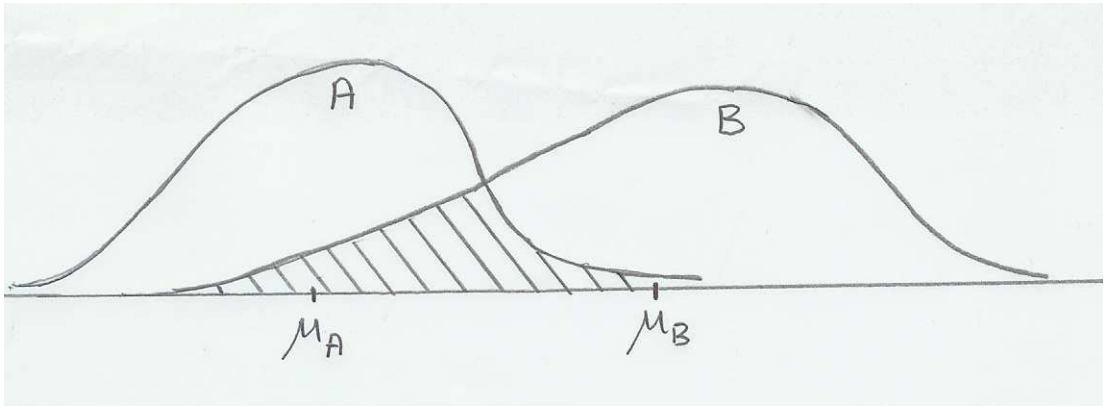
HOOFSTUK 8
Effekgroottes en Groepsoorvleueling

8.1 Inleiding

Wanneer twee populasies vergelyk wil word ten opsigte van ‘n kontinue responsveranderlike (soos bv. IK, diastoliese bloeddruk, toetspunte van persone), is die gestandaardiseerde verskil δ ‘n effekgrootte-indeks (kyk Hoofstuk 4). So kan ρ_{pb}^2 as die proporsie variansie toe te skryf aan populasielidmaatskap (kyk Hoofstuk 5), ook as indeks gebruik word. In die geval waar meer as twee populasies vergelyk word, kan δ vir kontraste of η^2 as veralgemening van ρ_{pb}^2 gebruik word as effekgrootte-indekse (kyk Hoofstuk 6). Al hierdie indekse is gebaseer op die aanname van homogeniteit van variansies van die populasies en met die uitsondering van die indekse $\Delta_1, \overline{\Delta_2}, \overline{\Delta_m}, \delta_g$ en δ_c (kyk paragraaf 4.3) is daar geen ander effekgrootte-indekse beskikbaar as die variansies van populasies heterogeen is nie. Dieselfde probleem ontstaan ook wanneer na meerveranderlike populasies gekyk word: al die indekse (soos in Hoofstuk 7 bespreek) aanvaar dat die populasies dieselfde kovariansie-matrikse besit.

‘n Moontlike oplossing sou wees om ‘n indeks te verkry wat op populasie-oorvleueling gebaseer is. Volgens figuur 8.1 is die ingekleurde gedeelte die oorvleueling tussen populasie-verdelings A en B. Let op dat die populasies nie noodwendig normaal met gelyke variansies hoef te wees nie. Dis duidelik dat die oorvleueling omgekeerd eweredig aan die verskil in lokaliteit van die twee verdelings is, soos bv. weergegee deur $\mu_B - \mu_A$. Dit beteken ook dat as die *nie-oorvleueling* van die populasieverdelings groot is, verskil μ_A en μ_B baie en as dit min is, verskil die populasiesgemiddeldes min.

Figuur 8.1: Oorvleueling van twee populasieverdelings



In hierdie hoofstuk bespreek ons hoe die oorvleueling tussen populasies omgesit kan word in ‘n effekgrootte-indeks. Daar sal eers na die geval van homogene variansies gekyk word in die twee en meer-populasie gevalle met een en meer veranderlikes. Omdat daar reeds effekgrootte-indekse in hierdie gevalle bestaan, sal probeer word om verbande tussen die nuwe indeks en die bestaandes te bepaal. Daarna sal die indeks in die gevalle van ongelyke variansies of kovariansie-matrikse ingevoer word en bespreek word. Vervolgens word eers gekyk na die klassifikasie van waarnemings en wat ‘n trefkoers is.

8.2 **Afstand en klassifikasie**

Gestel ‘n p-veranderlike populasie g het ‘n gemiddelde vektor of *sentroïed* van $\boldsymbol{\mu}_g = \left(\mu_{1g}, \mu_{2g}, ..., \mu_{pg}\right)$ en kovariansiematriks $\boldsymbol{\Sigma}_g$, dan kan die sogenaamde *Mahalanobis-afstand* van ‘n vektor van waarnemings $\boldsymbol{x}_u = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_s\right)$ van ‘n objek u (bv. persoon) vanaf die sentroïed van g, geskryf word as

$$\Delta_{ug} = \left[\left(\boldsymbol{x}_u - \boldsymbol{\mu}_g\right)\boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{x}_u - \boldsymbol{\mu}_g\right)'\right]^{1/2} \qquad (8.1)$$

In die eenveranderlike geval met p = 1, herlei dit tot

$$\Delta_{ug} = \frac{x_u - \mu_g}{\sigma_g} \quad ,$$

waar μ_g en σ_g die gemiddelde en SA van populasie g is.

Om nou $\boldsymbol{x_u}$ te klassifiseer as behorende by een van k populasies, word van voorspellende diskriminantontleding ("predictive discriminant analysis", PDA) gebruik gemaak. Volgens Huberty (1994:45) is die doel van PDA as volg:

Gestel dat ons ewekansige steekproewe trek uit k populasies van groottes $n_g, g = 1,...,k,$ wat bestaan uit metings op elk van $N\left(=\Sigma n_g\right)$ objekte. Deur van hierdie N x p data-matriks gebruik te maak, wil ons bepaal watter een van die k populasies is die mees moontlike een waaruit die (N + 1)-de objek getrek kon word.

Om hierdie populasie te bepaal waaruit 'n toekomstige objek, met waarneming $\boldsymbol{x_u}$, kom, word aanvaar dat die populasies elk meerveranderlik normaal verdeel is en dan word die maksimum-aanneemlikheidsmetode gebruik. In Huberty (1994: hoofstuk IV) word die agtergrond en metode in diepte bespreek. Vir die doel van ons bespreking is dit genoeg om met die volgende te volstaan.

8.2.1 A priori waarskynlikhede

Laat π_g die proporsie objekte wees in die k populasies gesamentlik, wat uit populasie g kom. As 'n objek dus ewekansig gekies word uit die populasies gesamentlik, dan gee π_g die waarskynlikheid dat dit uit populasie g kom. Hierdie waarskynlikheid word "a priori" genoem omdat dit vooraf bekend is – voordat enige steekproewe getrek word.

Indien die k populasies se groottes nie bekend is nie, kan π_g op twee wyses verkry word:

(a) Kies dit volgens goeie oordeel en vooraf kennis: die navorser weet bv . uit ervaring dat objekte uit populasie 1 twee keer soveel voorkom as uit populasie 2. In so 'n geval kies hy/sy $\pi_1 = \frac{2}{3}$ en $\pi_2 = \frac{1}{3}$.

(b) Maak die aanname dat die steekproewe se groottes proporsioneel is aan dié van die populasies, dan kan $\pi_g = p_g = n_g / N$ geneem word.

(c) Kies $\pi_g = \frac{1}{k}$, gelyk vir al k populasies.

