KM 3(1981)
pag 9-16

ENIGE OPMERKINGEN OVER VERDELINGSVRIJE METHODEN IN DE TOEGEPASTE STATISTIEK [1]

Ph. van Elteren [2]

Omstreeks het jaar 1950 werd de Statistische Afdeling van het Mathematisch Centrum betrokken bij een luguber experiment met een primitieve diersoort: poliepen. Deze dieren werden onthoofd en kregen vervolgens gelegenheid hun kop in diverse milieus te regenereren. De vraag was in welke milieus de regeneratie significant sneller of trager verliep dan in het controle-milieu: fysiologisch zout. Een probleem hierbij was dat de diverse maten, die voor de regeneratiesnelheid werden voorgesteld verre van normaal verdeeld waren en dat er geen adekwate normaliserende transformaties konden worden bedacht. Daarom achtte men toepassing van de toets van Student voor twee steekproeven niet verantwoord.

De toenmalige souschef van de afdeling, Jan Hemelrijk stelde voor een enige jaren te voren beschreven methode van Mann en Whitney [1] toe te passen, daar deze geen eisen zou stellen aan het type verdeling van de waarnemingen. Het hoofd van de afdeling Prof. van Dantzig had daar weinig fiducie in omdat de vervanging van de waarnemingen door rangnummers, inherent aan de toets, op een te groot verlies van informatie zou neerkomen. Niettemin werd de toets tijdens de eerstvolgende vakantie van Van Dantzig met succes op de poliepen-gegevens toegepast. Dit was voor van Dantzig aanleiding een nadere studie van de toets van Mann en Whitney, later in Nederland meer bekend als de toets van Wilcoxon voor twee steekproeven, te entameren. Dit werd het begin van een periode van een tiental jaren, waarin de Statistische Afdeling van het M.C. het onderzoek vooral richtte op verdelingsvrije methoden en deze methoden in de praktijk bij voorkeur toepaste.

De verdelingsvrije toetsen waren destijds om de volgende redenen aantrekkelijk:

1. Bij de toen beschikbare reken-hulpmiddelen konden ze relatief gemakkelijk worden toegepast
2. Aan de hand van deze toetsen konden de beginselen van de toetsingstheorie worden gedemonstreerd zonder dat veel kennis over schattingstheorie of standaardverdelingen (t, χ^2, F) bekend moest worden veronderdersteld.
3. De verdelingsvrije methoden waren een aantrekkelijk studieobject voor

---

[1] Dit is gebaseerd op een lezing gehouden op de "Mathemedistica", een gezamenlijke bijeenkomst van de Mathematisch-Statistische en de Medisch-Biologische Secties van de VVS, op 18 maart 1981, te Amsterdam, met als onderwerp "Verdelingsvrije methoden nog steeds actueel?"

[2] Mathematisch-Statistische Adviesafdeling
Katholieke Universiteit Nijmegen

promovendi. De promoties van Nederlandse statistici in de vijftiger en begin zestiger jaren geven daar blijk van (Hemelrijk [3], van Eeden [4], Terpstra [5], Doornbos [6]).

De eerste aanzet voor de verdelingsvrije methoden was al gegeven door Fisher, die opmerkte dat, bijvoorbeeld in het geval van twee steekproeven uit dezelfde populatie, alle mogelijke permutaties van de gevonden waarnemingen over beide steekproeven even waarschijnlijk waren. Op grond van dit permutatiebeginsel kon hij een exacte toets voor twee steekproeven construeren gebaseerd op de gevonden waarden. Het bleek dat het resultaat van deze toets, wanneer de waarnemingen niet al te onregelmatige verdelingen hadden, redelijk overeenkwam met dat van de standaard t-toets. Dergelijke permutatie-toetsen zijn uitvoeriger bestudeerd door Pitman [6], maar het bleven methoden van theoretisch belang (ter rechtvaardiging van standaard toetsen bij twijfel aan de normaliteitsonderstelling) tot dat men er toe kwam de waarnemingen te vervangen door rangnummers (en later ook andere scores) naar opklimmende grootte. Dit had het voordeel, dat het mogelijk werd de permutatieverdeling van de toetsgrootheid eens en voor altijd te tabelleren. Een derde belangrijke ontwikkeling voor de verdelingsvrije methoden was de publikatie van limietstellingen voor de verdeling van de toetsgrootheden zowel onder de getoetste als onder alternatieve hypothesen. Daardoor is het mogelijk de permutatieverdeling van de toetsgrootheid te benaderen door standaardverdelingen met als parameters (in het geval van rangnummer scores) eenvoudige functies van steekproefgrootten, het aantal steekproeven en eventueel de omvang van de knopen.

