

SAMENVATTING

De t-toets wordt gebruikt om op basis van een aselechte steekproef een statistische uitspraak te doen over de lokatieparameter van een normaal verdeelde populatie. Een experiment wordt in de praktijk veelal zo opgezet en uitgevoerd dat op basis daarvan uitspraken kunnen worden gedaan over meer dan 1 lokatieparameter en over meer dan 1 variabele. Het gebruiken van een generalisatie van de t-toets voor het evalueren van de resultaten ligt dan vanuit wiskundig oogpunt zeer voor de hand. De relevante verdelingstheorie is daarvoor beschikbaar: de F-verdeling, de verdeling van Hotelling's T^2 statistic en de verdeling van bijvoorbeeld Hotelling's gegeneraliseerde T^2 (c.q. Lawley-Hotelling's Trace statistic). In de statistische literatuur is zeer veel aandacht besteed aan de wijze waarop men, hiervan uitgaande, komt tot uitspraken over afzonderlijke contrasten met betrekking tot elk van de variabelen of lineaire combinaties daarvan. Deze zogenaamde simultane uitspraken waren en zijn aanleiding tot onevenredig veel controverse en verwarring in verhouding tot hun praktische bruikbaarheid. In zeer veel praktijksituaties in experimenteel onderzoek zal men bij een juiste probleemformulering geen behoefte hebben aan simultane toetsingsprocedures. Dit rapport bevat de aantekeningen over deze materie gemaakt ten behoeve van een voordracht voor een gezamenlijke bijeenkomst van de Landbouwkundige en Medisch-Biologische Sectie van de Vereniging van Statistiek op 14 oktober 1988.

1) Groep Landbouwwiskunde
Dienst Landbouwkundig Onderzoek
Postbus 100
6700 AC Wageningen, tel. 08370 - 19100

1. INLEIDING

De Student-toets en de daarvoor afgeleide betrouwbaarheidsintervallen kunnen worden gebruikt voor het doen van uitspraken over de lokatieparameter van 1 populatie of over het verschil tussen de lokatieparameters van 2 populaties, op basis van steekproeven uit die populaties en op basis van meting van 1 variabele. In de onderzoekspraktijk wordt meestal meer dan 1 variabele tegelijkertijd bestudeerd, terwijl het aantal te vergelijken populaties ook meestal meer dan 2 bedraagt. Het meest voor de hand liggend voor het doen van uitspraken in dergelijke situaties is de Student-toetsen toe te passen op elke variabele en op elk tweetal te vergelijken situaties afzonderlijk. Ook voor de hand liggend is het, zeker voor in statistische theorie geïnteresseerden, om gegeneraliseerde procedures te ontwikkelen, waarmee meer populaties en/of meer variabelen tegelijk kunnen worden beschouwd. Op basis daarvan worden vervolgens gevolgtrekkingen gemaakt over afzonderlijke variabelen en/of verschillen tussen populaties.

Helaas voeren beide methoden tot verschillende resultaten. De gegeneraliseerde methoden blijken tot voorzichtiger en derhalve minder scherpe uitspraken aanleiding te geven. In het oordeel van sommigen is dit een voordeel omdat men zichzelf op deze wijze beschermt tegen het extraheren van meer informatie dan de gegevens bevatten. Door het kiezen van een onbetrouwbaarheid voor het toetsen van een samengestelde hypothese bereikt men immers dat voor alle afgeleide uitspraken over deelhypothesen een simultane onbetrouwbaarheid geldt die ten hoogste gelijk is aan de gekozen onbetrouwbaarheid voor het toetsen van de samengestelde hypothese. Dit laatste is niet het geval als men voor iedere deelhypothese afzonderlijk een onbetrouwbaarheid kiest, onafhankelijk van het aantal te toetsen hypothesen. Het wel of niet toepassen van simultane procedures is sinds lang en nog steeds een controversieel punt in de statistische beroepsuitoefening, waarover overigens slechts weinig wordt gepubliceerd. Een recente referentie met daarin verdere verwijzingen is Perry (1986); door Carmer & Walker (1982) wordt het dilemma op een ludieke manier geschetst in de vorm van een statistisch sprookje. In tegenstelling hiermee is er wel een omvangrijke literatuur over simultane procedures en de onderlinge vergelijking daarvan, waaraan nog jaarlijks vele artikelen worden toegevoegd.

2. DE GENERALISATIES VAN DE T-TOETS

Beschouwd wordt een situatie waarin gegevens geklassificeerd zijn in M klassen ($M \geq 1$; elke klasse vertegenwoordigt 1 populatie) en bestaan uit waarnemingen aan p variabelen ($p \geq 1$). Alle gebruikelijke veronderstellingen van het klassieke model zijn van toepassing. Het aantal waarnemingen per klasse is constant en gelijk aan r . De gegevens kunnen worden samengevat in een tabel Y met Mp gemiddelden en een matrix S die een schatting is van de covariantiematrix van de p variabelen berustend op n vrijheidsgraden.