8.2.2 Gelyke populasie-kovariansiematrikse

As aanvaar word dat $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, word die afstand in (8.1) Δ^*_{ug}, waar $\boldsymbol{\Sigma}_g$ met $\boldsymbol{\Sigma}$ vervang word. Hierdie afstand word dan beraam deur:

$$D^*_{ug} = \left[\left(\boldsymbol{x_u} - \bar{\boldsymbol{x}}_{\boldsymbol{g}} \right) \boldsymbol{S}^{-1} \left(\boldsymbol{x_u} - \bar{\boldsymbol{x}}_{\boldsymbol{g}} \right)' \right]^{\frac{1}{2}} \qquad (8.3)$$

waar $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ en $\mathbf{S}$ die steekproef-sentroïed en saamgevoegde kovariansie matriks van al die steekproewe is. Deur die maksimum-aanneemlikheidsmetode te gebruik, lewer dit die volgende klassifikasie-reël (Huberty, 1994: 61-62):

Deel objek u by populasie g in as

$$D^{*2}_{ug} - 2\ell n(\pi_g) < D^{*2}_{ug'} - 2\ell n(\pi_{g'}), \qquad (8.4)$$

vir alle $g \neq g'$.

Hierdie heet die *lineêre klassifikasiereël* .

8.2.3 Ongelyke populasie-kovariansiematrikse

Hier word Δ_{ug} beraam met:

$$D_{ug} = \left[\left(\boldsymbol{x_u} - \boldsymbol{x_g} \right) \boldsymbol{S_g^{-1}} \left(\boldsymbol{x_u} - \boldsymbol{\bar{x}_g} \right)' \right]^{1/2}, \tag{8.5}$$

waar $\boldsymbol{S_g}$ die steekproef-kovariansiematriks van populasie g is. Die maksimum aanneemlikheidsmetode lewer in hierdie geval die *kwadratiese klassifikasiereël*:

Deel objek u by populasie g in as:

$$\ell n\left|S_g\right| + D_{ug}^2 - 2\ell n(\pi_g) < \ell n\left|S_{g'}\right| + D_{g'}^2 - 2\ell n(\pi_{g'}), \tag{8.6}$$

vir alle $g \neq g'$.

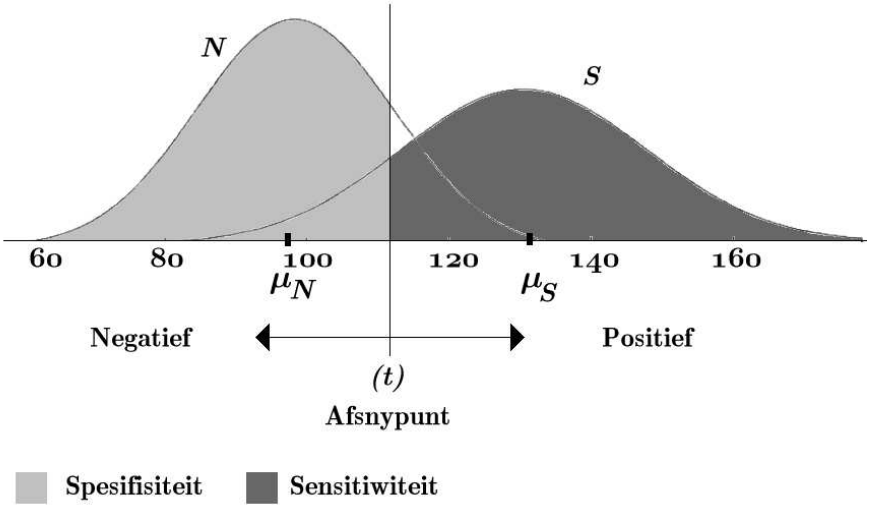
8.2.4 Twee eenveranderlike populasies: klassifikasie met ROC-analise

Om klassifikasie van objekte in hierdie geval te doen, kan van die metodes by die sg. *ontvanger-bewerkingskarakteristieke* ("receiver-operating characteristic") kromme, in kort die ROC-kromme gebruik gemaak word.

Gestel 'n sekere siektetoestand of abnormaliteit word bestudeer en persone word gekategoriseer as positief as hulle die siekte of abnormaliteit het en negatief andersins. Ander voorbeelde is ‘n kliniese sielkundige wat persone kan indeel as depressief of normaal, ‘n bank wat ‘n kliënt as ‘n risiko-geval of nie klassifiseer by die oorweging van ‘n aansoek om ‘n lening. Hierdie indelings word dalk volgens bestaande "goue standaard" diagnostiese toetse gedoen wat relatief duur en tydsaam mag wees. As daar 'n metode, ook ‘n siftingstoets genoem, wat makliker en goedkoper is om die siekte/abnormaliteit/risiko te identifiseer, sou dit belangrik wees om te weet hoe betroubaar dit is as voorspeller vir die siekte/abnormaliteit/risiko. In verdere bespreking sal ons na die populasies van siekes (S) en nie-siekes (N) verwys om persone met of sonder die siekte, abnormaliteit of risiko aan te dui. Tipies is die metings wat hierdie siftingstoets oplewer kontinu (kan varieer oor 'n sekere interval) en word

'n drumpelwaarde of afsnypunt baiekeer gebruik: bv. waardes bokant die afsnypunt dui op die siekte, en daaronder die afwesigheid daarvan. Figuur 8.2 gee 'n voorstelling van die verdelings van die populasies S en N se siftingstoets-metings.

Figuur 8.2:



Indien persone geklassifiseer word volgens die werklike status (volgens goue standaard) sowel die siftingstoets, lewer dit die volgende 2x2 – frekwensietabel op.

Tabel 8.1: frekwensies volgens werklike status en siftingstoets se indeling.

	Werklike status		
Siftingstoets	Siekes (S)	Nie-siek (N)	Totaal
+	A (ware pos.)	B (vals pos.)	A + B (toets +)
-	C (vals neg.)	D (ware neg.)	C + D (toets -)
Totaal	A + C (siekes)	B + D (nie-siekes)	N=A+B+C+D

In Tabel 8.1 is daar A persone wat siek is en positief reageer op die siftingstoets. As proporsie van al die siekes $(A+C)$, gee dit die *sensitiwiteit*, d.i.

$$\text{Sensitiwiteit} = \frac{A}{A+C}, \tag{8.7}$$

die proporsie korrek-geklassifiseerde positiewes (hier siek mense met toets positief).

So is D die korrek-geklassifiseerde persone wat nie siek is nie en is die proporsie daarvan t.o.v. nie-siekes die *spesifisiteit*, d.i.

$$\text{Spesifisiteit} = \frac{D}{B+D}, \tag{8.8}$$

die proporsie korrek-geklassifiseerde negatiewes (hier nie-siekes met toets negatief).

'n Goeie siftingstoets behoort dan ‘n hoë sensitiwiteit sowel as ‘n hoë spesifisiteit te besit, omdat die teendeel baie nadelig kan wees. Dit wil sê om persone met die siekte as nie-siek te klassifiseer (‘n getal van C persone) is nadelig asook om persone wat nie siek is, as siek te klassifiseer (‘n getal van B persone).

In die populasies (soos aangedui in Figuur 8.2) gee die oppervlakte regs van die afsnypunt onder die S-verdeling (hier die siekes), die *sensitiwiteit* en die oppervlakte links van afsnypunt onder die N -verdeling, die *spesifisiteit*. Die ideaal sou wees om die twee verdelings volkome te skei met 'n afsnypunt sodat beide die sensitiwiteit en spesifisiteit 1 is.

Keuse van optimum afsnypunt:

Figuur 8.2 gee maar een afsnypunt uit baie moontlike afsnypunte. Vir elke so ‘n afsnypunt *(t)* kan soos in Tabel 8.1 die proporsie ware positiewes *wp(t)* of

sensitiwiteit asook die proporsie vals positiewes *vp(t)*, d.i. 1 – spesifisiteit, bereken word. Die vraag ontstaan egter of daar nie 'n optimumwaarde vir t is nie? Een metode is om die *t*-waarde te gebruik wat ooreenkom met die Youden indeks-waarde YI, waar

$$\begin{aligned} \mathrm{YI} &= maks_t\left(wp(t)-vp(t)\right) \\ &= maks_t\left(wp(t)+wn(t)-1\right)\ , \end{aligned} \qquad (8.9)$$

d.i. die maksimum waarde van die som van die sensitiwiteit (wp) en spesifisiteit (wn) minus 1. Die *optimum waarde* van die *afsnypunt* t *word dan verkry waar die som* $wp+wn$ *'n maksimum is.*

Omdat by gegewe t die verdelingsfunksies van X_N en X_S gegee word deur:

$$F(t)=\mathrm{P}\left(\mathrm{X}_N \le t\right)=wn(t)$$

en

$$G(t)=\mathrm{P}\left(\mathrm{X}_S \le t\right)=1-wp(t) \quad ,$$

volg dat

$$YI = maks_t\left(F(t)-G(t)\right)\ . \qquad (8.10)$$

Dus $\mathrm{YI}>0$ impliseer dat $F(t)\ge G(t)$ vir elke t, wat beteken dat X_S se verdeling grotendeels regs van die verdeling van X_N lê (kyk bv. Figuur 8.2). Indien $\mathrm{YI}\le 0$ beteken dit dat die siftingstoets nie beter is as wanneer individue ewekansig as positief en negatief geklassifiseer word nie.

