TEST UTILITY:
HET NUT VAN TESTS

Denny Borsboom en Klaas Visser

TEST UTILITY

- Is het de moeite waard om te testen?
- Kosten
  - Economisch
  - Niet-economisch
- Baten
  - Economisch
  - Niet-economisch
- De verhouding tussen kosten en baten wordt voor een groot deel bepaald door de *psychometrische eigenschappen van de gebruikte procedure*
- Deze eigenschappen zijn empirisch na te gaan en je kunt er verstandig over redeneren
- Dit college legt uit hoe dat werkt

VOORSPELLEND TESTGEBRUIK

- Vaak wil een gebruiker op basis van testscores *voorspellen* hoe iemand het gaat doen in de toekomst
- Datgene wat je wil voorspellen heet het *criterium*
- Bijvoorbeeld: selectie van succesvolle sollicitanten of goede studenten
- Bij voorspellen is belangrijk hoe sterk de *correlatie* tussen testscore en criterium is
- Dit heet de *predictieve validiteit* of *criteriumvaliditeit*

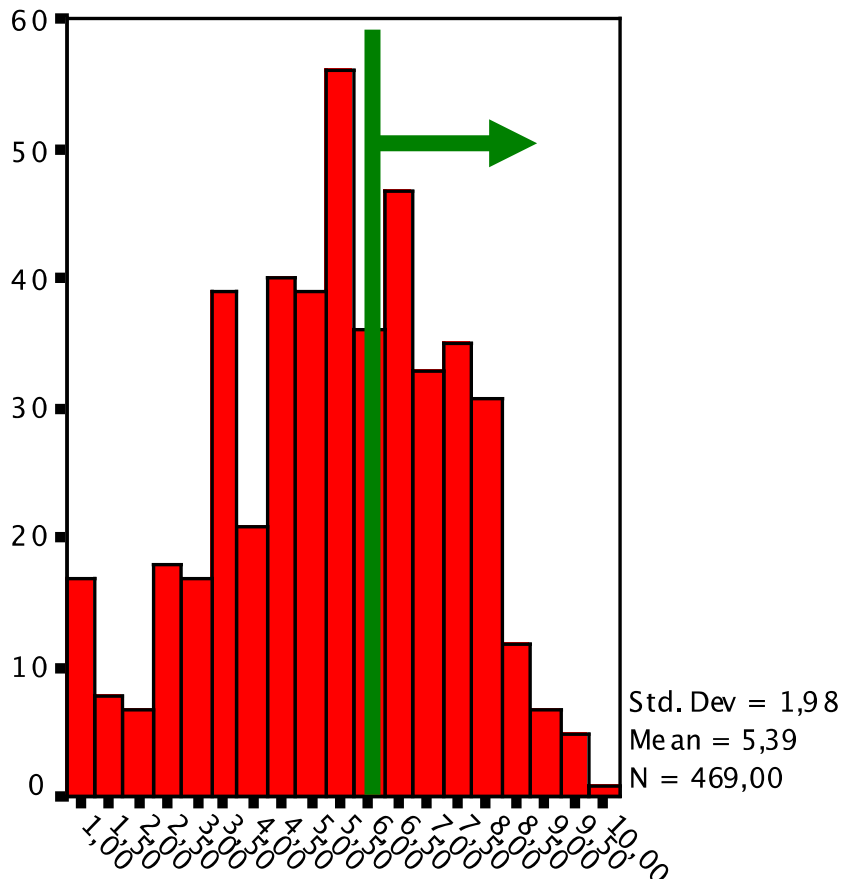
CASE STUDY: SELECTIE AAN DE POORT

- Testweekdata van de UvA:
  - Scores op een motivatietest
  - Studieresultaten
- De data zijn van enkele jaren geleden, voordat er geselecteerd werd: iedereen is dus aangenomen
- Daardoor kunnen we de efficientie van een eventuele selectieprocedure nagaan
- We kunnen namelijk uitrekenen hoeveel beter de resultaten zouden zijn, als we zouden selecteren
- Deze vergelijking geeft aan welke winst we kunnen behalen, en welke prijs daar tegenover staat

CRITERIUM: STUDIERESULTATEN

- Studiesucces wordt geoperationaliseerd als het gemiddelde cijfer op de eerste drie tentamens
- We gaan er even van uit dat de Universiteit alleen mensen wil hebben die op dit gemiddelde zes of hoger scoren
- Dit is dus voor nu onze definitie van het criterium
- Merk echter op dat een andere definitie van het criterium een andere waarde van de criteriumvaliditeit geeft en dus tot een ander oordeel over selectie kan leiden
- De definitie van het criterium is dus zeer belangrijk en het is aan te bevelen die van tevoren vast te leggen, want achteraf kun je altijd wel een criterium bedenken dat “werkt”