Gevraagd wordt een betrouwbaarheidsinterval voor de verwachtingswaarde van een gegeven lineaire combinatie van de variabelen en van de klassegemiddelden:

$$\hat{v} = a'Yb$$

waarin a en b willekeurig gekozen vaste vectoren zijn. De standaardafwijking van $\hat{v}$, $s_{\hat{v}}$, kan worden geschat op basis van S . Deze bedraagt:

$$s_{\hat{v}} = (a'a \cdot b'Sb/r)^{\frac{1}{2}}.$$

Als a en b inderdaad gegeven vaste vectoren zijn, heeft $\hat{v}/s_{\hat{v}}$ een t -verdeling en kan op basis daarvan een betrouwbaarheidsinterval worden opgesteld. Als a en/of b door de waarnemingen worden ingegeven en derhalve in feite stochastisch zijn kan een simultaan betrouwbaarheidsinterval voor $E\hat{v}$ worden gegeven door

$$\hat{v} \pm k s_{\hat{v}}$$

waarin k afhangt van α , M , p en n .

In veel praktijksituaties zal men er de voorkeur aan geven de mogelijke keuzen van de vector a op een of andere wijze te beperken. Dan hangt k af van $m = M - q$, waarin q een aantal onafhankelijke lineaire beperkingen voor a is; bijvoorbeeld als men zich beperkt tot contrasten tussen de klassen is $q=1$ en $m = M - 1$, of als a een gegeven vaste vector is wordt $m=1$.

Met Y kunnen mp onafhankelijke lineaire combinaties $\hat{v}$ worden gevormd. Om aan de eis van een simultane onbetrouwbaarheid ter grootte van α te voldoen kan men daarom voor grote n kiezen:

$$k = \sqrt{\chi_{mp}^2(1-\alpha)}$$

Onderstaande tabel geeft voor een aantal situaties de waarden voor k bij $n=\infty$. Vermeld is tevens bij welk toetsingskriterium men deze waarde van k vindt bij $n=\infty$ en tussen haakjes de naam waaronder $ks_{\hat{v}}$ bekend staat.

p	m	pm	$\chi^2_{pm}(.95)$	k	toetsingskriterium
1	1	1	3.84	1.69	Student's t (LSD)
1	4	4	9.49	3.08	Snedecor's F (Scheffé's S)
4	1	4	9.49	3.08	Hotelling's T^2
2	2	4	9.49	3.08	Lawley-Hotelling's Trace T of Hotelling's generalized T^2
3	3	9	16.9	4.11	
4	4	16	26.3	5.13	
5	5	25	37.7	6.14	

Voor kleine n moet men voor k de tabellen van de genoemde verdelingen raadplegen, zie bijvoorbeeld Anderson (1984). De waarden van k bij kleine n zijn groter en nemen sneller toe met m en p dan bij $n=\infty$.

De bovengenoemde procedures zijn gebaseerd op toetsen die gevoelig zijn voor algemene alternatieven. Er zijn ook vele andere toetsingskriteria en daarop gebaseerde k-waarden gangbaar. De bekendste daarvan zijn

- Tukey's T voor paarsgewijze vergelijkingen tussen gemiddelden, gebaseerd op de verdeling van de "Studentized range",
- Dunnett's many-one statistics voor vergelijking van een aantal behandelingsgemiddelden met 1 controle,
- Roy's Maximum Root statistic die gevoelig is voor paarsgewijze vergelijkingen tussen gemiddelden t.a.v. een hoofdrichting van de variabelen,
- Wilks' Λ , evenals Lawley-Hotelling's Trace T geschikt voor toetsen tegen algemene alternatieven; voor deze toets is geen algemeen recept voor het berekenen van k bekend.

Voor een overzicht van de methoden kan men raadplegen Anderson (1984) voor multivariate situaties en Miller (1966) of Hochberg & Tamhane (1987) voor univariate situaties.

3. GEBRUIKSEIGENSCHAPPEN VAN DE SIMULTANE PROCEDURES

De gedachte achter de simultane procedures is dat deze de gezamenlijke onbetrouwbaarheid van alle uitspraken over een meer dan 1 parameter omvattende nulhypothese beperken tot een vooraf te kiezen waarde α . Hieronder wordt verstaan: de kans dat bij 1 of meer van de uitspraken tot het bestaan van een afwijking van de hypothese omtrent EY wordt besloten, terwijl in werkelijkheid de hypothese over EY geldig is, is ten hoogste gelijk aan α . Alle genoemde simultane procedures, maar ook nog vele andere niet genoemde procedures, voldoen hieraan. De prijs die voor deze protectie betaald moet worden is een

beperking van de scherpte van de uitspraken: de lengten van de op simultane procedures gebaseerde intervallen voor afzonderlijke contrasten nemen snel toe met p_m (k is ruwweg gelijk aan $1 + \sqrt{p_m}$). Anders gezegd: simultane uitspraken gaan ten koste van het onderscheidingsvermogen. Bij wijze van voorbeeld: zoals uit de tabel in de vorige paragraaf blijkt, is de efficiëntie van het toetsen van een contrast onder condities met $p_m=25$ ongeveer $9 \times$ zo klein dan onder condities met $p_m=1$!