Die optimum *t* kan dus beraam word waar die beraamde verskil $\hat{F}(t)-\hat{G}(t)$ 'n maksimum is. Daar word vier metodes deur Krzanowski & Hand (2009), paragraaf 9.4 bespreek om die optimum t, aangedui deur t^*, te bepaal:

(a) Twee-normaalmetode:

Met F en G beide normaal is met gemiddeldes μ_N, μ_S en standaardafwykings σ_N, σ_S:

$$\text{YI} = \text{maks}_t \left[\Phi\left(\frac{t-\mu_N}{\sigma_N} \right) - \Phi\left(\frac{t-\mu_S}{\sigma_S} \right) \right],$$

wat indien die eerste afgeleide na t aan nul gelyk gestel word, dit lewer:

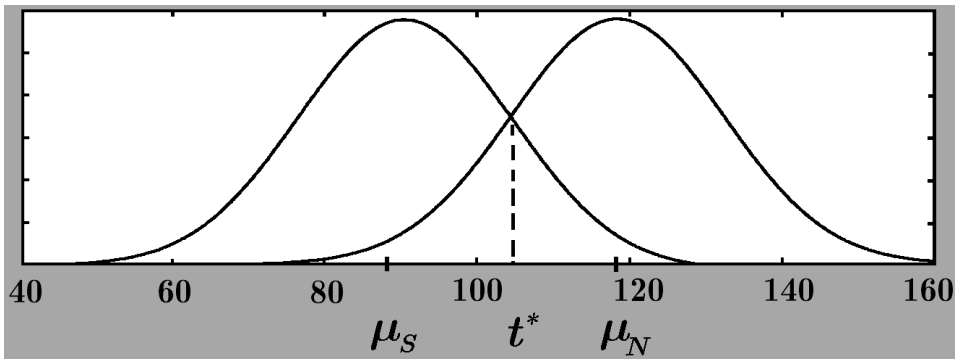
$$t^* = \frac{\mu_S \sigma_N^2 - \mu_N \sigma_S^2 - \sigma_N \sigma_S \sqrt{(\mu_N - \mu_S)^2 + (\sigma_N^2 - \sigma_S^2) \ln(\sigma_N^2 / \sigma_S^2)}}{(\sigma_N^2 - \sigma_S^2)} \quad . \qquad (8.11)$$

Indien $\sigma_N^2 = \sigma_S^2 = \sigma^2$, is

$$t^* = \tfrac{1}{2}(\mu_N + \mu_S) \ , \qquad (8.12)$$

d.i. die optimum afsnypunt lê halfpad tussen die gemiddeldes van die verdelings en waar die normaal-digtheidsfunksies mekaar sny (kyk Figuur 8.3).

Figuur 8.3



Om t^* te beraam met $\hat{t}^*$, word die beramers $\hat{\mu}_N$, $\hat{\mu}_S$, $\hat{\sigma}_N^2$ en $\hat{\sigma}_S^2$ in (8.11) vervang.

(b) Getransformeerde-normaal metode:

Omdat die aanname dat $G(t)$ en $F(t)$ normaal is, soms onrealisties is, kan 'n geskikte monotone transformasie (Box-Cox-transformasie) *Y=h(X)* op X uitgevoer word om normaliteit te bewerkstellig. Omdat onder monotone transformasies, YI onveranderd bly kan t^* bepaal word soos in (a), maar op Y se verdelings, waarna dit terug getransformeer word i.t.v. X .

(c) Empiriese metode:

F en G kan beraam word met hulle empiriese verdelingsfunksies

$$\hat{\mathrm{F}}(\mathrm{t}) = \mathrm{n}'_{\mathrm{N(t)}} / \mathrm{n}_{\mathrm{N}}$$

(8.13)

$$\hat{\mathrm{G}}(\mathrm{t}) = \mathrm{n}'_{\mathrm{S(t)}} / \mathrm{n}_{\mathrm{S}} \ ,$$

waar $\mathrm{n}'_{\mathrm{A(t)}}$ die getal individue uit populasie A is waarvan X se waarde kleiner of gelyk is aan t.

Deur nou uit 'n reeks t-waardes dié een te kies waarvoor $\hat{F}(t) - \hat{G}(t)$ 'n maksimum waarde het, word $\hat{t}^*$ verkry.

(d) Kernberamermetode:

Hier word $F(t)$ en $G(t)$ bepaal deur van kernberaming van die digtheidfunksies f_s en f_g gebruik te maak.

Effekgroottes van onderskeibaarheid van twee eenveranderlike verdelings:

Uit die ROC-analise volg twee maatstawwe wat as effekgroottes gebruik kan word, nl. (a) die Youden-indeks en (b) die oppervlakte onder die ROC-kromme.

- Youden-indeks: Volgens die definisie daarvan in (8.10) is dit duidelik dat die waarde daarvan tussen 0 en 1 kan varieer, met waarde 0 as die

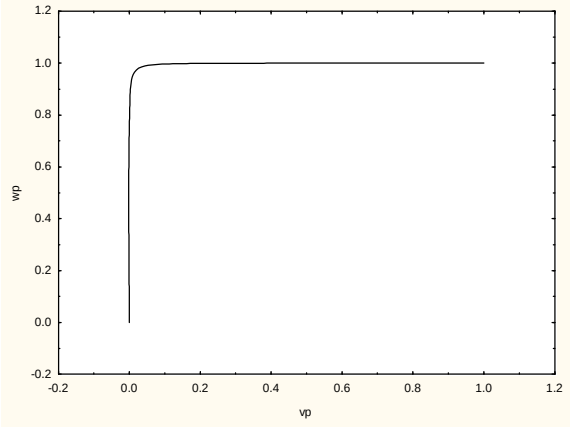
verdelings van die twee populasies volkome ooreenstem en waarde 1 as daar hoegenaamd geen oorvleueling is nie.

- Oppervlakte onder die ROC-kromme (AUC):

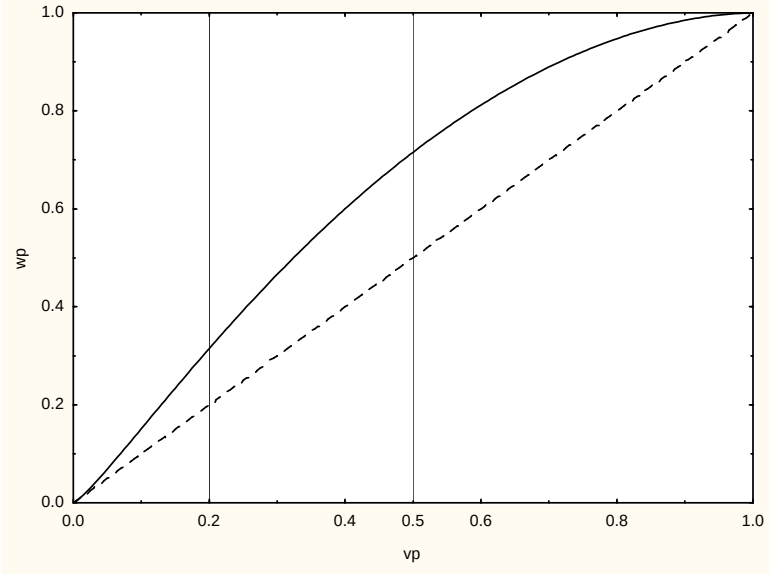
Beskou die ROC-krommes in Figuur 8.4:

Figuur 8.4:

(a)



(b)



Figuur 8.4(a) is verkry as die populasies in Figuur 8.2 beide normaal is en vir populasie S geld dat die gemiddelde en standaardafwyking $\mu_S = 4, \sigma_S = 1,$ terwyl vir populasie N $\mu_N = 0$ en $\sigma_N = 1$. Hier is die populasies vir alle praktiese doeleindes volkome onderskeibaar deurdat digtheidsfunksie van populasie S amper geheel en al regs van dié van N lê. Hierdie ROC-kromme is naby die beste moontlike kromme en die optimum afsnypunt kan maklik tussen die twee verdelings gekies word. Die ander uiterste word gegee in Figuur 8.4(b) as die diagonaallyn (stippellyn) waar die twee verdelings beide $N\ (0;1)$ geneem word – dus glad nie onderskeibaar is nie. Hier kan lede van elke populasie ewekansig ingedeel word as positief of negatief. Figuur 8.4(b) gee ook 'n ROC-kromme wat in 'n situasie soos in Figuur 8.2 verkry word (vollyn) en daaruit op 'n geskikte optimum afsnypunt besluit kan word.

As ons na Figuur 8.2 kyk, is dit duidelik dat die oppervlakte onder die ROC-krommes in (a) amper 1 is, by (b) tussen 0,5 en 1, in die geval van vollyn, en presies 0,5 by die stippellyn. Hierdie oppervlakte ("area under the curve") word deur AUC aangedui en is 'n *maatstaf om die onderskeibaarheid van die verdelings* S en N aan te dui. Hoe groter AUC, hoe meer onderskeibaar. Vir AUC die waarde 0,5 is daar geen onderskeibaarheid tussen S en N nie.

AUC kan ook as volg geïnterpreteer word:
Gestel 'n individu word ewekansig uit elk van die populasies S en N gekies en hulle siftingtoets se tellings is X_S en X_N, dan geld dat

$$\text{AUC} = \text{P}(X_S > X_N)\ ,$$

wat beteken dat AUC die waarskynlikheid is dat X_S groter is as X_N. In terme van Figuur 8.4 is hierdie waarskynlikheid naby 1 as die twee populasies grootliks onderskeibaar is (Figuur 8.4(a)), terwyl die

waarskynlikheid 0,5 as die twee populasieverdelings geheel en al oorvleuel (Figuur 8.4(b)). Die AUC is ook bekend as die Gini-indeks.
Verdere beskrywings van hoe o.a. AUC uit ROC-analises beraam kan word, word in die dokument "Gebruik van ROC - analises om goeie siftingstoetse, gebaseer op kontinue veranderlikes, te bepaal" wat by www.nwu.ac.za/af/p-statcs/index_a.html afgelaai kan word.

8.3 **Trefkoers en beraming daarvan**

'n *Trefkoers* is die proporsie korrekte klassifikasies van objekte oor al die populasies heen.

Huberty (1994: Hoofstuk VI) onderskei tussen drie soorte trefkoerse:

(a) *Optimale* trefkoers $P^{(o)}$: dit is die trefkoers as die klassifikasiereël op die bekende populasie sentroïedes en kovariansiematrikse gebaseer word (d.i. $\boldsymbol{u_g}$ en Σ_g bekend).

(b) *Werklike* trefkoers $P^{(a)}$: die verwagte trefkoers van 'n toekomstig steekproef (of toets-steekproef) waar die klassifikasiereël op die modelberamingsteekproef (Engels: "training sample") gebaseer is. Dit word ook die voorwaardelike trefkoers genoem.

(c) *Verwagte* trefkoers $P^{(e)}$: die verwagte proporsie korrekte klassifikasies oor alle moontlike steekproewe van grootte $N = \Sigma_g n_g$. Nou is $P^{(e)} = E\left(P^{(a)}\right)$. Hierdie trefkoers heet ook die **onvoorwaardelike** trefkoers en is van belang voordat enige steekproef getrek word.

Vervolgens word aandag aan die beraming van die trefkoers in verskillende gevalle gegee.

8.3.1 Twee eenveranderlike normaal-populasies met variansies homogeen

Cohen neem die maatstaf U_2 as die proporsie van populasie B wat groter is as dieselfde proporsie van populasie A (die proporsie van die aangeduide oppervlakte relatief tot A se totale oppervlakte, in Figuur 8.1). Met die effekgrootte $\delta = \left|\mu_B - \mu_A\right| / \sigma$ in terme van die twee populasies se gemiddeldes en gemeenskaplike SA, kan die verdelings as normaal $N(0;1)$ en $N(\delta;1)$ geneem word sonder verlies aan algemeenheid. Dit beteken dus dat

$$U_2 = \Phi(\delta / 2), \tag{8.14}$$

waar $\Phi(x)$ die kumulatiewe verdelingsfunksie is van 'n $N(0;1)$- verdeling is.

Voorbeeld 8.1

Beskou Voorbeeld 4.2 waar $\mu_B = 111$, $\mu_A = 105$, $\sigma = \sigma_A = \sigma_B = 10$ die gemiddelde IK's en SAs van populasies A en B was. Met $\delta = \left|\mu_B - \mu_A\right| / \sigma = \left|111 - 105\right| / 10 = 0{,}6$, volg dat

$$U_2 = \Phi\left(\frac{0{,}6}{2}\right) = 0{,}618 \quad ,$$

wat beteken dat 'n proporsie van 0,618 van populasie B groter IK's het as dieselfde proporsie van A (kyk weer na figuur 8.2).

Opmerking: Cohen (1969, 1977, 1988) se tabel 2.2.1 gee vir geselekteerde waardes van δ waardes vir U_2. Vir die riglynwaardes $\delta = 0{,}2$ $0{,}5$ en $0{,}8$ vir klein, medium en groot effekte, is die ooreenstemmende waardes van $U_2 = 0{,}54$ $0{,}60$ en $0{,}66$.

Volgens Huberty & Holmes (1983) is U_2 (hulle noem dit P_c) die waarskynlikheid van 'n korrekte klassifikasie en kan beraam word deur:

$$P_c = \Phi\left(\hat{\delta}/2\right), \tag{8.15}$$

waar

$$\hat{\delta} = \left|\overline{x}_A - \overline{x}_B\right| / s \quad ,$$

die steekproef-effekgrootte (kyk Hoofstuk 4).

Om P_c te maksimeer, kan die volgende klassifikasiereël gebruik word vir $\overline{x}_A < \overline{x}_B$:

Deel objek *u* toe aan populasie A, indien $x_u < \frac{1}{2}\left(\overline{x}_A + \overline{x}_B\right)$ andersins deel dit toe aan B. Hierdie reël gebruik dieselfde afsnypunt as die optimum *t* in (8.12), waar normaliteit en homogene variansies aanvaar was.

Voorbeeld 8.2

Vir voorbeeld 4.3 was die steekproewe uit A en B se gemiddeldes 11 en 13 met variansies 5 en 7,5, terwyl die steekproefgroottes 5 was. Onder die aanname van homogeniteit van variansies, was $\delta = 0{,}8$ en dus is $P_c = \Phi\left(\hat{\delta}/2\right) = \Phi\left(0{,}4\right) = 0{,}66$. Die klassifikasiereël word dan:

Deel persoon *u* toe aan A indien $x_u < \frac{1}{2}\left(11+13\right) = 12$, andersins aan B.

As die variansies nie as homogeen aanvaar word nie en die steekproefvariansies en gemiddeldes in (8.11) vervang word, is die optimum afsnypunt

$$t^* = \frac{13\times 5 - 11\times 7{,}5 - \sqrt{5\times 7{,}5}\sqrt{\left(13-11\right)^2 + \left(5-7{,}5\right)\ln\left(5/7{,}5\right)}}{\left(5-7{,}5\right)} = 12{,}48,$$

wat verskil van die gemiddelde van 11 en 13.