CRITERIUM: STUDIERESULTATEN

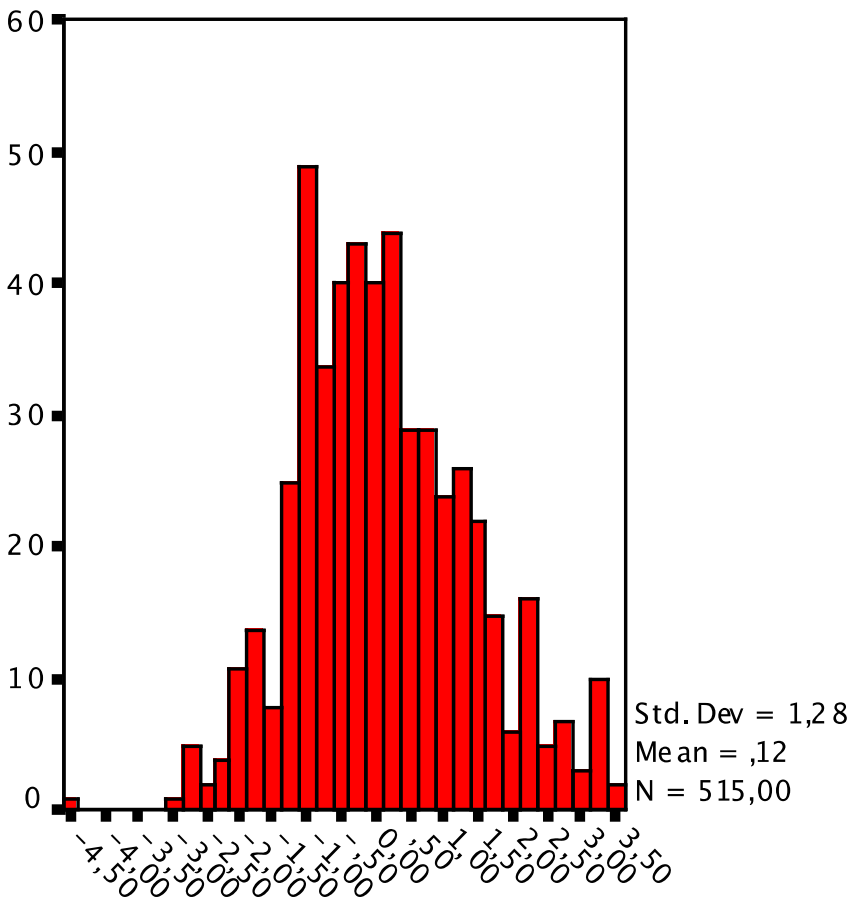


Gemiddeld tentamencijfer

PREDICTOR: MOTIVATIE

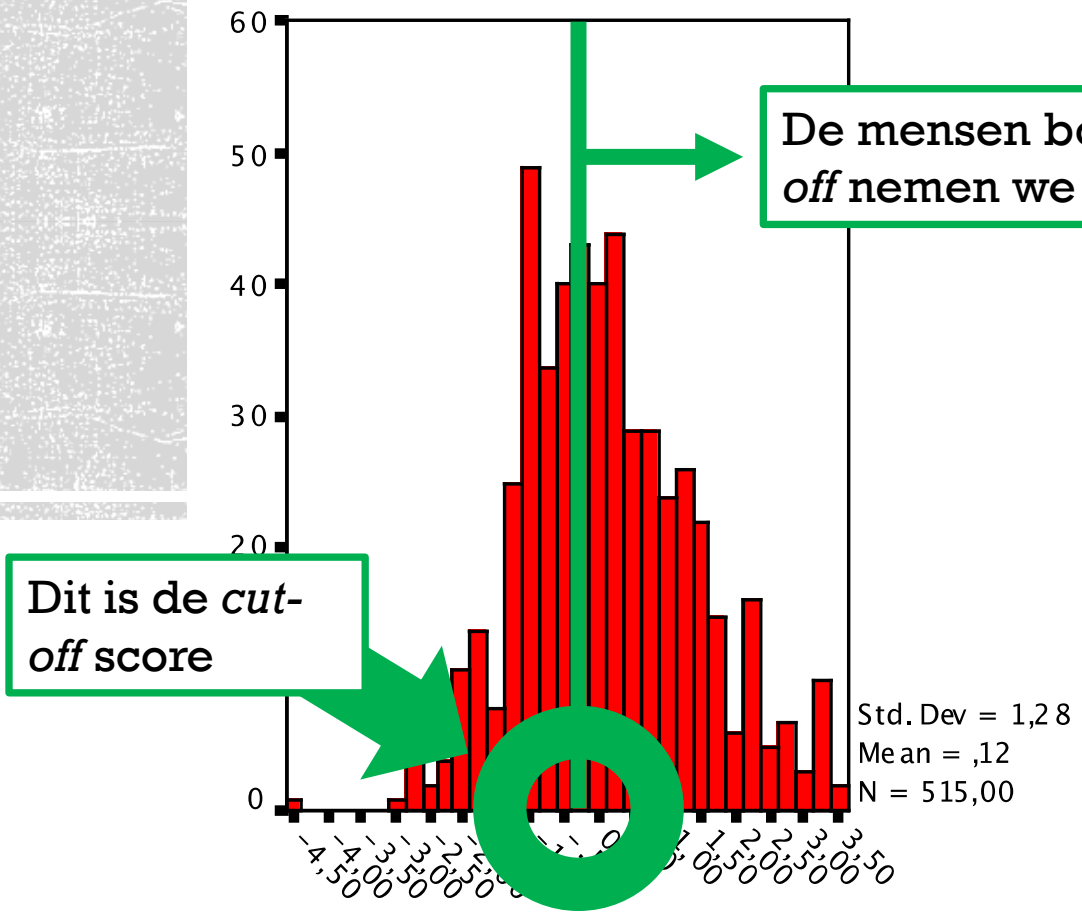
- Geoperationaliseerd als de score op een testje dat bestaat uit items als:
  - ‘Ik begin meestal uit mezelf te studeren’
  - ‘Ik heb een hekel aan hard werken voor mijn studie’
  - ‘Ik studeer zo goed mogelijk’
- Overigens zie hieraan al meteen een groot probleem met selectie op niet-cognitieve factoren: ze zijn makkelijk te *faken*
- Dit wordt heel gemakkelijk vergeten door voorstanders van selectie, maar selectie op criteria die te *faken* zijn is niet alleen inefficiënt maar holt ook het gezag van een instelling uit