Om deze in de praktijk ongewenste conservatieve eigenschap zoveel mogelijk te ontgaan zijn allerlei constructies bedacht, die soms bizar en theoretisch moeilijk te rechtvaardigen zijn. De bekendste hiervan zijn de breedtetoets (Student-Newman-Keuls-test) en Duncan's New Multiple Range Test. Dergelijke procedures hebben met elkaar gemeen dat ze in hun prestaties wat meer neigen naar die van een LSD maar toch onder de vlag van een vooraf in te stellen simultane onbetrouwbaarheid varen. Om te zien waar deze procedures in meerdere of mindere mate te kort schieten kan wellicht het best de volgende liberale procedure voor het vergelijken van een aantal gemiddelden worden beschouwd, die aan de gestelde eisen voldoet:

- voer eerst een F-toets uit met onbetrouwbaarheid α ,
- als de F-toets wordt verworpen worden met de LSD de afzonderlijke contrasten tussen de gemiddelden getoetst, elk met onbetrouwbaarheid α .

Het is duidelijk dat de geclaimde simultane onbetrouwbaarheid voor deze als FSD aan te duiden procedure slechts gerealiseerd wordt als de te toetsen hypothese omtrent EY strikt geldt. Zodra op een beperkte deelverzameling van de parameters in EY een voldoende grote afwijking optreedt om de F-toets met grote kans te verwerpen is er voor het toetsen van de rest van de parameters geen wezenlijke beperking meer van de gezamenlijke onbetrouwbaarheid. In grote lijnen wordt dit ook bevestigd door de resultaten van een simulatiestudie van Carmer & Swanson (1973); zie onderstaande tabel met de aan deze studie ontleende resultaten voor het toetsen van de verschillen tussen 10 behandelingsgemiddelden, die onder H_0 alle gelijk zijn ($\mu = 100$) en onder H_1 uiteenvallen in 2 groepen van 5 gelijke gemiddelden ($\mu_1 = 112$, $\mu_2 = 88$). In deze tabel zijn de gerealiseerde waarden van de onbetrouwbaarheid gegeven bij een ingestelde waarde $\alpha = 0.05$.

	gerealiseerde onbetrouwbaarheid per contrast		per experiment	
	H ₀	H ₁	H ₀	H ₁
Scheffé	.00		.00	.00
Tukey	.00		.05	.03
Student-Newman-Keuls	.00		.05	.11
Duncan	.03		.37	.33
FSD	.01	.05	.05	.45
LSD	.05		.58	.45

De procedures van Duncan en Student-Newman-Keuls nemen duidelijk een positie in tussen Scheffé/Tukey aan de conservatieve kant en FSD/LSD aan de liberale kant.

Ruwweg kan gesteld worden dat de procedures minder te rechtvaardigen zijn als simultane procedure naarmate deze liberaler worden met betrekking tot het onderscheidingsvermogen per contrast. De conclusie lijkt derhalve te moeten zijn dat, als een toets van een hypothese betreffende een vector van parameters gewenst wordt, het daarbij bijbehorende slechtere onderscheidingsvermogen voor specifieke contrasten moet worden geaccepteerd. Daarom is het van groot praktisch belang na te gaan in welke situaties het toetsen van samengestelde hypothesen en vervolgens het gebruiken van simultane procedures wenselijk of noodzakelijk is.

4. MOTIVERING VOOR HET AL DAN NIET GEBRUIKEN VAN SIMULTANE PROCEDURES

De beschikbaarheid van mathematisch statistische theorie voor het toetsen van hypothesen over parametervectoren is op zichzelf niet voldoende rechtvaardiging voor het gebruiken daarvan in alle situaties waarin sprake is van meer dan 1 parameter. Ook het al of niet tegelijkertijd in een experiment verkrijgen van informatie over een aantal parameters is in principe niet van belang voor de keuze van de te toetsen hypothesen. Alleen doel en vraagstelling van het onderzoek dienen bepalend te zijn voor de statistische analysemethoden die worden toegepast, uiteraard binnen de beperkingen die de experimenteertechniek en de aard van de gegevens opleggen. Dit brengt met zich mee dat de bespreking aan de hand van voorbeelden dient te geschieden.