□

8.3.2 Twee meerveranderlike normaalpopulasies met gelyke kovariansie-matrikse

Huberty (1994: 83-86) veralgemeen U_2 van Cohen met die optimum trefkoerse vir populasie A en B as:

$$P_A^{(O)} = 1 - \Phi\left(\frac{\Gamma - \frac{1}{2}\Delta^2}{\Delta}\right) \text{ en } P_B^{(O)} = 1 - \Phi\left(\frac{-\Gamma - \frac{1}{2}\Delta^2}{\Delta}\right) \tag{8.16}$$

waar $\Gamma = \ell n(\pi_B / \pi_A)$, π_g die a priori waarskynlikhede van lidmaatskap tot g, Δ die Mahalanobis-afstand gedefinieer as

$$\Delta^2 = (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)' . \tag{8.17}$$

Deur Γ en Δ te beraam uit twee ewekansige steekproewe met $K = \ell n(p_B / p_A)$ en $\hat{D}$ waar $\hat{D}^2 = \frac{n-m-3}{n-2}D^2 - \frac{mn}{n_A n_B}$, (8.18)

met m die aantal veranderlikes en

$$D^2 = (\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)' \quad , \quad n = n_A + n_B . \tag{8.19}$$

D^2 kan maklik bereken word volgens (7.6) in Hoofstuk 7 (daar aangedui met $\hat{D}^2$) uit Wilks se Λ vir die steekproewe. $\hat{D}^2$ in (8.18) is dan ekwivalent aan (7.7).

Dan volg dat:

$$\hat{P}_A^{(O)} = 1 - \Phi\left(\frac{K - \frac{1}{2}\hat{D}^2}{\hat{D}}\right) \ , \ \hat{P}_B^{(O)} = 1 - \Phi\left(\frac{-K - \frac{1}{2}\hat{D}^2}{\hat{D}}\right) \tag{8.20}$$

Hieruit volg die totale populasie-trefkoers as

$$\hat{P}^{(O)} = p_A\hat{P}_A^{(O)} + p_B\hat{P}_B^{(O)} . \tag{8.21}$$

Vir die spesiale geval waar $p_A = p_B$ herlei (8.21) na:

$$\hat{P}^{(O)} = \hat{P}_A^{(O)} + \hat{P}_B^{(O)} = \Phi\left(\hat{D}/2\right) , \tag{8.22}$$

wat ‘n veralgemening is van (8.14).

Voorbeeld 8.3

Beskou Voorbeeld 7.1: Hier was die beraamde $\hat{D}$ om die eksperimentele- en kontrolegroepe se BDI voor-, na- en opvolgtoetse se gemiddeldes te vergelyk: $\hat{D} = 4{,}42$. Omdat die twee groepe ewe groot gekies is, kan die aanname $p_E = p_K$ gemaak word, sodat

$$\hat{P}^{(O)} = \hat{P}_E^{(O)} = \hat{P}_K^{(O)} = \boldsymbol{\Phi}\left(\hat{D}/2\right) = \boldsymbol{\Phi}\left(2{,}21\right)$$

$$= 0{,}986 \ .$$

Dit is amper seker dat persone korrek geklassifiseer kan word in die twee groepe as van BDI se voor-, na- en opvolgtoetse gebruik gemaak word.

Deel persoon *u* toe aan populasie g as

$$D_{ug}^{*2} < D_{ug'}^{*2} \text{ vir } g \neq g = 1{,}2.$$

Hier is die aanname van gelyke kovariansie-matrikse gemaak en

$$\bar{\boldsymbol{x}}_1' = \bar{\boldsymbol{x}}_{\boldsymbol{E}}' = \left(13{,}04 \;\; 8{,}72 \;\; 6{,}72\right) \quad , \quad \bar{\boldsymbol{x}}_2' = \bar{\boldsymbol{x}}_{\boldsymbol{K}}' = \left(11{,}56 \quad 15{,}56 \quad 16{,}36\right)$$

en

$$\boldsymbol{S} = \begin{pmatrix} 35{,}76 & 12{,}57 & 14{,}97 \\ 12{,}57 & 58{,}37 & 48{,}90 \\ 14{,}97 & 48{,}90 & 80{,}17 \end{pmatrix}$$ is die saamgevoegde kovariansiematriks.

Die lineêre klassifikasiereël met hierdie data herlei na: klassifiseer in groep E as: -4,153 + 0,178Voor + 0,89Na + 0,17Opvolg < -3,239 + 0,346Voor + 0.113Na - 0.061Opvolg,

en andersins in groep K. Elke persoon se voor-, na- en opvolgtellings van BDI word links en regs in bogenoemde formule vervang en as die linkerkant wel kleiner is as die regterkant, word die persoon in groep E geklassifiseer, anders in groep K.

Dit bring mee dat persone 1 en 9 uit groep E verkeerdelik in groep K ingedeel word (dus 23 uit 25 persone korrek geklassifiseer), terwyl persone 10, 14-16, 18-

21 uit groep K verkeerdelik by groep E ingedeel word (17 uit 25 korrek geklassifiseer). ▱

8.3.3 Meer as twee meerveranderlike populasies

In die praktyk het ons baie keer meer as twee meerveranderlike populasies waarin ons objekte wil klassifiseer. Onder aanname van normaalpopulasies, kan die klassifikasiereël in paragraaf 8.2 gebruik word. Die probleem is egter dat geen beramer $\hat{P}^{(O)}$ soos in (8.14) bestaan vir die trefkoers nie. Volgens Huberty (1994) kan PDA op twee wyses gebruik word : intern en ekstern.

(a) *Interne ontleding* beteken dat die klassifikasiereëls op dieselfde data waarop dit gebaseer is, toegepas word en die objekte dus hergeklassifiseer word. Die proporsie objekte wat korrek geklassifiseer word, gee ‘n beraming van die trefkoers, die *skynbare of hersubstitusie trefkoers* genoem. Hierdie trefkoers is sydig vir enige van $P^{(o)}$, $P^{(a)}$ of $P^{(e)}$ en *oorberaam* die trefkoers. Hierdie metode word deur die bekende statistiese rekenaarpakkette Statistica, SPSS, BMDP en SAS uitgevoer, sodat die beraamde trefkoers omsigtig geïnterpreteer moet word.

(b) *Eksterne ontleding* kan opgedeel word in die sogenaamde uithou (“holdout”) – metode, die laat-een-weg (“leave-one-out” afgekort L-O-O) metode en maksimum-a-posteriori-waarskynlikheidsmetode.

(i) By die uithou-metode word ‘n sogenaamde toetssteekproef uitgehou deur dit ewekansig te kies uit die oorspronklike steekproef. Die oorblywende data (die sogenaamde modelberamingsteekproef) word gebruik om die klassifikasiereël mee op te stel, waarna die elemente binne die toetssteekproef geklassifiseer word om daaruit die trefkoers te bepaal. Hierdie metode gee slegs ‘n goeie beraming van

$P^{(a)}$, maar slegs in gevalle waar die toetssteekproef dieselfde grootte is as die modelberamingsteekproef. Bogenoemde pakkette kan almal hierdie trefkoers bereken.

(ii) Die laat-een-weg (L-O-O) metode behels dat een objek weggelaat word en ‘n klassifikasiereël gebaseer word op die oorblywende N - 1 objekte se waarnemings. Die objek word dan geklassifiseer in een van die g populasies. Die proses word herhaal vir elkeen van die N objekte oor al die populasies heen en die trefkoers word daarna bepaal. Hierdie trefkoers is egter streng gesproke nie ‘n beraming vir enige van $P^{(o)}$, $P^{(a)}$ of $P^{(e)}$ nie, omdat dit op ‘n steekproefgrootte van N - 1 in plaas van N gebaseer is. Vir N nie te klein nie, kan dit wel as ‘n beramer vir $P^{(a)}$ gebruik word. Die DISCRIM-prosedure van SAS bepaal hierdie trefkoers deur die CROSSVALIDATE-opsie te gebruik, terwyl SPSS die ‘Leave on out’ opsie het. STATISTICA het egter nie hierdie opsie nie.

