KM 30(1989)
pag 21 - 32

Multiple Comparisons met SAS

Jan B. Dijkstra

Samenvatting

Naast enkele minder bekende methoden biedt SAS voor het simultaan vergelijken van gemiddelden de volgende toetsen: (1) Fisher's Least Significant Difference test, (2) Paarsgewijze t-toetsen met aangepaste onbetrouwbaarheid volgens Sidak of Bonferroni, (3) Tukey's Honestly Significant Difference test, (4) Duncan's Multiple F-test en (5) De Multiple Range test van Newman, Duncan en Keuls. Aandacht wordt besteed aan de robuustheid van deze methoden tegen verschillen in de varianties en tegen de aanwezigheid van een enkele uitschieter in de waarnemingen. Er worden twee aanpakken vergeleken voor ongebalanceerde designs.

1. Fisher's Least Significant Difference test (1935)

Methoden voor enkelvoudige variantie-analyse richten zich op de nulhypothese:

$$H_0: \mu_1 = \cdots = \mu_k$$

De waarnemingen zijn x_{ij}. Hierbij doorloopt i de verzameling van groepen en identificeert j de waarnemingen binnen iedere groep. Er geldt $i = 1, \cdots, k$ en $j = 1, \cdots, n_i$. Het aantal groepen is dus k en n_i is de i-de groepsgrootte. De gebruikelijke modelveronderstelling is $x_{ij} = \mu_i + e_{ij}$ waarbij de fouten e_{ij} normaal verdeeld zijn met verwachting 0 en konstante variantie σ^2. Ook wordt van de fouten verondersteld dat deze onderling onafhankelijk zijn.

De eerste stap in Fisher's Least Significant Difference test is een toetsing van H_0 met een klassieke F-toets met gekozen onbetrouwbaarheid, bijvoorbeeld $\alpha = 0.05$. Indien H_0 niet verworpen wordt is hiermee het proces voltooid. Maar anders wordt ook nog nagegaan welke gemiddelden significant verschillen. Daartoe wordt als tweede stap de hypothese $H_0: \mu_i = \mu_j$ getoetst voor ieder paar ongelijke i en j uit de collectie van 1 tot en met k. Hierbij wordt een t-toets gebruikt met dezelfde onbetrouwbaarheid α. De varianties van alle steekproeven worden bij de berekening van de toetsingsgrootheden samengevoegd.

Als de verzameling van paarsgewijze vergelijkingen niet was voorafgegaan door de simultane F-toets, dan zou de feitelijke onbetrouwbaarheid α^* aanzienlijke waarden kunnen aannemen. Duncan (1951) heeft uitgerekend dat voor een nominale α van 0.05 de waarde van α^* kan oplopen tot 0.1223 voor 3 steekproeven, 0.2034 voor 4 steekproeven en zelfs

Rekencentrum Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven, telefoon: 040-474535/474089

0.9183 voor 20 steekproeven. Hierbij stelt α^* de kans voor om tenminste één paar populatiegemiddelden uit de verzameling verschillend te noemen terwijl ze dat in feite niet zijn. Als alle lokatieparameters gelijk zijn, dan biedt Fisher's strategie voldoende bescherming, zodat $\alpha^* = \alpha$. Maar dit gaat niet op als bijvoorbeeld $\mu_1 = \cdots = \mu_m$ en $\mu_{m+1} = \cdots = m_k$, terwijl μ_1 en μ_{m+1} onderling verschillend zijn.

Hayter (1986) heeft de maximale kans op een onterechte verwerping door Fisher's Least Significant Difference test berekend. Hierbij is het maximum genomen over alle mogelijke configuraties van de werkelijke waarden van de lokatieparameters. Hayter's resultaat:

$$\alpha^*(k,\nu,\alpha) = P[q_{k-1,\nu} > \sqrt{2}t_\nu(\frac{\alpha}{2})]$$

Hierbij geldt $\nu = N-k$ met $N = \sum_{i=1}^{k} n_i$. De stochastische grootheid $q_{k-1,\nu}$ volgt de Studentised range verdeling met parameters $k-1$ en ν. De kritieke waarde $t_\nu(\frac{\alpha}{2})$ wordt gegeven door de inverse van Student's t-verdeling met ν vrijheidsgraden en tweezijdige onbetrouwbaarheid α. Als er meer dan twee groepen zijn, dan geldt de gelijkheid alleen voor gebalanceerde designs. Bij ongelijke steekproefgroottes geeft α^* een bovengrens aan. Op grond van dit resultaat is eenvoudig in te zien dat Fisher's test kan worden verbeterd door bij de tweede stap gebruik te maken van $q_{k-1,\nu}(\alpha)/\sqrt{2}$ in plaats van $t_\nu(\frac{\alpha}{2})$.