4.1 Multivariate analyses in experimenteel onderzoek

In een veel geciteerd artikel wordt door Finney (1956) krachtig stelling genomen tegen het gebruik van multivariate variantie-analyse bij landbouwkundige experimenten. Zijn betoog komt er op neer dat de interesse van de onderzoeker uitsluitend ligt in conclusies over de door deze te definiëren variabelen als zodanig. Het door hem gekritiseerde voorbeeld betrof de opbrengsten in 2 opeenvolgende jaren van dezelfde veldjes in een proef met een meerjarig voedergewas. Analyses van som en verschil van deze beide opbrengsten liggen zonder meer voor de hand en zijn direct interpreteerbaar. Tevens is duidelijk dat het belang van de analyses op de som en op het verschil niet gelijk is: de analyse van de som geeft de hoofdresultaten van het experiment, de analyse van het verschil heeft een meer diagnostische betekenis. Er is dan ook geen houdbare motivering te bedenken voor het doen van uitspraken over sommen en verschillen tegelijk, met een gezamenlijke onbetrouwbaarheid. Het behoeft ook nauwelijks betoog dat andere lineaire combinaties van de beide variabelen (zoals die gesuggereerd door de beide hoofdassen) eerder het zicht op relevante resultaten belemmeren dan verhelderend werken.

Het voorbeeld van Finney is een speciaal geval van wat ook wel profiel-analyse of analyse van herhaalde waarnemingen wordt genoemd. Daarbij wordt aan elke experimentele eenheid een aantal in de tijd opeenvolgende waarnemingen verricht zoals bijvoorbeeld groeigegevens of gegevens over voedselopname of over de toestand van een dier of gewas. En de onderzoeksvraag kan zijn: "verschillen de behandelingen met betrekking tot hun profiel?" zonder verdere specificatie. In dat geval zou een multivariate toets op zijn plaats kunnen zijn en kan men bij het nader bestuderen van door de gegevens gesuggereerde aspecten van profielen, waarin de behandelingen verschillen terecht een simultane procedure gebruiken. Tegelijkertijd kan het nuttig zijn zich af te vragen of men zich, gegeven de weinig specifieke vraagstelling, niet eerder in een exploratieve fase van het onderzoek bevindt dan in een confirmatieve en of men zich derhalve niet onnodig voorzichtig opstelt door een simultane onbetrouwbaarheid te hanteren. Het ligt overigens veel meer voor de hand te verwachten dat a priori enkele specifieke aspecten van het profiel kunnen worden aangegeven waarvoor op grond van de onderzoekdoelstelling meer belangstelling bestaat dan voor andere. Het is in dat geval zeker niet doelmatig het onderscheidingsvermogen voor die aspecten te beperken door vast te houden aan een simultane procedure voor alle mogelijke niet vooraf gespecificeerde aspecten.

Een ander voorbeeld betreft proeven met mestvarkens. Belangrijke variabelen daarbij zijn het aantal dagen (n) dat een varken nodig heeft om van 25 kg tot 100 kg te groeien (vaak gegeven als groei per dag = $75/n = y_1$) en de hoeveelheid voer die daarvoor wordt gebruikt (gegeven als de voederconversie = hoeveelheid voer in kg per 75 kg groei = y_2). Deze variabelen zijn elk voor zich van belang: y_1 is een maat voor de benutting van de stalcapaciteit, y_2 is een maat voor de directe kosten. Economisch kan ook een gegeven combinatie van y_1 en y_2 betekenis hebben, bijvoorbeeld $c = C_1 \cdot n + C_2 \cdot y_2$ met C_1 en C_2 gegeven kosten per eenheid. Er is echter geen reden voor een multivariate analyse van y_1 en y_2 . Deze zou immers niets toevoegen aan de beschrijving van de verschillen tussen de behandelingen met betrekking tot de genoemde variabelen afzonderlijk of tot de combinatie c . Ook dient te worden opgemerkt dat meestal niet zozeer het toetsen van het bestaan van verschillen tussen de behandelingen dan wel het schatten van verschillen met hun precisie van belang is. Bij dit laatste is een multivariate analyse van geen enkel nut.

4.2 Analyse van kwantitatieve factoren

Bij de analyse van de responsie van kwantitatieve factoren ligt het voor de hand gebruik te maken van regressiemodellen en daarbij behorende schattings- en toetsingsmethoden. Paarsgewijze vergelijken van de instellingen van een kwantitatieve factor is niet doelmatig en dus zeker ook niet het gebruiken van daarop gerichte simultane toetsingsprocedures. Dit inzicht wordt blijkbaar niet algemeen geaccepteerd, zie bijvoorbeeld Peterson (1977) en Dawkins (1983). Afgezien daarvan blijft wel de vraag actueel of men bij het gebruik van regressiemethoden wel of niet simultane procedures zou moeten toepassen. Het antwoord op die vraag moet zijn dat dit afhangt van het doel van het onderzoek.