(iii) Die maksimum a-posteriori-waarskynlikheidsmetode (“Maximum-posterior-probability method” afgekort M-P-P metode) gee ‘n verdere alternatiewe metode om $P^{(a)}$ te beraam. Dit word bereken as die gemiddeld van al die objekte se maksimum a-posteriori-waarskynlikhede:

$$\hat{P}^{(a)} = \frac{1}{N}\sum_{u=1}^{N} maks\left\{\hat{P}\left(1\middle|\boldsymbol{x_u}\right), \hat{P}\left(2\middle|\boldsymbol{x_u}\right), \ldots, \hat{P}\left(k\middle|\boldsymbol{x_u}\right)\right\}, \qquad (8.23)$$

waar $\hat{P}\left(g\middle|\boldsymbol{x_u}\right)$ die beraamde a posteri-waarskynlikheid is dat objek u in populasie g val. Hierdie waarskynlikheid kan via ‘n interne- of eksterne ontleding beraam word. Beide SAS se DISCRIM-prosedure en SPSS se DISCRIMINANT bepaal met behulp van die interne metode die waardes van $\hat{P}\left(g\middle|\boldsymbol{x_u}\right)$, waarby die beramer $\hat{P}^{(a)}$ dan as die M-P-P/I beramer aangedui word. SAS se DISCRIM beraam onder

die CROSSVALIDATE-opsie $\hat{P}(g \mid \boldsymbol{x}_u)$ met die eksterne L-O-O-metode en die beramer $\hat{P}^{(a)}$ staan dan bekend as M-P-P/L-O-O. Volgens Huberty (1994) is hierdie metode verkieslik as meerveranderlike normaliteit aanvaar kan word. Indien hierdie aanname in die weegskaal is, is L-O-O beter.

Tabel 8.2 gee die verskillende klassifikasies deur van bogenoemde metodes gebruik te maak vir die data in Hoofstuk 3, Voorbeeld *F* met die 3 aktiwiteitsgroepe die populasies waaruit steekproewe van $n_1 = 694$, $n_2 = 227$ en $n_3 = 441$ $(N = 1362)$ getrek is.

Tabel 8.2: Klassifikasies met verskillende metodes:

		Klassifikasie in groepe			
(a) Intern / Lineêr		1	2	3	Totaal
	1	549 (78,1)	1 (0,1)	151 (21,8)	694
Groepe	2	96 (42,3)	0 (0,0)	131 (57,7)	227
	3	169 (38,3)	0 (0,0)	272 (61,7)	441
	Totaal	807 (59,3)	1 (0,1)	554 (40,7)	1 362
	% Fout	21,9	100,0	38,3	40,2
(b) L-O-O / Lineêr					
	1	540 (77,8)	1 (0,1)	153 (22,1)	
Groepe	2	96 (42,3)	0 (0,0)	131 (57,7)	
	3	169 (38,3)	0 (0,0)	272 (61,7)	
	Totaal	805 (59,1)	1 (0,1)	556 (40,8)	
	% Fout	22,2	100,0	38,3	40,4
(c) Intern / Nie-lineêr					
	1	207 (29,8)	66 (9,5)	421 (60,7)	
Groepe	2	25 (11,0)	20 (8,8)	182 (80,2)	
	3	27 (6,1)	14 (3,2)	400 (90,7)	
	Totaal	259 (19,0)	100 (7,3)	1 003 (73,7)	
	% Fout	70,2	91,2	9,3	54,0
(d) L-O-O / Nie-lineêr					
	1	206 (29,7)	66 (9,5)	422 (60,8)	
	2	26 (11,5)	13 (5,7)	188 (82,8)	
	3	30 (6,8)	19 (4,3)	392 (88,9)	
	Totaal	262 (19,2)	98 (7,2)	1 002 (73,6)	
	% Fout	70,3	94,3	11,1	55,1
A priori-waar-skynlikhede		0,51	0,17	0,32	

8.4 **Effekgrootte-indeks vir korrekte klassifikasie**

Die trefkoers soos met behulp van een van die metodes in die vorige paragraaf beraam kan word, gee ons ‘n indeks om die suksesvolheid van korrekte klassifikasie te beoordeel oor al die populasies heen. Om hierdie trefkoers egter te beoordeel is dit eers nodig om dit met die sg. kansklassifikasie se waarskynlikheid te vergelyk. Dit is die waarskynlikheid van toevallige klassifikasie wanneer glad nie na enige data gekyk word nie, en word die **kans-trefkoers** genoem. Volgens Huberty (1994) is daar gewoonlik twee wyses om kans-trefkoers te bepaal, nl.

(a) die proporsionele kans-kriterium en

(b) die maksimum kans-kriterium.

Vervolgens word die twee metodes behandel.

8.4.1 Proporsionele kans-kriterium

Met k populasies van dieselfde grootte waaruit steekproewe van gelyke grootte *n* getrek word, is die kans-trefkoers duidelik $1/k$ vir elke populasie en die verwagte treffrekwensie per populasie $n = n\,1/k + n\,1/k + ... + n\,1/k$. In die algemene geval van populasies met ongelyke groottes, kan die gelyke kanstrefkoerse van $1/k$ vervang word met p_g, die beraamde a priori waarskynlikheid van lidmaatskap tot populasie *g*. Met die steekproewe ook nie van gelyke grootte nie, maar n_g, $g = 1,...,k$, word die verwagte treffrekwensie vir populasie *g*: $e_g = p_g n_g$, sodat die kans-treffrekwensie oor al die populasies dan

$$e = \sum_{g=1}^{k} e_g = \sum_{g=1}^{k} p_g n_g \,. \qquad (8.24)$$

Neem ons $N = n_1 + n_2 + ... + n_k$, dan is die kans-trefkoers oor al die populasies:

$$H_e = \frac{e}{N} = \frac{1}{N}\sum_{g=1}^{k} p_g n_g \qquad (8.25)$$

Voorbeeld 8.4

In Hoofstuk 3, Voorbeeld *F* vorm die 3 aktiwiteitsgroepe die populasies waaruit steekproewe van $n_1 = 694$, $n_2 = 227$ en $n_3 = 441$ $(N = 1362)$ getrek is. Gestel die navorser het vooraf die *a* priori waarskynlikhede as *0,5*, *0,25* en *0,25* gekies. Nou is die kans-treffrekwensies *0,5* x *694 = 347* vir aktiwiteitsgroep *1*, *0,25* x *227 = 56,75* en *0,25* x *441 = 110,25* vir groep 2 en 3 respektiewelik. Dit gee dus 'n totale kanstreffrekwensie van *347 + 56,75 + 110,25 = 514*, sodat die kans-trefkoers $H_e = 514/1362 = 0,377$ is oor al die groepe heen. As aanvaar word dat die *a* priori waarskynlikhede proporsioneel is aan die steekproefgroottes, geld nou dat die totale kans-treffrekwensie word:

$$\left(\frac{694}{1362}\times 694\right) + \left(\frac{227}{1362}\times 227\right) + \left(\frac{441}{1362}\times 441\right)$$
$$= 353,6 + 37,8 + 142,8$$
$$= 534,2,$$

sodat die totale kans-trefkoers nou $H_e = 534,2/1362 = 0,392$ is.

Dit beteken dat indien die *1362* mans op toevallige wyse sonder gebruikmaking van die data ingedeel sou word, 'n trefkoers van 38-40% verwag kan word. □

8.4.2 Maksimum-kans kriterium

Hier word die kans-trefkoers H_e bloot as die maksimum van die verskillende beraamde a priori waarskynlikhede geneem:

$$H_e = maks\left(p_1, p_2, ... p_k\right) \qquad (8.26)$$

Volgens Huberty & Lowman (2000) is hierdie kriterium veral van toepassing by 2 groepe wanneer die a priori-waarskynlikhede radikaal verskil.