2. De methoden van Bonferroni en Sidak

Een alternatieve aanpak voor het simultaan vergelijken van steekproefgemiddelden wordt gevormd door een verzameling paarsgewijze t-toetsen. Een F-toets voor de steekproeven tesamen wordt achterwege gelaten en voor de bescherming van α^* wordt bij iedere t-toets een aangepaste onbetrouwbaarheid γ gebruikt die een functie is van α en het aantal steekproeven. De resultaten van een dergelijke strategie kunnen worden weergegeven door middel van een verzameling simultane betrouwbaarheidsintervallen:

$$\mu_i - \mu_j \in [x_i - x_j \mp t_\nu(\frac{\gamma}{2}) s\sqrt{1/n_i + 1/n_j}]$$

Hierbij geldt $x_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ en $s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-x_i)^2/(N-k)$. De eenvoudigste formule voor γ is gegeven door Bonferroni en deze wordt genoemd in Miller (1966). Er geldt $\gamma = \frac{2\alpha}{k(k-1)}$. En dit is gelijk aan de gekozen simultane onbetrouwbaarheid gedeeld door het aantal paren $\frac{k(k-1)}{2}$. Sidak (1967) gaf een andere relatie:

$$\gamma = 1-(1-\alpha)^{\frac{2}{k(k-1)}}$$

Beide aanpakken geven aanleiding tot een conservatief resultaat, maar omdat geldt:

$$1-(1-\alpha)^{\frac{2}{k(k-1)}} > \frac{2\alpha}{k(k-1)}$$

kunnen we concluderen dat Sidak minder conservatief is dan Bonferroni. Deze relatie geldt voor meer dan twee groepen.

3. Multiple Range tests

De nu te behandelen aanpak is ontleend aan Newman (1939), Duncan (1951) en Keuls (1952). Stel $x_{(1)}, \cdots, x_{(k)}$ zijn de steekproefgemiddelden in monotoon niet-dalende ordening. Stel verder dat $\mu_1, \cdots, \mu_k$ overeenkomstig hernummerd zijn. Voorlopig veronderstellen we ook nog dat alle steekproefgroottes gelijk zijn, zodat $n_i = n$. De eerste stap in de analyse betreft de hypothese H_0: $\mu_1 = \cdots = \mu_k$. Deze wordt getoetst met:

$$\mu_1 - \mu_k \in [x_{(1)} - x_{(k)} \mp q_{k,\nu}(\alpha)\frac{s}{\sqrt{n}}]$$

Duncan (1951) merkte op dat deze methode een nadeel heeft ten opzichte van de eertste stap in Fisher's Least Significant Difference test waar de gebruikelijke F-toets wordt toegepast. Stel namelijk dat er twee hypotheses zijn en dat H_0^1 een grotere likelihood heeft dan H_0^2. Dan kan bij gebruik van een q-toets H_0^1 worden verworpen terwijl H_0^2 wordt geaccepteerd. Deze paradoxale situatie is bij gebruik van een F-toets niet mogelijk.

Als de eerste stap aanleiding geeft tot acceptatie van de hypothese dan stopt de analyse. Anders wordt als volgt verder gegaan. De tweede stap betreft toetsing van de volgende twee hypotheses: H_0^A: $\mu_1 = \cdots = \mu_{k-1}$ en H_0^B: $\mu_2 = \cdots = \mu_k$. En zo wordt doorgegaan totdat elke hypothese is geaccepteerd. Het eindresultaat kan worden weergegeven door de geordende steekproefgemiddelden op een horizontale lijn te plaatsen en deze figuur te voorzien van een aantal (mogelijk overlappende) onderstrepingen. De interpretatie is dan dat een paarsgewijze vergelijking aanleiding geeft tot verwerping als er voor de twee steekproefgemiddelden niet een onderstreping is die beide omvat. Als een splitsingscandidaat p gemiddelden bevat dat wordt in de formule voor het betrouwbaarheidsinterval $q_{k,\nu}(\alpha)$ vervangen door $q_{p,\nu}(\alpha_p)$. Newman en Keuls suggereerden $\alpha_p = \alpha$. Dit garandeert alleen beheersing van α^* (de kans om tenminste één paar populatiegemiddelden uit de verzameling verschillend te noemen terwijl ze dat in feite niet zijn) als alle lokatieparameters gelijk zijn en hierdoor wordt deze aanpak vergelijkbaar met Fisher's Least Significant Difference test. Duncan suggereerde $\alpha_p = 1-(1-\alpha)^{p-1}$. Dit geeft α een paarsgewijze betekenis en streeft dus een ander doel na dan in deze notitie wordt beoogd. Ryan (1960) vond de volgende α_p:

$$\alpha_p = 1-(1-\alpha)^{\frac{p}{k}}$$

Met deze keuze kan voor de Multiple Range test bewezen worden dat $\alpha^* \leq \alpha$ als de steekproefgroottes gelijk zijn.