Merkwaardigerwijze lijkt de praktijk hieraan veelal voorbij te gaan. Zodra een behandelingseffect kan worden opgesplitst in lineaire, kwadratische en eventuele hogere orde componenten wordt bij het gebruik van variantie-analyse meestal afgezien van simultane procedures. Het is zeer gebruikelijk die componenten zonder nadere motivering elk afzonderlijk met de t -toets (of met een F -toets met 1 vrijheidsgraad in de teller) met onbetrouwbaarheid α te toetsen. Consequenter is het onderscheid te maken tussen bijvoorbeeld de volgende 3 situaties (waarbij gemakshalve is aangenomen dat het gaat om een factor op 3 niveaus waarvoor een lineaire en een kwadratische component kunnen worden berekend):

- men weet al dat de responsie dient te worden beschreven met een 2^e graads polynoom, zeg $E_y = a + bx + cx^2$. In dat geval is er alleen behoefte aan schatten van de parameters met hun standaardafwijking en eventuele onderlinge correlatie. Het is denkbaar dat men voor deze situatie een betrouwbaarheidsinterval wenst op te geven voor de vector van regressiecoëfficiënten (a, b, c). Als het gaat om vergelijkend onderzoek, zoals meestal het geval is, is het zeer de vraag of dit zinvol is. Zeker de waarde van parameter a heeft slechts betekenis onder de omstandigheden van het experiment. Dit geldt ook, wellicht in mindere mate, voor b en c. Het opgeven van een betrouwbaarheidsinterval voor de vector (a, b, c) of eventueel (b, c) kan licht de suggestie wekken dat meer dan lokale betekenis moet worden gehecht aan de conclusies. Afgezien hiervan is het in deze situatie zeker niet interessant een betrouwbaarheidsinterval voor de parameters afzonderlijk op te geven met simultane onbetrouwbaarheid.
- men weet dat de responsie kan worden beschreven met tenminste een 1^e graads polynoom. De toets op de 2^e graads term heeft dan ofwel betekenis als een routinematige controle op de geldigheid van het 1^e graads model ofwel wordt gebruikt om per geval te beslissen of een 1^e of 2^e graads model zal worden aangepast. In beide gevallen ligt een afzonderlijke toets op de 2^e graads term, met een apart daarvoor vast te stellen onbetrouwbaarheid, voor de hand. Na het uitvoeren van deze toets wordt het probleem een schattingsprobleem: zie boven.
- men tast nog volledig in het duister omtrent de te verwachten responsie, zoals bijvoorbeeld het geval zou kunnen zijn bij ondergeschikte variabelen die als extra in het hoofdonderzoek zijn meegenomen. In dat geval zijn simultane procedures methodologisch zeer wel verdedigbaar als men tot het bestaan van effecten zou willen besluiten. Alleen als het al dan niet bestaan van een effect van geen belang geacht wordt en geen aanleiding zal geven tot nader onderzoek is deze situatie denkbaar. Als het praktisch wel van belang is wat het effect van de onderzochte variabele is zal men, in deze blijkbaar nog exploratieve fase, juist niet het onderscheidingsvermogen wensen te beperken door een simultane onbetrouwbaarheid te stellen.

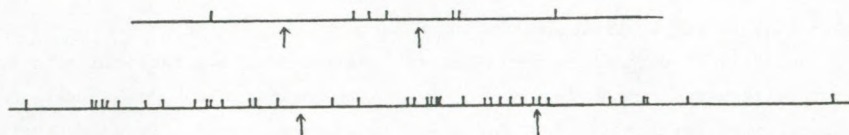
4.3 Analyse van kwalitatieve factoren

Duidelijke voorbeelden van situaties met kwalitatieve factoren zijn het vergelijken van rassen, het vergelijken van laboratoria of meetmethoden, het vergelijken van insecticiden, en dergelijke. Wanneer het gaat om het evalueren

van de responsies van bijvoorbeeld rassen of genotypen kunnen verschillende situaties aan de orde zijn:

- De vraag kan zijn of de genetische variatie in de genotypen voldoende groot is om daarin te gaan selecteren. In dat geval is men niet geïnteresseerd in afzonderlijke verschillen tussen de genotypen, maar in het schatten van slechts 1 parameter, namelijk de genetische variantie. Dit kan zonodig worden vooraf gegaan door een F-toets om vast te stellen dat deze tenminste een voor de praktijk interessante grootte overschrijdt. De strategie voor het vervolgens selecteren van de beste genotypen wordt bepaald door andere overwegingen dan de uitkomsten van significantietoetsen en betrouwbaarheidsintervallen.