Voorbeeld 8.5

In voorbeeld 8.4 word $H_e = 0,5$ by die voorafgekose a priori-waarskynlikhede en wanneer die proporsionele a priori-waarskynlikhede $694 / 1362 = 0,51$; $227 / 1362 = 0,17$ en $441 / 1362 = 0,32$ gebruik word, is $H_e = maks\ (0,51\ ;\ 0,17\ ;\ 0,32) = 0,51$.

□

8.4.3 Statistiese toetsing van tref-frekwensie

Indien die tref-frekwensie van populasie g in die klassifikasie frekwensietabel deur n_{gg} aangedui word, volg dat die totale tref-frekwensie gegee word deur:

$$o = \sum_{g=1}^{k} n_{gg} \ .$$

Onder die nulhipotese van toevallige klassifikasie, volg dat

$$z = \frac{o - e}{\sqrt{e(N-e)/N}} \sim N(0;1).$$

Die toetsing van hierdie nulhipotese lei tot die p-waarde: $p = P(Z \geq z)$, waar Z $N(0;1)$ verdeel is. □

Die onderste $(1-\alpha)100\%$ vertrouensgrens vir die werklike tref-frekwensie is dus (Huberty, 1994: 105):

$$o\text{-}z_{\alpha}\sqrt{e(N-e)/N}, \qquad (8.27)$$

waar z_{α} die $(1-\alpha)$-de persentiel van 'n $N(0;1)$ - verdeling is.

Vir populasie g volg soortgelyk dat $o_g - z_{\alpha}\sqrt{e_g\left(n_g - e_g\right)/n_g}$ (8.28)

die onderste $(1-\alpha)100\%$ vertrouensgrens vir die werklike tref-frekwensie gee.

Voorbeeld 8.6:

In Voorbeeld 8.4 was met die proporsionele kanskriterium $e = 534,2$ terwyl uit Tabel 8.1 (b) waar L-O-O / Lineêre metode gevolg is, volg $o = 540 + 0 + 272 = 812$.

$$z = \frac{812 - 534,2}{\sqrt{534,2\left(1362 - 534,2\right)/1362}} = \frac{277,8}{\sqrt{324,67}} = 15,42$$

sodat $p < 0,0001$. Die 95% vertrouens-ondergrens vir die werklike treffrekwensie is:

$$\begin{aligned} & 812 - 1,645\sqrt{534,2\left(1362 - 534,2\right)/1362} \\ =& 812\text{-}1,645\sqrt{324,67} \\ =& 812 - 29,6 = 782,4 \end{aligned}$$

Pas ons die gekose a priori-waarskynlikheid vir populasie 3 toe, is $e_3 = 0,25 \times n_3 = 0,25 \times 441 = 110,25$. Neem uit Tabel 8.1(b) die waargenome tref-frekwensie $o_3 = 272$, dan is $p < 0,0001$ die 99% vertrouensondergrens vir die werklike tref-frekwensie vir populasie 3:

$$\begin{aligned} & 272 - 2,33\sqrt{110,25\left(441\text{-}110,25\right)/441} \\ =& 272 - 2,33\sqrt{82,69} \\ =& 272 - 21,19 = 250,8. \end{aligned}$$

Dit beteken dus dat die totale tref-frekwensie so laag soos $782,4$ kan wees met 95% waarskynlikheid, terwyl die tref-frekwensie van populasie 3 so klein soos $250,8$ kan wees met 99% waarskynlikheid. □

8.5 **Effekgrootte-indeks: Verbetering-bo-kans**

Deur die werklike of waargenome trefkoers H_o te vergelyk met die kanstrefkoers H_e, word H_o gekorrigeer vir toevallige korrekte klassifisering van objekte. Die verskil tussen die kans foutkoers $1-H_e$ en die waargenome fout koers $1-H_o$, as proporsie van die kans foutkoers, word die effekgrootte indeks:

$$I = \frac{(1\text{-}H_e)-(1\text{-}H_o)}{1\text{-}H_e} = \frac{H_o - H_e}{1-H_e} \tag{8.29}$$

(kyk Huberty & Lowman, 2000, Huberty, 1994).

Uit die definisie van die effekgrootte-indeks I, kan dit ook beskryf word as 'n indeks vir die proporsionele verkleining van fout, of die verbetering-bo-kans indeks.

Let op dat die indeks afhang van die definisie van "kans" soos gereflekteer deur H_e, wat op sy beurt weer afhang van die beramings van a priori-waarskynlikhede.

Voorbeeld 8.7:

Vir Voorbeeld 8.4 se klassifikasie van persone binne die 3 populasies, gee Tabel 8.2 vier verskillende waargenome foutkoerse $(1\text{-}H_o)$. Gebruik ons die kansfoutkoers gebaseer op proporsionele a priori-waarskynlikhede $1-H_e = 1-0{,}392 = 0{,}608,$ volg die volgende waardes van I:

Metodes	p	$1\text{-}H_o$	$1\text{-}H_e$	I
(a) Intern/Lineêr	<0,0001	0,402	0,608	0,339
(b) L-O-O/Lineêr	<0,0001	0,404	0,608	0,336
(c) Intern/Nie-lineêr	<0,0001	0,540	0,608	0,112
(d) L-O-O/Nie-lineêr	<0,0001	0,551	0,608	0,094

□

Dit blyk dus uit Voorbeeld 8.7 dat metodes (a) en (b) beter klassifikasies was op grond van hoër waardes van I . Let op dat in al die gevalle $p < 0,0001$, wat beteken die nulhipotese van toevallige klassifikasie word deurgaans duidelik verwerp weens die groot steekproewe. Dit sê dus dat die klassifikasies nie bloot toevallig was nie, wat nie noodwendig beteken dit was goed nie. Om die suksesvolheid van die klassifikasies te beoordeel, kan die indeks I dus gebruik word. Die waarde $0,34$ by metode (a) in Voorbeeld 8.7 sê dat daar ‘n 34% verlaging in die foutkoers is by hierdie metode van klassifikasie in vergelyking met ‘n blote kansklassifikasie.

8.6 Verband tussen proporsie variansie $\left(\eta^2\right)$ en verbetering-bo-kans-indeks (I)

8.6.1 Homogene variansies of kovariansiematrikse

Om ‘n “gevoel” vir indeks I te verkry, het Huberty & Lowman (2000) vir die BISBEY-data in Huberty (1994) 6 van die 13 afhanklike veranderlikes beskou en groepe 1 en 2 vergelyk vir elkeen van die veranderlikes. Die variansies is deurgaans as homogeen aanvaar, sodat dit die eenveranderlike tweegroep geval lewer met homogene variansies.

Deur as a priori-waarskynlikhede $0,333$ en $0,667$ te gebruik en die maksimum-kans-kriterium te kies, is $H_e = 0,667$. In tabel 8.3 (Huberty & Lowman se Tabel 1) word die t-waardes van die t-toets op elkeen van die veranderlikes gegee en ook die trefwaarskynlikhede H_o as ‘n lineêre klassifikasiereël gebruik. Die proporsie variansie η^2 is beraam deur van (5.24) in Hoofstuk 5 gebruik te maak.

Tabel 8.3: Resultate van eenveranderlike 2-groep vergelykings met homogene variansies

t	p	$\hat{\eta}^2$	H_o	I
-1,07	0,286	0,010	0,698	0,09
-1,46	0,147	0,018	0,681	0,04
-3,58	0,001	0,101	0,707	0,12
-3,74	0,000	0,109	0,698	0,09
-6,29	0,000	0,258	0,810	0,43
-8,12	0,000	0,366	0,836	0,51

Daar is ‘n Pearson-korrelasie (r) van 0,90 tussen $\hat{\eta}^2$ en I verkry, terwyl die Spearman rangkorrelasie (r_s) van *0,81* dui op ‘n sterk monotone verband.

