

DISCRIMINANTANALYSE BIJ HET OPSPOREN VAN PATIENTEN MET TURNER-SYNDROOM

door E. MEULEPAS

Katholieke Universiteit te Leuven

Het Turner-syndroom is een samenloop van verscheidene afwijkingen bij sommige vrouwen, o.m. sexuele onvolwassenheid en dwerggroei; ook een korte nek (« webbed neck ») is typisch. Het werd door Turner in 1938 beschreven. De verklaring van het ontstaan van dit syndroom heeft nog wel een twintigtal jaren op zich laten wachten. Het gaat waarschijnlijk om een genetische stoornis.

Zoals men weet bevatten de celkernen van de mens 23 paren chromosomen, waarvan 22 paren zgn. autosomale chromosomen, gelijk in eenzelfde paar, en één paar geslachtschromosomen, van het type XX bij de vrouw en XY bij de man. Het Y-chromosoom is determinerend voor het mannelijk geslacht, één X-chromosoom minstens is nodig voor leefbaarheid.

Bij vrouwen kan het nu in zeldzame gevallen voorkomen dat één van beide X-chromosomen ontbreekt. Dit XO-type blijkt nu haast steevast het Turner-syndroom te vertonen. Men heeft nu anderzijds bij bepaalde genetische aandoeningen, bvb. bij mongolisme, vastgesteld dat afwijkingen in het huidlijstenpatroon (« vingerafdrukken ») voorkomen. Ook voor XO-afwijkingen werd dit nu in onderhavige studie nagegaan en vastgesteld.

Men gebruikte hiervoor gemakkelijk te kwantificeren kenmerken van het huidlijstenpatroon. Men kwam aldus tot zeven veranderlijken waarvan er vijf aantallen huidlijsten tussen bepaalde punten van de handpalm voorstellen, en de overige een deelverhouding en een hoek. Deze veranderlijken werden gemeten bij 300 vrouwen en 40 Turner-patiënten.

Op grond van deze waarnemingen kan men nu twee vragen beantwoorden, welke nauw samenhangen :

1° is er verschil in huidlijstenpatroon tussen normale vrouwen en Turner patiënten ?

2° indien er verschil is, hoe kan men van een vrouw uit haar huidlijstenpatroon haar genetisch type bepalen ?

We zullen in de nu volgende uiteenzetting deze problemen gradueel oplossen, en met de eenvoudige gevallen beginnen.

Eerste geval : één kwalitatieve dichotomische veranderlijke welke de groepsindeling verwekt.

Onderstel dat de classificatie betekent : toewijzing aan het XX- of het XO-type, waarbij de veranderlijke t resp. de waarde 0 of 1 aanneemt.

Is de veranderlijke t waargenomen, dan is de classificatieregels triviaal. Is ze niet waargenomen, maar kent men de populatieverdeling, waarin $P(XO) = P(Z = 1) = 1/10.000$ bvb., dan is de regel : aanzie de (niet onderzochte) vrouw als normaal. De kans op een foute conclusie is dan $1/10.000$.

Tweede geval : één kwalitatieve dichotomische veranderlijke, gecorreleerd met de indelings variabele.

Weze $Z = 1$ bij vrouwen met Turner-syndroom,
 $Z = 0$ bij vrouwen zonder Turner-syndroom.

Op basis van Z moet een vrouw ingedeeld worden bij het XX- of het XO-chromosomentype.

Veronderstel dat bij $m = 300$ XX-vrouwen en bij $n = 40$ XO-vrouwen Z werd waargenomen, waarbij we Z door X in de eerste en door Y in de tweede groep zullen voorstellen. Dan stellen de gemiddelden $\bar{X}$, $\bar{Y}$ de waargenomen proporties p_1 , p_2 in beide groepen afzonderlijk voor, en $\bar{Z} = p$ de proportie over beide groepen samen.

De eerste, preliminaire, vraag leidt nu tot de t -toets voor het verschil tussen twee proporties. Is er geen verschil, dan komt het Turner-syndroom in gelijke mate bij XX- en XO-vrouwen voor en kan het niet dienen als basis voor classificatie.