4. Ongebalanceerde designs

Fisher's Least Significant Difference test heeft door de keuze van de F-toets en de t-toets geen enkel probleem met ongelijke steekproefgroottes. Ook bij de paarsgewijze vergelijkingen van Bonferroni en Sidak zijn geen problemen te verwachten. Maar bij de Multiple Range test ligt de zaak heel anders. Miller (1966) suggereerde om in dit geval over te gaan tot de mediaan van de steekproefgroottes. Winer (1962) zag meer in het harmonisch gemiddelde:

$$H = \frac{1}{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}}$$

Deze keuze is in SAS geimplementeerd. Verderop zal worden getoond dat deze keuze niet in alle gevallen aanbevelenswaard is. Kramer (1956) paste de formule voor de betrouwbaarheidsintervallen aan aan deze situatie:

$$\mu_i - \mu_j \in [x_{(i)} - x_{(j)} \mp q_{p,\nu}(\alpha_p)s\sqrt{(1/n_i + 1/n_j)/2}]$$

Kramer's aanpak bevat een valkuil die als volgt kan worden geillustreerd. Veronderstel dat er vier groepen zijn met steekproefgroottes n_1 tot en met n_4 na ordening op de steekproefgemiddelden. Veronderstel verder dat n_1 en n_4 veel kleiner zijn dan n_2 en n_3. Bij de eerste stap spelen alleen de uiterste gemiddelden een rol en hier kan de hypothese geaccepteerd worden ten gevolge van de kleine steekproefgroottes. Aan een vergelijking van μ_2 en μ_3 komt de methode dan niet meer toe. Maar juist deze vergelijking kan door de grote steekproeven bij een paarsgewijze toetsing resulteren in significantie. In deze situatie is het gebruik van Kramer's modificatie dus gevaarlijk en verdienen paarsgewijze vergelijkingen of Fisher's aanpak de voorkeur.

5. Tukey's (1953) Honestly Significant Difference test

Dit is evenals de Multiple Range test een methode die gebaseerd is op de q-verdeling, maar de strategie is er een van paarsgewijze vergelijkingen. Voorlopig wordt verondersteld dat de steekproefgroottes gelijk zijn, zodat $n_i = n$. De toets kan nu gepresenteerd worden als een stelsel simultane betrouwbaarheidsintervallen:

$$\mu_i - \mu_j \in [x_i - x_j \mp q_{k,\nu}(\alpha)\frac{s}{\sqrt{n}}]$$

Tukey gebruikt $q_{k,\nu}(\alpha)$ waar in Hayter's modificatie van de Least Significant Difference test $q_{k-1,\nu}(\alpha)$ staat. Het verschil komt voort uit de bescherming die de F-toets in het tweede geval biedt. Bij ongelijke steekproefgroottes kan hier Kramer's modificatie worden toegepast:

$$\mu_i - \mu_j \in [x_i - x_j \mp q_{k,\nu}(\alpha)s\sqrt{(1/n_i + 1/n_j)/2}]$$

Omdat hier niet wordt gewerkt met de geordende steekproefgemiddelden kan de eerder genoemde valkuil van Kramer's aanpak hier niet optreden. Games en Howell (1976) hebben door een uitgebreid simulatie-onderzoek aangetoond dat $\alpha^* \leqslant \alpha$ met deze aanpassing. Dit vermoeden is in 1984 door Hayter analytisch bewezen.

We wisten reeds dat de methode van Sidak minder conservatief is dan de methode van Bonferroni. Tukey werkt ook met paarsgewijze vergelijkingen en daarom is het interessant om deze hiermee te vergelijken. Tamhane (1979) heeft aangetoond dat:

$$\frac{q_{k,\nu}(\alpha)}{\sqrt{2}} < t_{\nu}\left(\frac{\gamma}{2}\right) \text{ met } \gamma = 1-(1-\alpha)^{\frac{2}{k(k-1)}}$$

Deze relatie geldt voor meer dan twee groepen. Bij twee groepen wordt het een identiteit. Op grond van dit resultaat kunnen we konkluderen dat Tukey nog minder conservatief is dan Sidak.