Het is niet onmogelijk dat men in deze situatie geïnteresseerd is in het uitvoeren van een clusteranalyse op basis van de waargenomen opbrengsten zoals voorgesteld door Calinski & Corsten (1985) of Corsten (1988), hoewel dit in de praktijk wellicht eerder en beter kan gebeuren op basis van de bekende genetische herkomst van de rassen en/of op allerlei andere gebruikseigenschappen van de rassen. Als men wel het groeperen van rassen beoogt op basis van de waargenomen opbrengsten, zal dit eerder een exploratieve operatie zijn dan een confirmatieve. Het idee bij deze procedure is het vinden van een aantal clusters en een daarbij behorende optimale clusterindeling van de rassen zodanig dat de kwadraatsom van verschillen tussen de rasgemiddelden binnen de clusters niet meer significant is bij het uitvoeren van een F-toets met een gekozen waarde van α ; door de auteurs wordt $\alpha = 0,25$ gekozen. Een probleem met deze methode lijkt te zijn dat het aantal te vinden clusters altijd ruwweg gelijk zal zijn aan $\sqrt{F/F(\alpha)}$, ongeacht of er een herkenbare groepsstructuur in de behandelingsgemiddelden aanwezig is. In onderstaande figuur zijn de gemiddelden van 2 voorbeelden uit Corsten (1988) uitgezet samen met de door de auteur gesuggereerde verdeling in 3 clusters, aangegeven met $\uparrow$ als scheidingsteken. De praktische relevantie van zo'n resultaat zal vermoedelijk beperkt zijn.



Veel aantrekkelijker lijkt het toepassen van een grafische procedure waarin rasgemiddelden worden uitgezet tegen normal scores, zoals voorgesteld door Plackett (1971; zie ook Perry, 1986, waaraan onderstaande figuur is ontleend: de getrokken lijnen komen overeen met de standaardafwijking van de gemiddelden). Desgewenst kan men nog een formele F-toets berekenen als een numerieke bevestiging van het succes van de exploratief bepaalde clustering.

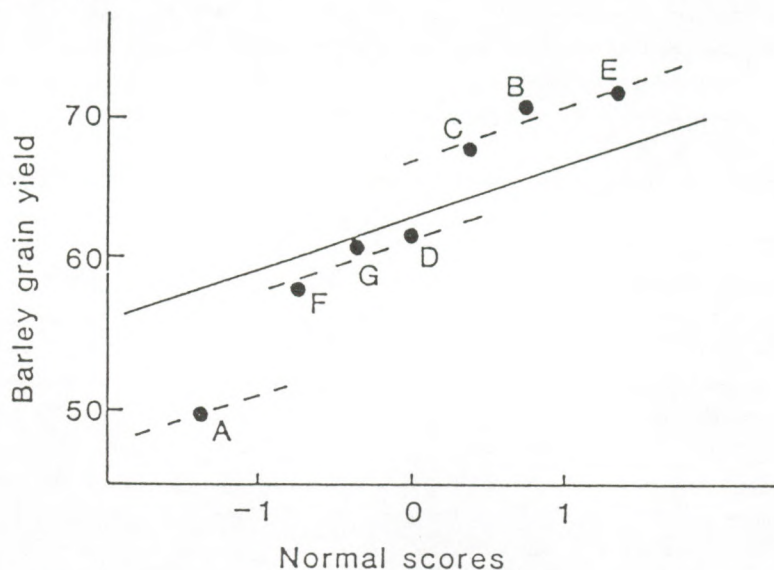


Fig. 2. Ranked mean barley grain yields in bushels per acre (Duncan 1955) plotted against normal scores

- Als het de bedoeling is vast te stellen of de onderzochte rassen voldoende verschillen van een of meer standaardrassen om te worden toegelaten, dient ieder nieuw ras op zijn eigen merites te worden beoordeeld onafhankelijk van of al dan niet tegelijk ook andere rassen worden onderzocht. In feite dient men deze situatie te benaderen met besliskundige methoden waarin verliesfuncties van producenten en/of gebruikers een belangrijke rol spelen. Als men te werk gaat met simultane procedures is het gevolg dat de afweging van de belangen van producenten en gebruikers afhankelijk gesteld wordt van het aantal te vergelijken rassen uitsluitend op basis van de gekozen statistische analyse en dat kan niet de bedoeling zijn.

- Als het er om gaat vast te stellen welke van een aantal met voorbedachte rade in het experiment opgenomen rassen het best geschikt is voor een bepaald doel zal men in principe geïnteresseerd zijn in bepaalde, zo niet alle, rasvergelijkingen afzonderlijk. De precisie van elke beoogde vergelijking is onafhankelijk van welke vergelijkingen nog meer worden gemaakt in hetzelfde experiment. Het is moeilijk in te zien waarom de grootte van een betrouwbaarheidsinterval wel daarvan zou moeten afhangen. Zie voor een humoristische bespreking van deze situatie Carmer & Walker (1982). In dit geval is zeker geen sprake van een situatie waarin men uitspraken wil doen over van te voren onbenoemde contrasten waarvan de coëfficiënten door de waarnemingen zelf worden ingegeven en derhalve stochastisch zijn.