Tegen het einde van de vijftiger jaren waren bij de statistische afdeling van het M.C. de volgende verdelingsvrije methoden gemeengoed geworden:

Toets voor twee steekproeven	: Wilcoxon
Toets voor $k \geq 3$ steekproeven, algemeen	: Kruskal-Wallis
tegen geordende alternatieven	: Terpstra, van Eeden
Toetsen voor gepaarde waarnemingen	: Tekentoets, Symmetrietoets van Hemelrijk en later: Rangtekentoets van Wilcoxon
Toetsen voor k steekproeven met een tweede klassificatie (één waarneming per cel) algemeen	: Friedman

tegen geordende alternatieven	: Combinatie van rangcorrelatie-toetsen van Kendall of Spearman
Rangcorrelatie methoden uit het boekje van Kendall [7]	: Kendall, Spearman, Paired comparisons, m rankings

Principe van lineaire combinatie van onafhankelijke toetsen.

Deze methoden werden gepropageerd in plaats de corresponderende standaard-toetsen. Een onderzoeker, die zijn waarnemingen al had samengevat in gemiddelden en standaardafwijkingen,werd verzocht om de individuele waarnemingsuitkomsten ter beschikking te stellen om daarop een verdelingsvrije toets te kunnen toepassen. Dat deze toetsen slecht aansloten bij vertrouwde grootheden uit de beschrijvende statistiek en de schattingstheorie werd nog niet als een nadeel gezien. Het toetsen van hypothesen was de hoofdzaak en dat ging uitstekend met verdelingsvrije methoden. Onderwijl werden in Engeland en Amerika en in Nederland o.a. bij Philips en te Wageningen overwegend standaard methoden, zoals variantieanalyse, bestudeerd en toegepast. In het M.C. groeide pas meer waardering voor deze methoden na een colloquium over een desbetreffend boekje van Mann [8] in de tweede helft van de vijftiger jaren. Toen bleek dat de manipulaties met kwadraatsommen en vrijheidsgraden op fraaie stellingen gebaseerd waren.

Er werd inmiddels ook getracht verdelingsvrije toetsen te construeren voor hypothesen vertrouwd uit de variantieanalyse . Het bleek in principe mogelijk hoofdeffecten verdelingsvrij te toetsen in een schema met twee ingangen en een willikeurig aantal waarnemingen per cel maar zonder interactie (Benard - van Elteren [9]). Maar een permutatie toets voor gelijkheid van varianties in variant voor verschillen in de locatie-parameters of een dergelijke toets voor afwezigheid van interactie invariant voor hoofdeffecten bleken niet mogelijk te zijn.
Na jaren slaagde de Statistische Afdeling van het M.C. erin een bijna onleesbaar geworden exemplaar van de lecture-notes van Pitman [10] te bemachtigen en te ontcijferen. Naar deze notes werd in veel publicaties over het onderscheidingsvermogen van verdelingsvrije toetsen verwezen. Uit de Pitman notes bleek dat een verdelingsvrije toets gebaseerd op rangnummers in veel gevallen slechts weinig minder efficiënt is dan de corresponderende standaardtoets ook in situaties waarin aan de modelveronderstellingen voor de standaardtoets is voldaan. Een en ander kon Pitman aantonen door berekening van de asymptotische relatieve doeltreffendheid van de verdelingsvrije toets ten op zichte van de standaardtoets en had

dus betrekking op situaties met grote steekproeven.