We zullen dit geval toch volledig uitwerken omdat we de essentie van discriminantanalyse reeds tot uiting kunnen brengen.

Voor ongepaarde steekproeven berekent men :

$$t^2 = \frac{p_1 - p_2}{\frac{m p_1 q_1 + n p_2 q_2}{m + n - 2} \left(\frac{1}{m} + \frac{1}{n} \right)} \quad (1)$$

We stellen dit nu in andere notaties voor :

men weet dat

$$s_w^2 = \frac{m p_1 q_1 + n p_2 q_2}{m + n - 2} \quad (2)$$

de binnenvariantie van Z voorstelt, d.i. het gewogen gemiddelde van de varianties $\frac{m p_1 q_1}{m - 1}$ en $\frac{n p_2 q_2}{n - 1}$ van resp. X en Y in beide groepen afzon-

derlijk (het moge wel vermeld worden, tegen veelverbreide formules in, dat de geschatte varianties niet $p_1 q_1$, $p_2 q_2$ zijn, en de geschatte variantie van een geschatte proportie niet $\frac{pq}{n}$ is, maar $\frac{pq}{n-1}$, d.i. $\frac{s^2}{n}$ mits $s^2 = \frac{npq}{n-1}$); verder kan

$$s_B^2 = m(p_1 - p)^2 + n(p_2 - p)^2, \quad (3)$$

de tussenvariantie, omgewerkt worden tot

$$s_B^2 = \frac{(p_1 - p_2)^2}{1/m + 1/n} \quad (4)$$

uit (1) tot (4) volgt dan

$$t^2 = \frac{s_B^2}{s_W^2} = F(1, m + n - 1) \quad (5)$$

Stelt men nu $d = p_1 - p_2 = \bar{X} - \bar{Y}$

dan noemt men $D^2 = \frac{d^2}{s_W^2} \quad (6)$

het afstandskwadraat van Mahalanobis tussen beide groepen, en

$$T^2 = \frac{D^2}{1/m + 1/n} = t^2 \quad (7)$$

de T^2 van Hotelling; in het speciale geval van één veranderlijke kan t zelf als een gemiddelde afstand beschouwd worden.

Men bepaalt nu een indelingsveranderlijke L , de zgn. (lineaire) discriminantfunctie door

$$L = aZ \quad (8)$$

met a bepaald uit $s_W^2 \cdot a = d \quad (9)$

Haar gemiddelden zijn

$$\bar{L}_X = a\bar{X} = ap_1 \quad \text{in de eerste groep}$$

$$\bar{L}_Y = a\bar{Y} = ap_2 \quad \text{in de tweede groep}$$

$$\text{en } \bar{L} = a\bar{Z} = ap \quad \text{voor beide groepen samen.}$$

Men bemerkt dat

$$\bar{L}_X - \bar{L}_Y = a(p_1 - p_2) = ad = \frac{d^2}{s_W^2} = D^2 \quad \text{wegens (9) en (6).}$$

Is nu voor een gegeven individu de L-waarde zó dat bvb. $\bar{L}_X < L < \bar{L}$ $< \bar{L}_Y$ dan behoort het tot de eerste groep, terwijl in het geval $L > \bar{L}$ met $\bar{L}_X < \bar{L} < \bar{L}_Y$ het tot de tweede behoort.

De gevallen $L < \bar{L}_X < \bar{L} < \bar{L}_Y$ en $\bar{L}_X < \bar{L} < \bar{L}_Y < L$ zijn evident.

Cijfervoorbeeld : $\bar{X} = p_1 = 0$, $\bar{Y} = p_2 = 0,9$

dan is

$$d = 0,9$$

$$p = \frac{300 p_1 + 40 p_2}{340} = \frac{36}{340}$$

$$s_w^2 = \frac{40 \times 0,9 \times 0,1}{338}$$

$$a = \frac{338}{4} = 84,5$$

zodat

$$\bar{L}_X = 0, \quad \bar{L}_Y = 76,5, \quad \bar{L} = 8,95;$$

een vrouw met Turner-syndroom ($Z = 1$) krijgt een L-waarde $L = 84,5 \times 1 = 84,5$ waardoor ze in de tweede groep valt.