6. Multiple F-tests

De nu te behandelen strategie is ontleend aan Duncan (1951). De aanpak lijkt op die van de Multiple Range test, maar er wordt een F-toets gebruikt in plaats van een q-toets, zodat de eerste stap een klassieke enkelvoudige variantie-analyse wordt. Duncan suggereerde voor iedere stap $\alpha_p = 1-(1-\alpha)^{p-1}$. Evenals bij de Multiple Range test geeft dit α een paarsgewijze betekenis en maakt de toets dus onaantrekkelijk binnen het kader van deze notitie. Petrinovich en Hardyck (1969) hebben aangetoond dat het gedrag bij deze α_p vergelijkbaar is met paarsgewijze t-toetsen waarbij voor ieder paar een onbetrouwbaarheid α wordt gebruikt. Later (1955) is Duncan ook meer in de sfeer van de Simultane Statistiek gaan denken en suggereerde hij $\alpha_p = 1-(1-\alpha)^{(p-1)/(k-1)}$. Hiervoor geldt $\alpha^* \leqslant \alpha$, maar het resultaat is conservatiever dan de eerder genoemde keuze van Ryan $\alpha_p = 1-(1-\alpha)^{p/k}$ en is dus verouderd. Welsch (1977) heeft de aanpak van Ryan nog iets kunnen verscherpen:

$$\alpha_p = 1-(1-\alpha)^{\frac{p}{k}} \text{ voor } p < k-1 \text{ en}$$

$$\alpha_p = \alpha \text{ voor } p \geqslant k-1$$

Dit is tot nu toe de scherpst gevonden grens en deze is dan ook terecht in SAS geimplementeerd. En niet alleen hier, maar ook bij de Multiple Range test. De F-toets staat ongelijke steekproefgroottes toe en daarom lijkt de Multiple F-test aantrekkelijker dan de Multiple Range test. Maar er is een probleem. Denk aan vier steekproeven met geringe steekproefgroottes voor de uiterste twee steekproefgemiddelden en zeer veel waarnemingen voor de middelste twee gemiddelden. Als de eerste stap aanleiding geeft tot verwerping is de onvermijdelijke conclusie dat μ_1 en μ_4 verschillend zijn. Maar deze conclusie komt uitsluitend voort uit de strategie en hoeft ook zeker niet in overeenstemming te zijn met het resultaat van een stelsel paarsgewijze vergelijkingen. Daarom is de Multiple F-test over het algemeen bij serieus ongelijke steekproefgroottes af te raden. Verderop zal hiervan een voorbeeld worden gegeven.

7. Een voorbeeld met SAS

Gegeven zijn waarnemingen uit vier groepen. Om de tot nu toe genoemde problemen te demonstreren zijn de steekproefgroottes verschillend gekozen. Een samenvatting is gegeven in onderstaande tabel: Toetsing van H_0: $\mu_1 = \cdots = \mu_4$ met een F-toets levert een overschrijdingskans van 0.0029 op zodat deze hypothese wordt verworpen. Op deze data zijn

nummer	n_i	x_i	s_i	$s_i/\sqrt{n_i}$
1	3	11.36	1.540	0.8891
2	10	11.45	1.191	0.3768
3	10	13.40	0.955	0.3020
4	3	13.61	1.593	0.9197

middels SAS alle tot nu toe genoemde methoden toegepast. Bij Fisher's Least Significant Difference test betreft dit de oude versie; SAS heeft Hayter's correctie (nog) niet aangebracht. Een zeer onplezierige eigenschap van SAS is dat zodra de Multiple F-test of de Multiple Range test is gevraagd, voor alle methoden wordt gewerkt met het harmonisch gemiddelde (hier 4.615) van de steekproefgroottes. En dus ook ook voor de toetsen die een dergelijke maatregel helemaal niet nodig hebben. Bij een gekozen onbetrouwbaarheid $\alpha = 0.05$ worden de volgende resultaten bereikt:

- De Least Significant Difference test vindt twee clusters, te weten $[\mu_1, \mu_2]$ en $[\mu_3, \mu_4]$. De Multiple F-test en de Multiple Range test komen tot hetzelfde resultaat.
- Volgens Sidak en Bonferroni is geen enkel verschil significant. Gezien het geringe verschil tussen deze methoden is deze overeenstemming niet verwonderlijk.
- Tukey's Honestly Significant Difference test heeft iets meer onderscheidend vermogen dan Sidak en Bonferroni zoals we reeds zagen. Deze methode vindt dan ook twee clusters: $[\mu_1, \cdots, \mu_3]$ en $[\mu_2, \cdots, \mu_4]$. Derhalve worden μ_1 en μ_4 als enig paar significant verschillend genoemd.