Motivatie



Z-score op motivatietest

De cut-off score



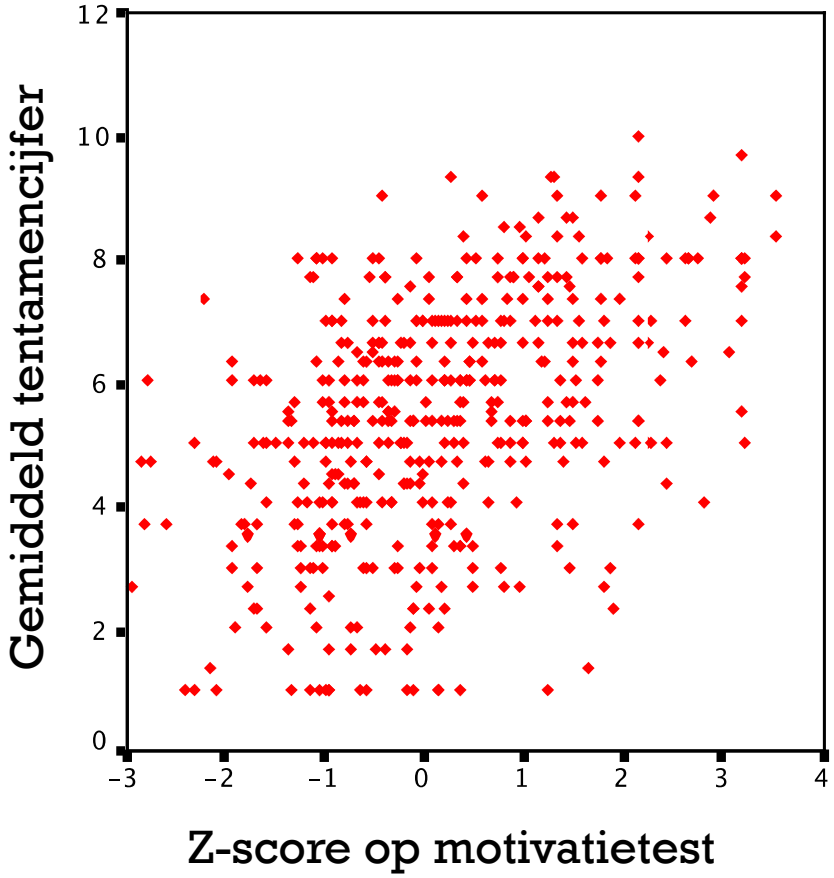
De mensen boven de *cut-off* nemen we aan

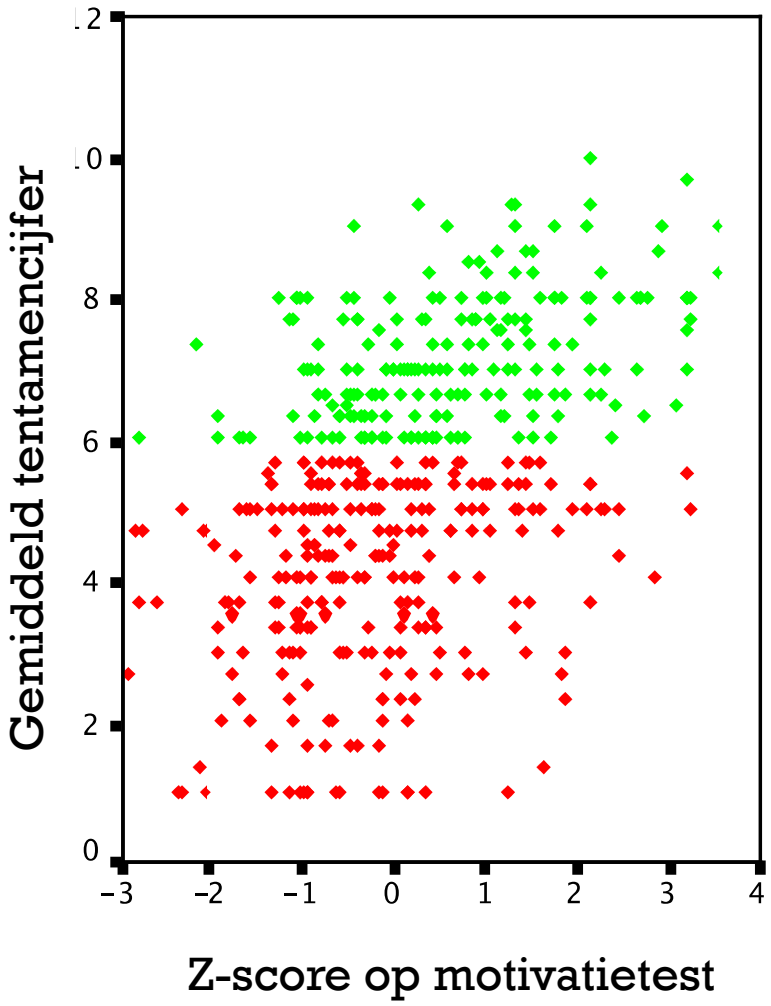
Dit is de *cut-off* score

SELECTIE

- Selectie op motivatiescores heeft alleen zin, als succes en motivatie *samenhangen*
- Er moet dus een *correlatie* zijn tussen deze variabelen: de predictor variabele moet over criteriumvaliditeit beschikken
- Als dat niet zo is:
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten
  - behaal je geen winst en kun je net zo goed loten