Bovenstaande discussie is beperkt gehouden tot het vergelijken van rassen en genotypen, omdat daarbij diverse aspecten kunnen worden belicht. In praktijksituaties met andere kwalitatieve factoren zal steeds een van de hier besproken mogelijkheden (of wellicht nog andere) aan de orde zijn. Bijvoorbeeld bij het vergelijken van meetmethoden of laboratoria kan het niet doelmatig zijn minder goed naar de verschillen te kijken naarmate er meer in het onderzoek worden betrokken. Vrijwel steeds hebben vergelijkende onderzoeken betrekking op met naam bekende preparaten, merken, soorten, werkwijzen, etc. Er lijkt dan nauwelijks plaats voor het uitvoeren van simultane toetsingsprocedures. Dit is ook de boodschap in een commentaar van Anscombe (1965) waaruit de volgende passage wordt genomen als illustratie van de gekunsteldheid van een situatie waarin mogelijk wel vruchtbaar gebruik gemaakt zou kunnen worden van simultane toetsen.

"The following has been suggested to me (by Richard Olshen) as a case for the defense. A panacea manufacturer advertises on television that trials have shown his product to be more effective than any other leading brand. Such trials (if they are not a downright fabrication) certainly seem to present a situation of the third type. Their objective is not to help the manufacturer reach a decision, but hopefully to permit him to make a multiple-comparison statement that will impress the public and boost sales. He could appropriately use the simultaneous confidence intervals of this paper; indeed, the Food and Drug Administration could appropriately require him to do so. The more equally ineffective other leading brands there were, the harder would it be for him to obtain the evidence he needed, and the more trials would he have to conduct and suppress before achieving a favorable one. Thus would Statistics and Economics go hand in hand to protect the public".

4.4 Analyse van gestructureerde behandelingen

De in 4.1 besproken situatie van kwantitatieve factoren is een voorbeeld waarin een zekere structuur in de behandelingen aanwezig is. In doelmatig opgezette experimenten worden zeer vaak bepaalde structuren aangebracht met het oog op de doelstellingen. Zeer gebruikelijke zijn factoriële proeven waarin alle combinaties van de instellingen van een aantal factoren tegelijk worden onderzocht. In dergelijke proeven zijn niet alle vergelijkingen op gelijke wijze interessant. Dit geeft aanleiding om bij de analyse rekening te houden met de aangebrachte structuur, bijvoorbeeld door de hoofdeffecten van de factoren en de interacties tussen de factoren elk afzonderlijk te bezien.

Een bespreking van de analyse van interacties kan niet veel nieuwe gezichtspunten aan de orde brengen, maar is wellicht toch gerechtvaardigd omdat het hier al gauw gaat om zeer grote aantallen parameters waarover uitspraken worden gedaan. Wanneer het interacties tussen kwantitatieve factoren betreft zal men trachten deze bijvoorbeeld in termen van polynomen te beschrijven. In dit geval is de discussie van 4.2 volledig relevant.

Ook in situaties met kwalitatieve factoren kan het zinvol zijn te trachten de interacties te modelleren, bijvoorbeeld door multiplicatieve modeltermen op te nemen en deze afzonderlijk te schatten en te toetsen. Zie bijvoorbeeld Corsten & van Eijnsbergen (1972) en Johnson & Graybill (1972), die overigens methoden beschrijven voor het geval zonder herhalingen. Er is echter geen reden dergelijke modellen niet te gebruiken als wel herhalingen beschikbaar zijn; bijvoorbeeld voor het bestuderen van milieu-genotype-interactie lijken deze modellen relevant.

Het toetsen van alle interactie-parameters samen met een F-toets is alleen toelaatbaar als een routinematige aanpassingscontrole beoogd wordt van een model waarin de effecten van de factoren additief zijn verondersteld. Zodra het gewenst is een interactie min of meer exploratief te gaan onderzoeken is het niet doelmatig dit te doen met simultane procedures: de kans dat relevante effecten niet worden ontdekt stijgt snel met het aantal parameters: zie de tabel in paragraaf 2.

Een bijzonder voorbeeld is wellicht het volgende. In een zaai- en oogsttijdenproef voor spinazie worden een aantal combinaties van zaaitijden en oogsttijden beproefd. Opsplitsing van de waargenomen effecten in hoofdeffecten

en interacties is in zo'n geval niet erg zinvol omdat men alleen in combinaties is geïnteresseerd. Er is eigenlijk ook niet een toetsingsprobleem: men wenst slechts te weten wat bij iedere combinatie de spinazie-opbrengst is als uitgangspunt voor bedrijfseconomische planning en beschouwingen. Of een bepaalde combinatie meer opbrengt dan een andere is van ondergeschikt belang omdat de combinaties meestal geen alternatieven zijn waaruit men vrijelijk kan kiezen. Het opgeven van schattingen met hun precisie voor afzonderlijke combinaties lijkt alles wat men zich als samenvatting van de proefresultaten zou kunnen wensen.