Derde geval : één continue veranderlijke.

Het vorige geval gaat onverminderd door, mits de binnenvariantie s_w^2 op de geëigende wijze te berekenen; men kan weer een L-veranderlijke bepalen, wat echter overbodig is (zoals overigens in het vorige geval) omdat ze slechts op een schaalfactor na van Z verschilt.

Vierde geval : twee ongecorrelleerde continue veranderlijken.

Het is pas nu dat discriminantanalyse nuttig wordt, omdat ze de indeling op basis van meerdere veranderlijken vervangt door deze op basis van één L-veranderlijke.

Men bekomt deze L uit twee veranderlijken Z_1, Z_2 met gemiddelden $\bar{Z}_1, \bar{Z}_2$ over beide groepen samen, en $\bar{X}_1, \bar{X}_2$ in de eerste, $\bar{Y}_1, \bar{Y}_2$ in de tweede groep, en binnenvarianties $(11)_w, (22)_w$ door te stellen

$$L = L_1 + L_2$$

waarin L_1 en L_2 beide volgens (8) en (9) uit Z_1 resp. Z_2 bepaald worden, zodat

$$L = a_1 Z_1 + a_2 Z_2 \quad (10)$$

Hieruit volgen de gemiddelden

$$\bar{L}_X = a_1 \bar{X}_1 + a_2 \bar{X}_2$$

$$\bar{L}_Y = a_1 \bar{Y}_1 + a_2 \bar{Y}_2$$

$$\bar{L} = a_1 \bar{Z}_1 + a_2 \bar{Z}_2$$

waarmede classificatie op basis van $L \geq \bar{L}$ kan gebeuren zoals in het tweede geval.

De proportie P van misclassificaties wordt gegeven door

$$P = 1 - F(D/2)$$

waarin $F(x) = P(X \leq x)$ de cumulatieve Gaussverdeling voorstelt, en D volgt uit

$$D^2 = \bar{L}_X - \bar{L}_Y.$$

De preliminaire toets voor verschil steunt nu op

$$T^2 = t_1^2 + t_2^2 \quad (11)$$

De critische waarde voor T^2 volgt uit de betrekking

$$\frac{T^2}{m+n-2} \cdot \frac{m+n-p-1}{p} = F(p, m+n-p-1) \quad (12)$$

waarin $p = 2$ het aantal veranderlijken voorstelt; vermits nu

$$F(2, m+n-3) < F(1, m+n-2) < 2 F(2, m+n-3)$$

$$\text{is} \quad \frac{T^2}{2} \cdot \frac{m+n-3}{m+n-2} < t^2 < T^2 \cdot \frac{m+n-3}{m+n-2}$$

$$\text{zodat} \quad t_a < T_a < t_a \sqrt{2} \quad (13)$$

In een (t_1, t_2) -assenstelsel ligt het critisch gebied voor de marginale toetsen op basis van t_1 en t_2 samen buiten een vierkant met middelpunt in de oorsprong en halve zijde t_a , maar voor de gezamenlijke toets op basis van T^2 buiten de cirkelomtrek $t_1^2 + t_2^2 = T_a^2$ welke wegens (13) de zijden van het vierkant snijdt. Een zeer belangrijk gevolg hiervan is dat T^2 significant kan zijn terwijl t_1 en t_2 het beide niet zijn en omgekeerd.

Vijfde geval: twee continue gecorrelleerde veranderlijken.

L wordt nog steeds door (10) bepaald, maar a_1 en a_2 zijn nu de oplossingen van het stelsel

$$(11) a_1 + (12) a_2 = d_1$$

$$(21) a_1 + (22) a_2 = d_2$$

waarin (12) de binnencovariantie tussen Z_1 en Z_2 voorstelt.