Op dieselfde wyse gaan Huberty & Lowman (2000) voort om uit dieselfde data die verbande te bepaal tussen die beraamde η^2 en I-indeks in die volgende gevalle:

(a) Eenveranderlike, 3-groep vergelykings met homogene variansies: $r = 0,97$

(b) Meerveranderlike, 2-groep vergelykings met homogene kovariansie-matrikse: $r = 0,95$

(c) Meerveranderlike, 3-groep vergelykings met homogene kovariansie-matrikse: $r = 0,98$

Hoewel hierdie resultate nie noodwendig in die algemeen geld nie, gee dit tog goeie aanduidings dat daar ‘n positiewe lineêre bestaan tussen η^2 en I in die geval van homogene variansies en kovariansiematrikse.

8.6.2 Heterogene variansies of kovariansiematrikse

Die beraamde proporsie variansie η^2 wat in die vorige paragraaf as effekgrootte-indekse in verband gebring is met I kan nou nie meer gebruik word nie, omdat homogene variansies of kovariansiematrikse aanvaar moet word. Huberty & Lowman (2000) het daarom I probeer korreleer met toetsstatistieke wat gebruik word om nulhipoteses van gelyke gemiddeldes te toets waar heterogene variansies of kovariansiematrikse aanvaar word. Vir 'n gekose datastel het hulle in die volgende gevalle die korrelasie bepaal:

(a) Eenveranderlike, 2-groep vergelykings met heterogene variansies: Toetsstatistiek J (James se 2de orde toets) met I (die nie-lineêre / L-O-O-metode met maksimum-kanskriterium) het die korrelasie $r = 0,89$ opgelewer.

(b) Eenveranderlike, 3-groep vergelykings met heterogene variansie: $r = 0,88$

(c) Meerveranderlike (4 veranderlikes), 2-groep vergelykings met heterogene kovariansiematrikse Toetsstatistiek T (van Yao) met I : $r = 0,97$

(d) Meerveranderlike 3 veranderlikes), 3-groep vergelykings met heterogene kovariansiematrikse: Toetsstatistiek S (van Johansen) met I: $r = 0,84$

8.7 **Riglynwaardes vir indeks *I***

Wanneer twee populasies vergelyk word, stem Huberty & Holmes (1983) saam met Cohen (1969) dat die gestandaardiseerde verskil van $\delta = 0,2$ op 'n klein effek dui (kyk paragraaf 4.5, Hoofstuk 4). Volgens hulle tabel 2 is die verwagte trefkoers $P^{(e)}$ dan omtrent $0,55$. Aanvaar ons gelyke a priori-

waarskynlikhede vir die **twee populasies**, volg dit dat $H_e = 0,5$, terwyl as $P^{(e)}$ met H_o beraam word as $0,55$, is $I = \frac{0,55 - 0,5}{0,5} = 0,1$. Huberty & Holmes voel egter dat 'n medium effek by klassifikasie met 'n verwagte trefkoers van $0,65$ (d.i. $I = 0,3$) behoort ooreen te stem, wat ekwivalent is met $\delta = 1,0$. Verder vereis hulle as 'n groot effek $P^{(e)} = 0,75$ (d.i. $I = 0,5$), waar $\delta = 1,5$.

Huberty & Lowman (2000) stel egter die volgende riglyne voor, gebaseer op voorlopige analises vir eenveranderlike 2-groep klassifikasie met homogene variansies:

Effek	
Klein	$I < 0,1$
Medium	$0,15 < I < 0,25$
Groot	$I > 0,3$

Vir die k-groep geval, stel Huberty & Lowman (2000) dieselfde riglyne voor. Hulle stel ook riglyne vir die ander gevalle voor. Tabel 8.4 gee 'n opsomming van hulle riglyne:

Tabel 8.4: Riglyne vir Verbetering-bo-kans-indeks

		Aantal	Effek		
Eenveranderlike		**Populasies**	**Klein**	**Medium**	**Groot**
	Homogeen	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
Variansies		k	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
	Heterogeen	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
		3	$I < 0,05$	$0,10 < I < 0,20$	$I > 0,25$
Meerveranderlike					
	Homogeen	2	$I < 0,15$	$0,2 < I < 0,3$	$I > 0,35$
Kovariansie-		3	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
Matrikse	Heterogeen	2	$I < 0,1$	$0,15 < I < 0,25$	$I > 0,3$
		3	$I < 0,05$	$0,10 < I < 0,20$	$I > 0,25$

Om op te som, is Huberty & Lowman (2000) se aanbeveling dat oor al die gevalle geneem, $I \leq 0,1$ as ‘n klein effek beskou word, en $I \geq 0,35$ as ‘n groot effek geneem kan word. Hulle waarsku egter dat hulle voorstelle gegrond is op ‘n beperkte aantal datastelle en dat hulle nie eintlik na die gevalle van meer as 3 tot 4 populasies gekyk het nie.

8.8 **Gebruike van indeks I**

Die effekgrootte-indeks I kan op twee wyses gebruik word. Eerstens kan dit as ‘n indeks vir die suksesvolheid van die klassifikasiereël wat in die diskriminantontleding gebruik word. Dit sou belangrik wees wanneer primêr belang gestel word in hoe ‘n mate die klassifikasiereël toekomstige waarnemings korrek gaan klassifiseer. ‘n Tweede gebruik van I is as ‘n effekgrootte-indeks in die plek van η^2 (en spesiale gevalle daarvan, bv. δ) wanneer heterogeniteit van variansies voorkom. Waar η^2 deur die getal groepe en aantal waarnemings (N)

beïnvloed kan word, word I se waarde nie direk deur N beïnvloed nie (Huberty & Lowman, 2000).

Wanneer heterogeniteit van variansies voorkom, kan die indekse $\Delta_1, \overline{\Delta_2}, \overline{\Delta_m}, \delta_g$ en δ_c in die eenveranderlike, 2-groep geval gebruik word. Vir enige ander situasie (kyk Hoofstukke 6 en 7 van die handleiding), word homogeniteit van variansies egter aanvaar. Hier sou I dus met vrug gebruik kon word, wanneer dit op die kwadratiese klassifikasiereël gebaseer is.

Normaliteit word gewoonlik aanvaar wanneer statistiese toetsing van gemiddelde vektore gedoen word en opgevolg word met die beraming van die effekgrootte η^2. In wese is die normaliteitsaanname nie nodig by die gebruik van klassifikasiereëls nie, maar spruit gewoonlik daaruit voort. Huberty (1994: Hoofstuk X) gee egter metodes hoe om diskriminantontleding te doen as normaliteit nie aanvaar word nie. As die veranderlikes kontinu, maar nie-normaal is, kan bv. rangtransformasies en naaste-buurman ontledings gedoen word. Hierdie metodes is beskikbaar by bv. SAS.

Vir kategoriese, nominale veranderlikes is daar volgens Huberty twee moontlikhede. Eerstens kan daar vir elkeen van sulke veranderlikes c - 1 skynveranderlikes (met waardes *0* en *1*) geskep word, waar daar c kategorieë is. Die probleem is dat op die wyse die finale hoeveelheid skynveranderlikes onhanteerbaar baie raak. Hiervoor kan egter die gewone klassifikasiereëls gebruik word. ‘n Tweede moontlikheid is om ‘n Fisher-Lancaster ontleding (Huberty, 1994: 153) te doen waarby by elke veranderlike-kategorie ‘n telling bepaal word. Hierdie getransformeerde data kan opnuut onderwerp word aan die gewone klassifikasiereëls.

Omdat bogenoemde metodes bykans alle situasies dek (o.a. heterogeniteit van variansie en nie-normaliteit), is die bepaling van die verbetering-bo-kans-indeks I ‘n baie meer algemene indeks as bv. η^2.

Meer navorsing is nodig om die verband tussen I en η^2 te ondersoek in ‘n groter verskeidenheid van situasies as deur Huberty en Lowman (2000). Ook is riglynwaardes vir I nog baie tentatief en behoort duideliker te word soos die gebruik daarvan meer algemeen word.

‘n SAS-program om I te bereken (*Groepsoorvleueling.sas*) is op die webblad van die handleiding beskikbaar.