Het rekening houden met structuur in de behandelingen is niet beperkt tot gebalanceerde situaties met kwantitatieve factoren. Perry (1986) geeft het volgende voorbeeld m.b.t. insectenvangsten met feromonen-vallen:

<u>preparaat + variant</u>	<u>onduidelijke samenvatting</u>		<u>alternatief</u>	
	<u>gemiddelde</u>	<u>Duncan</u>	<u>gem.</u>	<u>stafw</u>
A zonder anti-oxidant	1.380	a	}	1.395 .031
A met anti-oxidant	1.410	a		
B erg verweerd, snelle afgifte	1.524	a b	}	1.525 .031
B erg verweerd, normale afgifte	1.527	a b		
B matig verweerd	1.627	b c		1.627 .044
B niet verweerd	1.817	c		1.817 .044

Het is evident dat bij veronachtzamen van de structuur in de behandelingen de beschikbare informatie zelfs niet met de liberale Duncan-procedure duidelijk kan worden gepresenteerd.

5. BESLUIT

In zeer veel situaties in experimenteel onderzoek zal men bij een juiste probleemformulering geen behoefte hebben aan simultane toetsingsprocedures. Veelal geeft het schatten van parameters met een indicatie van de bijbehorende standaardafwijking een afdoende samenvatting van de onderzoeksresultaten. Toetsen zijn vaak alleen van belang bij het vergelijken van modellen of bij het vaststellen of een model wel past. Zelden treft men een situatie aan waarin een toets van een hypothese over een parametervector gewenst is en waarin het tevens gewenst is vervolgens een simultane procedure te gebruiken voor het toetsen van afzonderlijke contrasten. Het lijkt daarom beter als standaardprocedure geen gebruik te maken van simultane toetsingsprocedures (en dus zelfs niet van F-toetsen!) en daar alleen toe over te gaan als een onontkoombare motivatie daarvoor gegeven kan worden. In de regel kan volstaan

worden met het opgeven van standaardafwijkingen van verschillen, of desgewenst LSD-waarden, als een indicatie voor de orde van grootte van verschillen die een nadere beschouwing waard zijn.

6. LITERATUUR

- Anderson, T.W.; An introduction to multivariate statistical analysis. 2nd ed., Wiley, 1984.
- Anscombe, F.J.; Comments on Kurtz-Link-Tukey-Wallace Paper. *Technometrics* 7, 167-168, 1965.
- Calinski, T & Corsten, L.C.A.; Clustering means in ANOVA by simultaneous testing. *Biometrics* 41, 39-48, 1985.
- Carmer, S.G. & Swanson, M.R.; An evaluation of ten pairwise multiple comparison procedures by Monte-Carlo methods. *JASA* 68, 66-74, 1973.
- Carmer, S.G. & Walker, W.M.; Baby Bear's dilemma: a statistical tale. *Agronomy Journal* 74, 122-124, 1982.
- Corsten, L.C.A.; Clustering treatments in non-standard design. In: H.H. Bock (Ed.) *Classification and related methods of data analysis*. Elsevier, 1988, p. 187-194.
- Corsten, L.C.A. & van Eijnsbergen, A.C.; Multiplicative effects in two-way analysis of variance. *Statistica Neerlandica* 26, 61-68, 1972. (Addendum in 28, 51, 1974).
- Dawkins, H.C.; Multiple comparisons misused: why so frequently in response curve studies? *Biometrics* 39, 789-790, 1983.
- Finney, D.J.; Multivariate analysis and agricultural experiments. *Biometrics* 12, 67-71, 1956.
- Hochberg, Y. & Tamhane, A.C.; Multiple comparison procedures. Wiley, New York, 1987.
- Johnson, D.E. & Graybill, F.A.; Estimation of σ^2 in a two-way classification model with interaction and no replication. *JASA* 67, 388-394, 1972.
- Miller, R.G.; Simultaneous statistical inference. McGrawHill Inc., New York, 1966.
- Plackett, R.L.; Contribution to discussion of paper by O'Neill and Wetherill. *JSSR B* 36, 242-244, 1971.
- Peterson, R.G.; Use and misuse of multiple comparison procedures. *Agronomy Journal* 69, 205-208, 1977.
- Perry, J.N.; Multiple-comparison procedures: a dissenting view. *J. Economic Entomology* 79, 1149-1155, 1986.

Ontvangen: 14-10-1988
Geaccepteerd: 25-03-1990