SAS biedt een optie waarmee de resultaten van alle genoemde methoden behalve de Multiple Range test en de Multiple F-test worden gegeven in de vorm van simultane betrouwbaarheidsintervallen. Bij deze aanpak wordt gewerkt met de oorspronkelijke steekproefgroottes. Dit resulteert in de volgende uitspraken:

- De resultaten van Fisher's Least Significant Difference test zijn dezelfde als bij gebruikmaking van het harmonisch gemiddelde van de steekproefgroottes: de paren μ_1, μ_2 en μ_3, μ_4 zijn niet significant verschillend en alle andere paren wel. Dit grote aantal gevonden verschillen is een gevolg van het feit dat hier bij de paarsgewijze vergelijkingen steeds dezelfde ongecorrigeerde onbetrouwbaarheid α wordt gebruikt.
- Sidak, Bonferroni en Tukey komen bij gebruikt van de juiste steekproefgroottes tot dezelfde conclusie: alleen het paar μ_2, μ_3 is significant verschillend. En dat is in overeenstemming met de wijze waarop deze kunstmatige steekproeven zijn geconstrueerd.

Uit het voorgaande kunnen de volgende conclusies getrokken worden: (1) De keuze van SAS, om voor alle methoden met het harmonisch gemiddelde van de steekproefgroottes te werken als de Multiple Range test of de Multiple F-test bij de gevraagde toetsen zitten, is onjuist. (2) De Multiple Range test en de Multiple F-test zijn alleen bruikbaar als de steekproefgroottes (bijna) gelijk zijn, en (3) Het wordt tijd dat SAS Hayter's correctie op Fisher's Least Significant Difference test aanbrengt.

8. Robuustheid ten aanzien van ongelijkheid in de varianties

De eerste stap van Fisher's Least Significant Difference test is toepassing van de klassieke F-toets. Hiervan is reeds door Brown en Forsythe (1974) en Ekbohm (1976) aangetoond dat de controle over de gekozen onbetrouwbaarheid zeer onbevredigend is als de varianties binnen de verschillende groepen (grote) verschillen vertonen. Er zijn echter een aantal alternatieven gepubliceerd voor deze situatie en daarvan is de tweede orde methode van James (1951) de beste keuze, zoals uit een studie van Dijkstra en Werter (1981) blijken moge. Deze methode gebruikt de volgende toetsingsgrootheid:

$$J = \sum_{i=1}^{k} w_i (x_i - \bar{x})^2$$

Hierbij geldt $w_i = n_i / s_i^2$, $\bar{x} = \sum_{i=1}^{k} w_i x_i / W$ en $W = \sum_{i=1}^{k} w_i$. Voor de berekening van de kritieke waarde wordt een tweede orde Taylor benadering van een analytische oplossing gebruikt. De daarbij behorende formules zijn dermate ingewikkeld dat ze hier achterwege worden gelaten.

Aanpassing van de t-toets voor ongelijke varianties is ook niet bijzonder problematisch. Welch (1949) gebruikt de volgende toetsingsgrootheid:

$$t = \frac{x_i - x_j}{\sqrt{s_i^2/n_i + s_j^2/n_j}}$$

Bij benadering is deze toetsingsgrootheid onder de nulhypothese van gelijke lokatieparameters verdeeld als $t_{\nu_{ij}}$, waarbij ν_{ij} gevonden wordt middels onderstaande uitdrukking:

$$\nu_{ij} = \frac{(s_i^2/n_i + s_j^2/n_j)^2}{\frac{s_i^4}{n_i^2(n_i-1)} + \frac{s_j^4}{n_j^2(n_j-1)}}$$

Als ν_{ij} geen geheel getal getal is dan wordt de waarde vervangen door het dichtsbijzijnde gehele getal. In een uitgebreide simulatiestudie heeft Wang (1971) aangetoond dat bij deze toets de controle over de gekozen onbetrouwbaarheid zeer bevredigend is, ongeacht de waarde van de nuisance parameter $\theta = \sigma_i^2 / \sigma_j^2$.