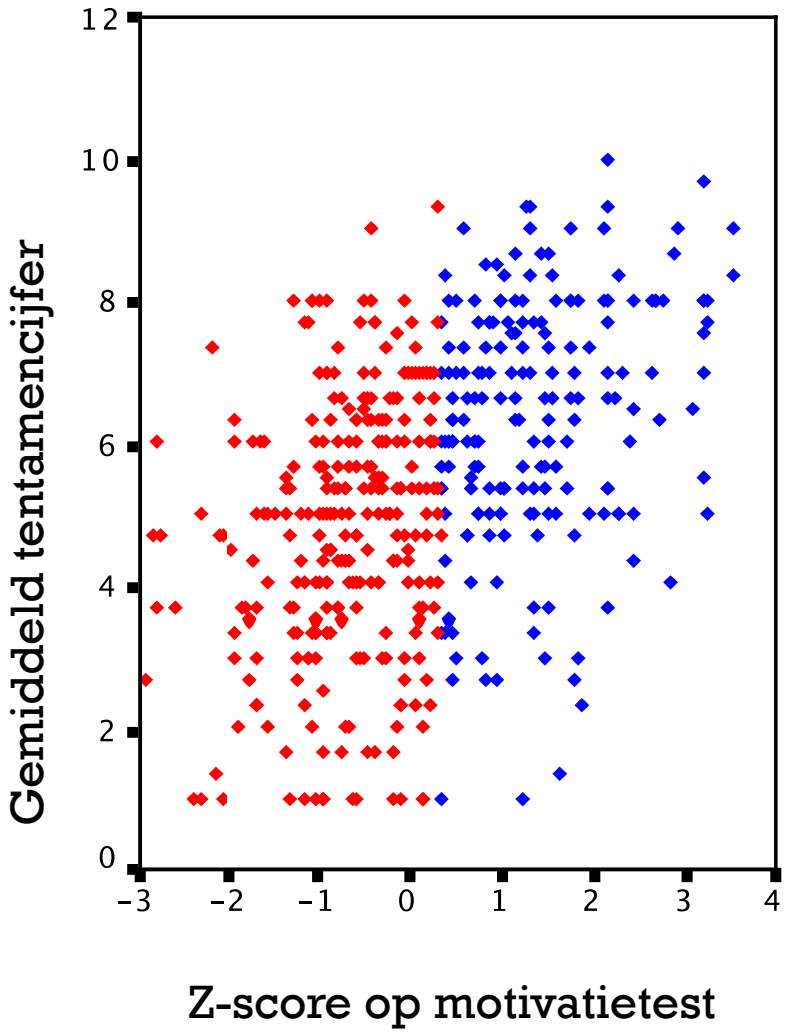
Met een
correlatie van
nul,
is selectie
flauwekul


Gemiddeld tentamencijfer
12
10
8
6
4
2
0
−3
−2
−1
0
1
2
3
4
Z-score op motivatietest



Voldoendes

Onvoldoendes

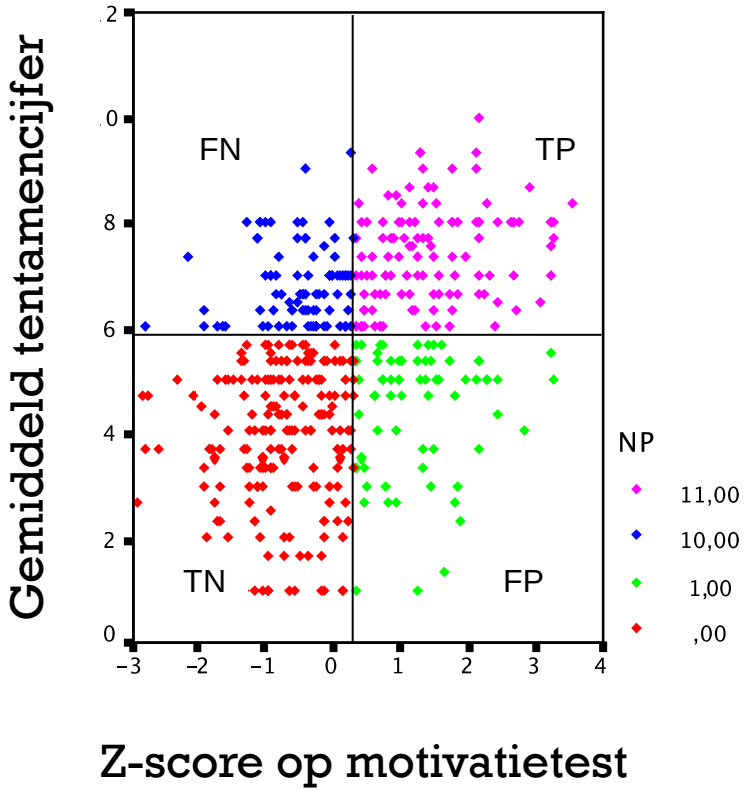


Laag gemotiveerd

Hoog gemotiveerd

POSITIVES & NEGATIVES

- True positive: aangenomen en geschikt
- True negative: niet aangenomen en niet geschikt
- False positive: aangenomen en niet geschikt
- False negative: niet aangenomen en geschikt



TP=true positive

TN=true negative

FP=false positive

FN=false negative

DEZELFDE SITUATIE WEERGEGEVEN IN EEN KRUISTABEL

		Succes		
		nee	ja	
Aangenomen	nee	223	85	308
	ja	86	121	207
		309	206	515

HOE BEPALEN WE DE EFFICIENTIE VAN SELECTIE?