Nu gaat (11) over in

$$T^2 = \frac{1}{1-r^2} (t_1^2 - 2r t_1 t_2 + t_2^2) \quad (14)$$

Het kritisch gebied voor de T-toets ligt nu buiten een ellips met vergelijking $T^2 = T_a^2$ waarin T^2 door (14) gegeven is; deze ellips heeft assenlengten $2 T_a \sqrt{1 \pm r}$ in de richtingen $\pm 45^\circ$ ingeschreven in een vierkant met zijden $2 T_a$.

Algemeen geval : p continue veranderlijken.

Notaties : $\underline{Z}$ kolomvector der veranderlijken

$\underline{Z}$ » » gemiddelden

$\underline{d}$ » » gemiddelde verschillen

$\underline{a}$ » » a-coëfficiënten

S binnen-variantie-covariantiematrix.

Dan is

$$\underline{a} = S^{-1} \underline{d}$$

$$\underline{L} = \underline{a}' \underline{Z}$$

$$\underline{D}^2 = \underline{a}' \underline{d} = \underline{d}' S^{-1} \underline{d}$$

$$T^2 = \frac{\underline{D}^2}{1/m + 1/n}$$

$$F(p, m+n-p-1) = \frac{m+n-p-1}{p} \cdot \frac{T^2}{m+n-2}$$

$$\frac{\frac{T^2}{m+n-2} / p}{1/(m+n-p-1)}$$

De laatste vorm voor F is analoog aan de F-toets voor het door p veranderlijken gezamenlijk verklaarde gedeelte van de kwadraatensom van een afhankelijke veranderlijke in meervoudige regressie, waarin de residuele

variantie aan 1 gelijk zou zijn (wat door schaalverandering trouwens steeds te bereiken is). Ook andere toetsen in meervoudige regressie kunnen door een soort dualiteitsprincipe omgewerkt worden tot toetsen voor discriminant-analyse.

Deze toetsen kan men aanwenden om bepaalde belangrijke vragen te beantwoorden. Het algemeen geval roept nl. nog verscheidene problemen op.

Een eerste volgt uit het algemeen wetenschappelijke principe dat overtollige of nutteloze gegevens buiten beschouwing moeten gelaten worden. Daarom werden de berekeningen stapsgewijze uitgevoerd, waarbij in een voorafbepaalde volgorde (nl. van verwacht meest- tot minst-discriminerende) de veranderlijken in de analyse werden betrokken. Bij iedere stap werd een discriminantfunctie L berekend, en werd getoetst of weglaten van telkens één der reeds opgenomen veranderlijken verantwoord kon worden; ook toetsen voor het weglaten van telkens méér dan een veranderlijke zijn mogelijk, maar werden niet uitgevoerd.

Een tweede probleem van vereenvoudiging is, of de berekende a -coëfficiënten door afgeronde getallen mogen vervangen worden zonder significant verlies aan informatie. Men toetst dan een opgelegde L -functie.

Tenslotte zijn er nog twee problemen van meer technische aard: de variantie-covariantiematrices van beide groepen mogen niet verschillen, wat reeds op eerste zicht waar bleek te zijn, en de veranderlijken moeten normaal verdeeld zijn. Daarom werden normaliteitstoetsen uitgevoerd met behulp van lineaire functies van resp. de derde en vierde momenten, functies die zelf standaardnormaal verdeeld waren. Deze toetsen brachten geen significante afwijkingen van normaliteit aan het licht, wat op grond van het groot aantal waarnemingen wel te verwachten was.

BIBLIOGRAPHIE

- H. Van den Berghe, *Het zogenaamde syndroom van Turner*, Standaard-Wetenschappelijke Uitgeverij, 1966.
- E. Meulepas, *Discriminant analysis and consumer price index*, in Tijdschrift voor Economie (te verschijnen).
- T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.
- C.R. Rao, *Advanced Statistical Methods in Biometric Research*, Wiley, New York, 1952.
- M.G. Kendall, *The Advanced Theory of Statistics*, vol. II, Griffin, London, 1948.
- C.C.Li, *Introduction to Experimental Statistics*, Mc Graw-Hill, New York, 1964.
- A. Hald, *Statistical Theory with Engineering Applications*, Wiley, New York, 1952.