Voor de Multiple Range test ligt nu de volgende aanpassing aan ongelijke varianties voor de hand, waarbij gebruik is gemaakt van Welch's resultaten:

$$\mu_i - \mu_j \in [x_{(i)} - x_{(j)} \mp q_{p,\nu_{ij}}(\alpha_p) \sqrt{(s_i^2/n_i + s_j^2/n_j)/2}]$$

Een dergelijke strategie is echter niet aanbevelenswaard, zoals midddels een eenvoudig voorbeeld met vier steekproeven kan worden gedemonstreerd. Veronderstel dat de varianties behorende bij de uiterste twee steekproefgemiddelden (veel) groter zijn dan de varianties behorende bij de middelste twee. Het is dan goed mogelijk dat de strategie in de eerste stap eindigt, terwijl een paarsgewijze vergelijking van μ_2 en μ_3 tot een overtuigende verwerping aanleiding zou kunnen geven.

Analoog aan de Multiple F-test zou men kunnen denken aan een Multiple James test om het probleem van de ongelijke varianties te lijf te gaan. Maar ook dan onstaat er in bovenstaand voorbeeld een conflict met de strategie. Als namelijk in de eerste stap de hypothese, dat alle lokatieparameters gelijk zijn, wordt verworpen, dan wordt vervolgens geconcludeerd dat μ_1 en μ_4 verschillend zijn. Maar deze conclusie komt slechts voort uit de strategie en het is zeer wel mogelijk dat een paarsgewijze vergelijking niet tot verwerping zou leiden vanwege de grote varianties.

Aanpassing van Tukey's Honestly Significant Difference test middels Welch's resultaten levert geen problemen op:

$$\mu_i - \mu_j \in [x_i - x_j \mp q_{k,\nu_{ij}}(\alpha)\sqrt{(s_i^2/n_i + s_j^2/n_j)/2}]$$

Games en Howell (1983) hebben vastgesteld dat de performance van deze modificatie niets te wensen overlaat.

Gezien de ruime keuze aan mogelijkheden die op dit gebied geboden worden is het verrassend en betreurenswaardig dat hiervan niets in SAS is opgenomen. De noodzaak van een dergelijke uitbreiding kan gedemonstreerd worden aan de hand van het volgende voorbeeld:

nummer	x_i	s_i	n_i	$s_i/\sqrt{n_i}$
1	0.725	1.958	15	0.5056
2	0.759	0.343	15	0.0886
3	1.358	0.273	15	0.0705
4	1.999	1.722	15	0.4446

Voor deze situatie is het gebruik van Sidak-Welch, Bonferroni-Welch of Tukey-Welch aanbevolen, en deze drie methoden resulteren in dezelfde conclusie: μ_2 en μ_3 zijn verschillend en andere significanties worden niet gevonden. Dat is in overeenstemming met de wijze waarop dit kunstmatige voorbeeld is geconstrueerd. De niet-robuuste analyses uit SAS komen tot een heel ander resultaat. Fisher's ongecorrigeerde Least Significant Difference test en de Multiple F-test verklaren het paar μ_1, μ_4 significant verschillend en de overige vier methoden vinden geen enkel verschil. Deze uitkomsten kunnen als volgt worden verklaard:

- Bij alle strategiën uit SAS worden de varianties samengevoegd. Daardoor wordt de variantie van het verschil tussen μ_2 en μ_3 aanzienlijk overschat en dit maskeert de significantie. Daarom vinden de Multiple Range test alsmede de methoden van Sidak, Tukey en Bonferroni geen enkel verschil.
- De strategie van geordende steekproefgemiddelden komt in conflict met de ongelijkheid van de varianties bij de Multiple F-test. Er is voldoende verschil om de hypothese te verwerpen dat alle lokatieparameters gelijk zijn. Maar daaruit wordt dan geconcludeerd dat de uiterste twee verschillend zijn. En een dergelijke conclusie is bij heteroscedasticiteit niet verantwoord.

- Fisher's Least Significant Difference test resulteert om een andere reden in de uitspraak dat μ_1 en μ_4 verschillend zijn. Na verwerping van de gelijkheid van alle lokatieparameters middels de F-toets worden paarsgewijze t-toetsen toegepast met ongecorrigeerde onbetrouwbaarheid $\alpha = 0.05$ en samengevoegde varianties. De drempel voor verwerping wordt daardoor veel lager dan bij bijvoorbeeld Bonferroni of Sidak en het is dan ook niet verwonderlijk dat bij het grootste verschil tot significantie wordt geconcludeerd.

De conclusie van deze paragraaf is dat de door SAS aangeboden methoden niet robuust zijn ten aanzien van ongelijkheid in de varianties. Er zijn echter voldoende geschikte alternatieven uit de literatuur bekend en het wordt tijd dat SAS hier enkele van opneemt.