Bij kleine aantallen waarnemingen kunnen verdelingsvrije toetsen sterker in het nadeel zijn, dit o.a. omdat deze gebaseerd zijn op discreet verdeelde toetsgrootheden. Een toets met nominale onbetrouwbaarheidsdrempel α heeft dan in feite een onbetrouwbaarheid $\leq \alpha$ waardoor het onderscheidingsvermogen ongunstig kan worden beïnvloed. Dit bezwaar kan worden ondervangen door, als de toetsgrootheid een waarde juist buiten het kritieke gebied aanneemt, te loten of men de getoetste hypothese zal verwerpen. De kans op verwerpen wordt dan zo gekozen dat de onbetrouwbaarheid precies α wordt. In de praktijk heeft dat echter het nadeel dat bij dezelfde statistische gegevens met dezelfde toets niet steeds dezelfde conclusie wordt getrokken. Bij gevolg wordt deze randomseringsprocedure zelden toegepast.

Het is moeilijk waterdichte criteria op te stellen voor de keuze tussen standaard- en verdelingsvrije toetsmethoden. De statisticus in de praktijk kan bij de keuze worden geleid door de volgende overwegingen:

1. In groots opgezette onderzoekingen, waarbij de invloed van verschillende factoren gelijktijdig moet worden onderzocht en waar het verantwoord lijkt een gespecificeerd model te postuleren verdienen de standaardmethoden de voorkeur. In dergelijke situaties vormen deze methoden een overzichtelijk samenhangend systeem, waarop schattingen en betrouwbaarheidsintervallen voor de te onderzoeken parameters goed aansluiten.
2. In minder gecompliceerde onderzoekingen, waar de vraagstelling geen model met meer dan één factor vereist, komen verdelingsvrije toetsen in principe steeds in aanmerking. Ze verdienen dan de voorkeur als het gaat om variabelen die sterk en onregelmatige fluctuaties vertonen. In het medisch-biologisch toepassingsgebied is dat het geval met veel biochemische bepalingen (bijv. van hormonen).
3. Gedragswetenschappers stellen veelal dat bij variabelen met ordinale schalen alléén verdelingsvrije toetsen gebaseerd op rangnummers mogen worden toegepast. Hoewel dit niet zo exclusief kan worden gesteld dient verdelingsvrij toetsen bij dergelijke variabelen toch wel regel te zijn.
4. Verdelingsvrije toetsen zijn veelal aangewezen bij discrete of sterk geklassificeerde continue gegevens, vooral als de eindklassen open zijn. Rangnummers behoeven dan niet de beste scores te zijn maar de berekening van benaderende of exacte overschrijdingskansen is met de huidige hulpmiddelen ook bij toepassing van andere scores geen probleem meer.

In gevallen dat men de voorkeur zou moeten geven aan een verdelingsvrije methode wordt in de praktijk toch vaak een standaard-toets toegepast. Dit is gedeeltelijk te wijten aan de geringe aandacht besteed aan verdelingsvrije methoden in veel elementaire leerboeken. Redacties van tijdschriften zijn niet altijd op de hoogte van verdelingsvrije technieken en maken bezwaar tegen ingezonden artikelen met daarop gebaseerde statistische analyses.

Moderne electronische rekenapparatuur bevordert ook de toepassing van standaard-methoden. Een eenvoudig zak- of tafelrekenmachientje met enkele geheugens is voldoende om een t van Student, een correlatiecoëfficiënt van Pearson en regressiecoëfficiënten uit te rekenen met slechts één maal inslaan van de waarnemingen. Verscheidene machientjes bevatten zelfs toetsen voor het bepalen van overschrijdingskansen van t, χ^2 en F-verdelingen. Het rangschikken van waarnemingen naar grootte is met dergelijke apparatuur niet mogelijk en wordt dus een bewerking die vooraf "met de hand" moet gebeuren, waardoor verdelingsvrije methoden in het nadeel zijn .