- Hit rate: percentage juiste beslissingen:

$$\text{Hit rate} = \frac{\text{aantal true positives + aantal true negatives}}{\text{totaal aantal beslissingen}}$$

DE *HIT RATE*

		Succes: nee	Succes: ja	
Aangenomen	nee	**223**	**85**	308
	ja	**86**	**121**	207
		309	206	**515**

In totaal zouden we 223+121 juiste beslissingen genomen hebben met de voorgestelde selectieprocedure. De *hit rate* is het *percentage juiste beslissingen.* Hier is dat: (223+121)/515=67%

DE *HIT RATE* AFGEZET TEGEN DE *BASE RATE*

		Succes: nee	Succes: ja	
Aangenomen	nee	223	85	308
	ja	86	121	207
		309	**206**	**515**

De *base rate* is de *hit rate* die je hebt als je *niet selecteert.* Hier is dat: 206/515=40%. Dus 40% van de studenten deed het destijds sowieso al goed. Onze winst is dus 67% - 40% = 27% meer correcte beslissingen als we selecteren (want de hit rate van onze procedure = 67%).

HOE BEPALEN WE DE EFFICIENTIE VAN SELECTIE?

- Sensitiviteit: welk percentage van de *geschikte mensen* wordt ook daadwerkelijk *aangenomen*:

$$\text{Sensitiviteit} = \frac{\text{aantal true positives}}{\text{aantal geschikte kandidaten}}$$

DE *SENSITIVITEIT*

		Succes		
		nee	ja	
Aangenomen	nee	223	**85**	308
	ja	86	**121**	207
		309	**206**	515

In totaal zouden we 121 van de 206 gekwalificeerde mensen hebben aangenomen. De *sensitiviteit* is dan 121/206=0.59. Voor geschikte mensen is de procedure dus net iets beter dan een muntje gooien. Een brandmelder met zo'n lage sensitiviteit zou niemand ooit kopen en het feit dat 41% van de geschikte mensen afgewezen wordt is vanuit moreel oogpunt dubieus.

Hoe meer de sensitiviteit gaat dalen, hoe meer de studenten en hun ouders zullen balen

HOE BEPALEN WE DE EFFICIENTIE VAN SELECTIE?

- Specificiteit: welk percentage van de *ongeschikte mensen* wordt ook daadwerkelijk *afgewezen:*

$$\text{Specificiteit} = \frac{\text{aantal true negatives}}{\text{aantal ongeschikte kandidaten}}$$

DE *SPECIFICITEIT*

		Succes: nee	Succes: ja	
Aangenomen	nee	**223**	85	308
	ja	**86**	121	207
		309	206	515

In totaal zouden we 223 van de 309 niet gekwalificeerde mensen hebben afgewezen. De *specificiteit* is dus 223/309=0.72. Dit is iets beter dan de sensitiviteit, maar nog steeds aan de lage kant. De specificiteit is meestal iets minder moreel geladen, omdat onterecht aannemen minder erg is dan onterecht afwijzen.

DE WINST

- We zouden met een selectieprocedure een bescheiden winst kunnen behalen
- De vraag is of deze winst opweegt tegen de prijs: naast de kosten van een selectieprocedure het feit dat we 41% van de geschikte mensen onterecht zouden afwijzen
- Dit is een ingewikkelde vraag die niet met statistiek te beantwoorden is: het gaat om wat je belangrijk vindt

DUS: HOE KUN JE EFFICIENTIE BEPALEN?

1. Kies een predictor en een criterium
2. Bepaal de correlatie tussen predictor en criterium in een (ongeselecteerde!) populatie
3. Bekijk je efficientiewinst voor verschillende cutoff scores die je zou kunnen kiezen

Als je *niet* iedereen aanneemt kun je *niet* vaststellen hoe goed je procedure werkt

UTILITY ANALYSIS

- Wat als je geen geld of tijd hebt om empirisch onderzoek te doen?
- Dan kun Taylor-Russell tabellen raadplegen
- Deze tabellen geven de toename in het percentage goed functionerende mensen bij gebruik van een tests voor verschillende waarden van
  - de base-rate
  - de predictieve validiteit
  - de selectie-ratio