9. Robuustheid ten aanzien van uitschieters

Alle genoemde methoden maken gebruik van het steekproefgemiddelde en dat is niet robuust ten aanzien van uitschieters. Een enkele uitschieter is in principe voldoende om deze schatter over iedere denkbare grens heen te trekken. Zie hiervoor bijvoorbeeld Hampel, Ronchetti, Rousseeuw en Stahel (1986). Het steekproefgemiddelde wordt doorgaans berekend als $\bar{x} = \sum_{i=1}^{n} x_i / n$, maar in verband met een straks te behandelen robuuste generalisatie hiervan zal eerst een andere uitdrukking voor $\bar{x}$ gegeven worden:

$$\bar{x} = [\theta | \sum_{i=1}^{n} (x_i - \theta)^2 \ minimaal\ ]$$

Dit is voor de normale verdeling een Maximum Likelihood schatter. Een generalisatie hiervan wordt gevormd door de zogenaamde M-schatters, waarvan Huber (1973) een aantrekkelijk voorbeeld heeft gegeven:

$$x_H = [\theta | \sum_{i=1}^{n} \rho_H (\frac{x_i - \theta}{\sigma}) \ minimaal\ ]$$

De functie ρ_H is voor kleine absolute waarden van zijn argument kwadratisch zodat daar geen verschil is met de klassieke schatter. Maar zodra x_i verder van θ af ligt wordt overgegaan op een lineaire functie zodat de invloed van uitschieters aanzienlijk wordt verminderd:

$$\rho_H(r) = \frac{r^2}{2} \ voor\ |r| \leq H \ en$$

$$\rho_H(r) = H|r| - \frac{H^2}{2} \ voor\ |r| > H$$

Voor $H = 1.345$ wordt een efficiency van 95% bereikt voor de normale verdeling. In het algemeen zal σ onbekend zijn en moet deze waarde dus worden vervangen door een robuuste schatter:

$$\hat{\sigma} = 1.483[med_j |x_j - med_i x_i|]$$

Door gebruikmaking van de mediaan is deze schatter zeer ongevoelig voor uitschieters. De factor 1.483 zorgt ervoor dat de verwachting bij een normale bij benadering gelijk is aan

die van de klassieke standaarddeviatie. Voor de berekening is een iteratief proces en een beginschatter nodig. Goede resultaten zijn verkregen met iteratief herwogen kleinste kwadraten volgens Holland en Welsch (1977). De mediaan is zeer geschikt als beginschatter.

Alle genoemde methoden voor Multiple Comparisons zijn robuust te maken ten aanzien van uitschieters door het gemiddelde consequent te vervangen door de hierboven beschreven Huber-schatter. De berekening van de residuele variantie moet natuurlijk worden aangepast, maar daarvoor geeft Huber (1981) bruikbare principes. Dit alles resulteert in methoden met een zeer gering verlies aan onderscheidend vermogen ten opzichte van hun klassieke tegenhangers als de data normaal verdeeld zijn. Maar van uitschieters hebben ze weinig last, zoals beschreven is in Dijkstra (1987). Hierin wordt een voorbeeld behandeld met 6 steekproeven van 10 waarnemingen elk. Deze data stellen lengtes van personen voor. Ze zijn weergegeven in meters met twee cijfers achter de decimale punt. Een uitschieter wordt gesimuleerd door één keer de decimale punt te vergeten zodat een persoon 146 meter lang lijkt in plaats van 1.46 meter.

In een analyse met SAS is Fisher's Least Significant Difference test buiten beschouwing gelaten omdat daar Hayter's correctie niet is aangebracht. Zonder de uitschieter vinden alle methoden 12 verschillen en dat is in overeenstemming met de wijze waarop de data zijn gegenereerd. Maar met de uitschieter wordt geen enkele verschil gevonden. Zelfs niet bij de paarsgewijze methoden, omdat ook daar gebruik wordt gemaakt van de samengevoegde varianties.

Als de methoden worden aangepast met Huber-schatters vindt de Multiple Range test zonder de uitschieter dezelfde 12 verschillen. Bonferroni, Tukey, Sidak, de Multiple F-test en Fisher's Least Significant Difference test met Hayter's correctie vinden er slechts 11. Dit hangt samen met het geringe verlies aan onderscheidend vermogen. In tegenstelling tot bij de klassieke methoden verandert er hierbij echter niets in de conclusies zodra de uitschieter in de analyse wordt opgenomen.

De conclusie is dat de in SAS opgenomen klassieke methoden zeer gevoelig zijn voor uitschieters. En dat kan rampzalig zijn als de gebruiker de ingetoetste data niet eerst inspecteert. SAS biedt daartoe een goed stuk gereedschap in PROC UNIVARIATE, maar helaas zal niet iedereen dat gebruiken. Het zou prettig zijn als SAS wat meer robuuste methoden opnam, en niet alleen bij de Multiple Comparisons.