Bij een computer van enige omvang is het rangschikken geen probleem meer. Maar de beschikbare programmatuur werkt het toepassen van verdelingsvrije methoden niet in de hand. In statistische programmapaketten worden gewoonlijk wel een aantal verdelingsvrije toetsen opgenomen maar de uitvoer laat te wensen over.

Het SPSS-pakket bevat bijvoorbeeld de volgende verdelingsvrije methoden:
De rangtekentoets van Wilcoxon,
de toets van Wilcoxon voor twee steekproeven,
de toets van Kruskal-Wallis,
de toets van Friedman,
de rangcorrelatietoetsen van Kendall en Spearman ,
maar bv geen k-steekproeven toetsen tegen geordende alternatieven, geen verdelingsvrije simultane toetsen voor contrasten, geen toetsen gebaseerd op lineaire combinaties van onafhankelijke toetsgrootheden.

Exacte overschrijdingskansen worden alleen berekend bij de toets van Wilcoxon voor twee steekproeven (niet bijvoorbeeld bij de rangtekentoets waar dit toch goed mogelijk is). Bij elk van de verdelingsvrije toetsen wordt slechts één benadering voor de overschrijdingskans berekend. Bij Friedman en de rangtekentoets wordt geen correctie voor knopen toegepast. Bij de

toets van Wilcoxon voor twee steekproeven wordt steeds de kleinste van de twee mogelijke toetsgrootheden berekend. Bij gevolg is de opgegeven waarde van de gestandaardiseerde toetsgrootheid steeds negatief. De gebruiker, die dit niet opmerkt, zal gemakkelijk misleid worden ten aanzien van de richting van een geconstateerd verschil.

De computer kan echter ook van groot nut zijn voor de toepassing van verdelingsvrije methoden

1. Met behulp van een computer kunnen overschrijdingskansen en permutatieverdelingen van toetsgrootheden bij betrekkelijk kleine steekproeven exact worden berekend en bij grotere aantallen waarnemingen met voldoende nauwkeurigheid worden geschat, door het trekken van een steekproef van voldoende grote omvang uit alle te beschouwen permutaties. Met behulp van de binomiale verdeling kan gemakkelijk worden aangetoond dat een steekproef van 20000 populaties voldoende is om een overschrijdinskans in de buurt van de 5% met een zekerheid van 99,9% met geen grotere fout dan 0,5% (dus een relatieve fout van niet meer dan 10%) te schatten. Deze nauwkeurigheid is voor de praktijk zeker voldoende en 20000 simulaties behoeven voor de doorsnee computer geen affaire te zijn.
2. De computer maakt het mogelijk meer recente ontwikkelingen in de verdelingsvrije statistiek in toepassing te brengen. Voor het medisch-biologisch toepassingsgebied zijn bijvoorbeeld de verdelingsvrije technieken voor vergelijkend levensduur-onderzoek van groot belang. Deze technieken zijn reeds in populaire statistische programmapaketten doorgedrongen.

   Niet zonder betekenis zijn ook de ontwikkelingen op het gebied van multivariate verdelingsvrije methoden. Multivariate versies van de Wilcoxon, Kruskal-Wallis en Friedman toets zijn in principe niet moeilijk te construeren. Een computer is nodig om de voor de toetsgrootheden vereiste matrix inversies uit te voeren en permutatieverdelingen te simuleren.

Het volgende eenvoudige voorbeeld van toepassing van een multivariate toets is ontleend aan de praktijk van de Mathematisch-Statistische Adviesafdeling van de K.U. te Nijmegen.

Bij 22 infertiele en normaal fertiele vrouwen, geselecteerd op een aantal relevante kenmerken (o.a. geen pilgebruik) werd een aantal hormoonbepalingen verricht op een vijftal tijdstippen in de menstruele cyclus. Per