% dat bij huidige procedure succesvol is na aanstelling

percentage dat aangenomen wordt

predictieve validiteit

Table 7–1
Taylor-Russell Table for a Base Rate of .60

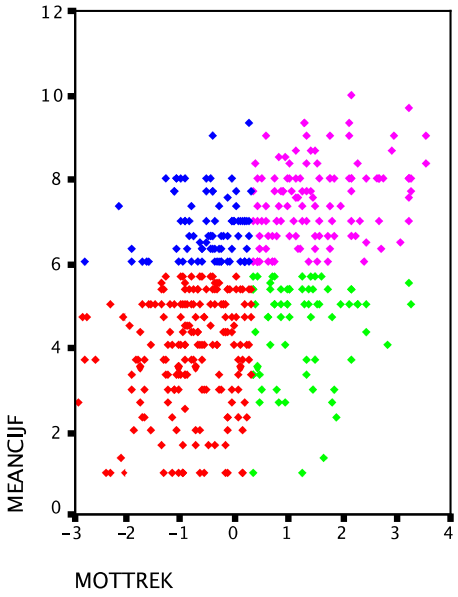
Validity (ρ_{xy})	Selection Ratio										
	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	.68	.67	.65	.64	.64	.63	.63	.62	.61	.61	.60
.15	.71	.70	.68	.67	.66	.65	.64	.63	.62	.61	.61
.20	.75	.73	.71	.69	.67	.66	.65	.64	.63	.62	.61
.25	.78	.76	.73	.71	.69	.68	.66	.65	.63	.62	.61
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
.35	.85	.82	.78	.75	.73	.71	.69	.67	.65	.63	.62
.40	.88	.85	.81	.78	.75	.73	.70	.68	.66	.63	.62
.45	.90	.87	.83	.80	.77	.74	.72	.69	.66	.64	.62
.50	.93	.90	.86	.82	.79	.76	.73	.70	.67	.64	.62
.55	.95	.92	.88	.84	.81	.78	.75	.71	.68	.64	.62
.60	.96	.94	.90	.87	.83	.80	.76	.73	.69	.65	.63
.65	.98	.96	.92	.89	.85	.82	.78	.74	.70	.65	.63
.70	.99	.97	.94	.91	.87	.84	.80	.75	.71	.66	.63
.75	.99	.99	.96	.93	.90	.86	.81	.77	.71	.66	.63
.80	1.00	.99	.98	.95	.92	.88	.83	.78	.72	.66	.63
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

Source: Taylor and Russell (1939).

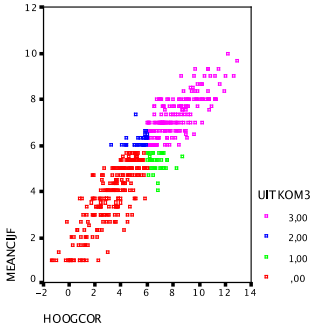
FACTOREN DIE DE *HIT RATE* BEINVLOEDEN

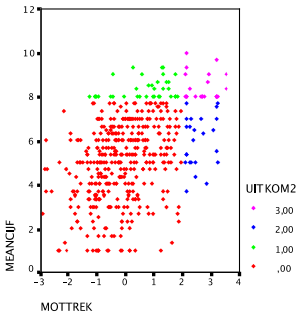
- Predictieve validiteit
  - Een hogere correlatie tussen predictor en criterium geeft een hogere *hit rate*
- De selectie-ratio (het percentage van de kandidaten dat je van plan bent aan te nemen)
  - Een scherpere selectie (minder mensen aannemen) geeft een hogere *hit rate*

TWEE MANIEREN OM EFFICIENTIE VAN SELECTIE TE VERBETEREN



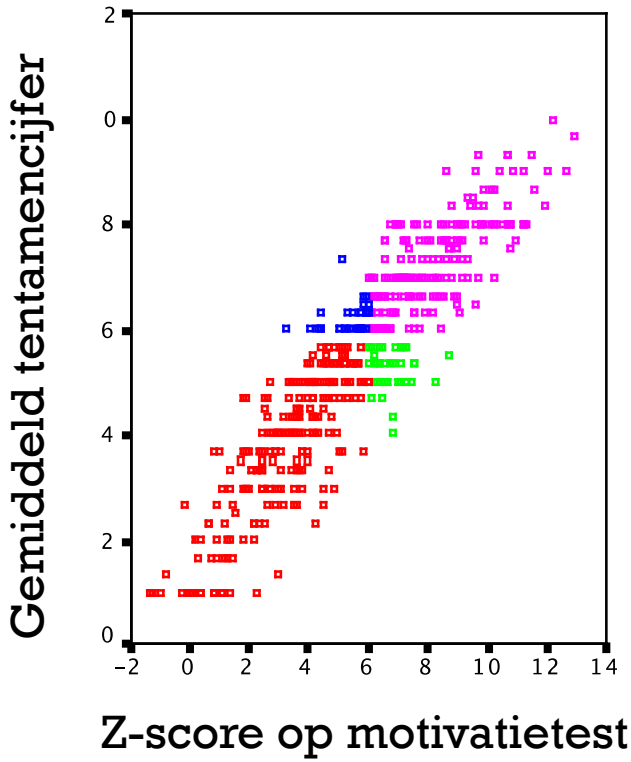
Verhoog de Predictieve validiteit





Kies een extremere selectieratio

I. HOGERE CORRELATIE (STERKERE PREDICTIEVE VALIDITEIT)



Bij de afgebeelde versterking van de preditieve valiteit is onze nieuwe *hit rate* = 87%, dus ruim een verdubbeling t.o.v. onze base rate van 40%

BEDREIGINGEN VAN DE PREDICTIEVE VALIDITEIT

- De criteriumvaliditeit kan worden bedreigd door alle factoren die de samenhang met het criterium verzwakken:
- Bijvoorbeeld:
  - Onbetrouwbaarheid: meer ruis betekent een zwakkere samenhang
  - Verkeerde keuze voorspeller: samenhang te laag
  - ‘Restriction of range’: beperking van de variantie in de testscore leidt tot zwakkere correlatie

VERSTERKING VAN DE PREDICTIEVE VALIDITEIT

- Predictieve validiteit kan worden versterkt door alle ingrepen die de samenhang versterken
- Bijvoorbeeld:
  - Hogere betrouwbaarheid: meer items afnemen
  - Betere voorspeller kiezen: hogere samenhang
  - Zorgen dat er voldoende variantie in de voorspeller is
- Soms lukt dit gewoon niet, ook niet als je heel hard denkt van wel

