranderd). Daarnaast wordt een compensatorische minimumeis gesteld van 70 % correct **gemiddeld** over alle tentamens (die dus tenminste ook afzonderlijk 55 % goed gemaakt moeten zijn).

In dit stelsel zou je kunnen zeggen dat het voorspellen van de uitslag van een tentamen de voorspelling is of de score onvoldoende is (beneden 55 %), voldoende (tenminste 70 %), of compenseerbaar (tussen beide grenzen in).

De techniek voor deze voorspelling is dezelfde als behandeld in de voorgaande paragraaf voor het voorspellen van de uitslag gezaktgeslaagd.

Voor de eerder al behandelde voorbeelden zijn de relevante kansen voor correcte voorspelling zonder veel extra moeite af te lezen uit de gegevens op blz. 11, omdat die al berekend werden voor de twee verschillende cesuren van 55 % en 70 %. De opmerkingen gemaakt op blz. 12 zijn dan ook in dit

geval relevant.

voorbeeld

toets 30 vragen, 55% grens is 16,5, 70% grens is 20,5, c+d=62

verwachte score	23	21	19	17	15	13
% juiste uitslag	81	56	46	42	67	86

toets 30 vragen, 55% grens is 16,5, 70% grens is 20,5, c+d=32

verwachte score	23	21	19	17	15	13
% juiste uitslag	79	51	41	38	66	83

Alléén voor de verwachte scores in het compenseerbare gebied liggen de percentages anders.

Vooral voor een tentamen waar het gemiddeld toetsresultaat in het compenseerbare gebied ligt, betekent het voorspellen van deze drieslag een aanzienlijke reductie in percentage succesvolle voorspellingen.

Vergelijking met de resultaten uit Hoogstraten en Vorst (1980)

Het ligt voor de hand om de hiervóór ontwikkelde

technieken toe te passen op de tentamens waarover Hoogstraten en Vorst (1980) percentages correcte voorspellingen (voldoende/compenseerbaar/onvoldoende) gerapporteerd hebben.

Om dat zinvol te kunnen doen zijn méér gegevens over de afzonderlijke tentamens nodig dan H & V rapporteren. Op grond van aantal vragen, gemiddeld resultaat, en SD van de tentamenresultaten kunnen de berekeningen uitgevoerd worden voor verschillende waarden van de 'nauwkeurigheid' c+d. Omdat H&V niet beschikken over de voorspelling vooraf van de toetsscore, zijn de resultaten van deze berekeningen niet zonder meer tegen de empirie af te zetten. Daarom moeten de berekeningen vervolgd worden met het berekenen van succeskansen bij de uitslagvoorspelling. Daarbij zal een wat minder globale benadering voor het overall succespercentage gehanteerd moeten worden dan ik hier in de voorgaande paragrafen gebruikt heb. Vergelijking van de aldus modelmatig berekende succespercentages onder verschillende aannamen voor c+d laten zich dan vergelijken met de empirisch verkregen resultaten.

Bij de vergelijking van modelvoorspelling met empirische resultaten kan een probleem ontstaan, wanneer zou blijken dat er sprake is van systematische onder- dan wel overschatting van de uitslag. Dergelijke misspecificaties kunnen het gevolg zijn van een toevallig iets moeilijker of makkelijker uitvallen van de toets. Maar het kan ook

zijn dat er Andere verklaringen voor over- of onderschattingente geven zijn. Het kan wel eens moeilijk blijken om deze twee verschillende situaties van elkaar te onderscheiden. (Toevallige fluctuaties in de moeilijkheid van de toets zullen zich uiten in een patroon van enkele over en onderschattingen dat varieert over tentamens, of over afnamen van hetzelfde tentamen over verschillende jaren; verklaringen uit de tweede categorie voorspellen een systematische afwijking die over meerdere tentamens heen ó6k stelselmatig zal zijn).

Het toevallig wat moeilijker of makkelijker uitvallen van een tentamen mag dan de groepsresultaten beïnvloeden, voor de vraag naar de voorspelbaarheid van studieresultaten geeft het geen problemen. De individuele voorspelling houdt immers met dat soort fluctuatie wel degelijk rekening; omdat in een situatie waarin niet iedere student een eigen random getrokken toets voorgelegd krijgt de toevallig makkelijkheid of moeilijkheid de meeste studenten treft, ontstaat er een verschuiving in **groeps**resultaat van de correctheid van de voorspellingen. Voor andere stelselmatige onder of overschattingen kan gecorrigeerd worden, en kan de student eventueel ook zichzelf leren corrigeren; hier is de voorspelbaarheid in beginsel al evenmin aangetast.

Tenslotte: het patroon van succespercentages uitslagvoorspellingen van H & V over de verschillende tentamens moet verklaarbaar zijn uit

de overwegingen die ik in de voorgaande paragraaf genoemd heb. Dat moet aan de hand van beschikbare gegevens uit te zoeken zijn. Voortaan niet meer vragen of de student denkt een voldoende te behalen of niet, maar wat hij denkt dat zijn slaag**kans** is; dat is een veel rijker gegeven, en dat laat sterkere gevolgtrekkingen op basis van de data toe.

==========

[NB Ik weet niet of ik deze brief ook heb verzonden]

Universiteit van Amsterdam
Centrum voor Onderzoek van het Wetenschappelijk Onderwijs
Oude Turfmarkt 149 /Telefoon 525 2835
10 12 GC Amsterdam
Ben Wilbrink

Johan Hoogstraten

Harry Vorst

AMSTERDAM, 17-10-80

Beste Johan en Harry,

hierbij een snel in elkaar getimmerde uiteenzetting over de mogelijkheden die er zijn om de voorspelbaarheid van studieresultaten uit te zoeken op basis van een geschikt model,voor dat soort voorspelling.

Het gegeven model moet in staat zijn de resultaten die jullie gerapporéeerd hebben, te voorspellen. Of~dat inderdaad zo is, kan ik op dit moment niet zeggen, omdat ik de daarvoor benodigde data niet tot mijn beschikking heb.

Ik heb begrepen dat de beschikbare voorspellingsgegevens (uitslagvoorspelling vóóraf) niet zo erg jofel zijn. Hopelijk zijn ze bruikbaar genoeg om een voorlopige indruk te krijgen hoe mijn tentamenmodel-achtige analyse zich verhoudt met de empirie.

Hier en daar is het stuk wel erg in telegramstijl geschreven, en daarom misschien niet al te makkelijk leesbaar. Dan moet ik een en ander maar eens mondeling komen toelichten.

Bekijken jullie of ik voldoende aanknopingspunten geef om nog eens een keer opnieuw tegen de beschikbare data aan te gaan (of om een secundaite analyse aan mij over te laten; daarvoor heb ik geloof ik alleen maar nodig per tentamen het.aantal vragen,

gemiddelde en standaard deviatèel en de voorspelling vooraf van de uitslag met de daarbij behorende succespercentages per categorie)

Met vriendelijke groet,