De kritiek op SAS betreffende het ontbreken van robuuste methoden kan worden gerelativeerd door een vergelijking te maken met andere statistische pakketten. Voor de hier behandelde soort applicaties kunnen we daarbij denken aan SPSSX, BMDP en GLIM. SPSSX heeft in dit opzicht ongeveer even weinig te bieden als SAS. BMDP bevat wat mogelijkheden voor ongelijke varianties (foutief geprogrammeerd indien er een steekproef bij is met slechts één element), maar dit beperkt zich tot de simultane vergelijking van k gemiddelden. GLIM biedt geen robuuste methoden die als zodanig zijn aangekondigd, maar sommige methoden (met name de uitschieter-resistente) zijn in de taal van dit pakket wel zeer eenvoudig uit te drukken.

10. Literatuur

Brown, M.B. en A.B. Forsythe (1974) The small sample behaviour of some statistics which test the equality of several means.
Technometrics (16) 129-132.

Dijkstra, J.B. en P.S.P.J. Werter (1982) Testing the equality of several means when the population variances are unequal.
Communications in Statistics. Simulation and Computation (B10-6) 557-569.

Dijkstra, J.B. (1987) Analysis of means in some non-standard situations.
Proefschrift Technische Universiteit Eindhoven.

Duncan, D.B. (1951) A significance test for differences between ranked treatments in an analysis of variances.
Virginia Journal of Science (2).

Duncan, D.B. (1955) Multiple Range and Multiple F-tests.
Biometrics (11) 1-42.

Ekbohm, G. (1976) On testing the equality of several means with small samples.
The Agricultural College of Sweden (Uppsala) 547-553.

Fisher, R.A. (1935) The design of experiments.
Oliver and Boyd. Edinburgh en London.

Games, P.A. en J.F. Howell (1976) Pairwise multiple comparison procedures with unequal N's and/or variances: a Monte Carlo study.
Journal of Educational Statistics (1) 113-125.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw en W.A. Stahel (1986) Robust statistics: the approach based on influence functions.
John Wiley & Sons. New York.

Hayter, A.J. (1984) A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative.
The Annals of Statistics (12) 61-75.

Hayter, A.J. (1986) The maximum familywise error rate of Fisher's Least Significant Difference test.
Journal of the American Statistical Association (81) 1000-1004.

Holland, P.W. en R.E. Welsch (1977) Robust regression using iteratively reweighted least squares.
Communications in Statistics (A6-9) 813-827.

Huber, P.J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo.
Annals of Statistics (1) 799-821.

Huber, P.J. (1981) Robust statistics.
John Wiley & Sons. New York.

James, G.S. (1951) The comparison of several groups of observations when the ratios of the population variances are unknown.
Biometrika (38) 324-329.

Keuls, M. (1952) The use of the studentized range in connection with an analysis of variance.
Euphytica (1) 112-122.

Kramer, C.Y. (1956) Extension of Multiple Range tests to group means with unequal numbers of replications.
Biometrics (12) 307-310.

Miller, R.G. (1966) Simultaneous statistical inference.
McGraw-Hill Book Company. New York.

Newman, D. (1939) The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of the standard deviation.
Biometrika (31) 20-30.

Petrinovich, L.F. en C.D. Hardyck (1969) Error rates for multiple comparison methods: some evidence concerning the frequency of erroneous conclusions.
Psychological Bulletin (71) 43-54.

Ryan, T.A. (1960) Significance tests for multiple comparisons of proportions, variances and other statistics.
Psychological Bulletin (57) 318-328.

Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions.
Journal of the American Statistical Association (62) 626-633.

Tamhane, A.C. (1979) A comparison of procedures for multiple comparisons of means with unequal variances.
Journal of the American Statistical Association (74) 471-480.

Wang, Y.Y. (1971) Probabilities of type I errors of the Welch tests for the Behrens-Fisher problem.
Journal of the American Statistical Association (66) 605-608.

Welch, B.L. (1949) Further note on Mrs. Aspin's tables and on certain approximations to the tabled function.
Biometrika (36) 293-296.

Welsch, R.E. (1977) Stepwise multiple comparison procedures.
Journal of the American Statistical Association (72) 566-575.

Winer, B.J. (1962) Statistical principles in experimental design.
McGraw-Hill Book Company. New York.

Ontvangen: 07-03-1988
Geaccepteerd: 12-09-1988