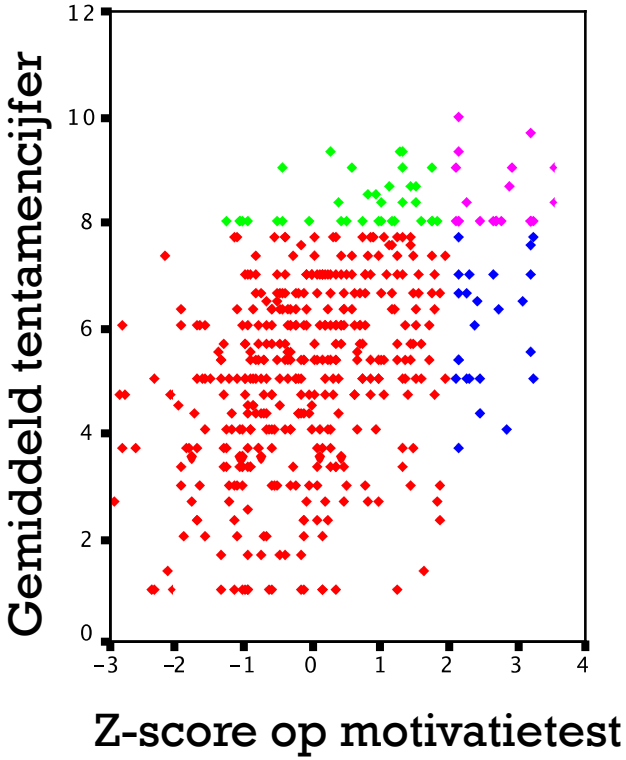
Heel hard willen dat

selectie werkt heeft

geen invloed op de

kwaliteit van selectie

II. ZWAARDERE SELECTIE (LAGERE SELECTIERATIO)



Bij de afgebeelde verzwaring van de selectie is onze nieuwe *hit rate* = 88%, dus ruim een verdubbeling t.o.v. onze base rate van 40%

FACTOREN DIE SELECTIE IN NEDERLAND BEMOEILIJKEN

- Correlatie tussen mogelijke predictoren en succes zijn niet spectaculair hoog (vaak gewoon laag)

- Universiteiten hebben een zeer ongunstige selectieratio (ze moeten veel mensen aannemen)

WAAROM SELECTIE IN DE VS ZO GOED WERKT

- Correlatie tussen mogelijke predictoren en succes is hoger
  - dit komt doordat er meer variantie is; in Nederland is er al selectie geweest op de middelbare school (VWO/HAVO/VMBO) en in de VS doet iedereen *high school*
  - In Nederland is er dus *restriction of range*
- Goede universiteiten hebben een zeer gunstige selectieratio
  - ze nemen heel weinig mensen aan; Harvard selecteert bv. *binnen* de 10% beste SAT scores)

DUS…

- Testgebruik voor selectie vereist een behoorlijke mate van samenhang tussen predictor en criterium
- Hoe hoger de samenhang, hoe beter de voorspelling, hoe groter de winst
- De efficiëntie van selectie kan *vooraf* bepaald worden in een studie waarin je *niet* selecteert; *als de selectie eenmaal is ingevoerd kun je niet meer bepalen hoe efficient je methode is*
- Als de criteriumvaliditeit van predictoren structureel beperkt wordt door bv. restriction of range dan kun je het efficiëntie-argument eigenlijk niet meer aanvoeren
- Dan moet je beleid dus rusten op andere, niet-psychometrische argumenten

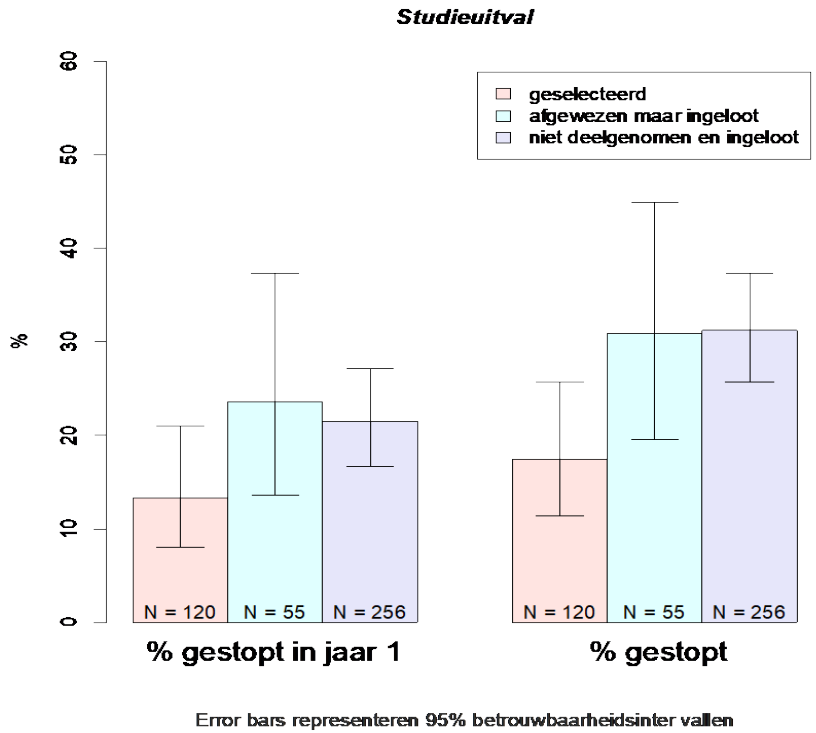
COHORT 2006: 4 JAAR GEVOLGD

3 Groepen

431 studenten	geselecteerd	niet geselecteerd	Niet deelgenomen aan selectie	totaal
voltijdinschrijvingen	120	55	256	431