hormoon werd dus een vector van vijf bepalingen verkregen. Het ging er om na te gaan of er significante verschillen bestonden tussen de beide groepen vrouwen met betrekking tot deze vectoren. Bij een vergelijkbaar onderzoek elders was per component een twee-steekproeven toets toegepast en besloten tot een verschil tussen beide groepen, als tenminste één van de vijf componenten een significant resultaat opleverde. Uiteraard is de simultane onbetrouwbaarheid van deze toets, indien per component op het 5%-niveau wordt getoetst, veel groter dan 5% en dienen de vijf componenten simultaan te worden getoetst. Boyett en Shuster [11] hebben een simultane permutatietoets beschreven gebaseerd op de grootste van de (in dit geval 5) toetsgrootheden van Student per component. Door het onregelmatige karakter van de verdelingen der hormoonbepalingen gaf deze toets onbevredigende resultaten. Daarom werd gekozen voor de permutatietoets gebaseerd op de grootste van de 5 toetsgrootheden van Wilcoxon voor twee steekproeven, waarmee inderdaad veel plausibeler resultaten werden verkregen. De overschrijdingskansen werden niet exact berekend, omdat daartoe $\binom{29}{7} = 1560780$ permutaties moesten worden onderzocht. In plaats daarvan werden de overschrijdingskansen geschat door na te gaan in welk percentage van 17000 aselecte trekkingen uit alle permutaties de grootste Wilcoxon-statistic tenminste gelijk was aan de daarvoor gevonden waarde. Tevoren kan worden gesteld dat de overschrijdingskans van de simultane toets groter moet zijn dan de kleinste overschrijdingskans van de Wilcoxon-toetsen voor de vijf componenten en kleiner dan vijfmaal deze overschrijdingskans. In een geval was deze kleinste overschrijdignskans $0{,}017 < 0{,}05$ terwijl $5 \times 0{,}017 = 0{,}085 > 0{,}05$. De schatting voor de simultane overschrijdingskans werd 0,06. Bij 17000 trekkingen zal deze schatting niet veel meer dan 0,5% afwijken, zodat in dit geval de getoetste hypothese bij $\alpha = 0{,}05$ niet kan worden verworpen. Voor nadere bijzonderheden over dit voorbeeld zie [12].

Misschien kan deze oppervlakkige excursie door de geschiedenis en de huidige praktijk van toepassing van verdelingsvrije methoden bijdragen tot een bevestigend antwoord op de vraag of verdelingsvrije methoden nog steeds actueel zijn.

Literatuur

[1] Mann,M.B. & Whitney,D.R., On a test whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18 (1947) 50-60.

[2] Hemelrijk,J. Symmetrietoetsen en andere toepassingen van de theorie van Neyman en Pearson. Ac. Proefschrift, Amsterdam 1950.

[3] Eeden,C. van, Testing and estimating ordered parameters of probability distributions, Ac. Proefschrift, Amsterdam 1958.

[4] Terpstra,T.J., Interprelations and Generalizations of Kendall's rank-correlation test, Proefschrift Groningen 1962.

[5] Doornbos R., Slippage tests, Ac. Proefschrift, Amsterdam 1966.

[6] Pitman,E.J.G., Significance tests which may be applied to samples from any populations. Suppl. J.R. Stat. Soc. 4 (1937) 119, II The correlation coefficient test Idem 4 (1937) 225,III The analysis of variance test Biometrika 29 (1938) 332.

[7] Kendall,M.G., Rank Correlation Methods, first edition Ch.Griffin and Co. Ltd. London (1948)

[8] Mann,M.B., Analysis and design of experiments, Dover Publications, New York (1949).

[9] Benard,A. and Elteren,Ph. van, A generalization of the method of m rankings, Proc. Kon. Ned. Ak. van Wet. A 56 en Indagationes Mathematicae 15 (1953) 359-369.

[10] Pitman,E.J.G., Lecture notes on nonparametric statistics (Lectures given for the University of North Carolina, Institute of Mathematics, 1948).

[11] Boyett,J.M. and Shuster,J.J., Non parametric one-sided tests in multivariate analysis with medical applications, J. Am. Stat. Ass. 72(1977) 665-668.

[12] Rounen,F.J.M.E., De fertiliteitsfunctie van de cervix uteri, Ac. Proefschrift Nijmegen (1980), in het bijzonder de statistische appendix van W. Doesburg en W. Lemmens.