RESULTATEN: UITVAL

Na 1 jaar (niet sign) en na 4 jaar (wel sign)



SAMENVATTING RESULTATEN

- Geen verschil tussen deelnemers en niet-deelnemers
- Wel verschil tussen geselecteerden en niet-geselecteerden
- Gecontroleerd voor andere variabelen als geslacht, vooropleiding en VWO cijfers

- Onder de geselecteerden:
  - Veel minder uitval (18% vs 30%)
  - Iets hogere cijfers
  - Minder studievertraging
  - Vaker een bachelordiploma binnen 4 jaar

Visser, et al: 2012

CONCLUSIES

- Selectietoets op basis van een curriculum sample werkt
- Het werkt beter dan loten
- Ook bewijs geleverd bij Psychologie Groningen, Geneeskunde Nijmegen, Psychobiologie UvA

- Maar:
- Er blijven foute voorspellingen, maar minder dan bij loten
- En misschien sluipt er toch bias in
- Mensen gaan zich voorbereiden, zeker bij geneeskunde als je drie keer aan een selectie mag meedoen.

Niessen, 2017, de Visser 2017

WAAROM WERKT HET

- 4 factoren
  - Inzet is beslissende factor in studiesucces (inzetbereidheid)
  - De toets meet combinatie van motivatie en intelligentie
  - Er vindt enige zelfselectie plaats (ook gevaarlijk!)
  - Het biedt veel extra voorlichting en informatie

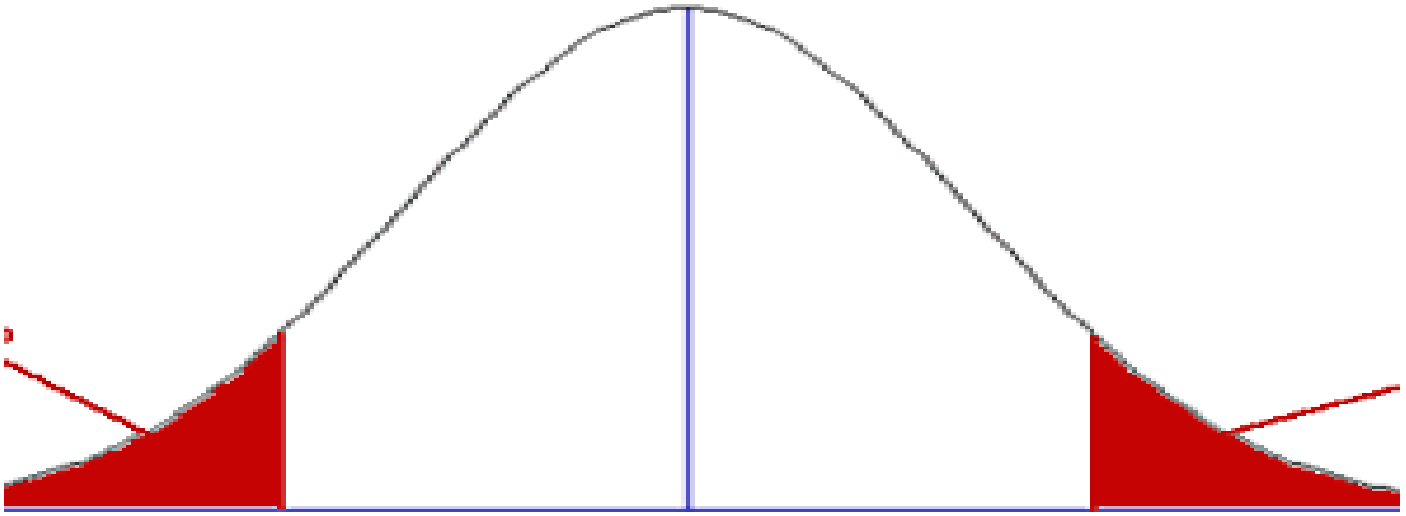
  Maar: studenten doen veel om te worden aangenomen

  Schripsema (2017), Niessen (2017): als je moet selecteren zijn VWO resultaten en een curriculum sample de beste voorspellers

Niessen, 2017, Schripsema, 2017; de Visser, 2016

UITDAGINGEN VOOR DE TOEKOMST

- Welke studenten gaan **uitvallen**, zijn **ongeschikt** en hoe kunnen we dat goed voorspellen (Tandheelkunde, Geneeskunde, etc.).
- Ontwikkeling nieuw instrumentarium
- Hoe bevorder je diversiteit zowel in afkomst als in geslacht?



CONCLUSIE

- Selecteren in de Nederlandse situatie: nauwelijks te doen
- Geen voorspellers met hoge predictieve validiteit
- Bij lotingstudies of selectieve opleidingen:
  - We kunnen een topgroep selecteren en als je heel veel aanmelders hebt en weinig plaatsen lukt dat gemakkelijker (res. mas)
    - Maar je wijst heel veel geschikte mensen af
  - Hoe groter het probleem (bijv. uitval) is, hoe beter het kan werken
  - Eindexamencijfers zijn de beste voorspellers
  - Curriculum sample methode voorspelt even goed en is deels onafhankelijk van die eindexamencijfers

TOT SLOT

- Veel interessant werk voor selectiepsychologen
  - Hoe kun je beter voorspellen wie de ***ongeschikte*** studenten zijn
  - Wat werkt als je niet alleen op VWO cijfers mag selecteren
  - Hoe kunnen de selectiekosten laag worden gehouden
  - Hoe hou je de bias zo klein mogelijk